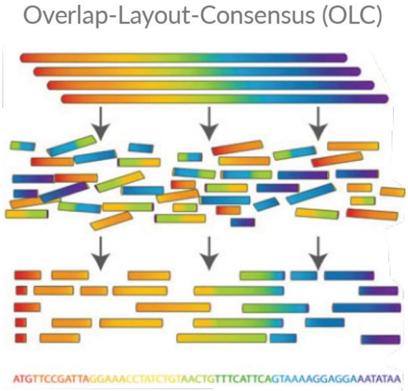
Overlap Layout Consensus (OLC)

Very viable strategy.

Strategy of choice for long read assembly.

Newbler, Celera, Canu

Canu is a fork of the Celera assembler.
Designed for high-noise long read technologies.

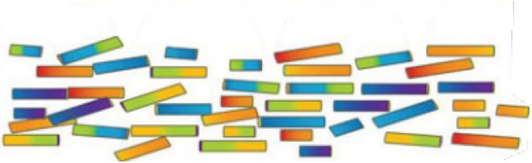

Overlap-Layout-Consensus (OLC)
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

Overlap Layout Consensus (OLC)

Overlap

Find all overlaps between reads

- All-vs-all pairwise read alignment
- Min overlap length enforced
- Min percent identity enforced



Heuristics

- Minhash
  (identify reads with possible overlap)
- Seed extend / seed chain align
  (kmer hits then DP alignment)

Overlap Layout Consensus (OLC)

Overlap

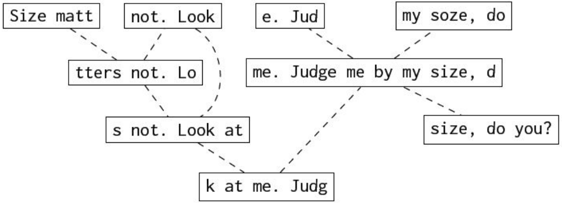
Construct overlap graph to represent identified overlaps

Nodes represent reads

- Nodes have attributes!
- Read id
- Read length
- Sequence

Edges represent overlaps

- Edges have attributes!
- Length of overlap
- Type of overlap (suffix-to-prefix or containment)



Size matt

not. Look

e. Jud

my soze, do

tters not. Lo

me. Judge me by my size, d

s not. Look at

size, do you?

k at me. Judg

Size matters not. Look at me. Judge me by my size, do you?

Overlap Graph Simplification

Want a hamiltonian path

...but overlap graphs have many edges & dead ends

Transitive Edge Reduction (TER)

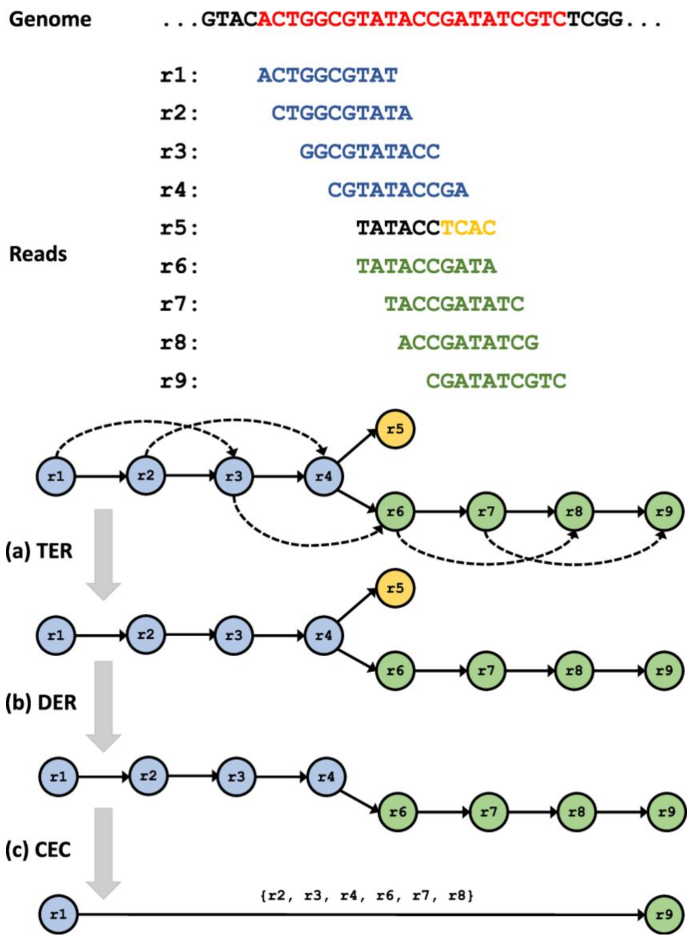
- Remove unnecessary edges

Dead-End Removal (DER)

- Remove short spurs / dead ends

Composite Edge Contraction(CEC)

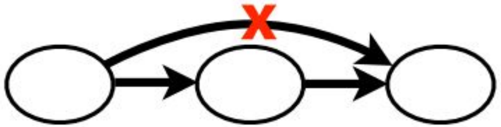
- Merges nodes in manner which does not lose information
- Quite complex. Not covering this.

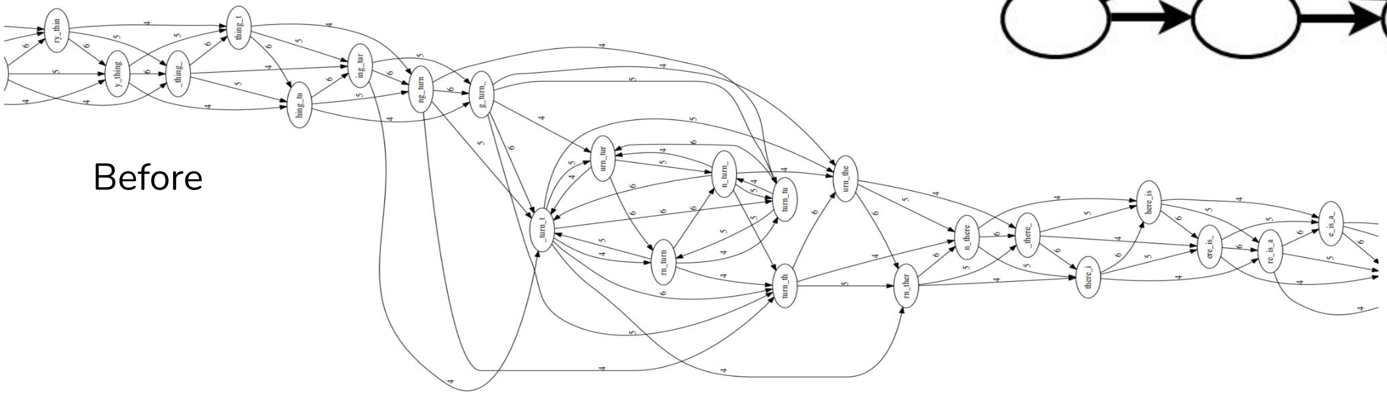


Paul et al (2019)

Transitive Edge Reduction

Process: (1) Remove edges which skip one node

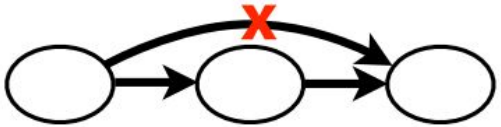


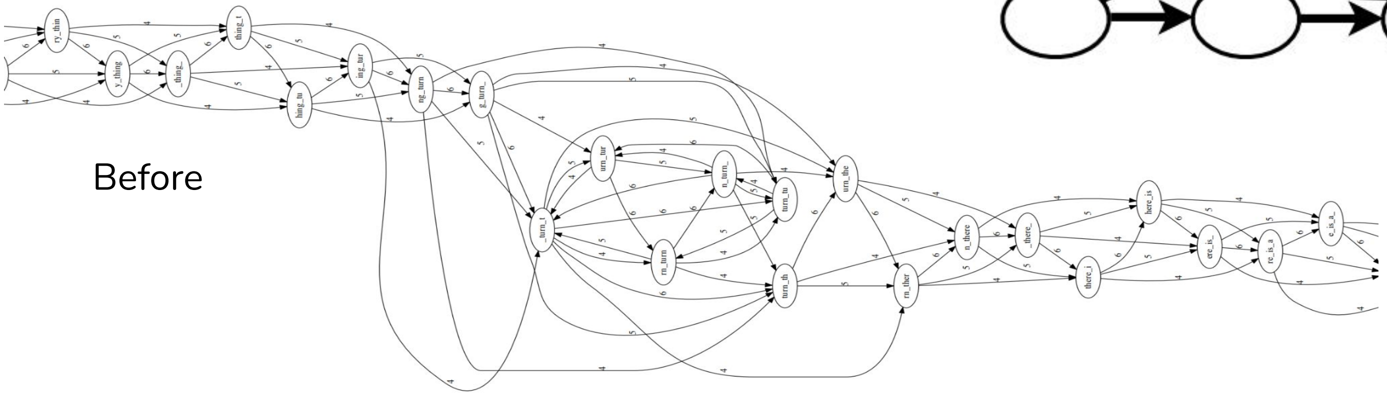


Before

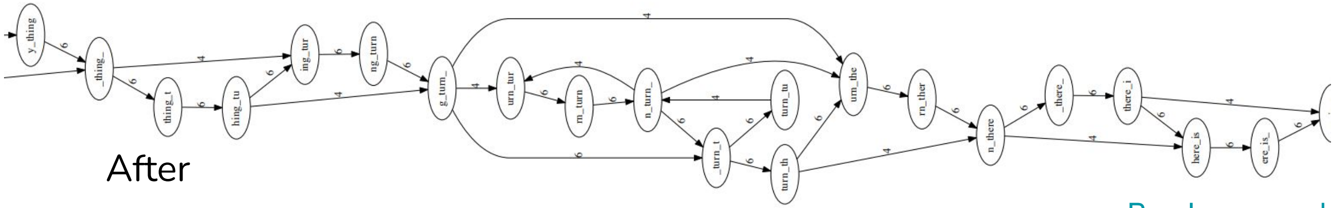
Transitive Edge Reduction

Process: (1) Remove edges which skip one node





Before

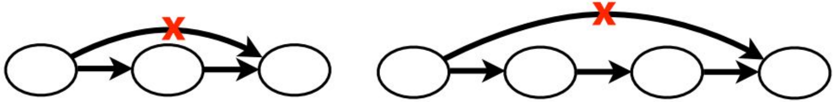


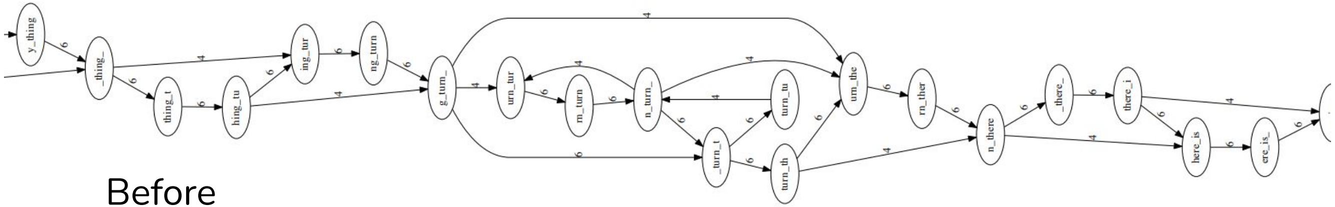
After

Ben Langmead

Transitive Edge Reduction

Process: (2) Remove edges which skip one or two nodes

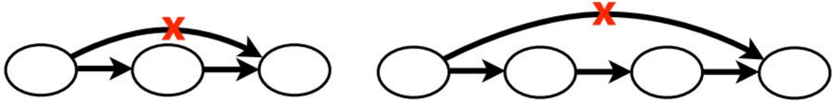


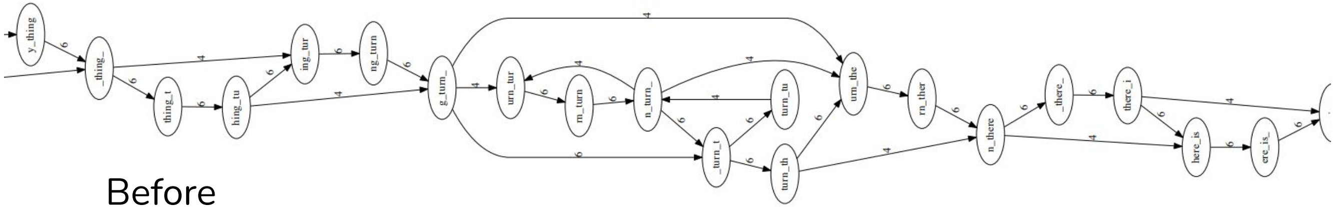


Before

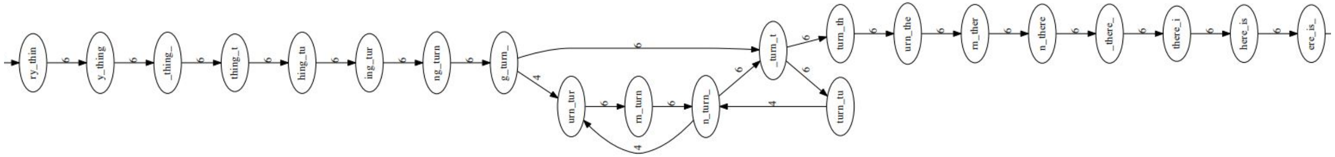
Transitive Edge Reduction

Process: (2) Remove edges which skip one or two nodes





Before



After

Ben Langmead

Dead-End Removal (DER)

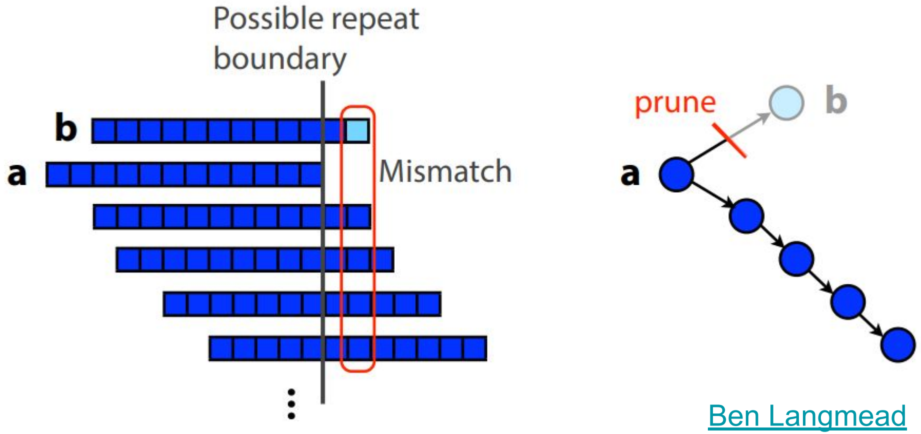
Remove short spurs / dead ends

Caused by sequencing errors

Caused by overlapping of chimeric sequences (repeats)

Simple to remove

- Identify, then prune
- Short length edges
- Low coverage (depth)



Ben Langmead

Overlap Layout Consensus (OLC)

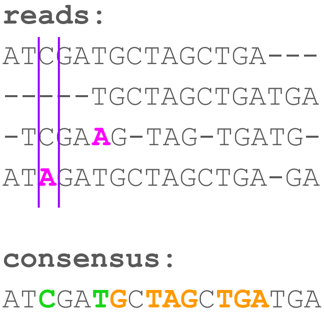
Consensus

Gather reads which make up a contig

Line them up (Multiple Sequence alignment)

Generate consensus sequence (eg voting)

Software can incorporate coverage and base-level quality scores of reads when generating consensus contigs.



```
reads:
ATCGATGCTAGCTGA---
-----TGCTAGCTGATGA
-TCGAAG-TAG-TGATG-
ATAGATGCTAGCTGA-GA

consensus:
ATCGATGCTAGCTGATGA
```