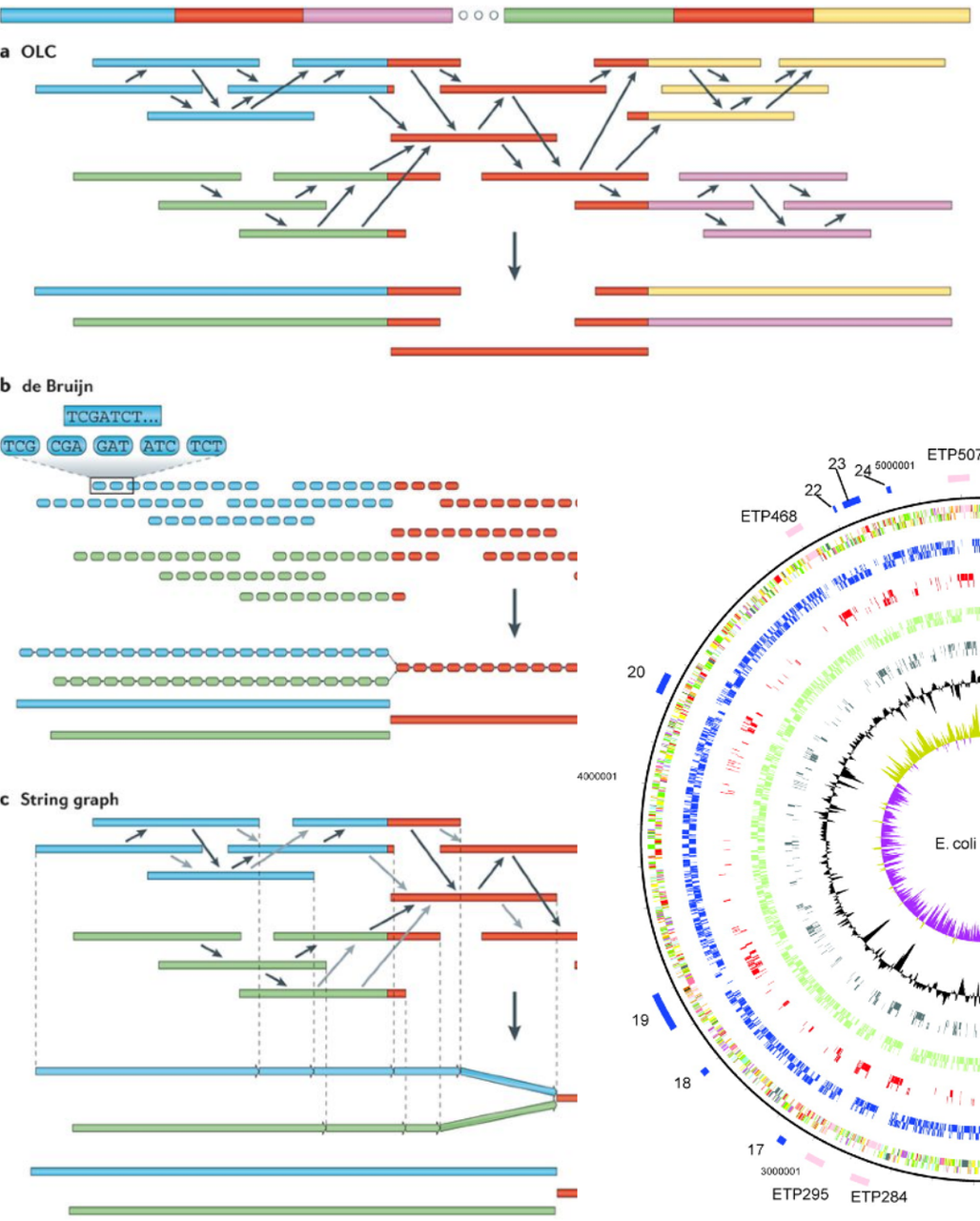
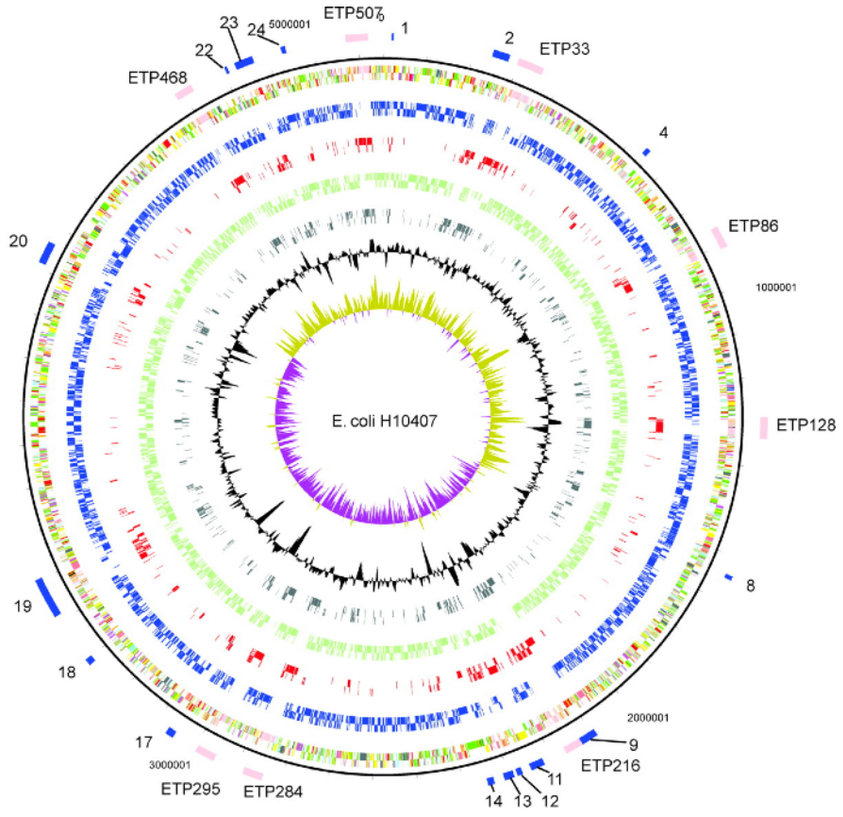
KAUST

Assembly Algorithms

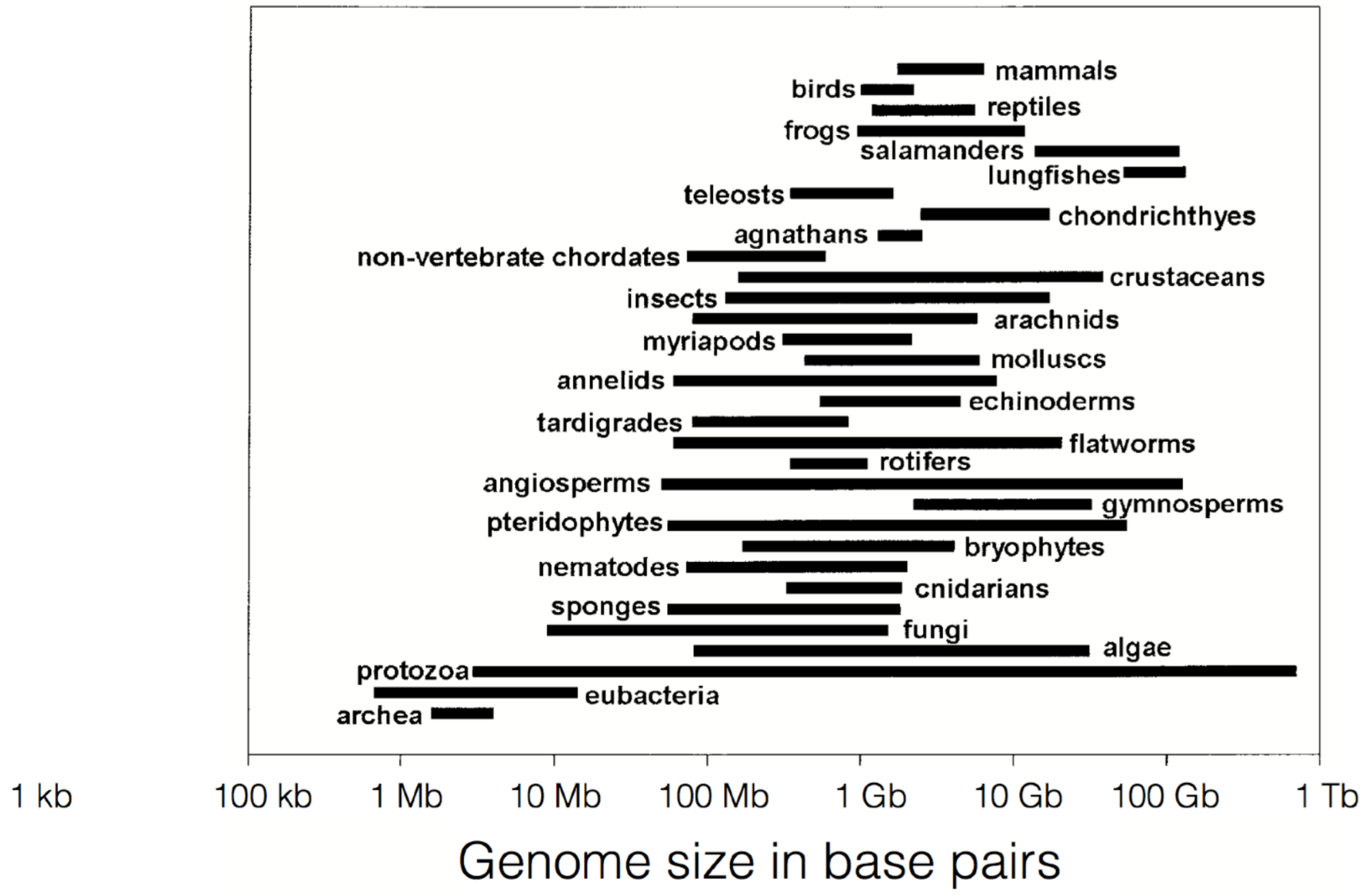




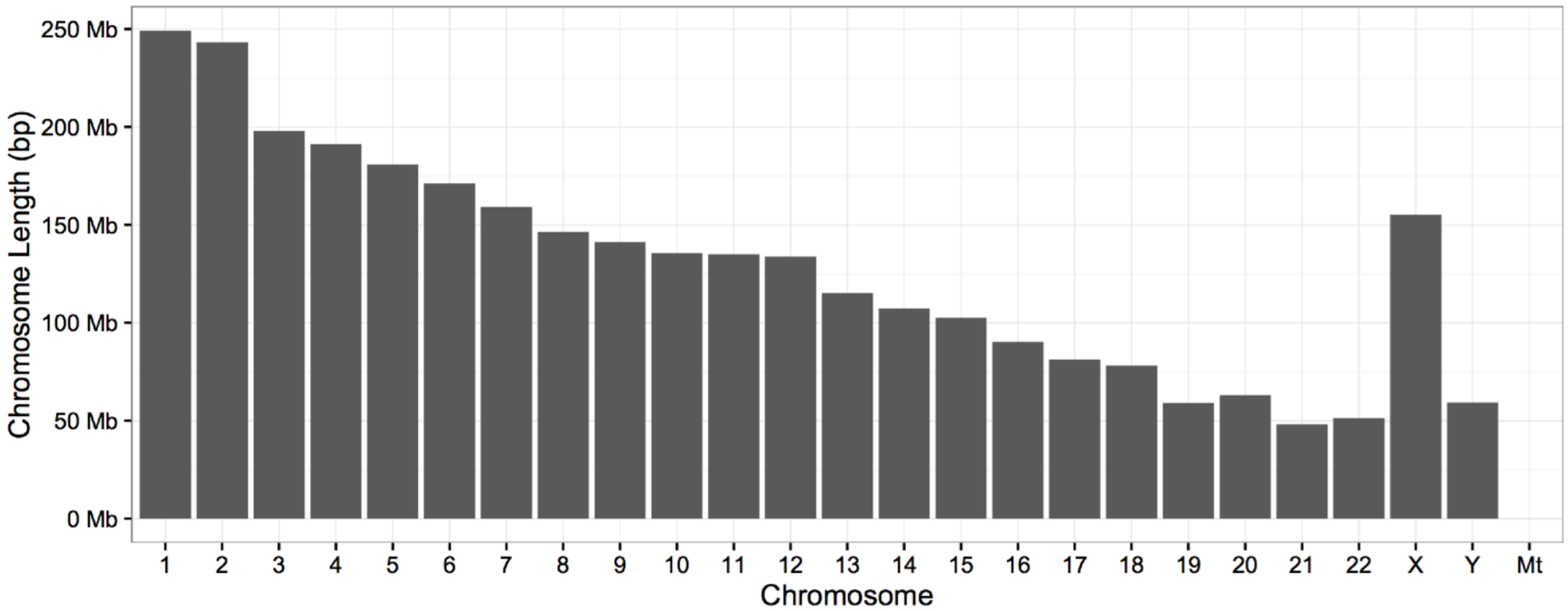
HTS Workshop Genomics & Transcriptomics KAUST 2019

Robert Lehmann
Octavio Salazar

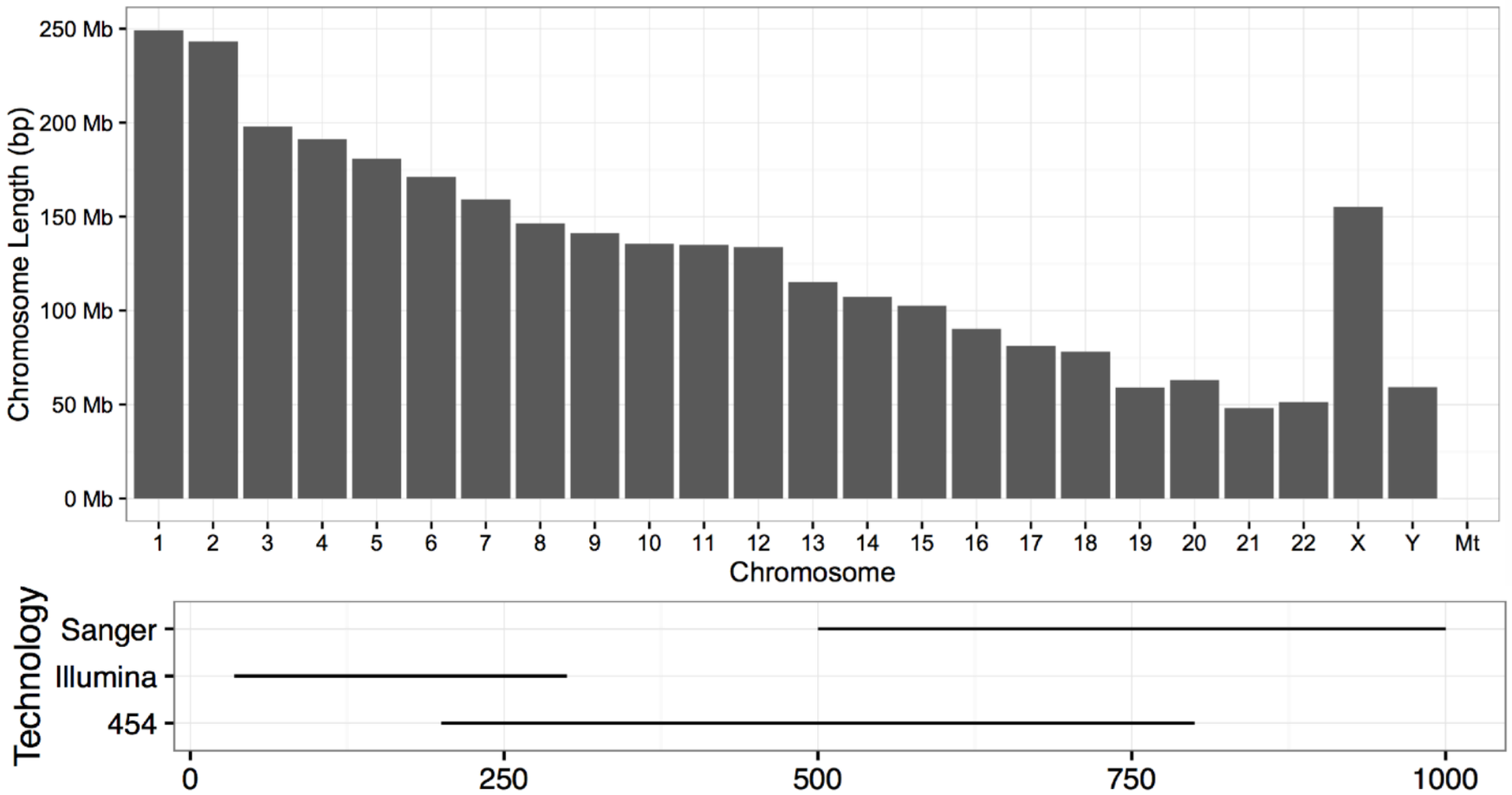
Complex organism = large genome?



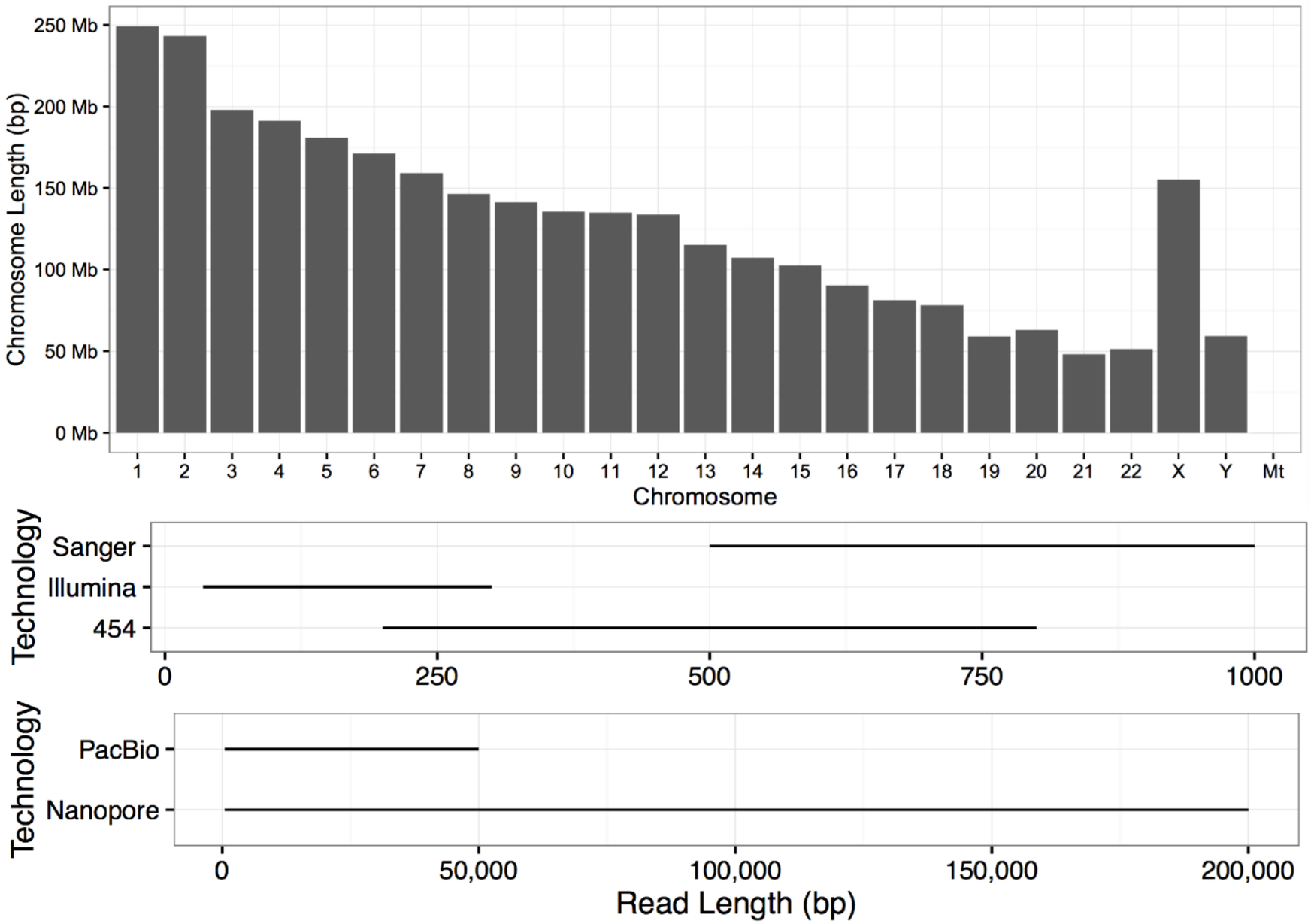
Read length and chromosome length



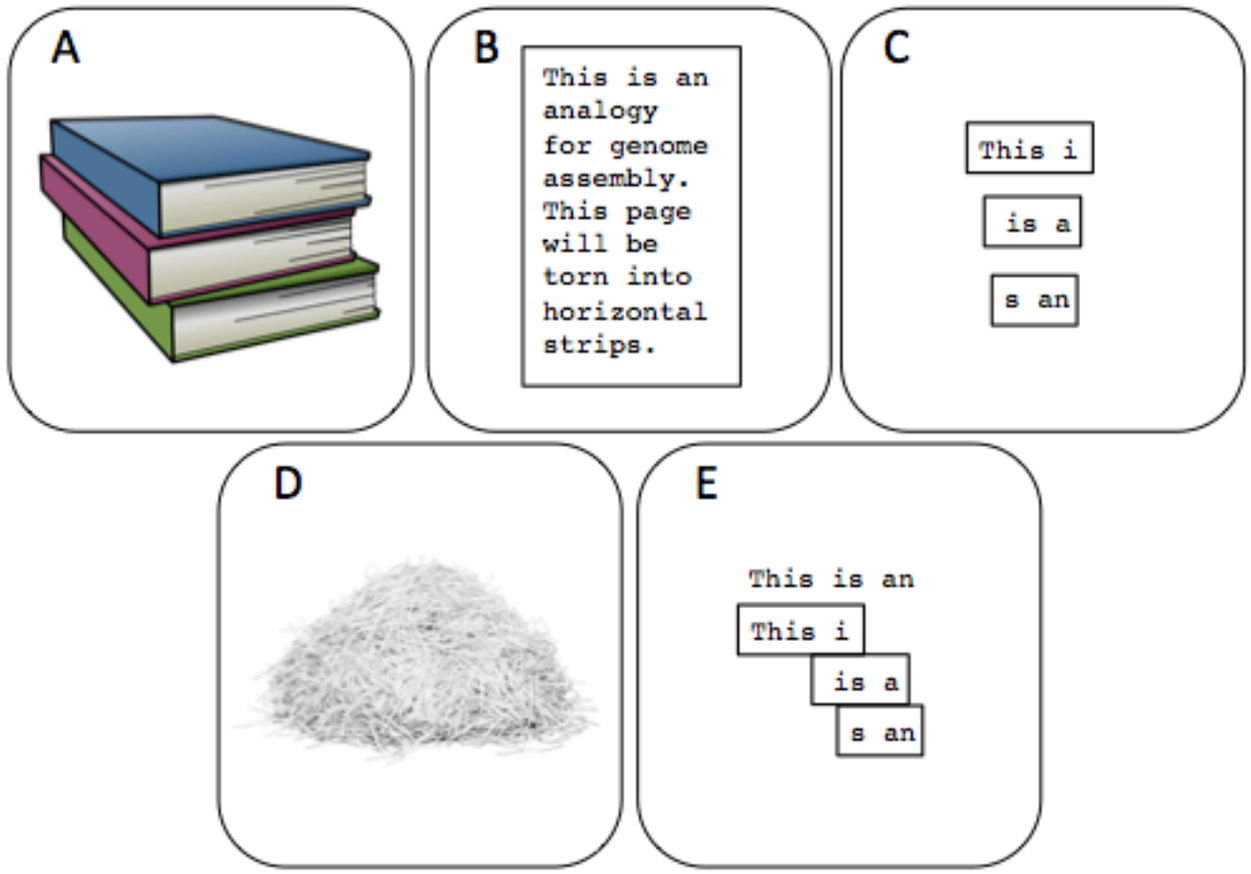
Read length and chromosome length



Read length vs. chromosome length

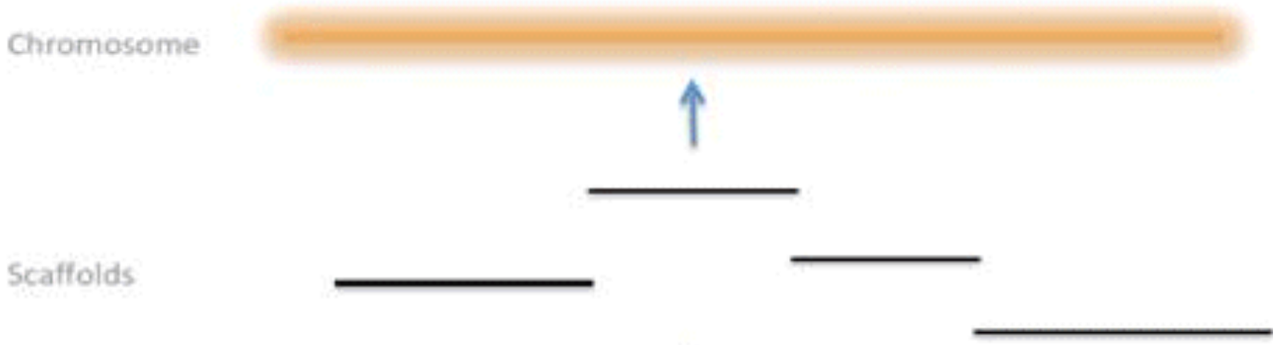


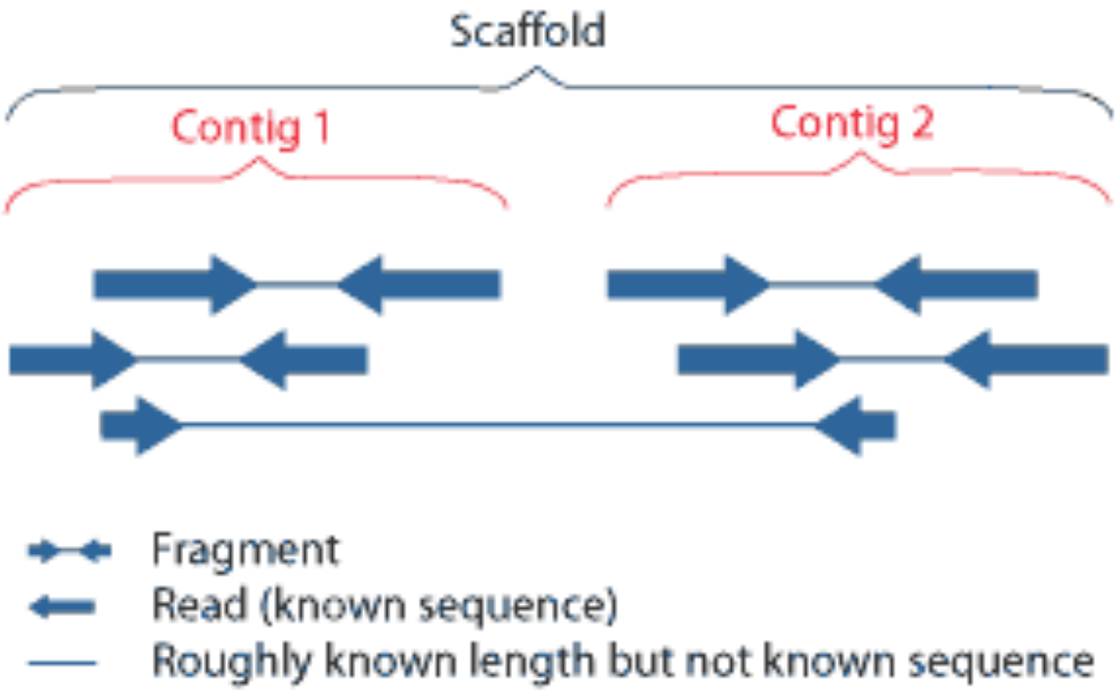
Assembling the pieces

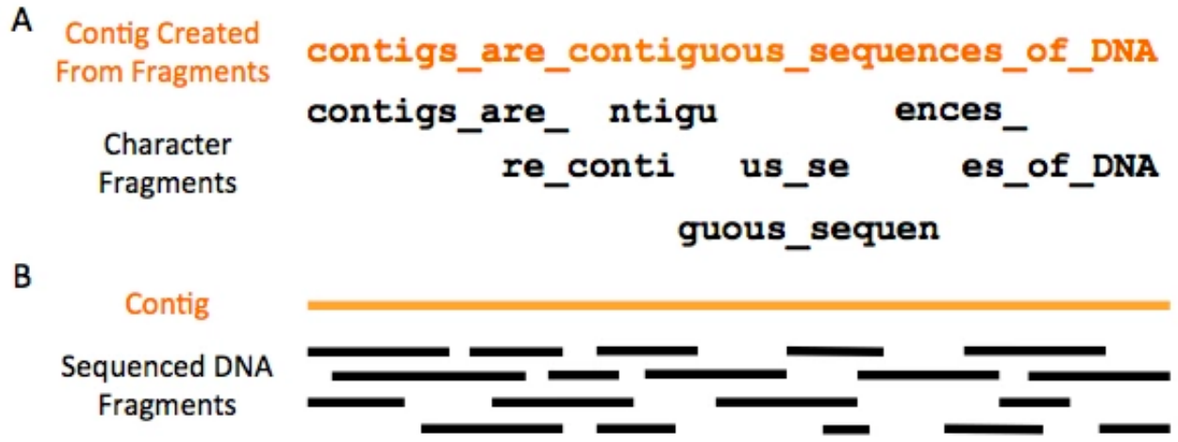


http://gcat.davidson.edu/phast/

Assembling the pieces

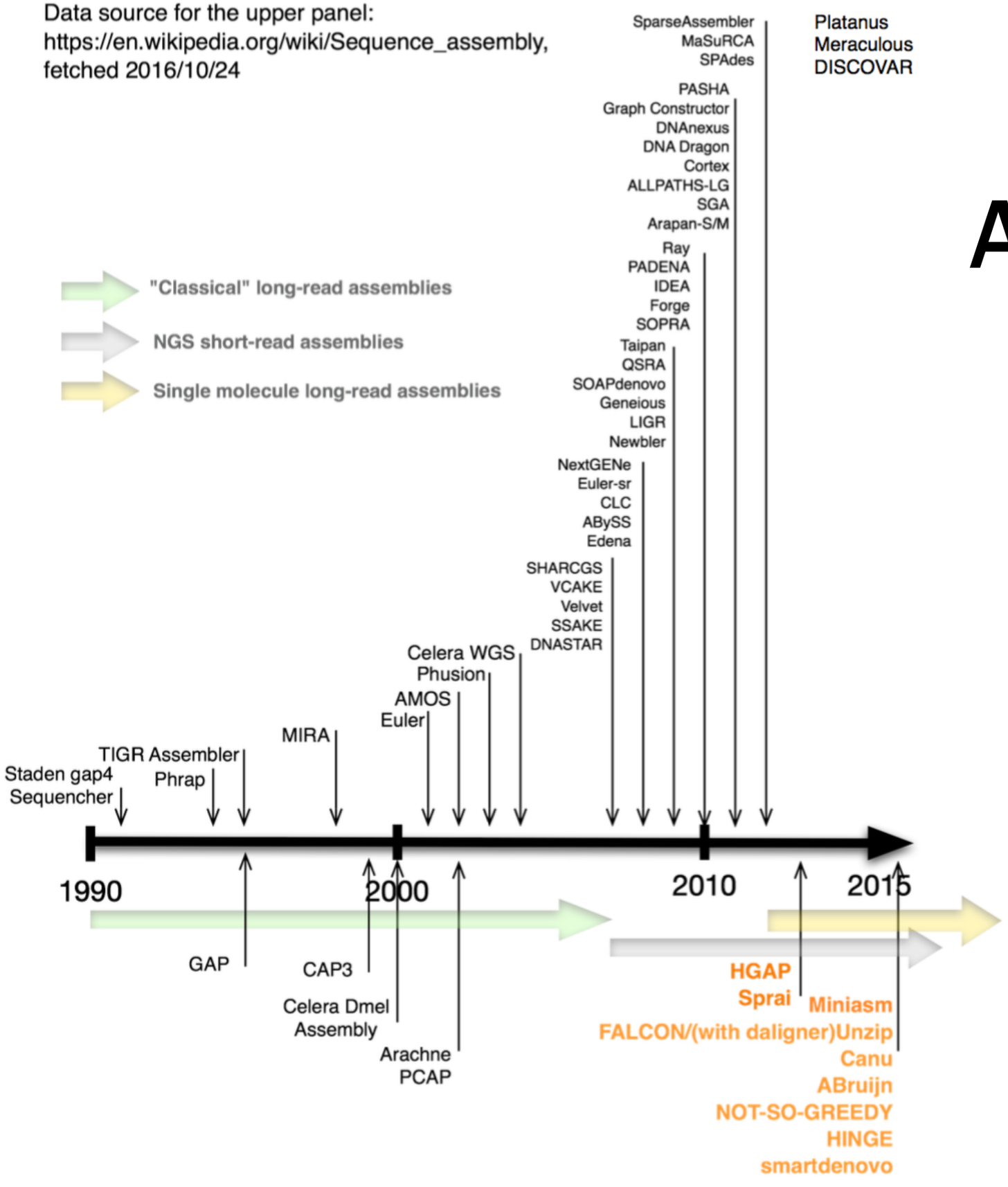






Lots of Assemblers developed

Data source for the upper panel:
https://en.wikipedia.org/wiki/Sequence_assembly,
fetched 2016/10/24



Chin J, https://speakerdeck.com/jchin/recent-progress-in-long-read-genome-assembly-theory-practice-and-future-challenges

Algorithm Feature	OLC Assemblers	DBG Assemblers
Modeled features of reads		
Base substitutions		Euler, AllPaths, SOAP
Homopolymer miscount	CABOG	
Concentrated error in 3′ end		Euler
Flow space	Newbler	
Color space	Shorty	Velvet
Removal of erroneous reads		
Based on K-mer frequencies		Euler, Velvet, AllPaths
Based on K-mer freq and QV		AllPaths
For multiple values of K		AllPaths
By alignment to other reads	CABOG	
By alignment and QV		
Correction of erroneous base calls		
Based on K-mer frequencies		Euler, SOAP
Based on Kmer freq and QV		AllPaths
Based on alignments	CABOG	
Approaches to graph construction		
Implicit		
Reads as graph nodes	CABOG, Newbler, Edena	
K-mers as graph nodes		Euler, Velvet, ABySS, SOAP
Simple paths as graph nodes		AllPaths
Multiple values of K		Euler
Multiple overlap stringencies		
Approaches to graph reduction		
Filter overlaps	CABOG	
Greedy contig extension		
Collapse simple paths	CABOG, Newbler	Euler, Velvet, SOAP
Erosion of spurs	CABOG, Edena	Euler, Velvet, AllPaths, SOAP
Transitive overlap reduction	Edena	
Bubble smoothing	Edena	Euler, Velvet, SOAP
Bubble detection		AllPaths
Reads separate tangled paths		Euler, SOAP
Break at low coverage		Velvet, SOAP
Break at high coverage	CABOG	Euler
High coverage indicates repeat	CABOG	Velvet
Special use of long reads	Shorty	Velvet
Graph partitions		
Partition by K-mers		ABySS
Partition by scaffolds		AllPaths
Uses for mate pairs		
Constrain path searches		Euler, Velvet, AllPaths
Guide path selection		Euler, Allpaths
Detect misassembled contigs	CABOG, Shorty	
Merge contigs or fill gaps	CABOG, Shorty	Velvet, ABySS, SOAP
Transitive link reduction	CABOG	SOAP
Detect, avoid repeat contigs	CABOG	Velvet, SOAP
Create scaffolds	CABOG, Shorty	Euler, Velvet, AllPaths, SOAP

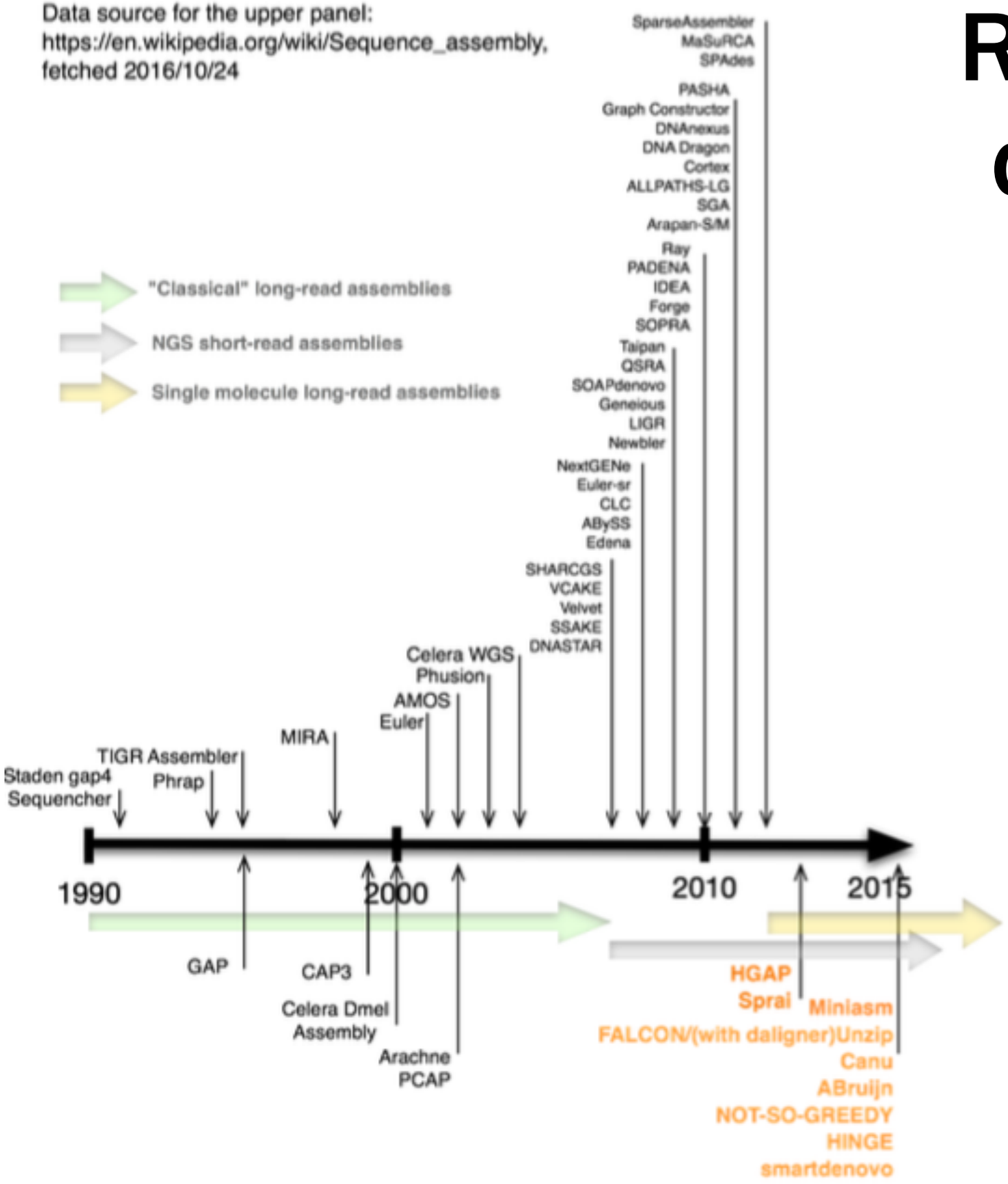
Different approaches to key problems

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data, 95(6), 315–327

Why so many different strategies?

- We don’t know the true sequence - which algorithm is right?
- Simulated data often doesn’t match reality
- Assemblers depend highly on input data
- Speed optimization vs. correctness
- …

Data source for the upper panel:
https://en.wikipedia.org/wiki/Sequence_assembly,
fetched 2016/10/24



Read length determines strategy

- **Classical long read:**
  - Sanger ~600 bp
  - Overlap - Layout - Consensus
- **NGS:**
  - Illumina ~ 150 bp
  - De-Bruijn graph
- **Single molecule long read:**
  - Pacbio/Nanopore/Minion ~ 10kb
  - Overlap - Layout - Consensus

Assembler paradigmes

Overlap - Layout - Consensus

CTAGGCCCTCAATTTTT

GGCGTCTATATCT

CTCTAGGCCCTCAATTTTT

TCTATATCTCGGCTCTAGG

GGCTCTAGGCCCTCATTTTTT

CTCGGCTCTAGCCCCTCATTTT

TATCTCGACTCTAGGCCCTCA

GGCGTCGATATCT

TATCTCGACTCTAGGCC

GGCGTCTATATCTCG

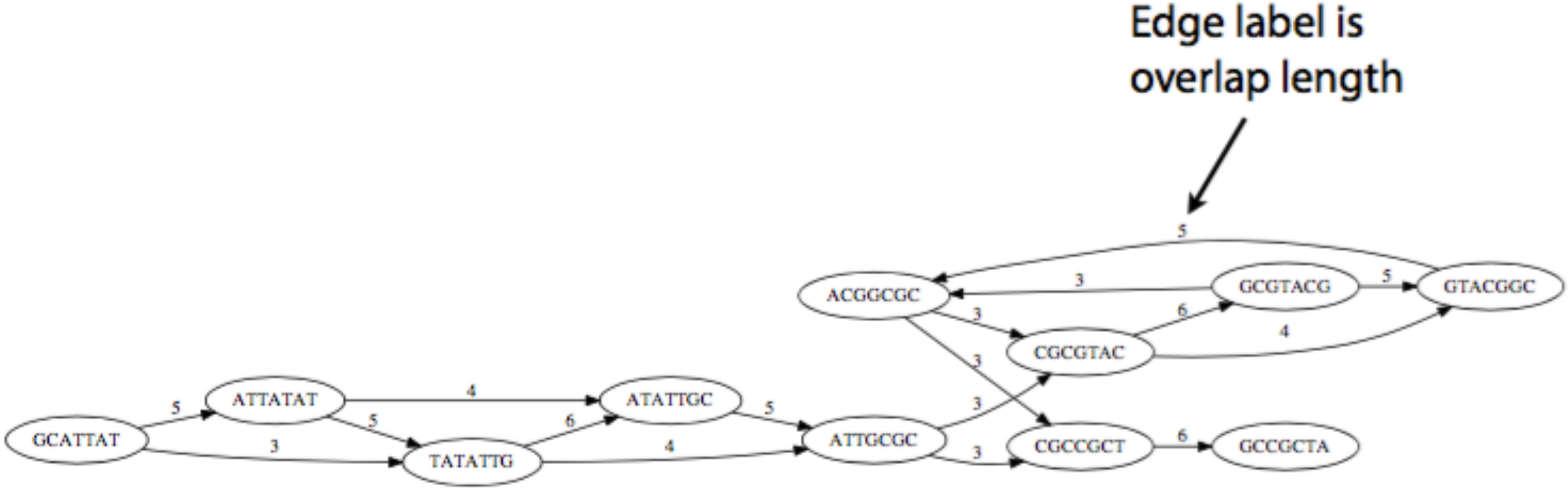
http://www.langmead-lab.org/teaching-materials/

Overlap - Layout - Consensus

Assembler paradigmes

Overlap - Layout - Consensus

Assembler paradigmes



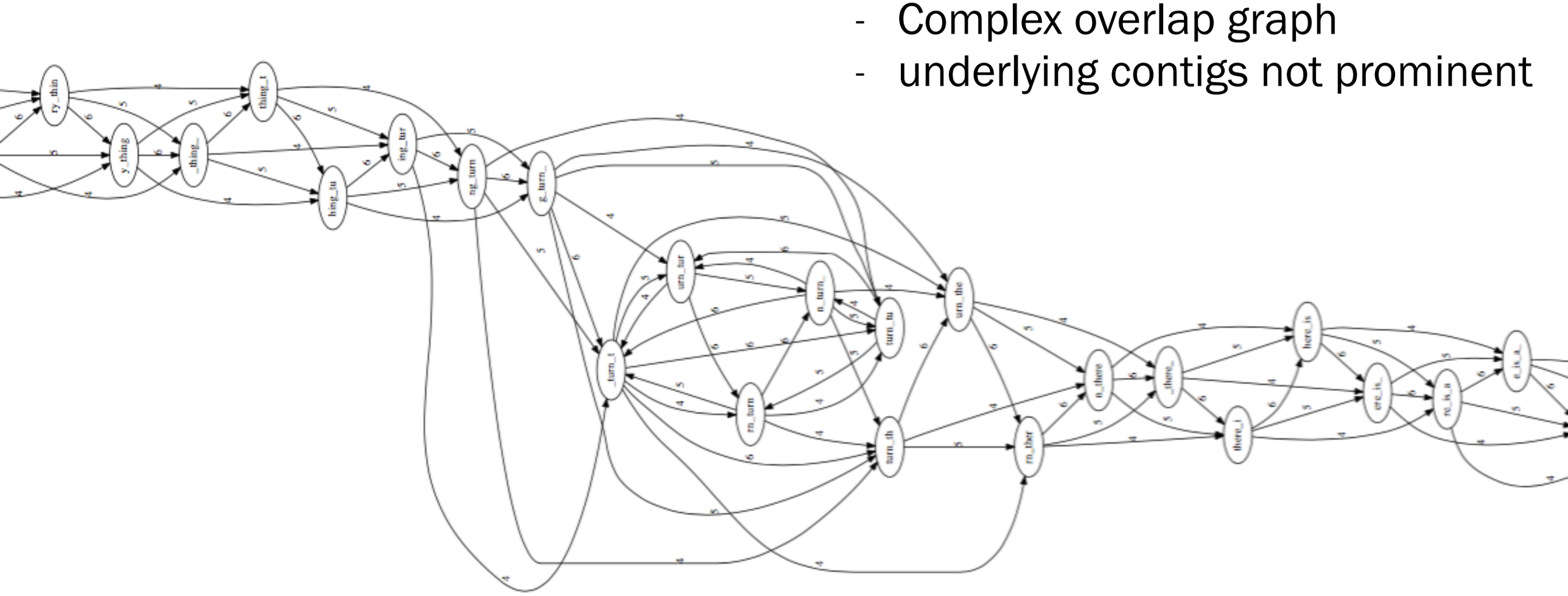
Original string: GCATTATATATTGCGCGTACGGCGCCGCTACA

Overlap - **Layout** - Consensus

Assembler paradigmes

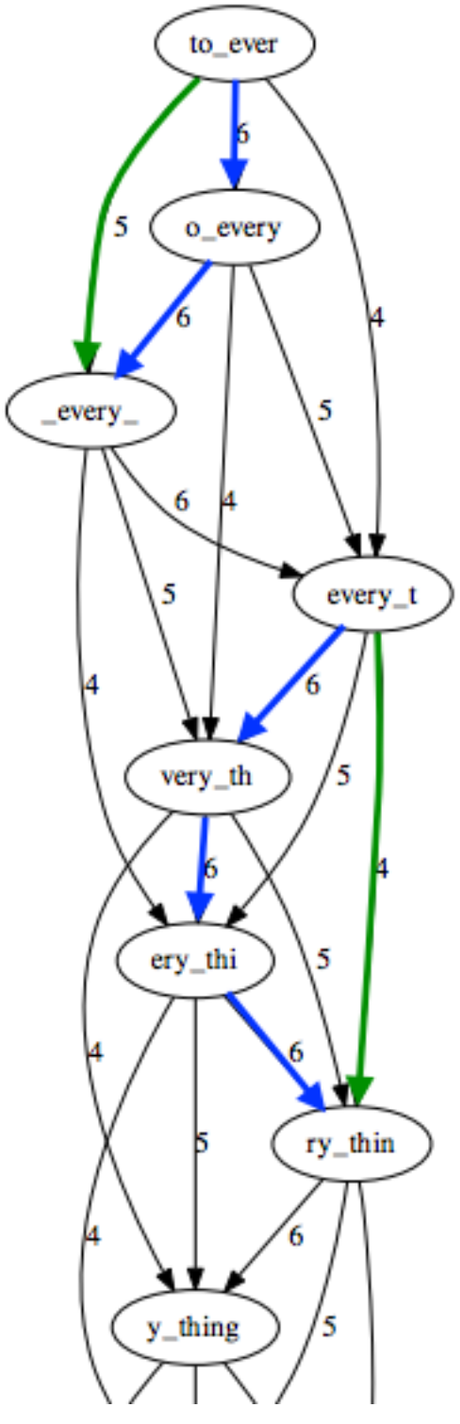
to_every_thing_turn_turn_turn_there_is_a_season

- Complex overlap graph
- underlying contigs not prominent



Overlap - **Layout** - Consensus

Assembler paradigmes

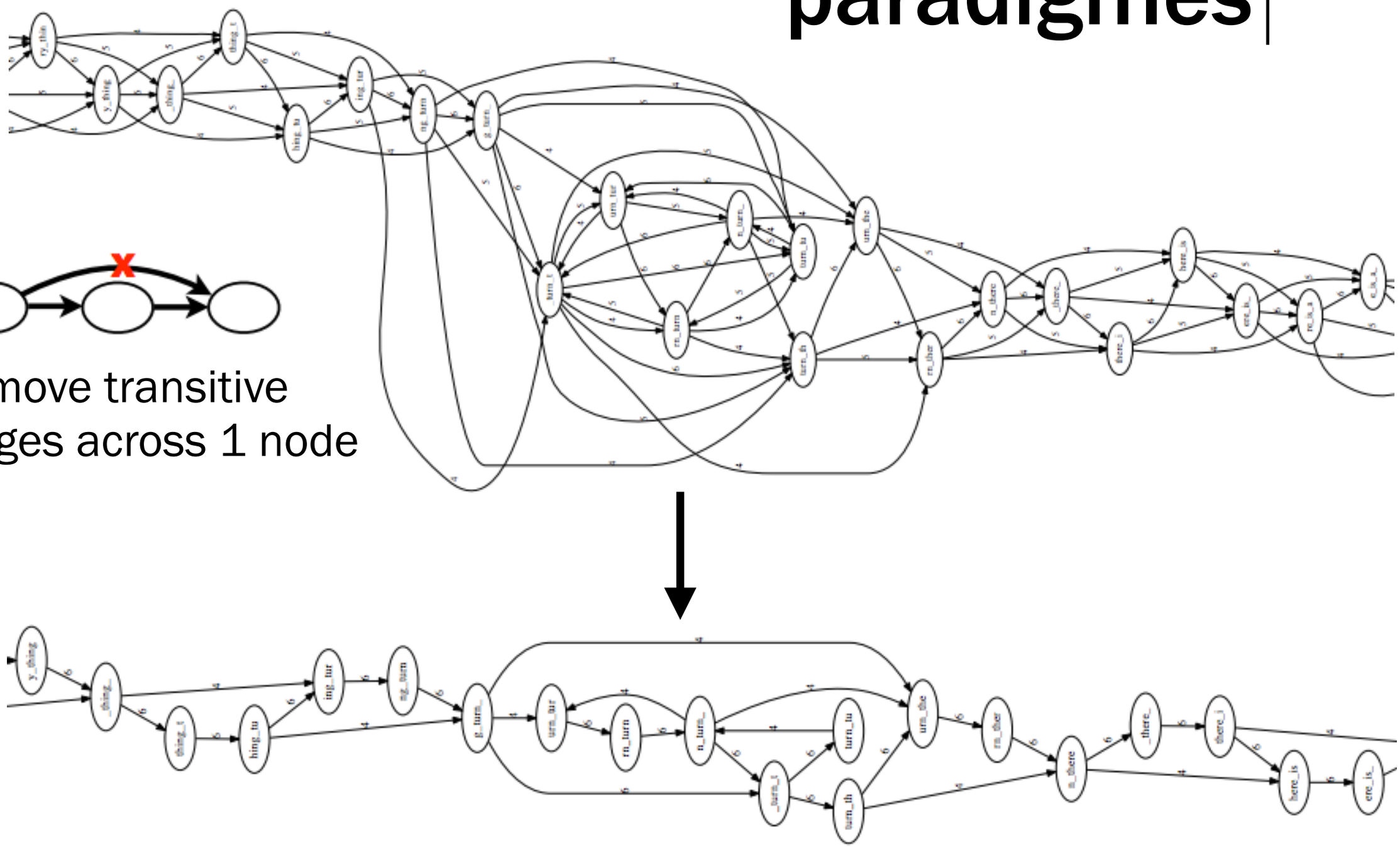


- Some edges are implied by others
- green edge can be inferred by blue

Overlap - **Layout** - Consensus

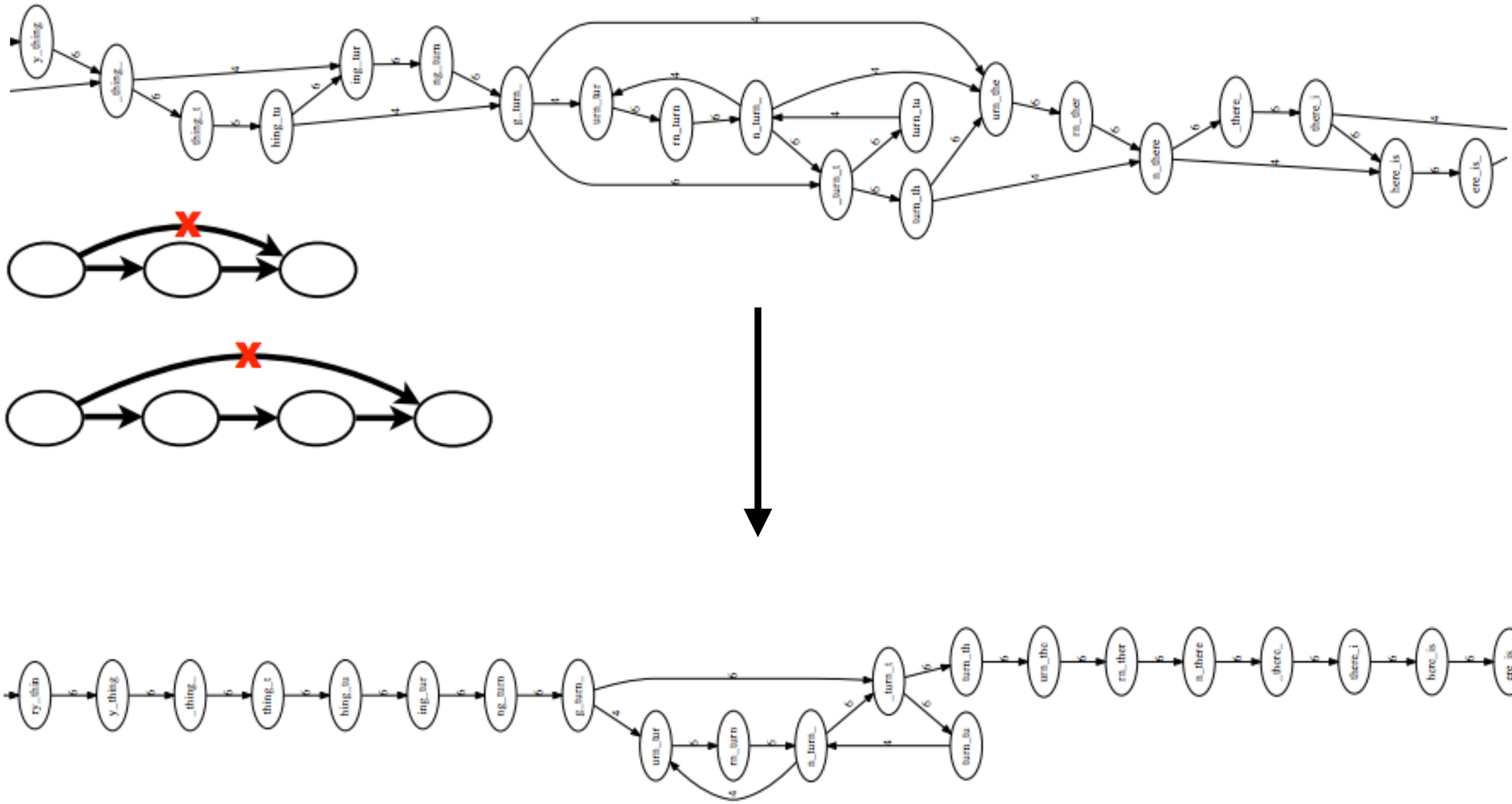
Assembler paradigmes

- remove transitive edges across 1 node



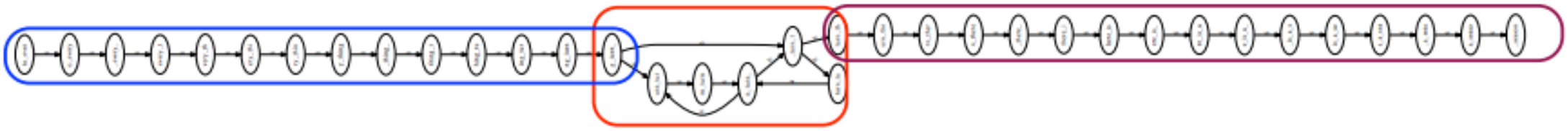
Overlap - **Layout** - Consensus

Assembler paradigmes



Overlap - **Layout** - Consensus

Assembler paradigmes



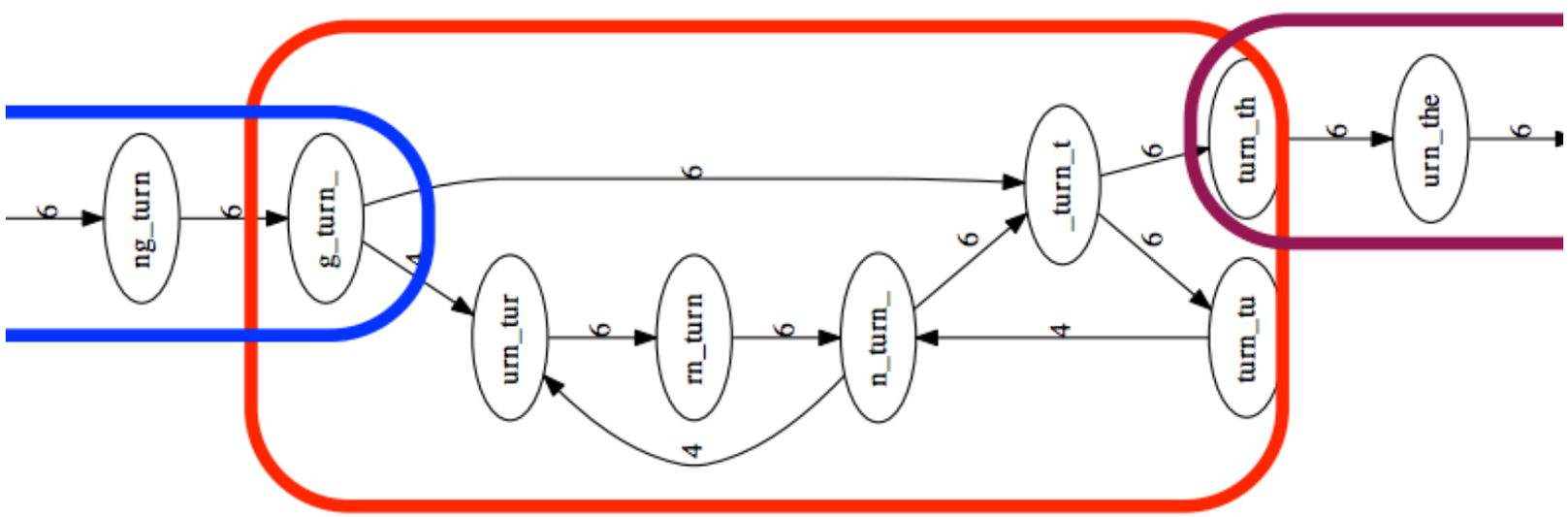
Contig 1

to_every_thing_turn_

Contig 2

turn_there_is_a_season

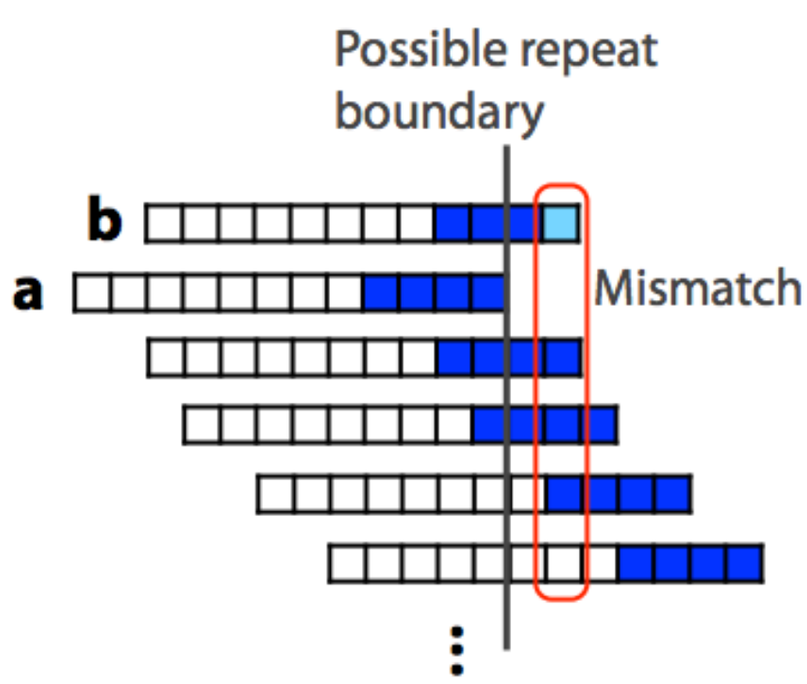
Unresolvable repeat

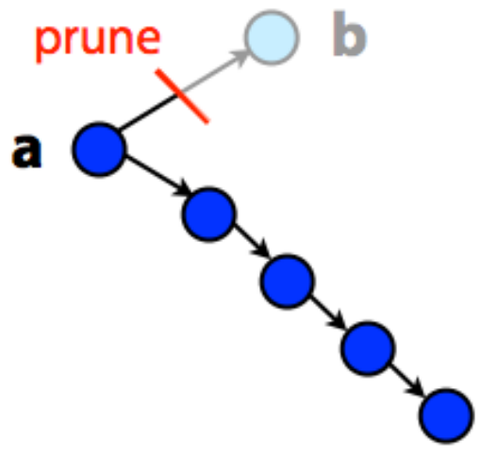


Overlap - **Layout** - Consensus

Assembler paradigmes

Must handle subgraphs that are spurious, e.g. because of sequencing error

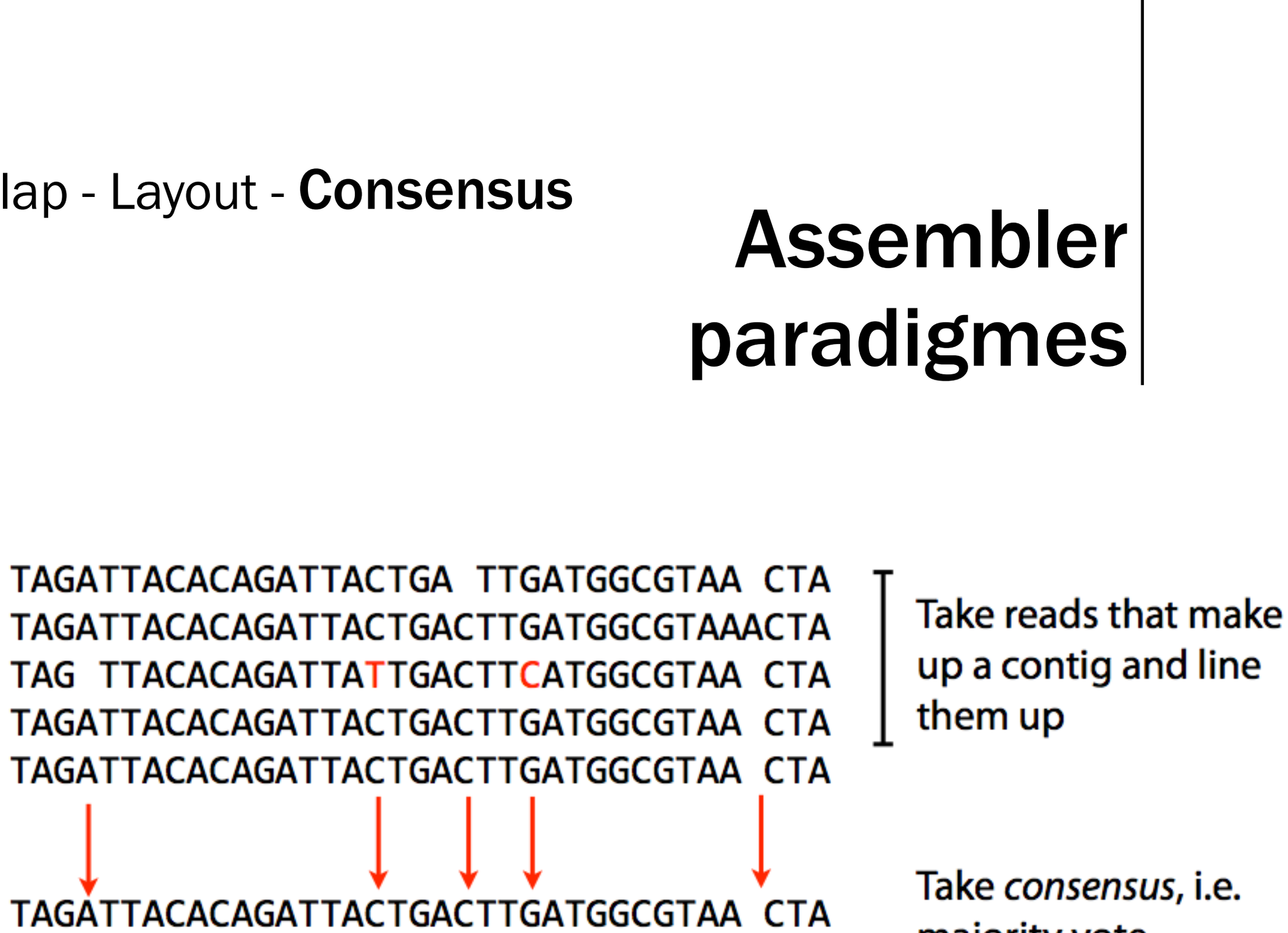




Mismatch could be due to sequencing error or repeat. Since the path through **b** ends abruptly we might conclude it's an error and prune **b**.

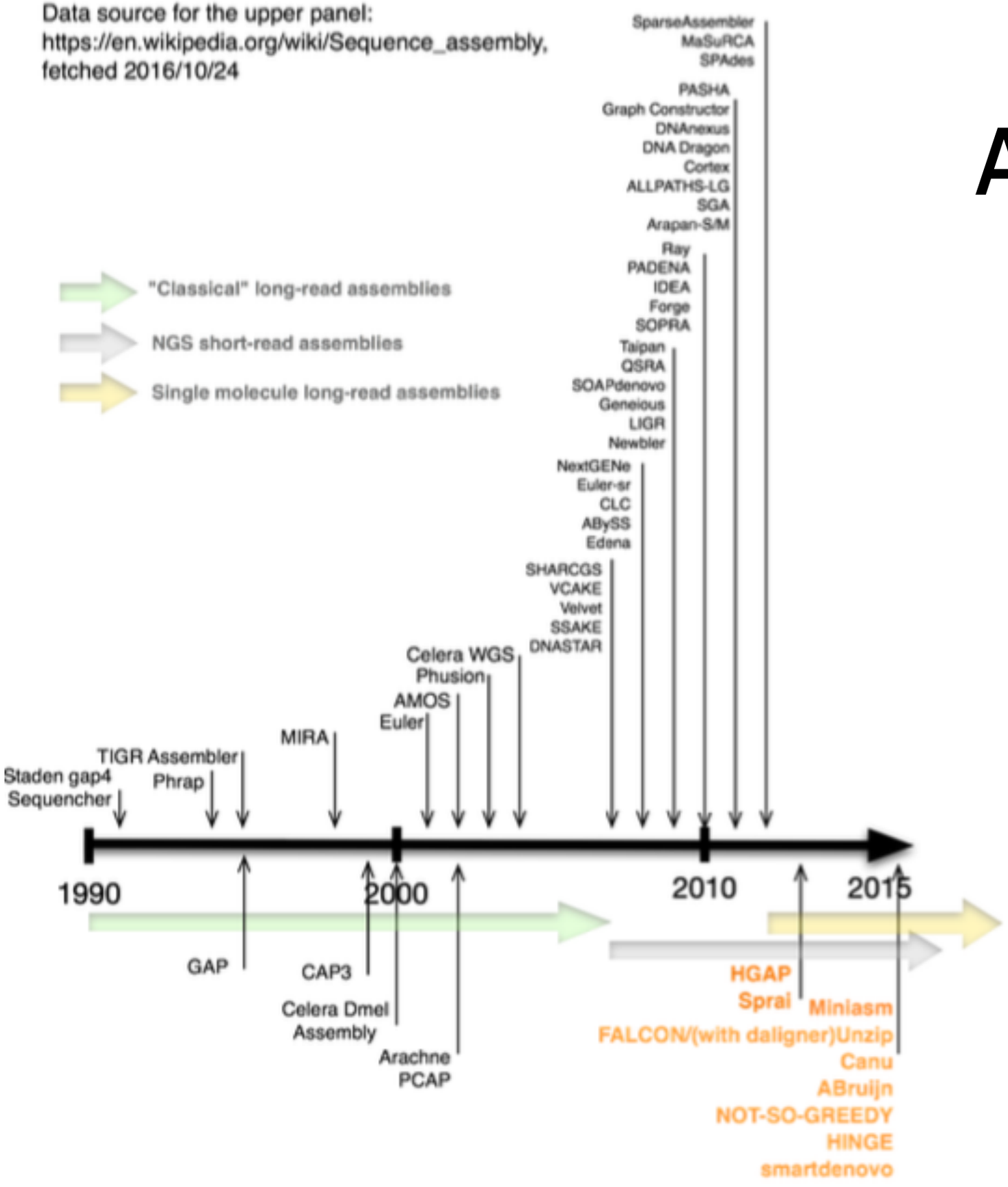
Assembler paradigmes

Overlap - Layout - **Consensus**



Complications: (a) sequencing error, (b) ploidy

Lots of Assemblers developed

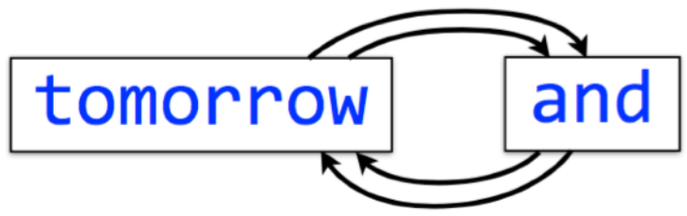


- **Classical long read:**
  - Sanger ~600 bp
  - Overlap - Layout - Consensus
- **NGS:**
  - Illumina ~ 150 bp
  - De-Bruijn graph
- **Single molecule long read:**
  - Pacbio/Nanopore/Minion ~ 10kb
  - Overlap - Layout - Consensus

Assembler paradigmes

How to represent complex string as ordered fragments?

"tomorrow and tomorrow and tomorrow"

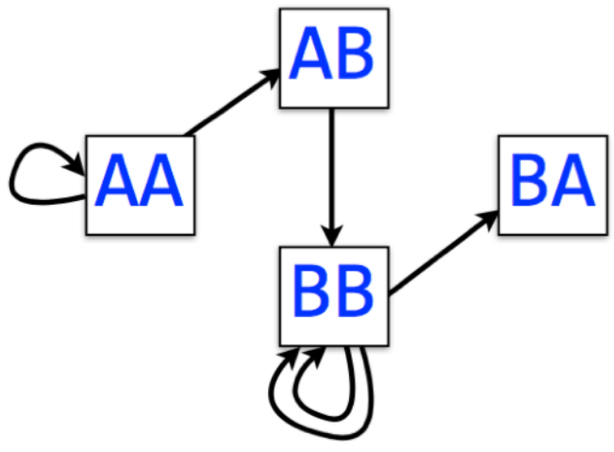


- graph of text fragments, e.g. words
- edge represents order of words in text
- Multigraph: multiple edges between nodes

Assembler paradigmes

De-Bruijn Graph

- Method to represent string as graph
- edge = k-mer
- node = k-1 mer



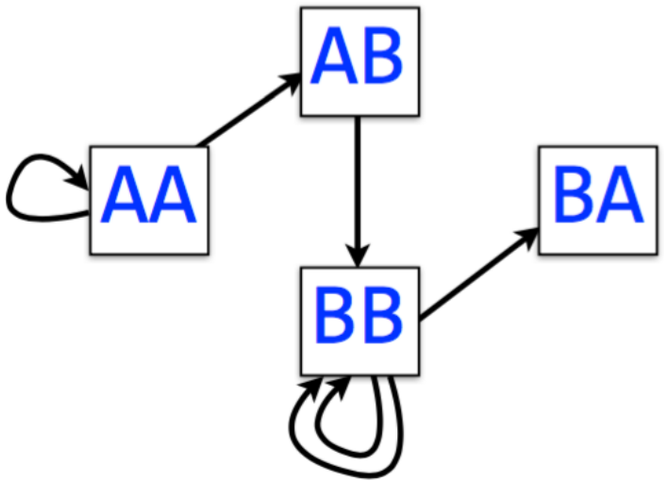
de Bruijn (1946). A Combinatorial Problem. Koninklijke Nederlandse Akademie v. Wetenschappen. 49: 758–764

Assembler paradigmes

- Node is balanced if indegree equals outdegree
- Node is semi-balanced if indegree differs from outdegree by 1

  Graph is connected if each node can be reached by some other node
- Eulerian walk visits each edge exactly once

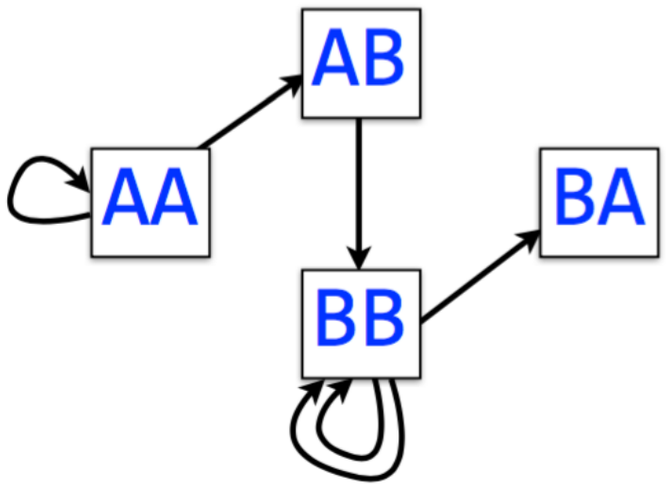
  Not all graphs have Eulerian walks. Graphs that do are Eulerian.
- A directed, connected graph is Eulerian if and only if it has at most 2 semi-balanced nodes and all other nodes are balanced



Jones and Pevzner. An Introduction to Bioinformatics Algorithms

Assembler paradigmes

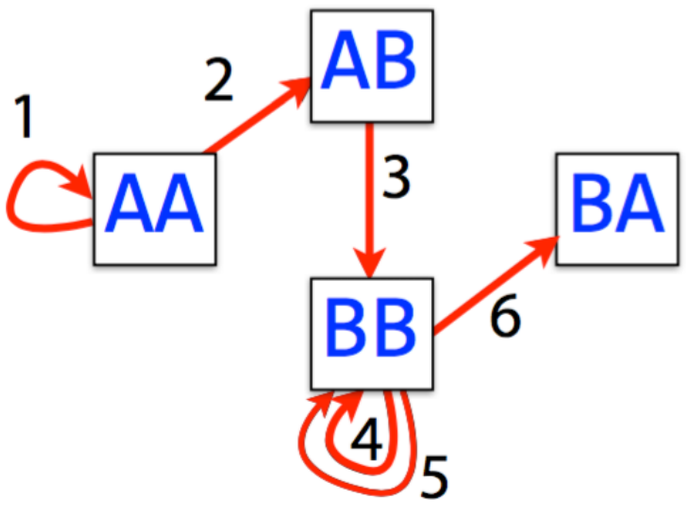
De-Bruijn Graph



- De-Bruijn graphs have Eulerian path
- walking along each edge once (in right direction) yields the original sequence

Assembler paradigmes

De-Bruijn Graph

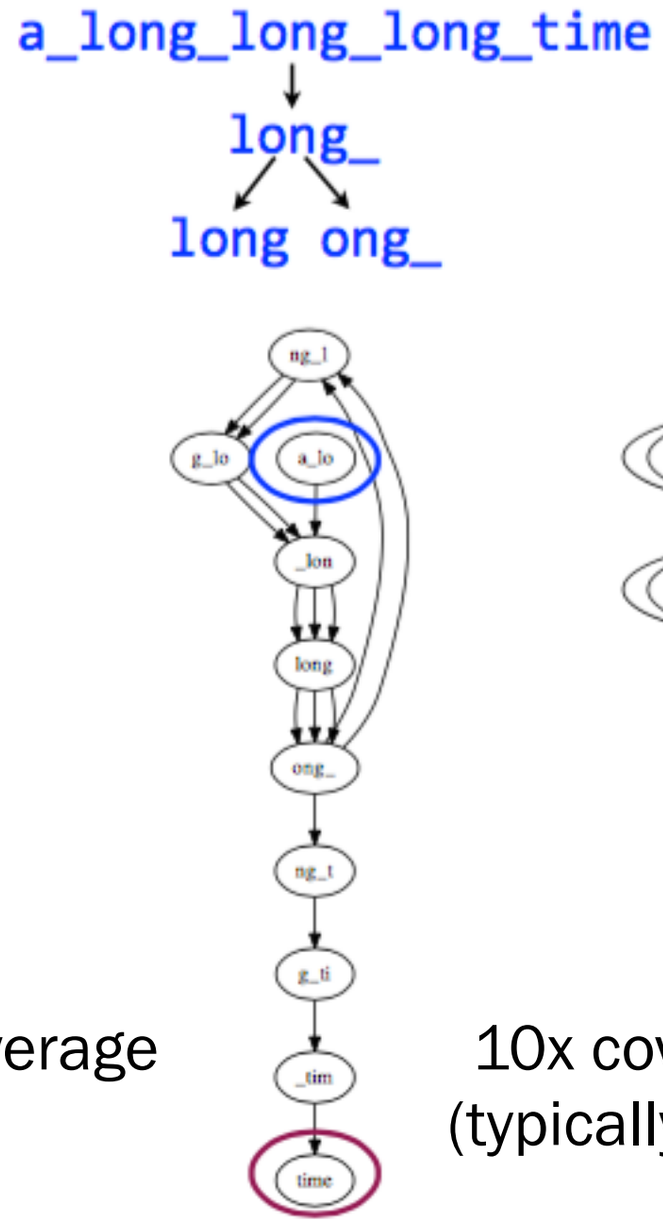


AAABBBBA

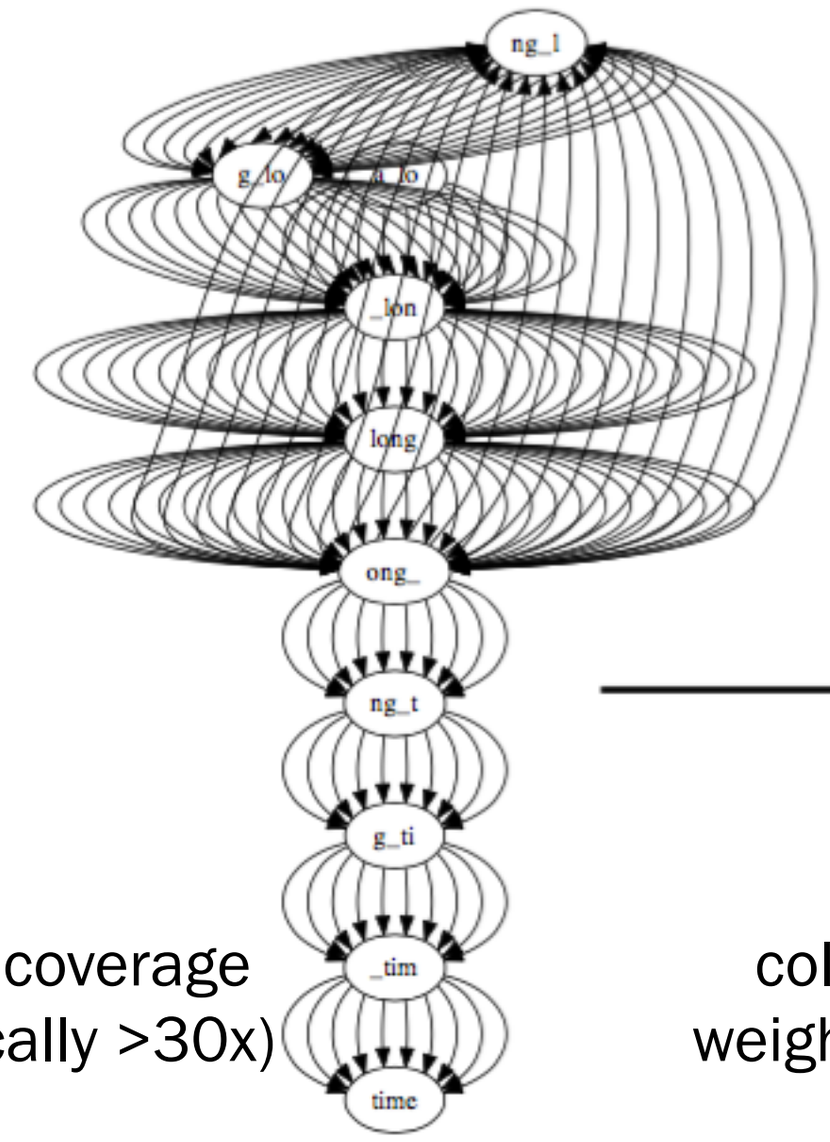
- De-Bruijn graphs have Eulerian path
- walking along each edge once (in right direction) yields the original sequence
- not possible with all graphs

Assembler paradigmes

De-Bruijn Graph

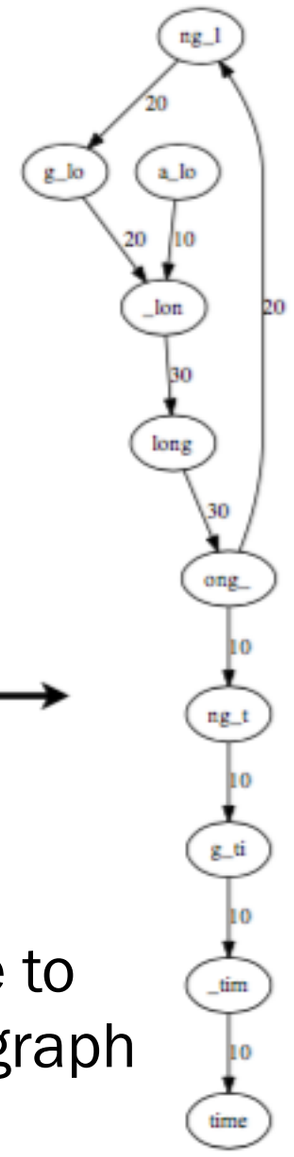


1x coverage



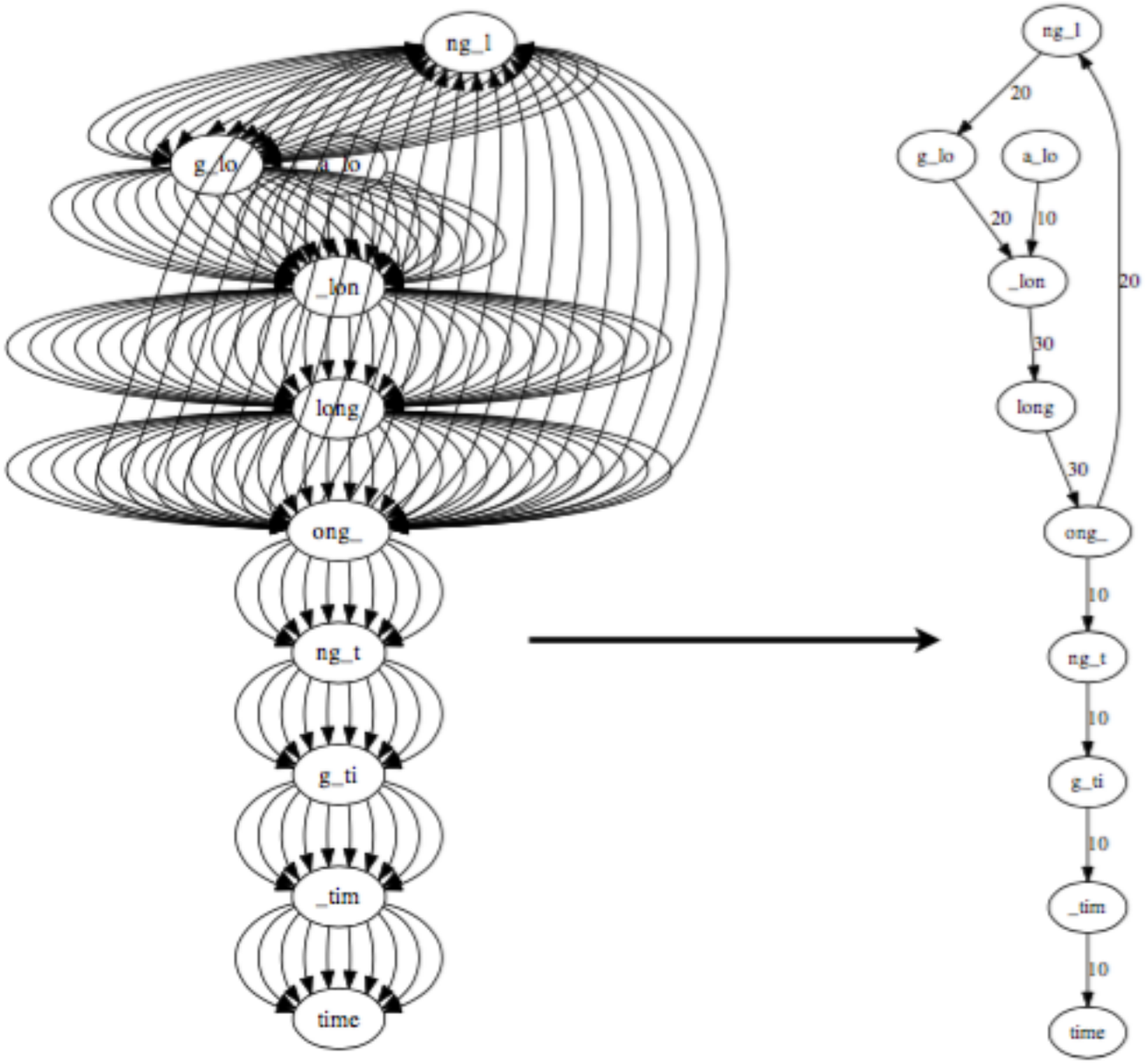
10x coverage
(typically >30x)

collapse to
weighted graph

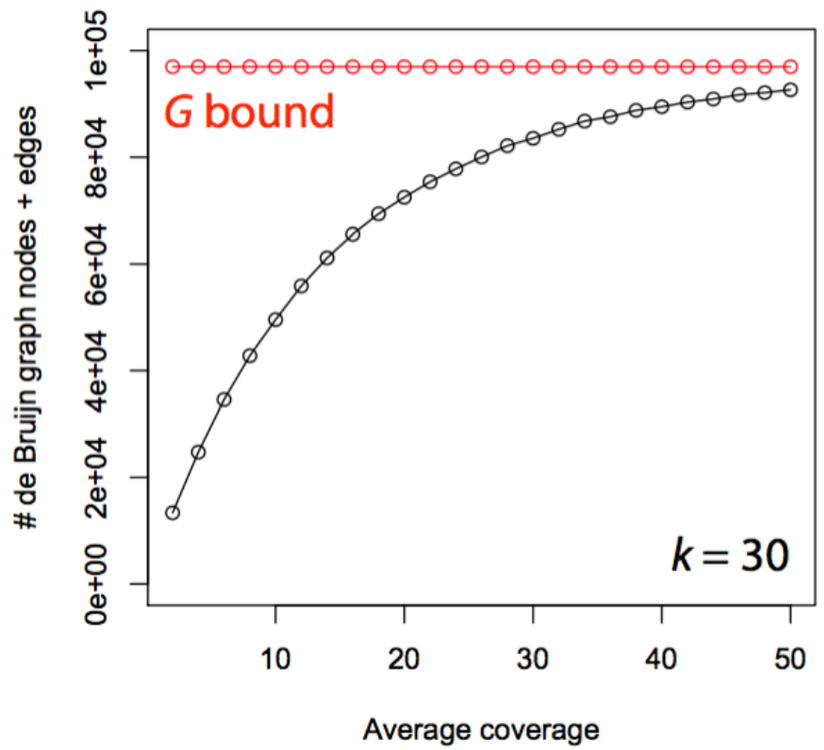


Assembler paradigmes

De-Bruijn Graph



Number of k-mer's depends on genome, not number of reads -> good for high coverage datasets (advantage over OLC)



Assembler paradigmes

De-Bruijn Graph

What about Errors?

a_long_long_long_time

long_

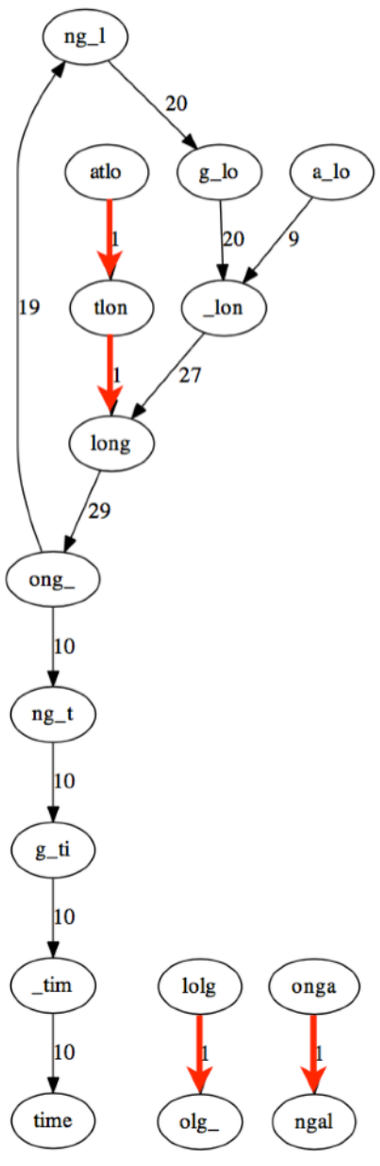
long ong_

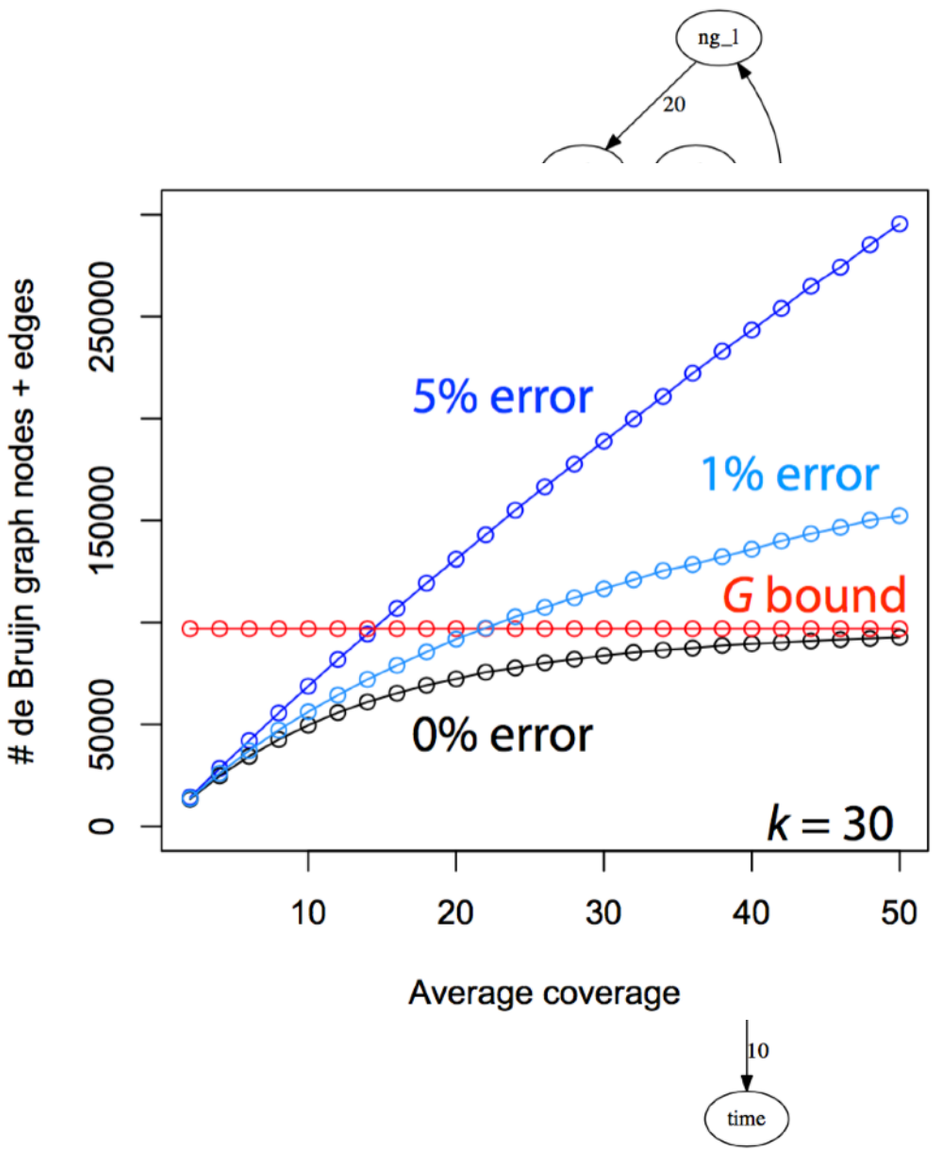
atlon

tlong

lolg_

ongal

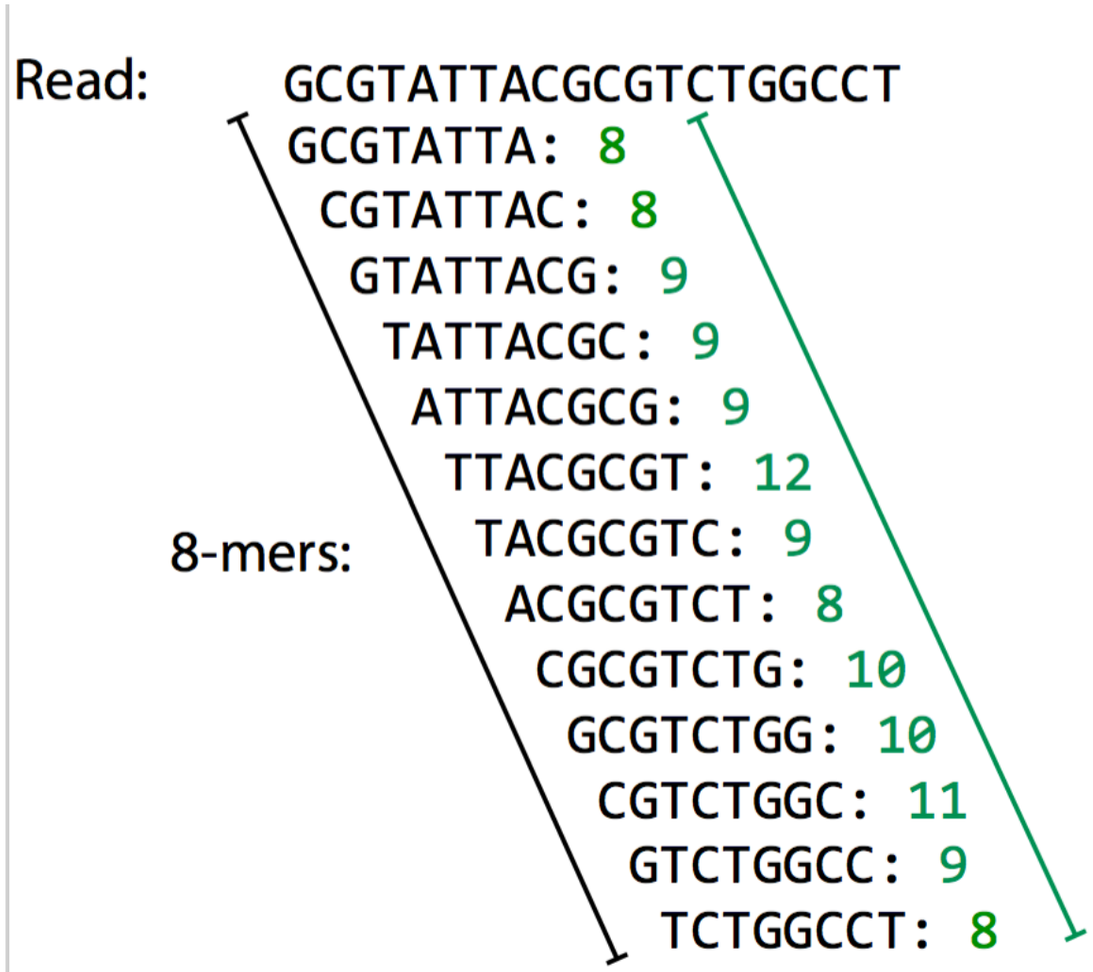


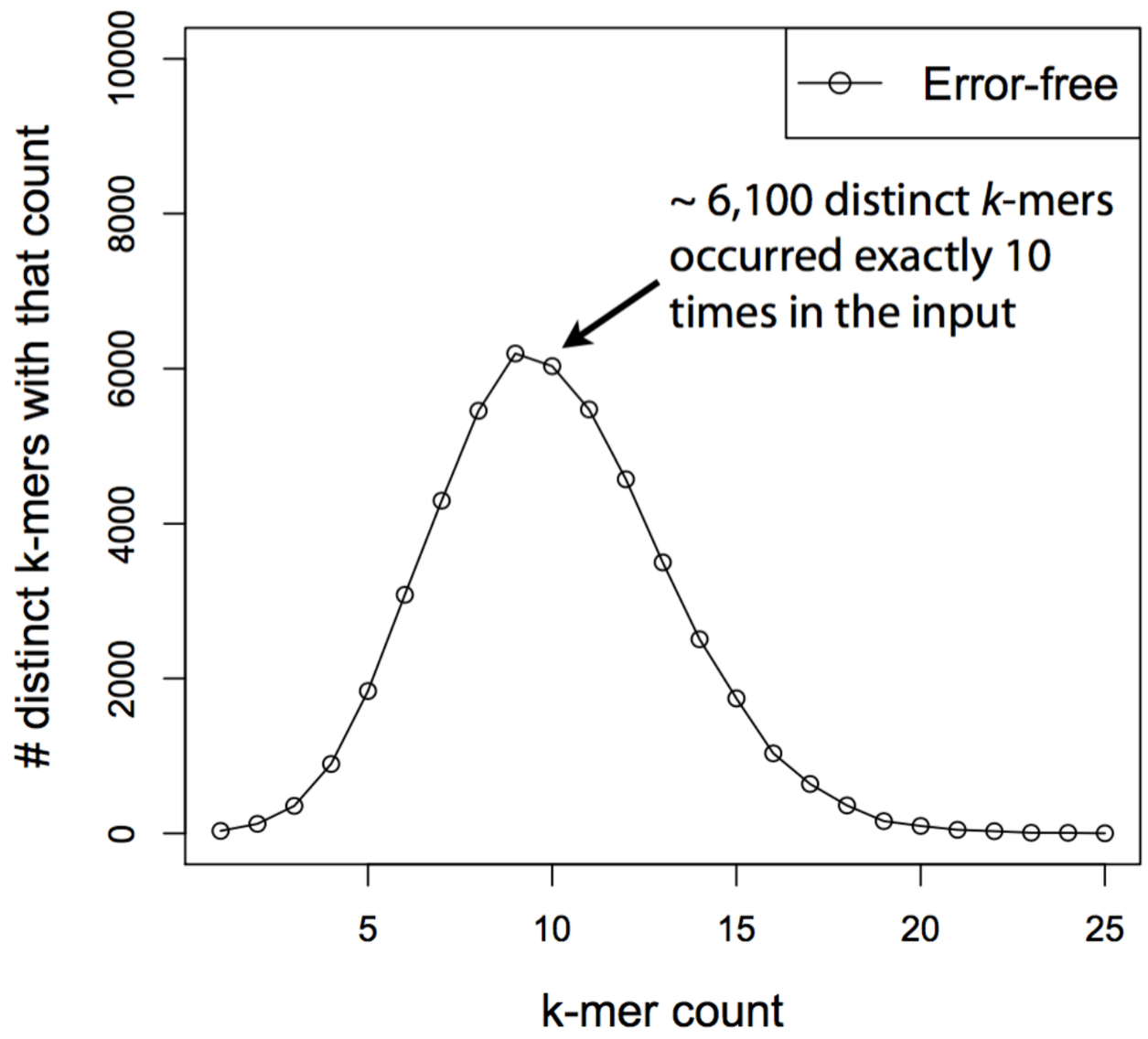


Assembler paradigmes

De-Bruijn Graph

How many sequencing errors?



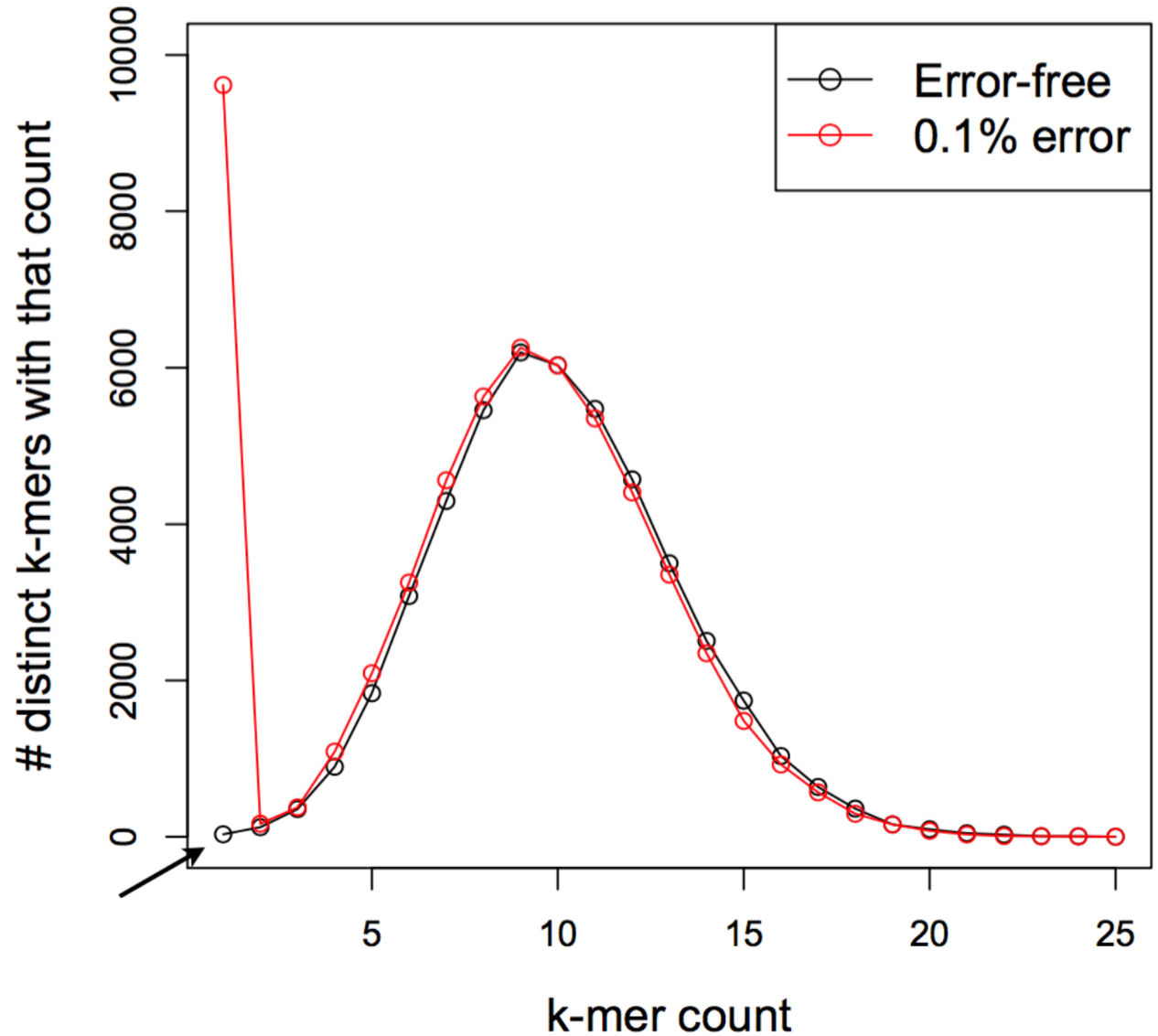


Assembler paradigmes

De-Bruijn Graph

How many sequencing errors?

```
GCGTATTACACGTCTGGCCT
GCGTATTA: 8
 CGTATTAC: 8
  GTATTACA: 1
   TATTACAC: 1
    ATTACACG: 1
     TTACACGT: 1
      TACACGTC: 1
       ACACGTCT: 2
        CACGTCTG: 1
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```



An assembly pipeline - SPAdes

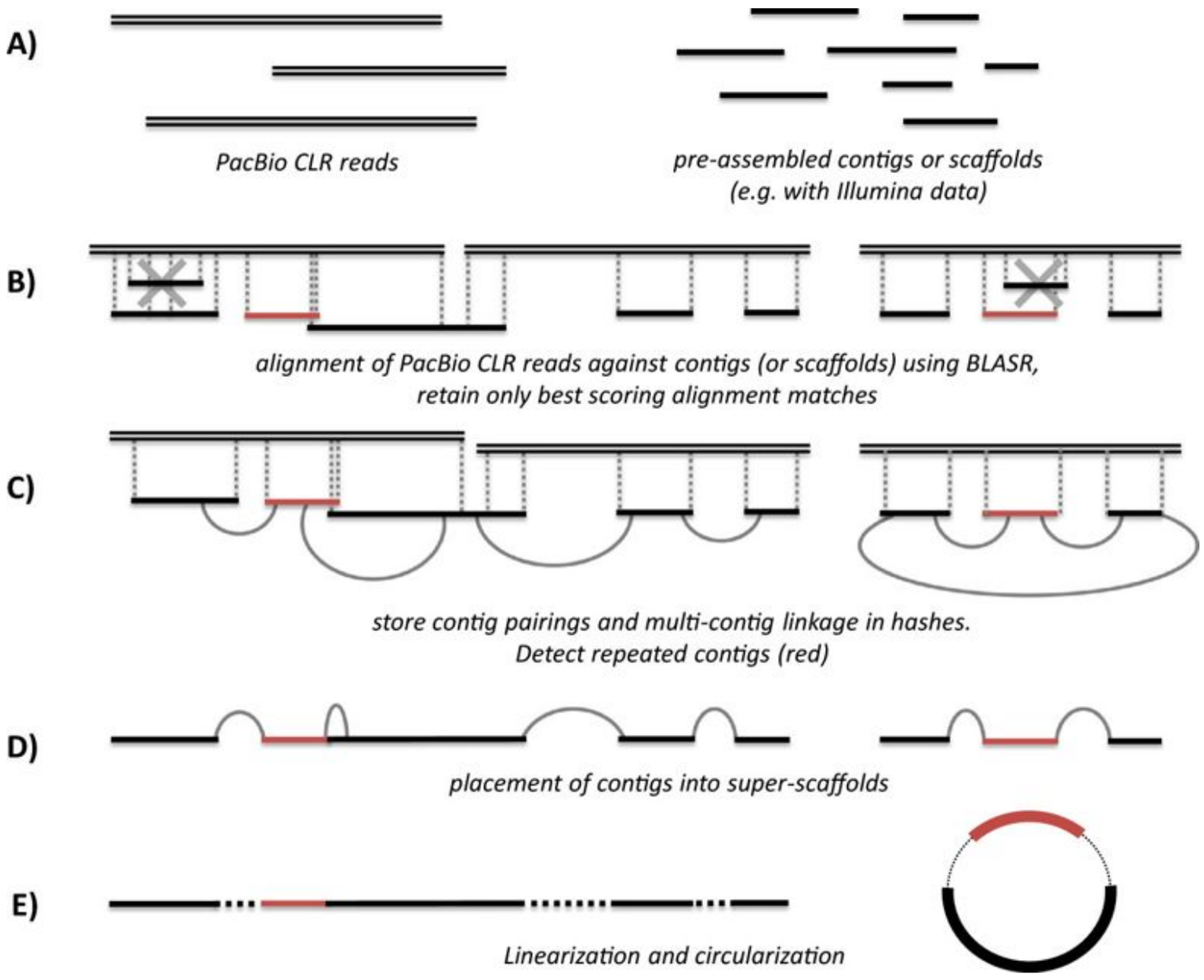
De-Bruijn graph based

1. Read error correction based on k-mer frequencies with BayesHammer
2. De Bruijn graph assembly at multiple k-mer sizes
3. Merging different k-mer assemblies (good for varying coverage)
4. combine Contigs into scaffolds using paired end/mate pair reads
5. Repeat resolution from paired end/mate pair data using rectangle graphs
6. Contig error correction based on aligning the original reads with BWA back to contigs

SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Bankevich A, ..., Pevzner PA. J Comput Biol. 2012 May;19(5):455-77

SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information Marten Boetzer and Walter Pirovano. BMC Bioinformatics201415:211

Scaffolding



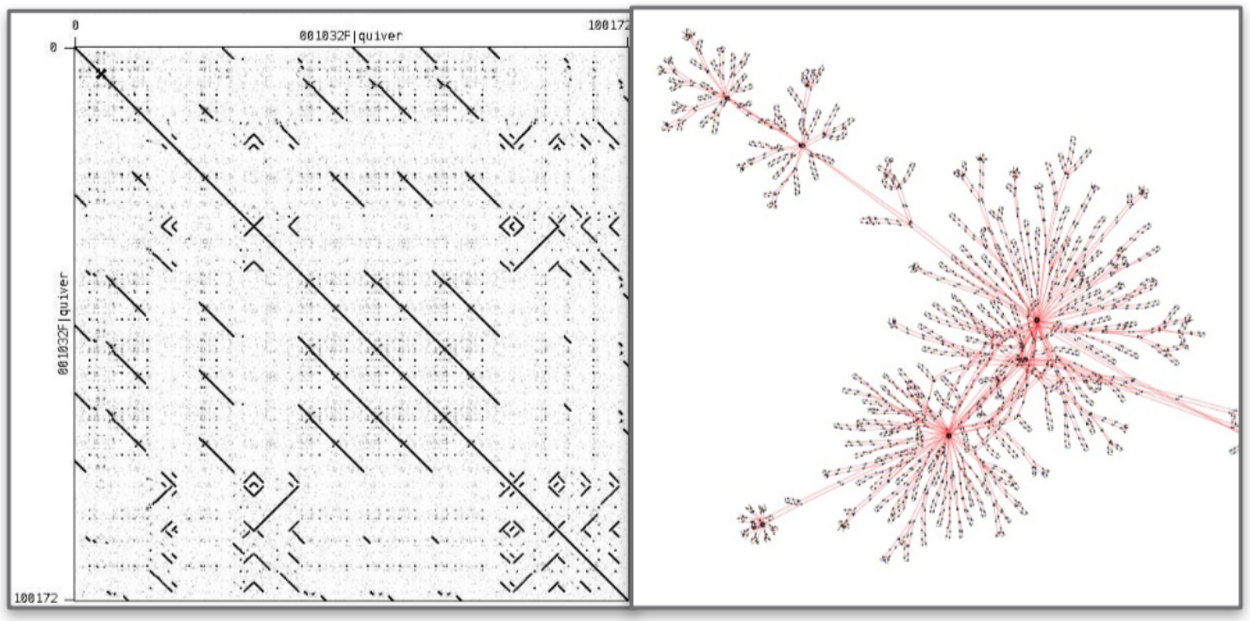
Open Problems: Repeats

- not enough known about true repeat structures across genomes

1984 Peltola, Söderlund, and Ukkonen. Nucleic Acids Res.

Repetitive structures within the DNA sequence can confuse any automatic assembly process which uses only the gel readings. Without computer assistance we have earlier determined the sequence of a cDNA (20) which turns out to be an interesting benchmark for automatically assembling. This cDNA contains three copies of two separate perfectly repeated segments both more than hundred bases in length

2019 - long read assembly - same problem



Open Problems: Repeats

Solution: longer Reads!

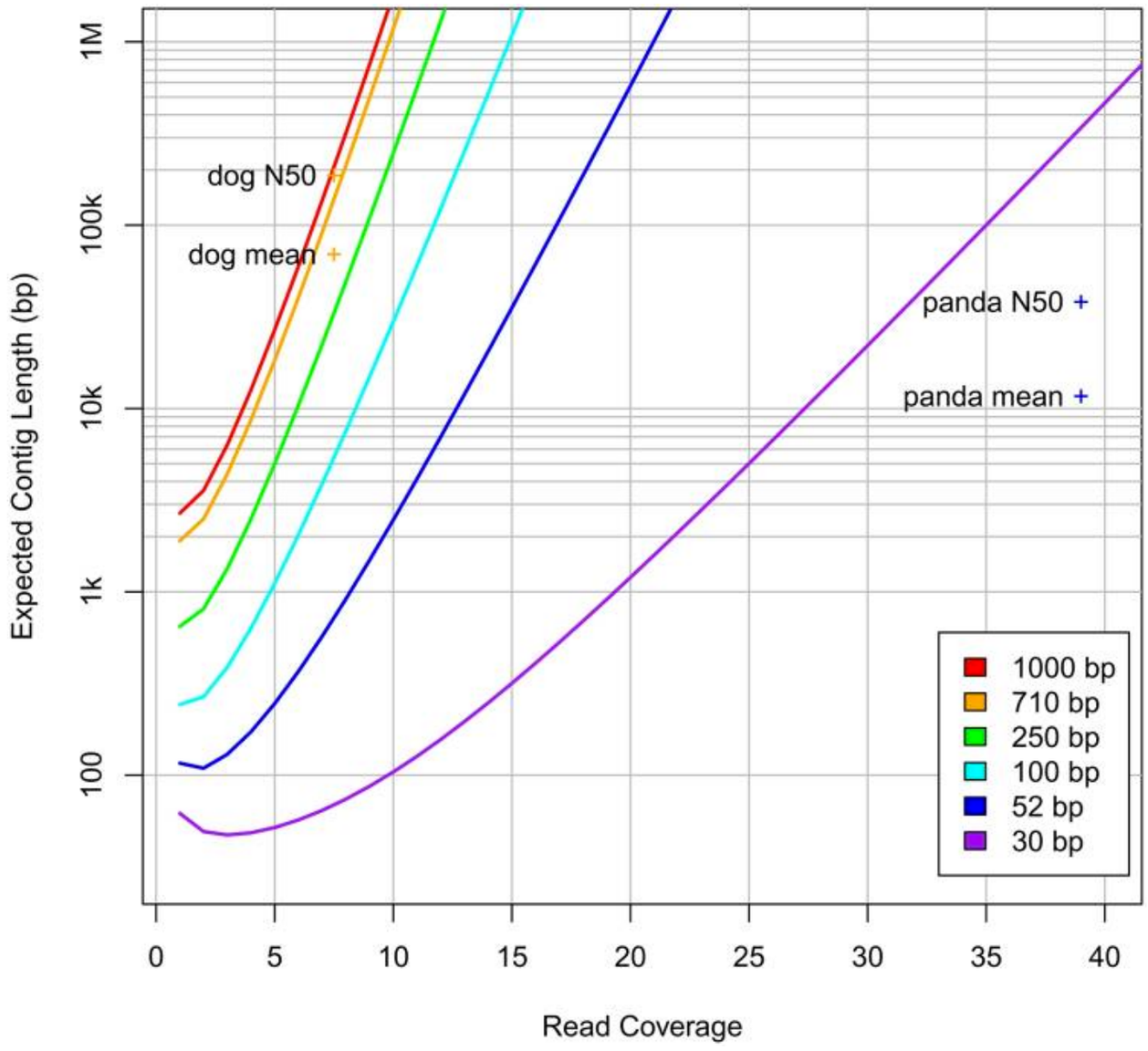
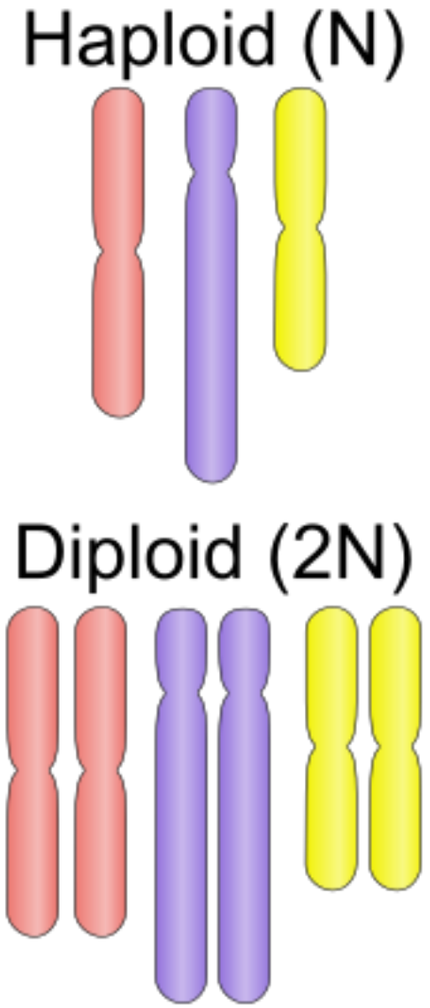


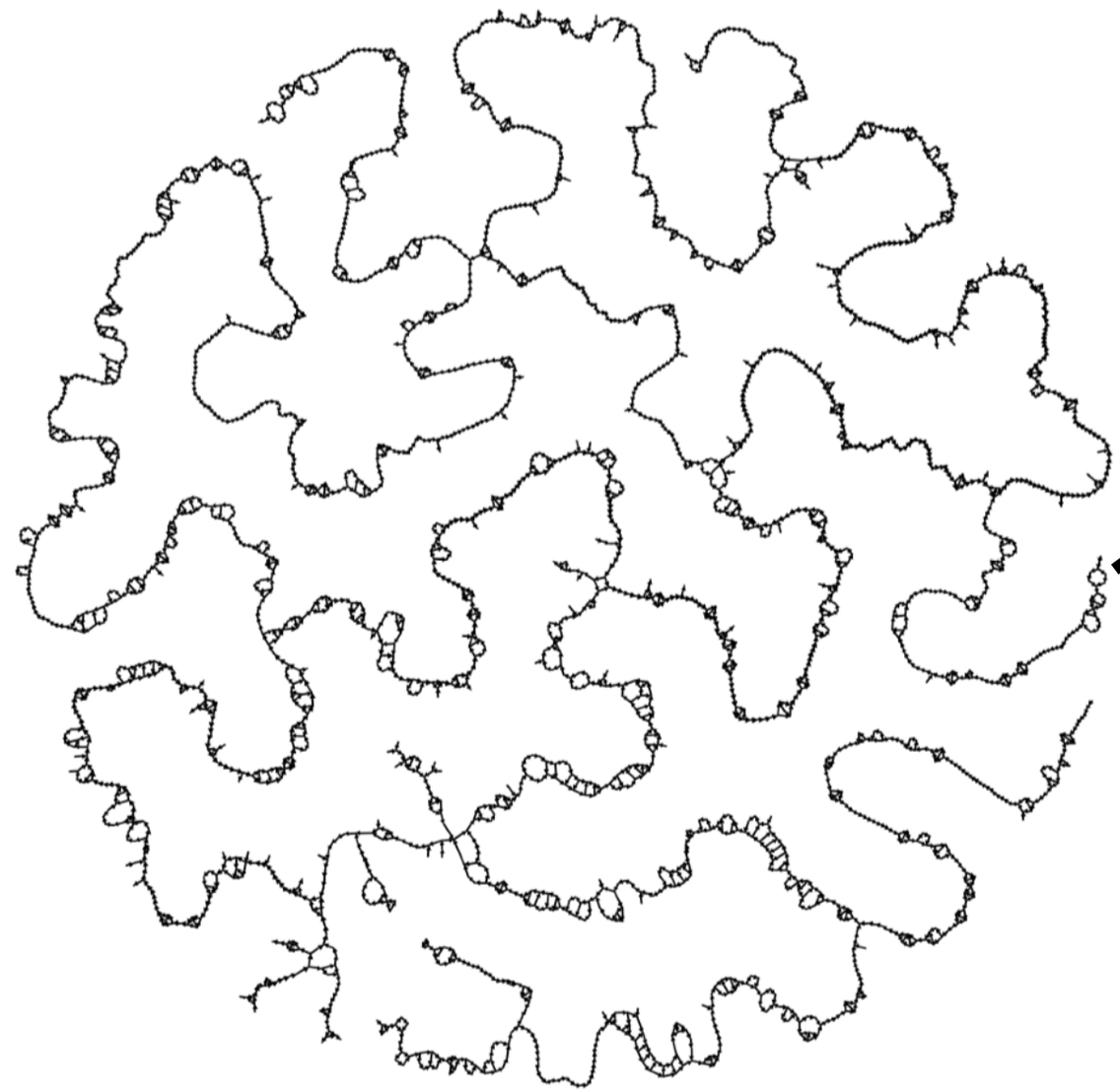
Fig. 3: Expected average of contig length for a range of different read lengths and coverage values. Also shown are the average contig lengths and N50 lengths for the dog genome, assembled with 710-bp reads, and the panda genome, assembled with reads averaging 52 bp in length.

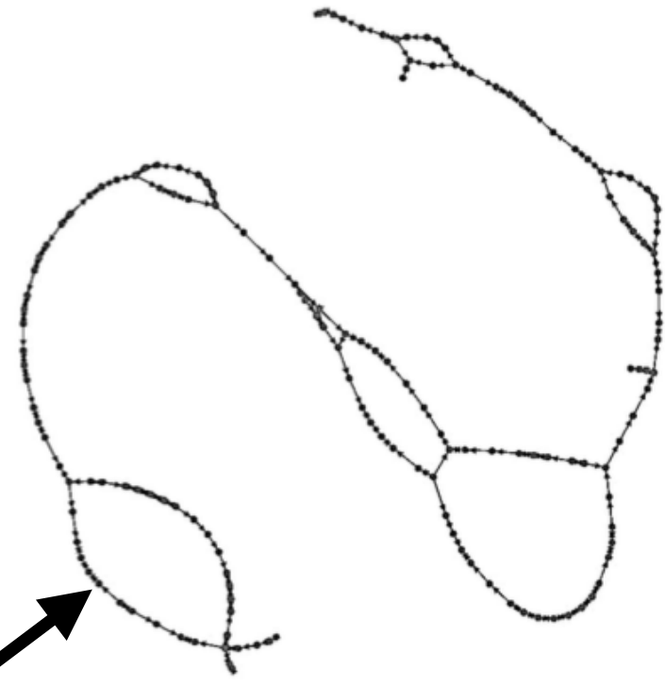
Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg. Genome Res. 2010 Sep; 20(9): 1165–1173.

Open Problems: Ploidy

- not enough known structural difference between haplotypes for species with diploid or polyploid genomes





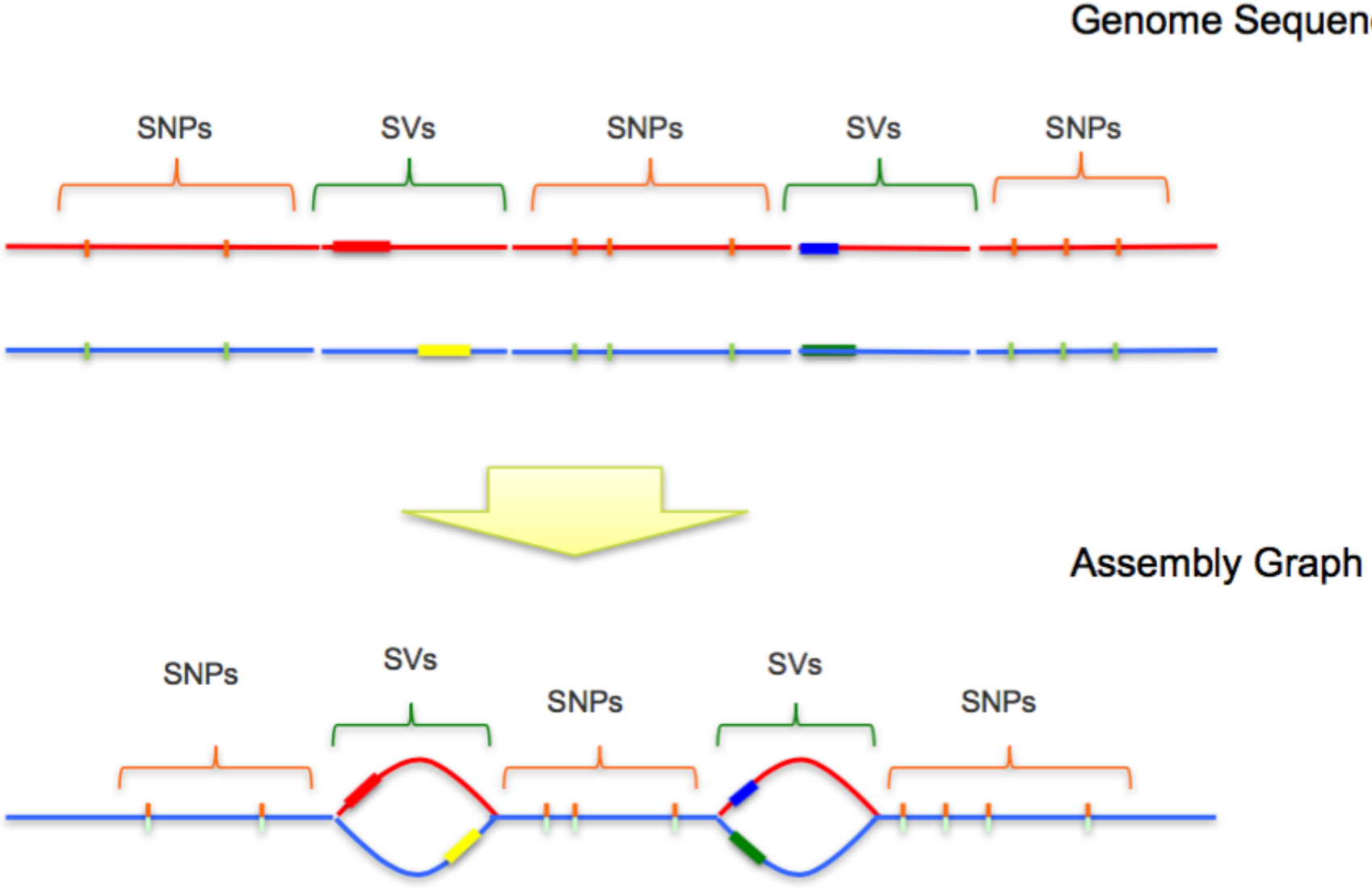


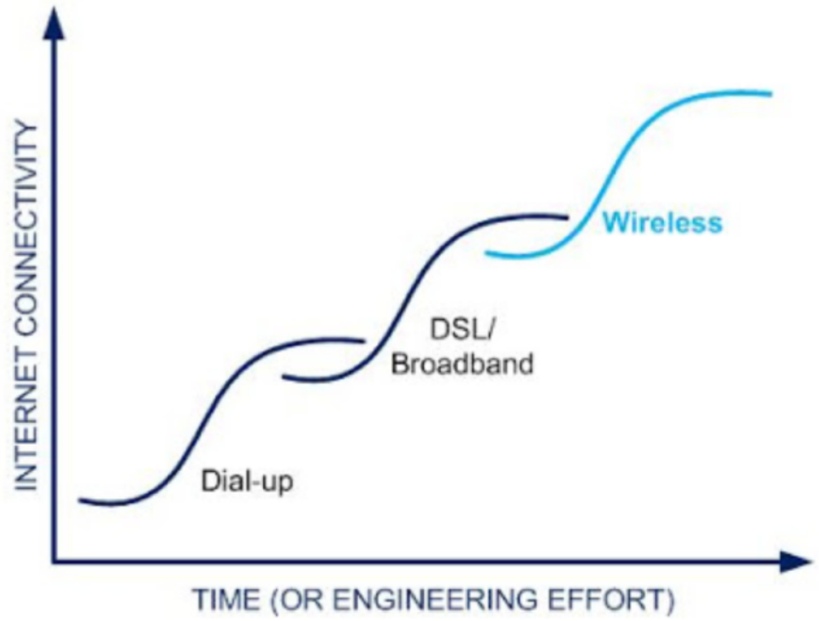
"Best Edge" Overlap Graph of all reads mapped to the region indicate two haplotypes observed.

Open Problems: Ploidy

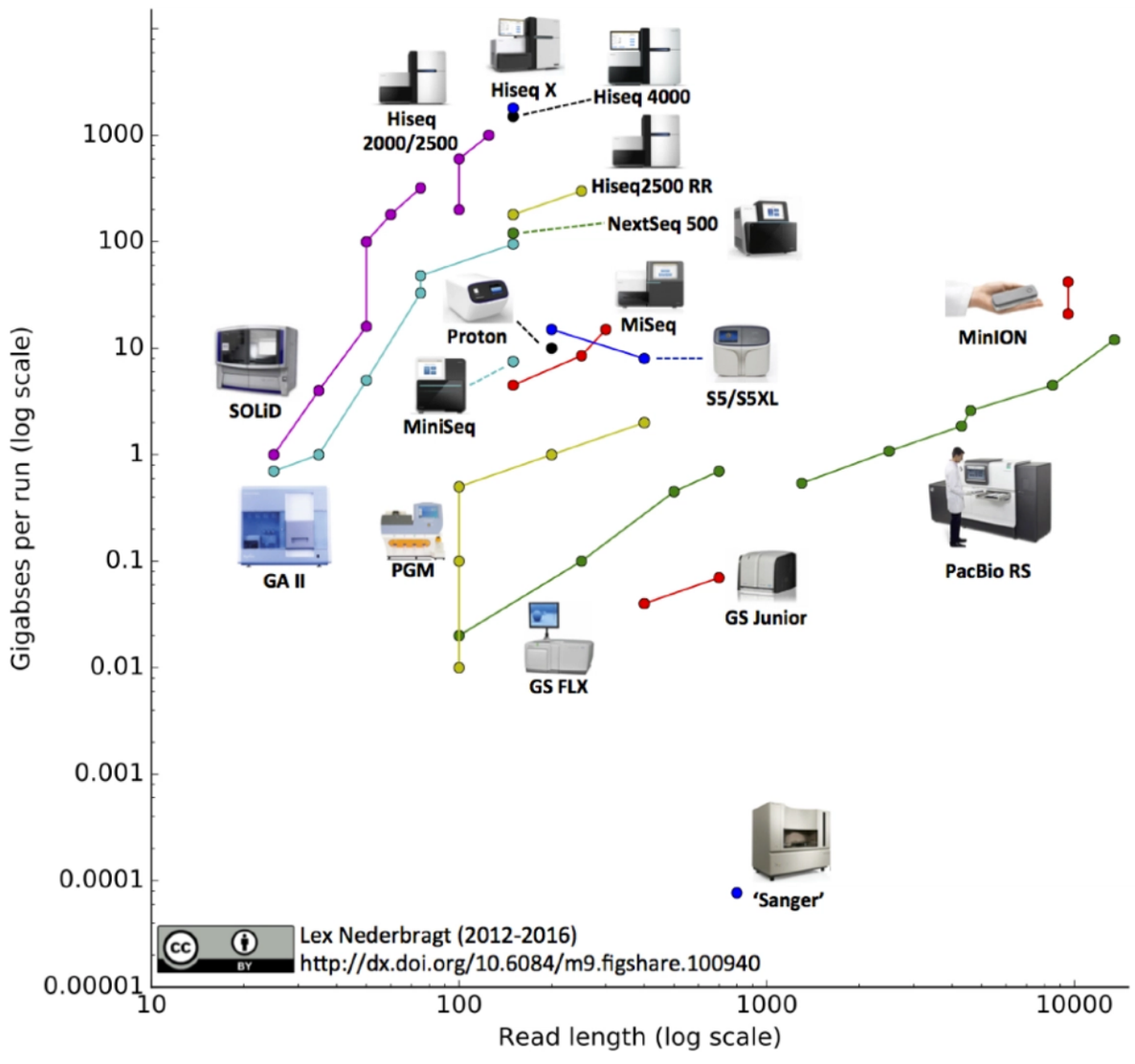
diploid-aware assembly algorithms, e.g. Falcon_unzip

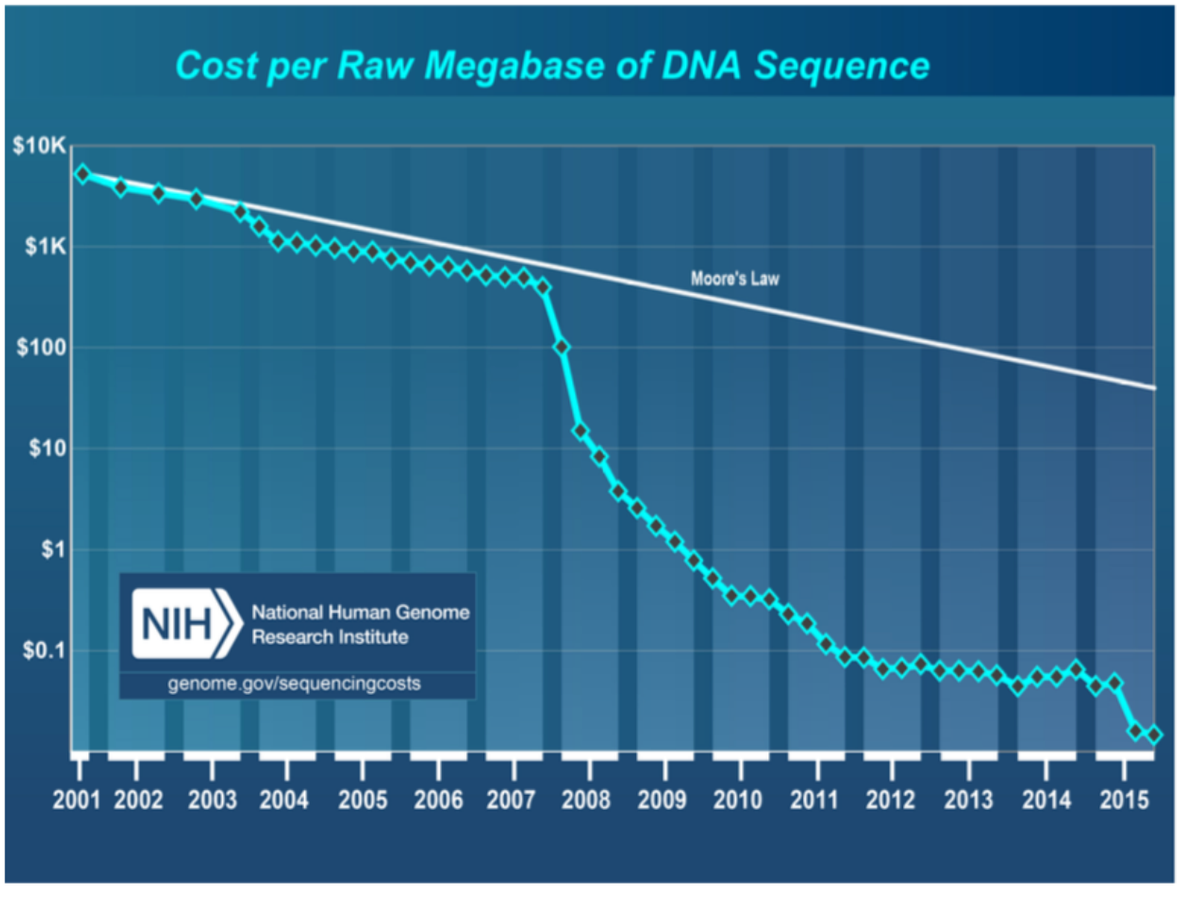
(Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016 Dec;13(12):1050-1054)





Reads are getting more and longer





https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/

Questions?