Velvet: Algorithms for de novo short read assembly using de Bruijn graphs

Daniel R. Zerbino and Ewan Birney

1/18/19

Justin Lee

CMSE890

Genome assembly problems

Next generation sequencing technologies:

- Pyrosequencing (454 Sequencing)
  - 400-500 basepairs(bp)
- Synthesis sequencing (Illumina, Solexa)
  - 35-50 bp
- SOLid sequencing
  - 35 bp

Current algorithms such as Atlas, Arachne, Celera, PCAP and Phusion are based on Sanger reads

Short read assemblies are available (SSAKE, SHARCGS, VCAKE)

Margulies et al., 2005
Bentley, 2006
www.appliedbiosystems.com
Havalk el al., 2004
Batzoglou et al., 2002
Mullikin and Ning, 2003

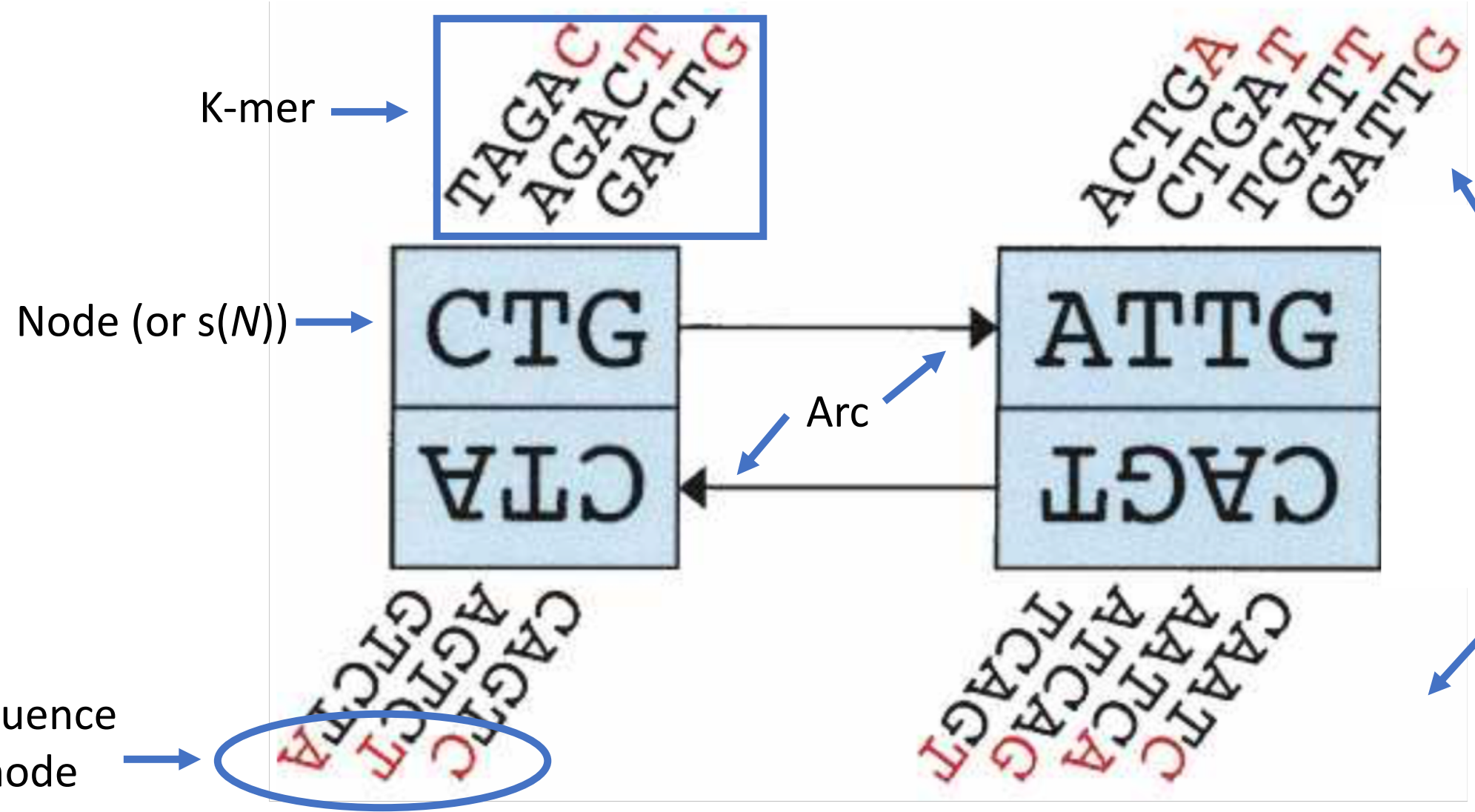
De Bruijn graphs

- Based on K-mer
  - Usually 21 bp k-mer for 25 bp length reads
  - Node = a series of overlapping k-mer

- Advantages
  - Overlapping algorithm is not required (setting length)
  - Saves time(~100M of reads)

- Disadvantages
  - Too short reads are not suitable for de Bruijn graphs

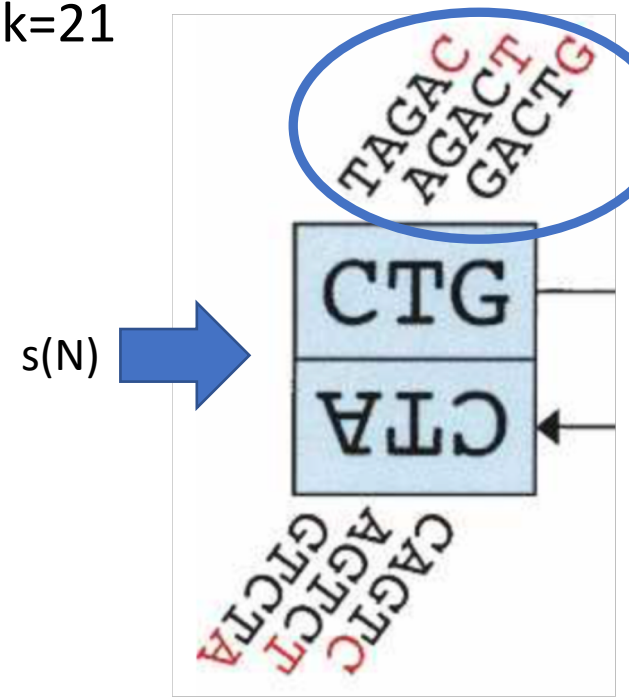
Structure of de Bruijn graph



Construction and simplification of de Bruijn graph

A three step process:

1. Reads are hashed according to a predefined k-mer length (usually for 25 bp reads) k=21
   - Smaller k-mers, increases connectivity of the graph (increases chance of overlap)
   - Each k-mer has an ID
   - Simultaneously recorded to its reverse complement



2. Original k-mers are overlapped by subsequent reads
   - Original k-mers of read is cut each time an overlap with another read begins or ends

3. Simplifying without any loss of information
   - Combine nodes (chains)

Velvet’s approach

1. Removes sequencing errors and handles variations
    - Merge sequences
    - Sequencing process or biological samples
        - Use topological features to overcome errors

2. Resolves repeats based on the available information
    - From low coverage long reads or paired shotgun reads
    - Separates paths sharing local overlaps

Erroneous data create structures

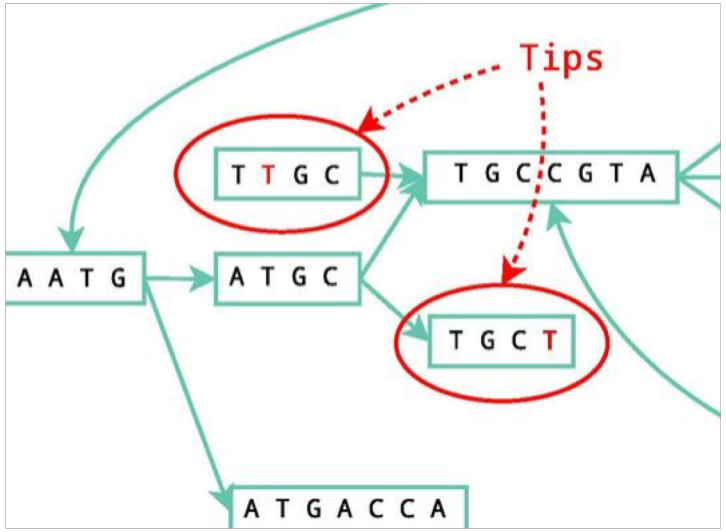
1. Tips
   - Errors at the edges of reads

2. Bulges
   - Internal read errors or to nearby tips connecting

3. Erroneous connections
   - Cloning errors or to distant merging tips

Removing tips

- A chain of nodes that is disconnected on one end
  - Affects only local and doesn't interrupt connectivity
- Two criteria required: Length and minority count
  - Tips gets removed only if it is shorter than 2k
  - A genuine sequence or errors if it's longer than 2k
  - Leading to the node has low multiplicity

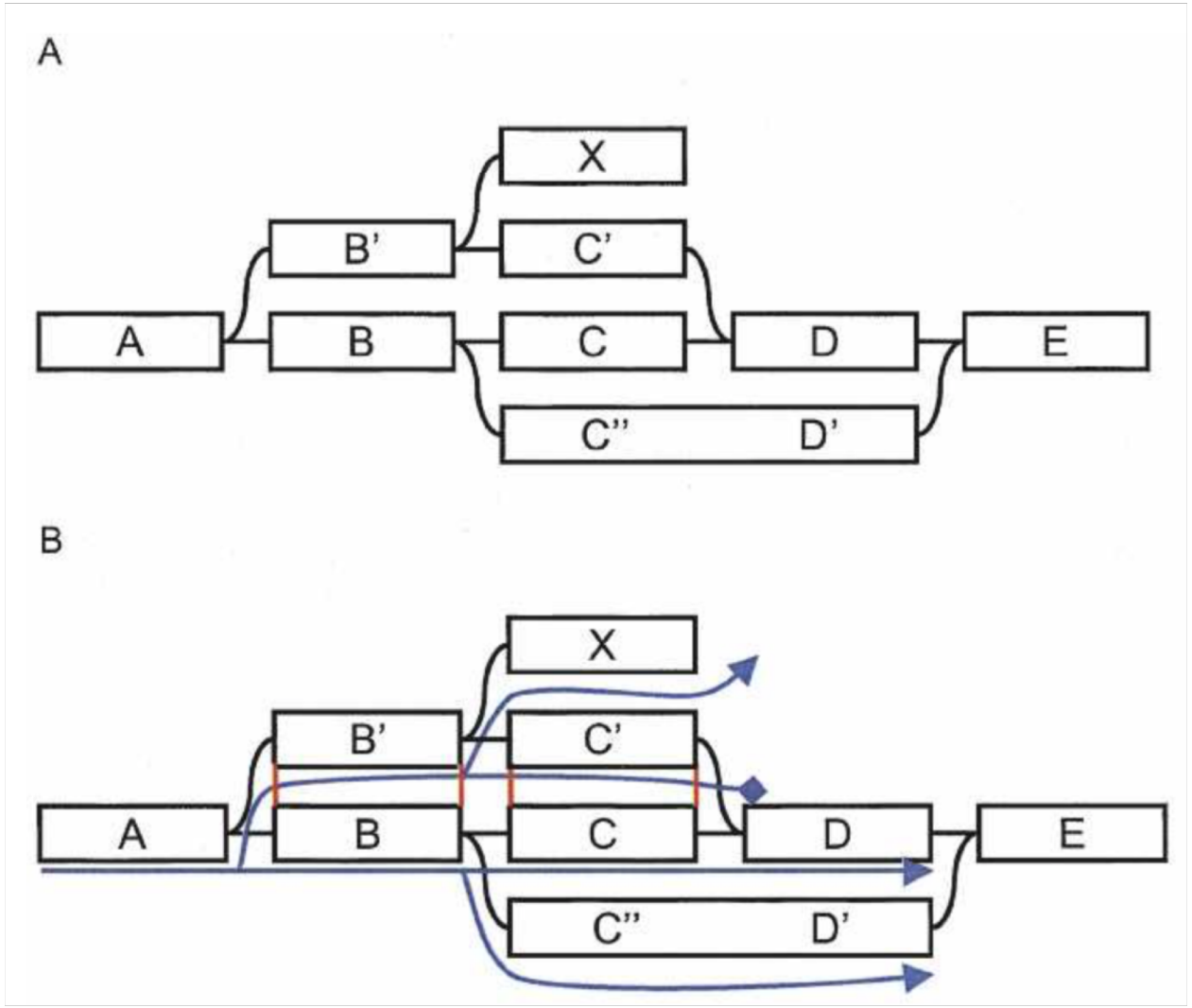


https://en.wikipedia.org/wiki/Velvet_assembler

“Tour Bus” algorithm removes bubbles

Original sequence

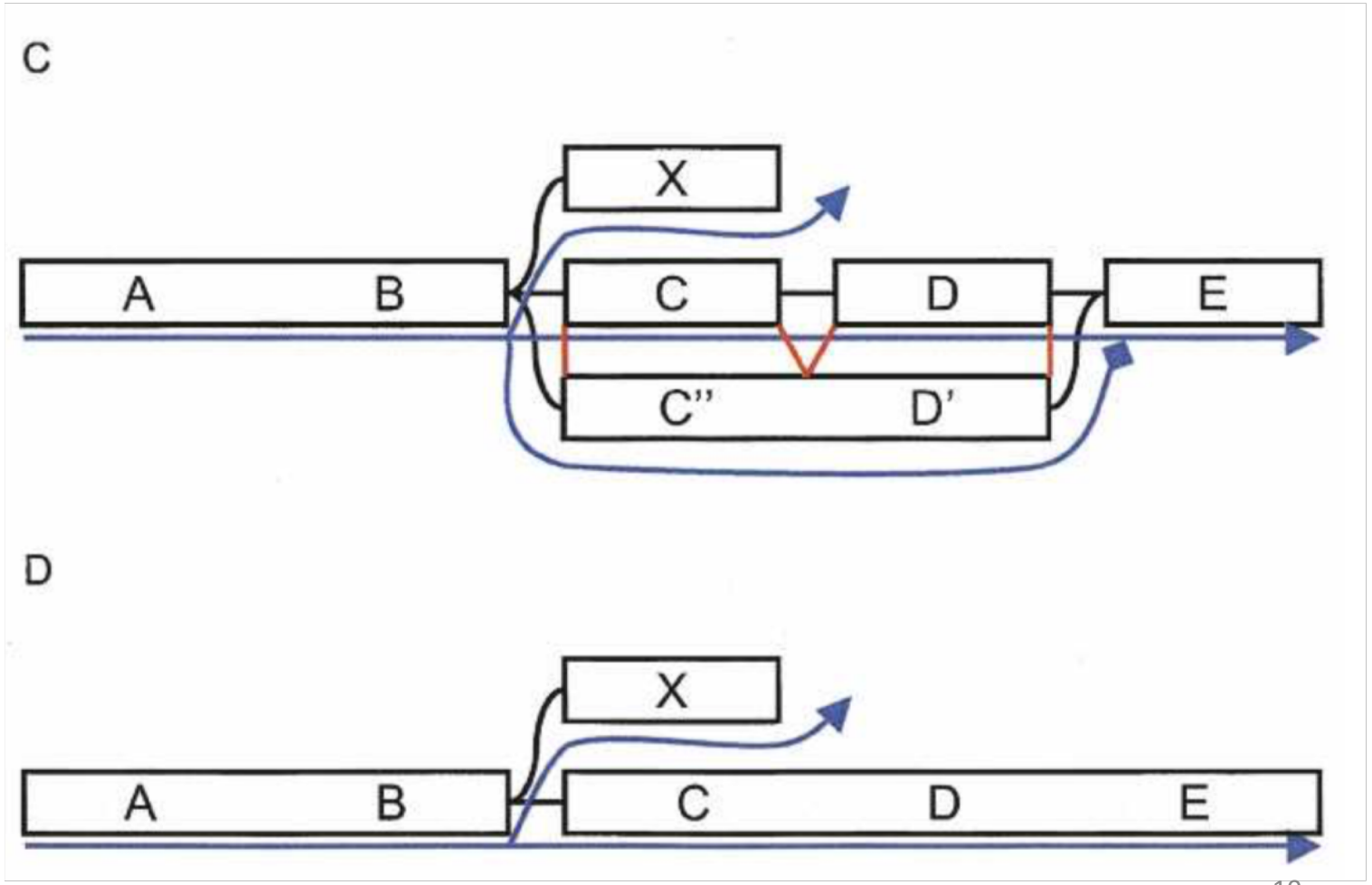
Dijkstra-like first search:
search for redundant paths



Tour Bus algorithm

- A good alignment get merged
- The longer sequence merges into the smaller
- Connectivity is conserved

Iterate

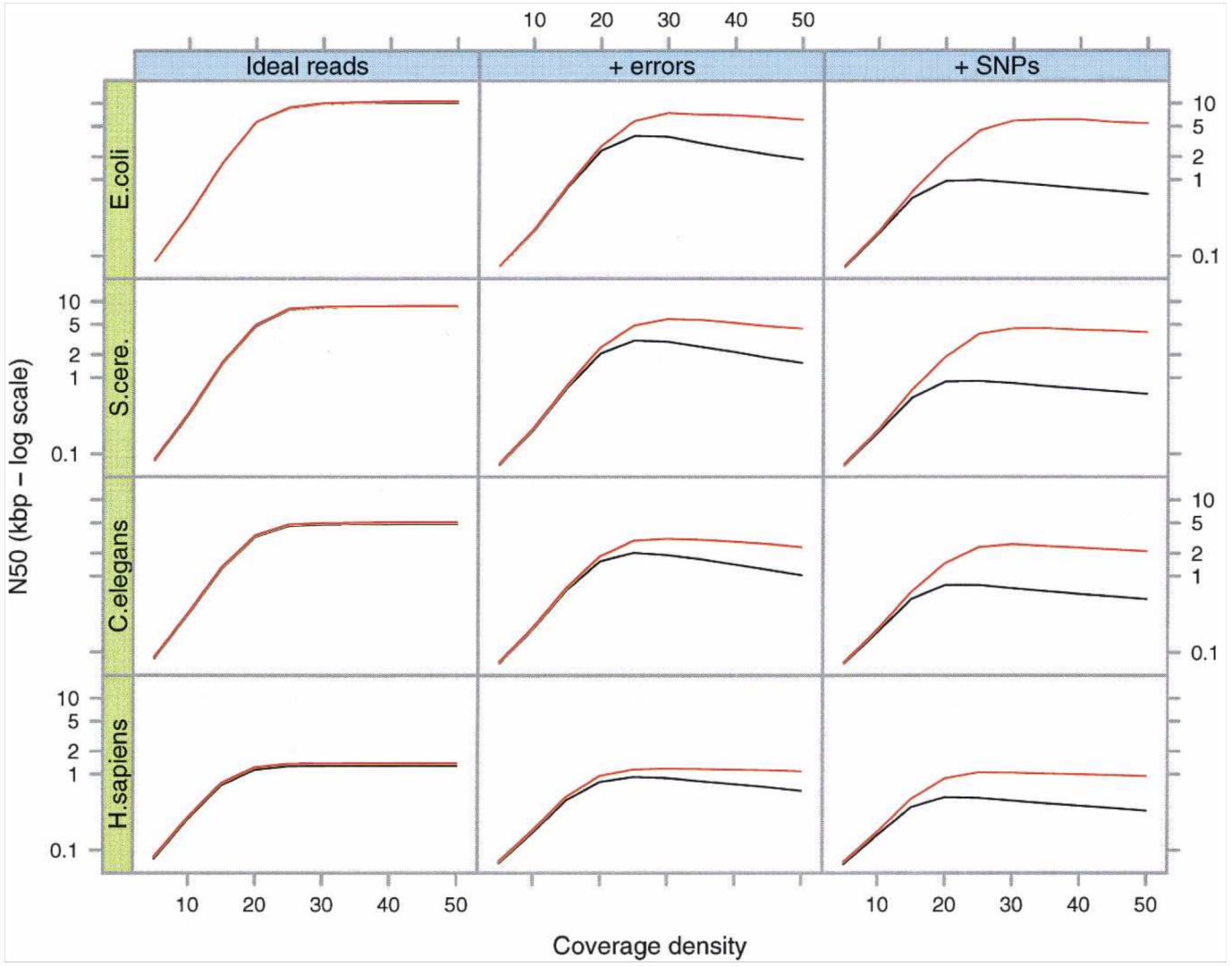


Removing erroneous connects

- Errors are removed after Tour Bus algorithm
- Do not generate correct paths or do not create any recognizable structures
- Removes them with a basic coverage cutoff
  - Set by the user (based on plots of node coverage after the removal of bubbles)

Velvet is not significantly affected by variations

- Complexity
- Black: After tip clipping
- Red: bubble smoothing



Error removal on experimental data

- 173,428 bp human BAC using Solexa sequencing
- 31-mer and 35 bp long
- 970X coverage

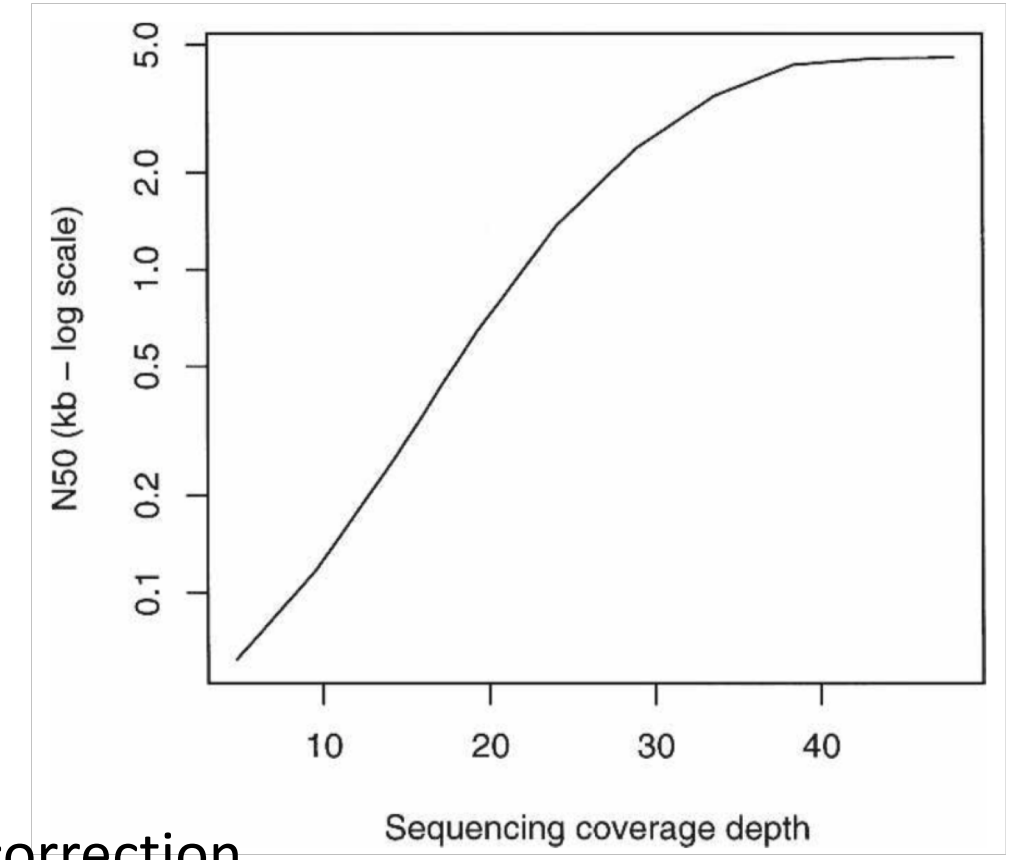
Table 1. Efficiency of the Velvet error-correction pipeline on the BAC data set

Step	No. of nodes	N50 (bp)	Maximum length (bp)	Coverage (percent >50 bp)	Coverage (percent >100 bp)
Initial	1,353,791	5	7	0	0
Simplified	945,377	5	80	4.3	0.2
Tips clipped	4898	714	5037	93.5	78.7
Tour Bus	1147	1784	7038	93.4	90.1
Coverage cutoff	685	1958	7038	92.0	90.0
Ideal	620	2130	9045	93.7	91.9

Error removal on experimental data

Table 2. Efficiency of the Velvet error-correction pipeline on the *Streptococcus* data set

Step	No. of nodes	N50 (bp)	Maximum length (bp)	Coverage (percent >50 bp)	Coverage (percent >100 bp)
Initial	3,621,167	16	16	0	0
Simplified	2,222,845	16	44	0.1	0
Tips clipped	15,267	2195	7949	96.2	95.4
Tour Bus	3303	4334	17,811	96.8	96.4
Coverage cutoff	1496	8564	29,856	96.9	96.5
Ideal	1305	9609	29,856	97.0	96.8



Significantly increased sensitivity and specificity of the correction

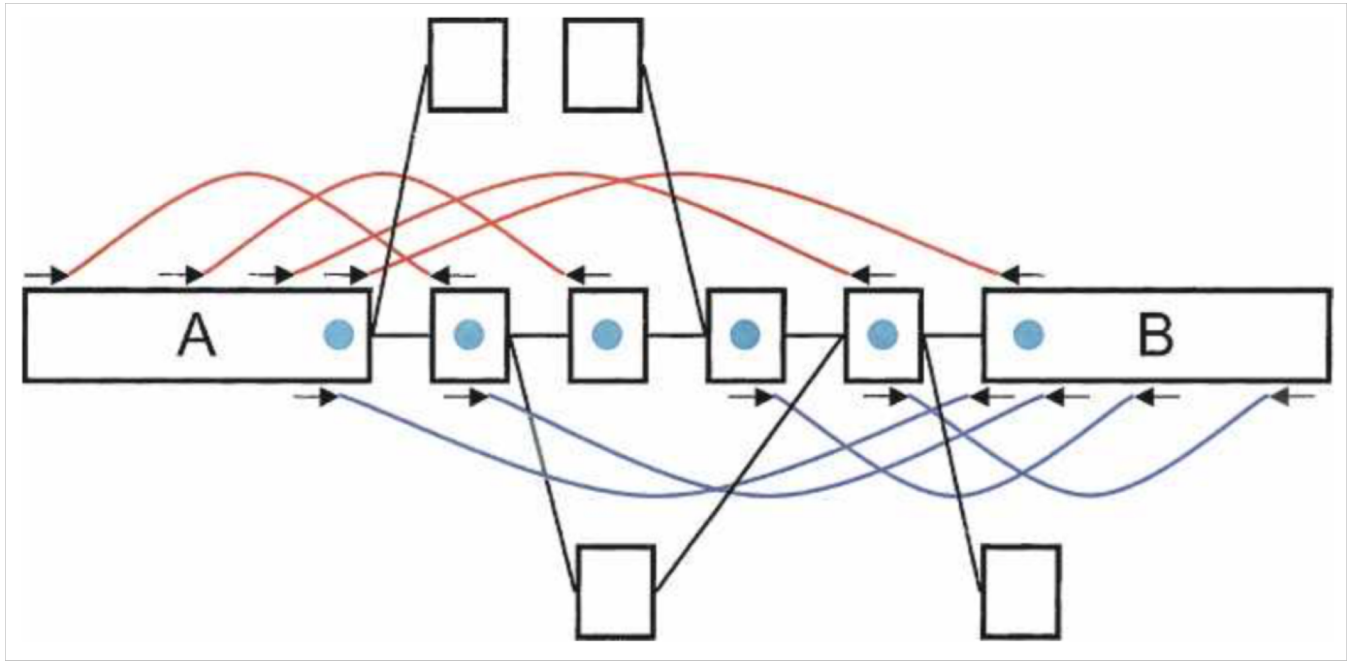
Maintained the integrity of the graph structure

Breadcrumb: resolution of repeats with short read pairs

Purpose: Correctly extend and connect contigs through repeated region

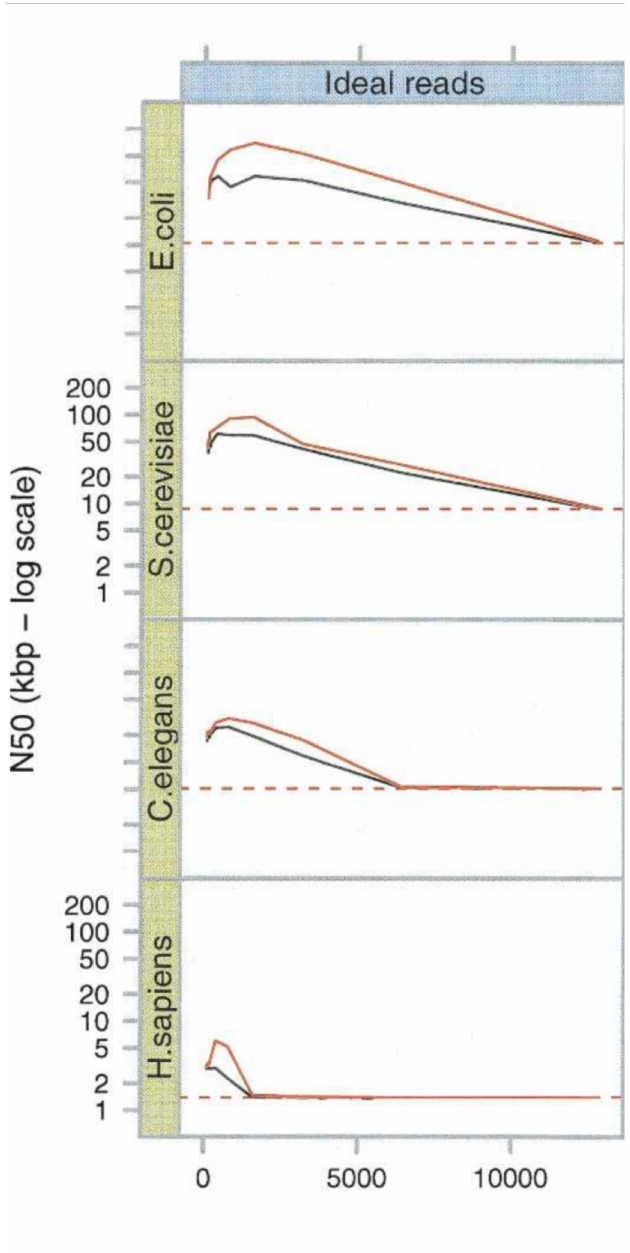
Pairs up the long nodes

- Flagged as ambiguous
- Eliminates duplicated nodes



Breadcrumb algorithm

- Broken black line: Tour Bus
- Broken red line: applying 4X coverage
- Black solid line : N50 after breadcrumbs
- Red solid line: N50 after super contigging



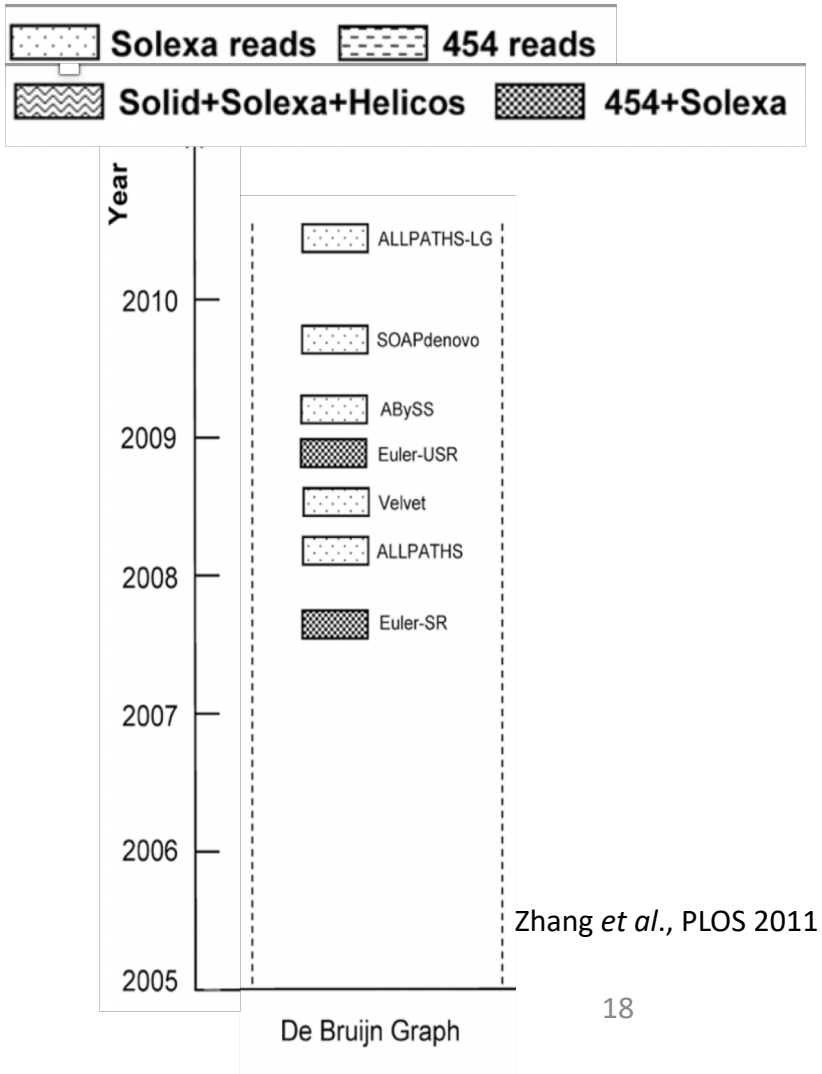
Comparison of different assemblers

Table 3. Comparison of short read assemblers on experimental *Streptococcus suis* Solexa reads

Assembler	No. of contigs	N50	Average error rate	Memory	Time	Seq. Cov.
Velvet 0.3	470	8661 bp	0.02%	2.0G	2 min 57 sec	97%
SSAKE 2.0	265	1727 bp	0.20%	1.7G	1 h 47 min	16%
VCAKE 1.0	7675	1137 bp	0.64%	1.8G	4 h 25 min	134%

Conclusions

- Velvet converts a de Bruijn graph into traditional assembly of contiguous sequence
- Hash k-mer, Tour Bus and Breadcrumb
- Algorithmic improvements
- Almost complete genome can be assembled with paired read information



Zhang *et al.*, PLOS 2011