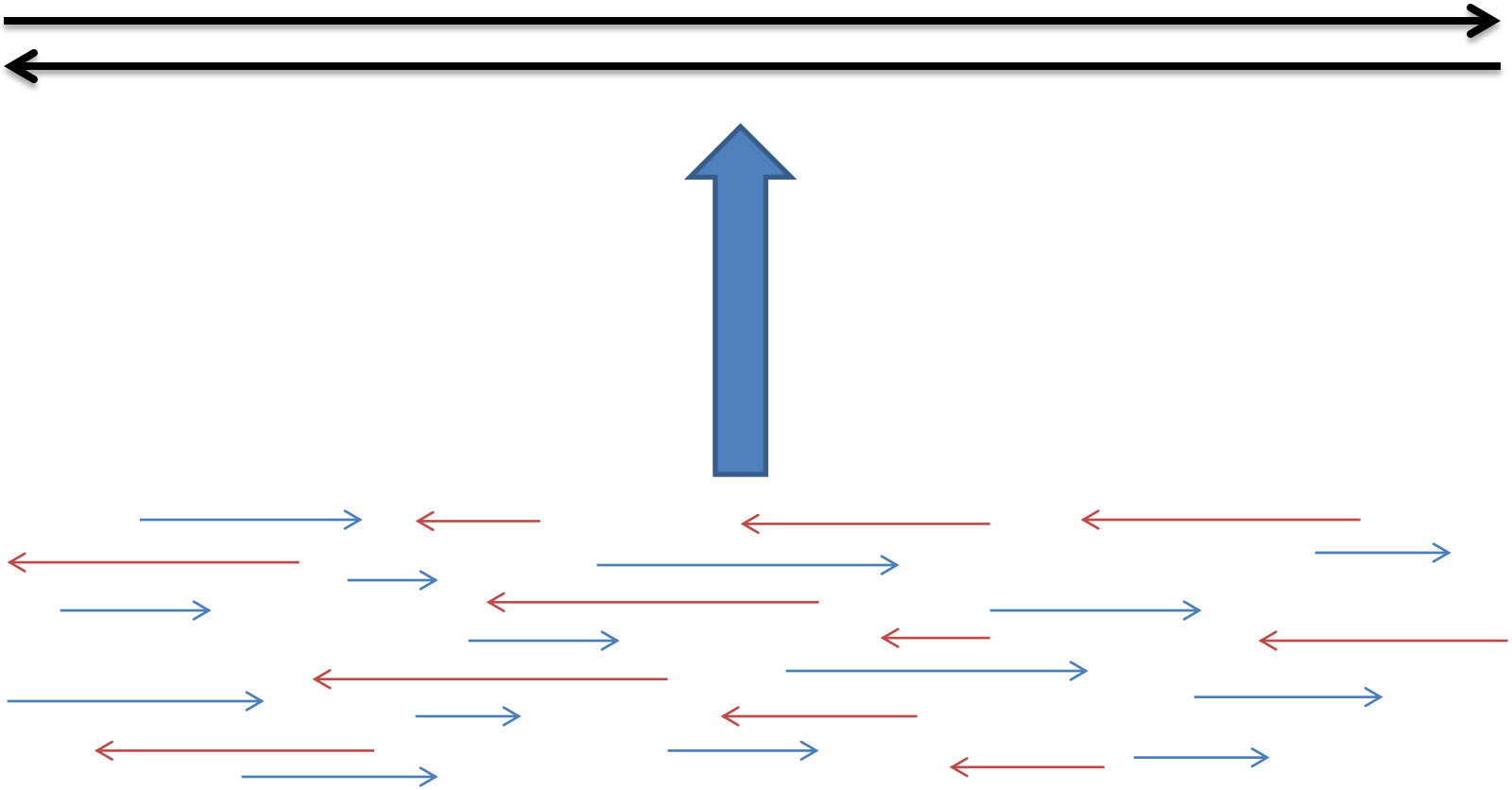
Velvet: Algorithms for de novo short read assembly using de Bruijn graphs

Matthew Tien

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Genome Assembly Problem



Overlap Graphs

- Construction
  - Nodes= reads
  - Edges= overlap between reads
- Disadvantages
  - Hard overlap problem with large amount of reads
  - Pair-wise computation too much
  - Hamiltonian Path Problem

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

De Bruijn Graphs

- K-mer approach
  - Define set length nucleotides of length k
- Construction
  - Node= k-mer
  - Edges= connect nodes by the path created from a read overlapping with the k-mer
- Advantages
  - Set length nodes, so no overlap algorithm needed
  - Eulerian path problem
- Disadvantages
  - Loss of information
  - Shorter contigs

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

De Bruijn Graphs useful for Short Reads

- Short reads (~100-200bp long)
  - 25-50 is too short for de Bruijn
- Short Reads
  - Many reads will have only a single or no base difference
  - More ambiguous connections

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

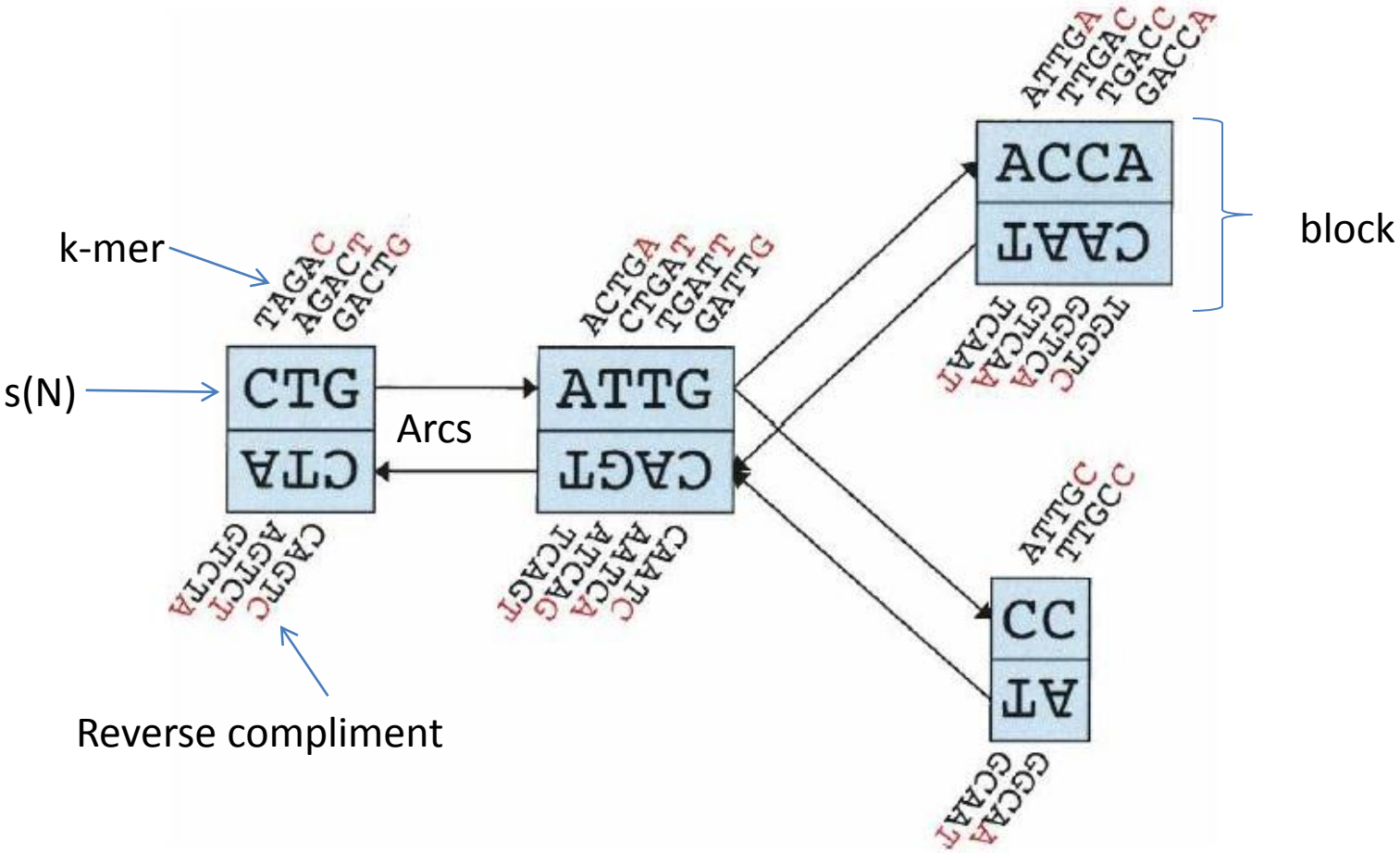
Velvet approach to de Bruijn graphs

- Error Correction Algorithm
  - Merge sequences that belong together

- Repeat solver
  - Separates paths sharing local overlaps

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Structure of de Bruijn Graph

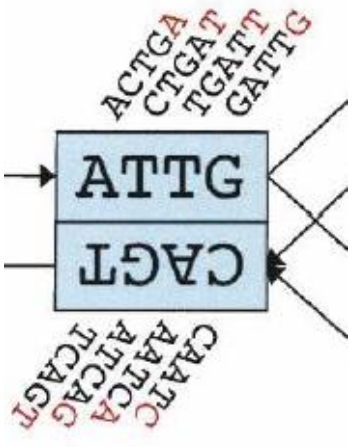


Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Construction (part 1)

- Two step process:
  1. Reads are hashed to a predefined k-mer length (k=21 or 25 bp)
     a. Each k-mer has an ID that maps the k-mer back to the read and its position in the read.
     b. Simultaneously recorded to its reverse complement

This is just to get a set of k-mers for the graph



Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

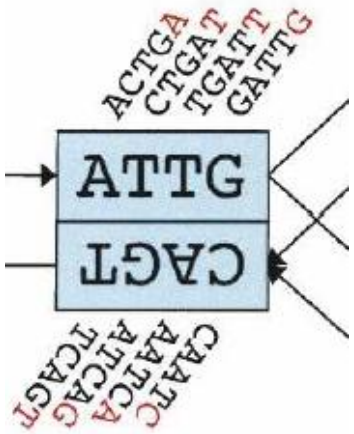
Construction (part 2)

- Two step process:

  2. For each read, it records which k-mer are overlapped by subsequent reads

     a. original set of k-mers is cut each time an overlap with another read beings or ends

This is to get the s(N) part of each k-mer



Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

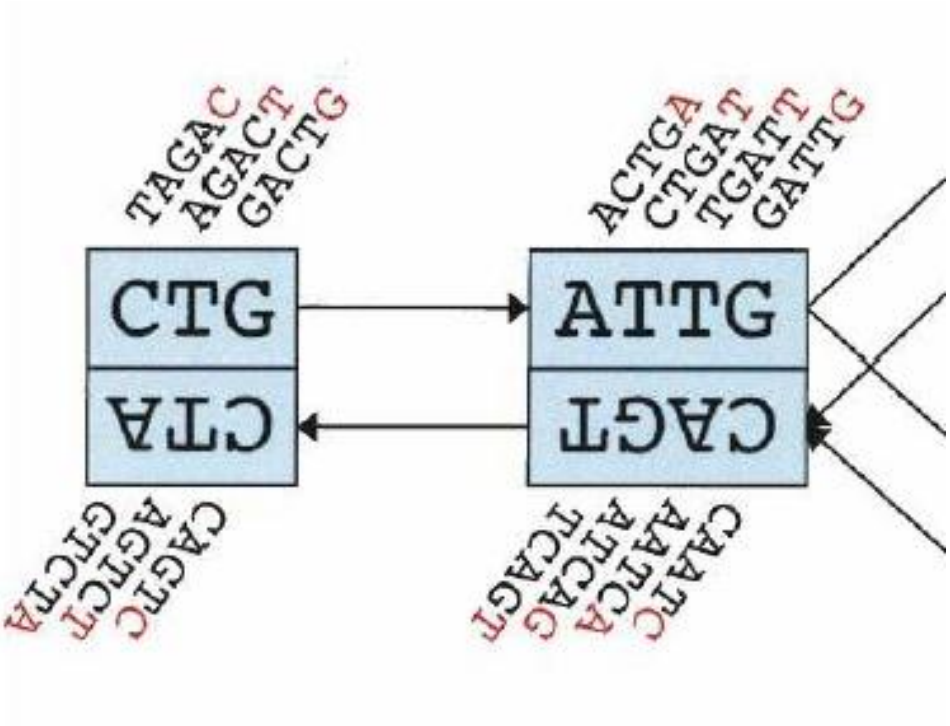
Construction (part 3)

- Use reads to add connections to the nodes

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Simplification

- Chains



Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Velvet approach to de Bruijn graphs

- **Error Correction Algorithm**
  - **Merge sequences that belong together**

- Repeat solver
  - Separates paths sharing local overlaps

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Error Removal

- Kinds of errors
  - Process: sequencing errors
  - Natural: single nucleotide polymorphisms
- Distinguishing the two is hard
  - Previous methods used to use a probabilistic chance of encounter such errors
  - Velvet: use topological cues to find errors and remove them

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Three kinds of Topological Errors

- Tips
  - Errors at the edge of reads
- Bulges
  - Internal read errors or nearby tips connecting
- Erroneous Connections
  - Cloning errors or to distant merging Tips

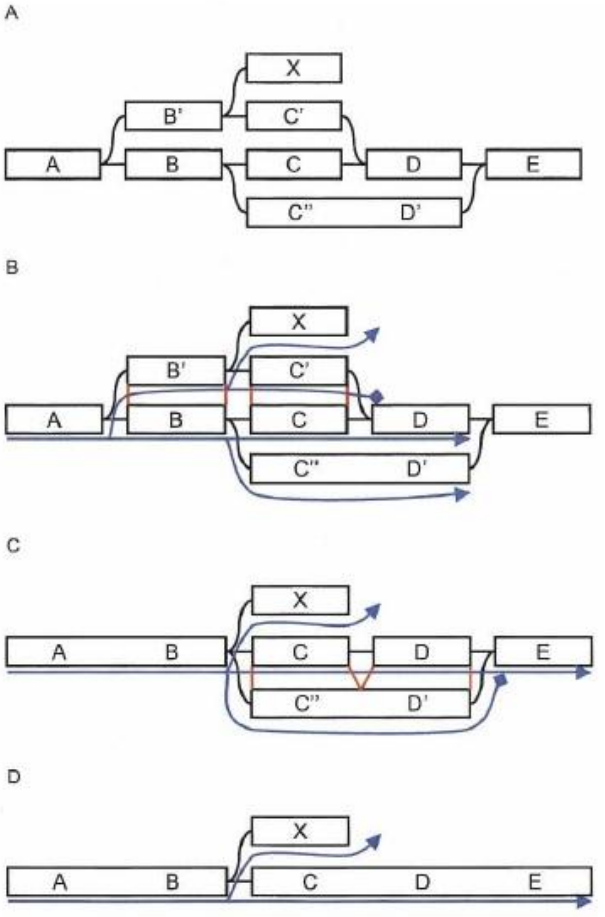
Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Removing Tips

- Tips is a chain of nodes that is disconnected at one end
- Two criterion Length and “Minority count”:
  - Remove read if it is shorter than 2k. Tips longer that 2k represent genuine sequence or an accumulation of errors
  - The tip is removed if a more common path is present in the node with the out going tip

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

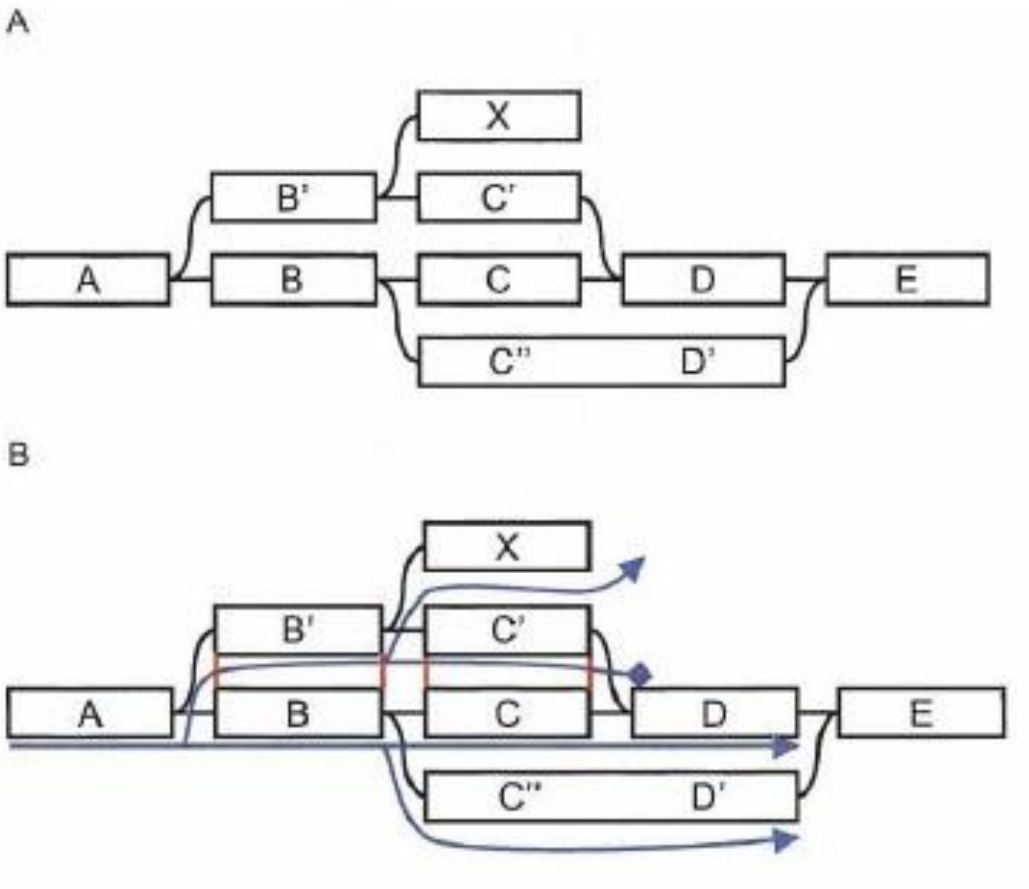
Removing bulges with “Tour Bus” algorithm



Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

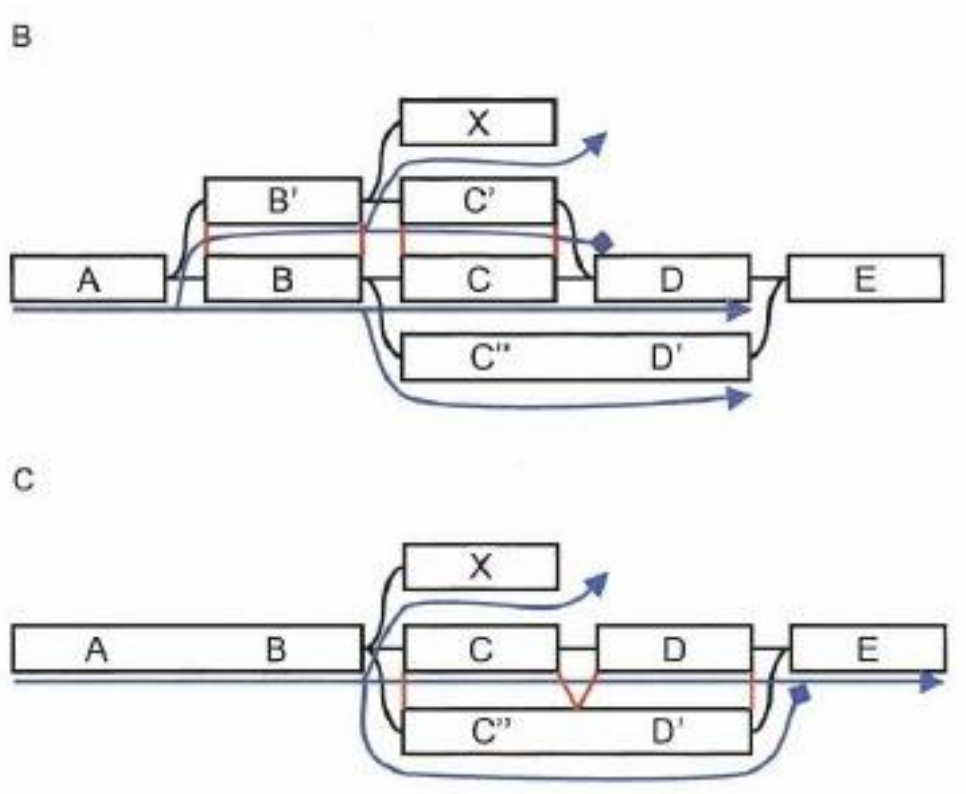
Dijkstra-like Depth First Search

Depth First Search



Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

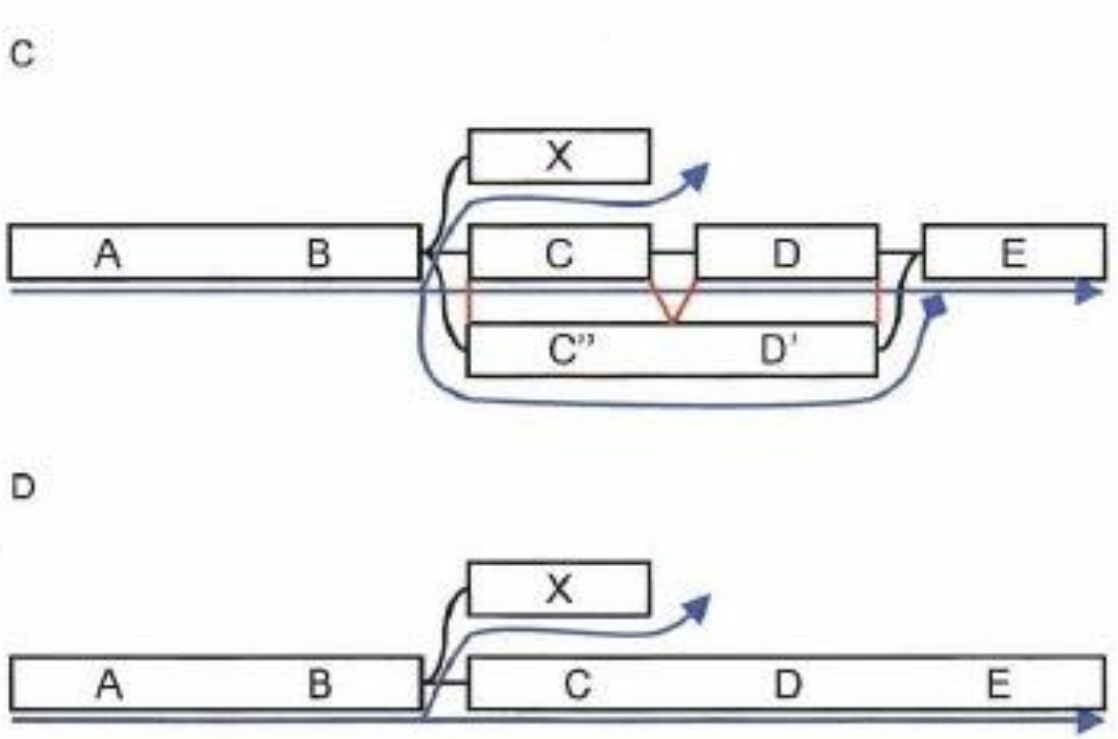
Converge Sequences when you hit a previously visited node



Take out BC and B'C' sequences, align them, if there is a good alignment then merge them with consensus sequence. The longer sequence is always merged into the smaller and connectivity is conserved

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Iterate through to get final graph



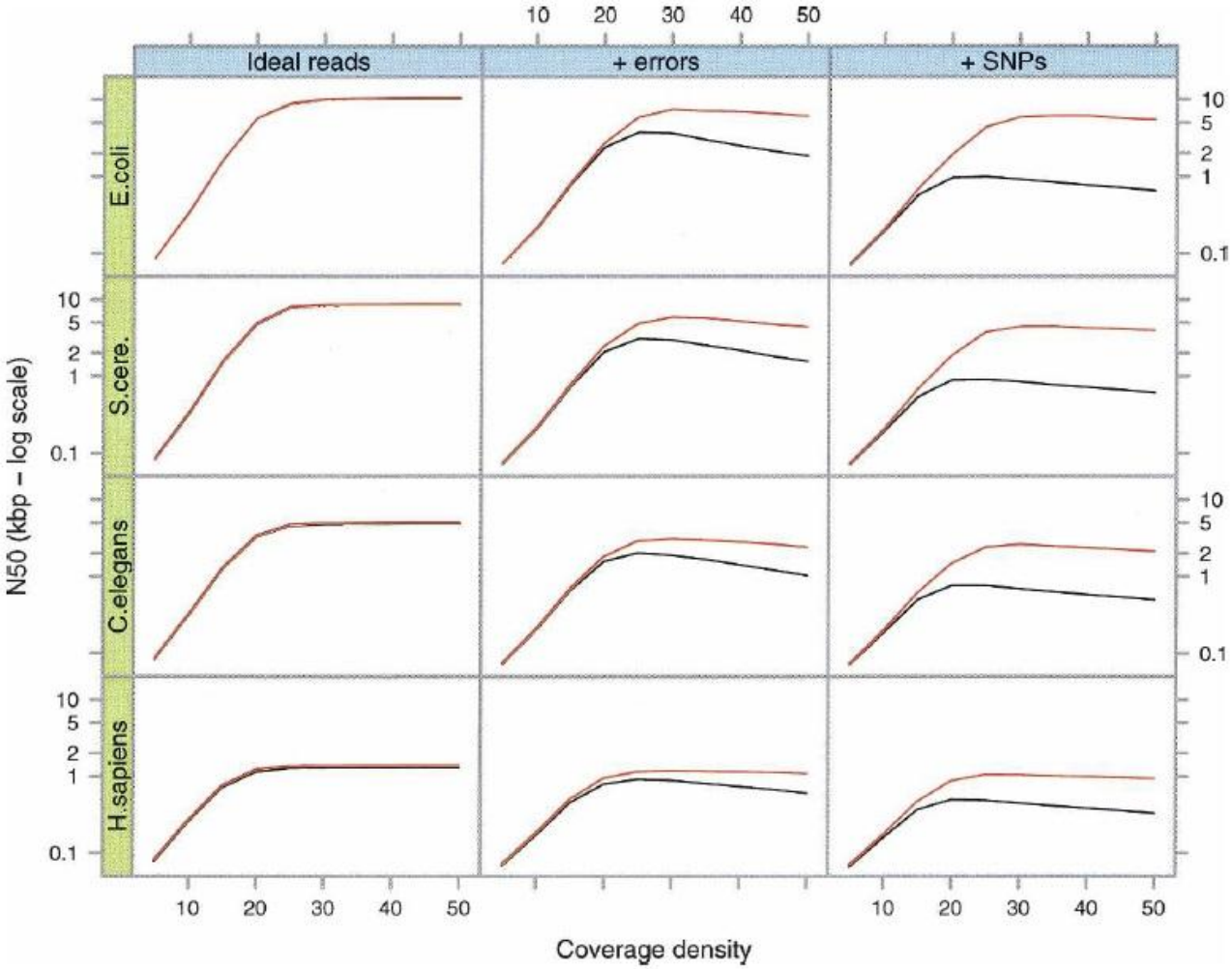
Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Removing Erroneous Connections

- Removal by a basic coverage cutoff set by the user based on plots of node coverage after the removal of tips and bubbles.

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Testing on Simulated Data Results



- 35 bp long
- different coverage values (5x – 50x)
- k=21
- Ideal reads- no errors
- +errors, 1% error rate
- +SNPs, 1% error rate on one strand, 1/500 bp on second strand. Fragments taken from both

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Testing on Experimental Data Results

173,428 bp, 970x coverage, Human. 35 bp long reads, k=31. Built a de Bruijn graph from a known finished sequence for comparison.

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

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Testing on Experimental Data Results

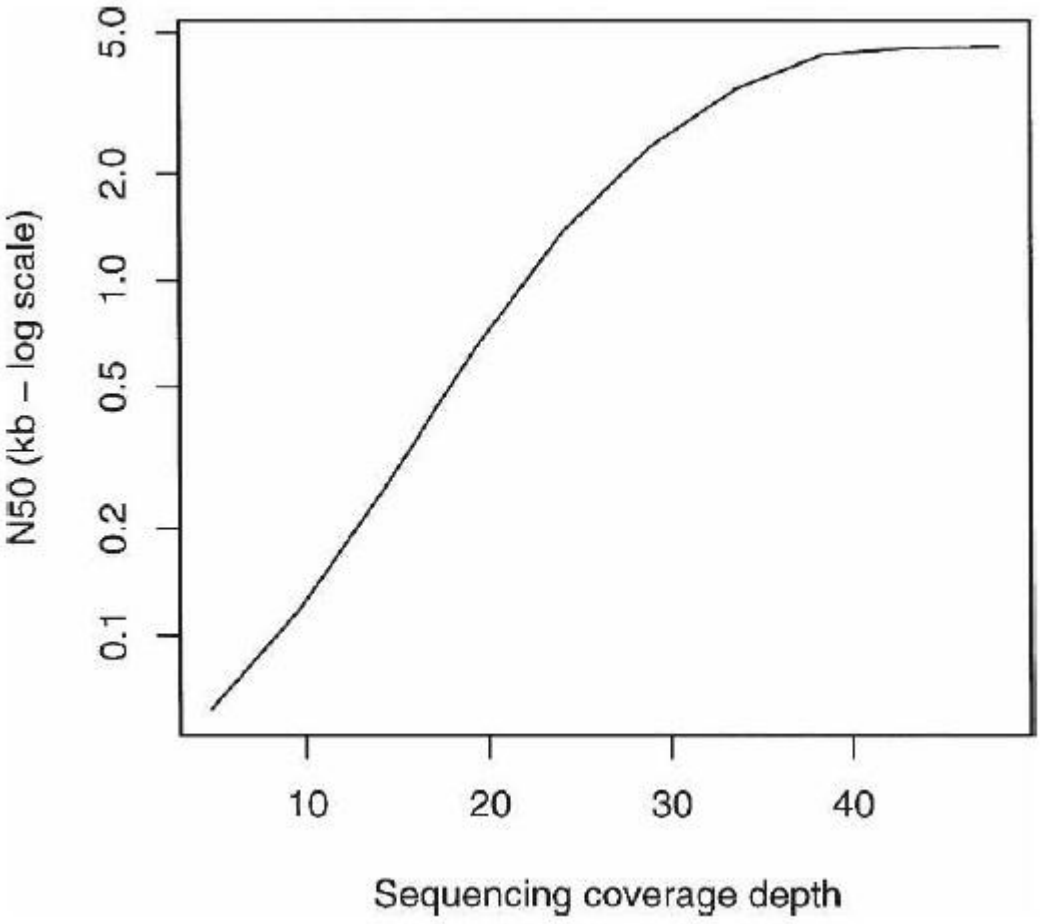
2,700,036 bp, 48x coverage*, Strep. suis.* 35 bp long reads, k=31. Built a de Bruijn graph from a known finished sequence for comparison.

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

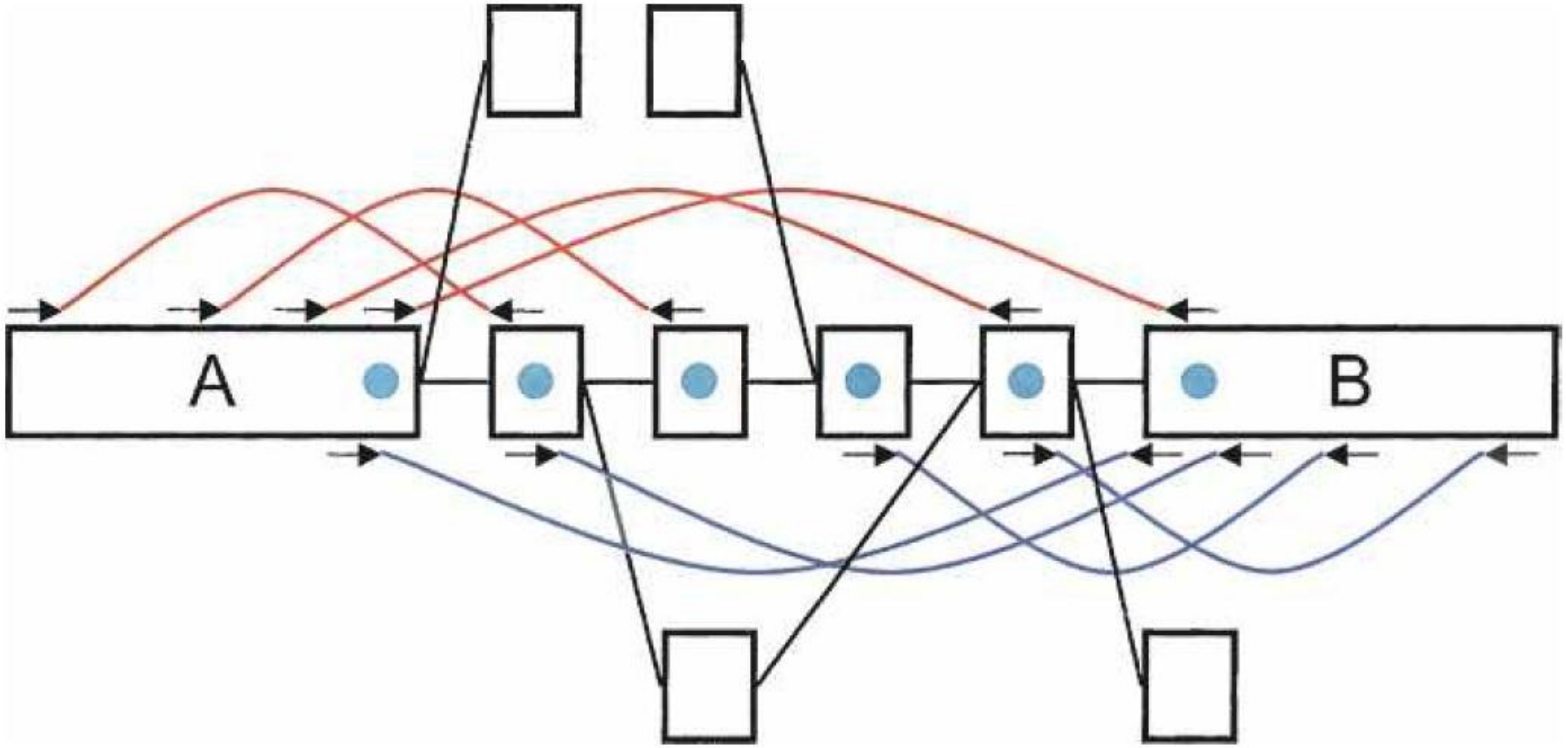
Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

N50 (average sequence length) from Velvet on Strep Data



Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Breadcrumbs: Resolution of repeats



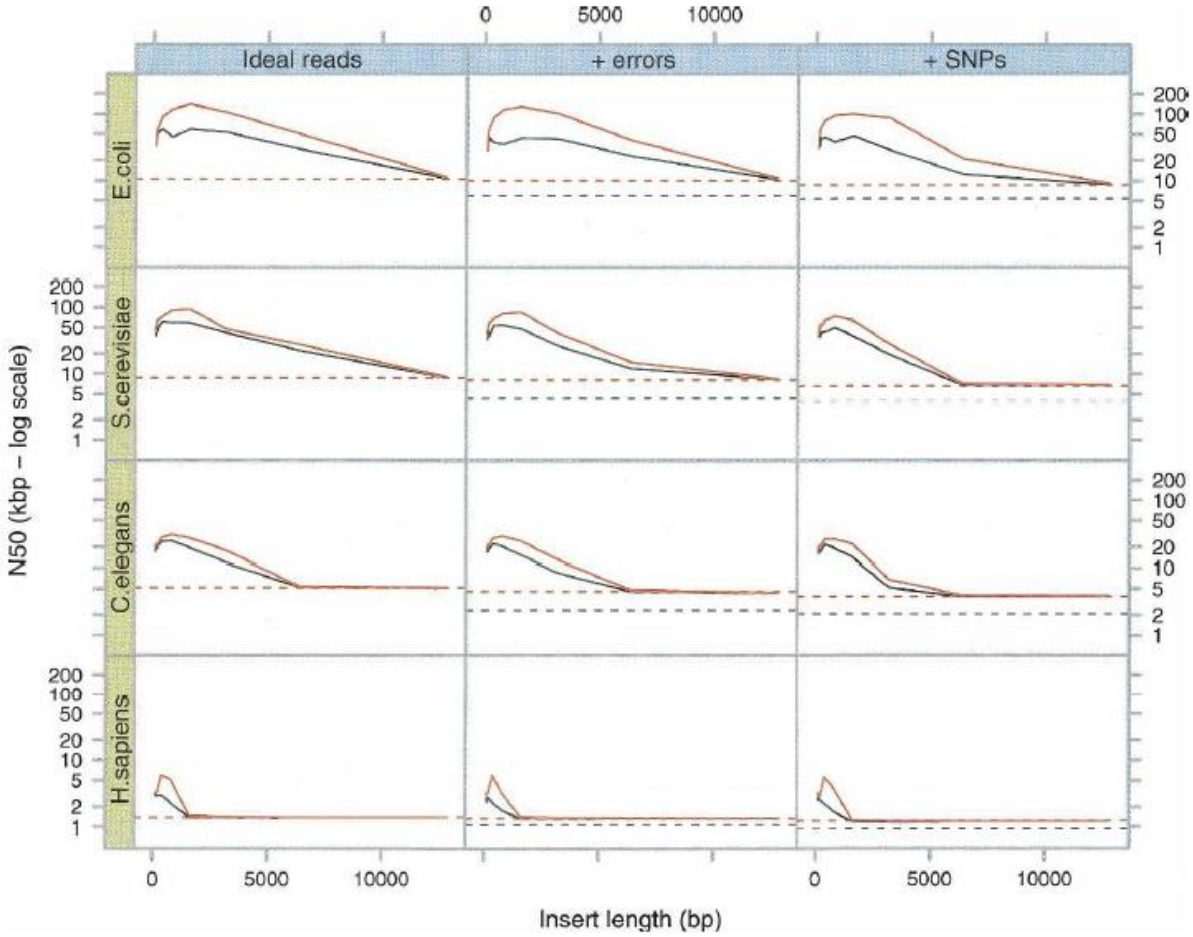
Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Breadcrumbs Algorithm

- Using the read pair (from sequencing data), find all long nodes.
  - Ignore the long contigs that pair up with the long nodes
- For each long node, breadcrumb flags all nodes containing the mate reads of the reads in that long node
- Find the simplest path and merge (hopefully)

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Breadcrumb results



- Same data set as before.
- Horizontal broken lines denote N50 lengths. Black indicates after Tour Bus. Red indicates after 4x coverage cutoff
- Black solid lines indicates N50 after breadcrumbs and Red indicates N50 after super-contigging

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Overview of Velvet

1. Hash k-mers
2. **Construct the graph**
3. Correct for errors
4. Resolve the repeats

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Comparison of short read assemblers on experimental *Strep.*

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%

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.

Picking k

E(X) = expected number of times a unique k-mer in a genome of length G is observed in a set of n read of length l. C is coverage.

Want to get E(x) of about 10-15. Helps pick k.

$$E(X) = \frac{n(l-k+1)}{G-k+1} \approx \frac{n}{G}(l-k+1) = C\frac{l-k+1}{l}$$

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Press 18 (2008): 821-29.