

Assembly of long, error-prone reads using repeat graphs

Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner

05.07.2021

Long read assembly

- ▶ error rate long read \leftrightarrow short read

Long read assembly

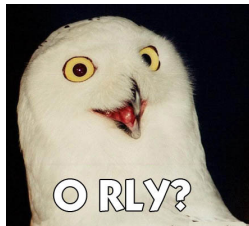
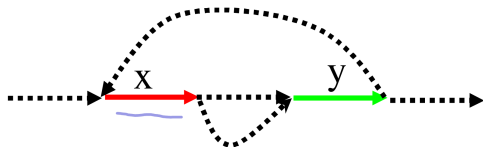
- ▶ error rate long read \leftrightarrow short read
- ▶ assembly fragmentation \rightarrow repeats

Long read assembly

- ▶ error rate long read \leftrightarrow short read
- ▶ assembly fragmentation \rightarrow repeats
- ▶ Flye should resolve these repeats correctly

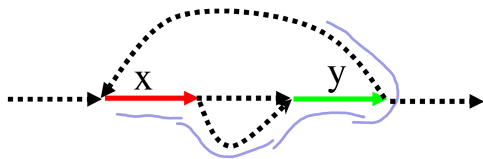
Disjointigs

- ▶ current assemblers use much time on correct contig assembly
- ▶ Flye uses a different approach:
- ▶ generate paths from overlapping reads without checking for correct assembly -> disjointigs



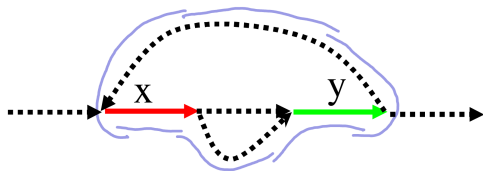
Disjointigs

- ▶ current assemblers use much time on correct contig assembly
- ▶ Flye uses a different approach:
- ▶ generate paths from overlapping reads without checking for correct assembly -> disjointigs

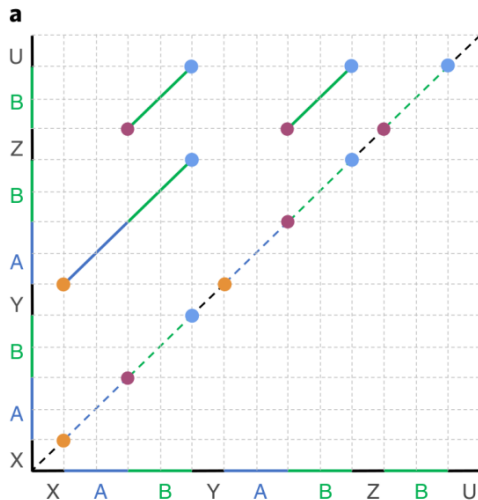


Disjointigs

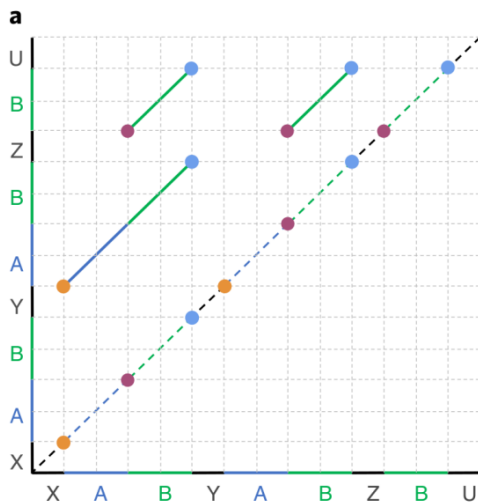
- ▶ current assemblers use much time on correct contig assembly
- ▶ Flye uses a different approach:
- ▶ generate paths from overlapping reads without checking for correct assembly -> disjointigs



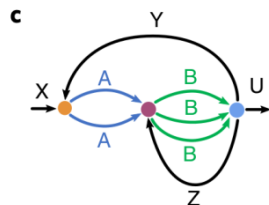
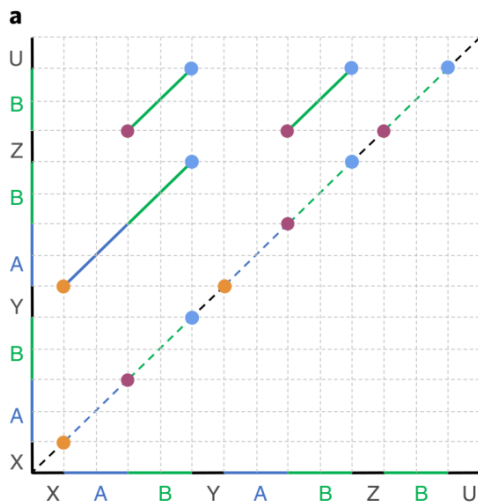
Repeat graph creation



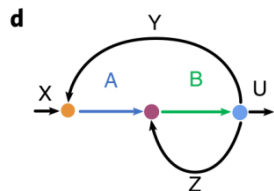
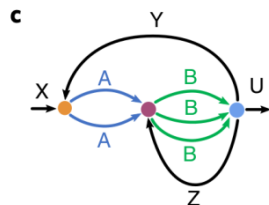
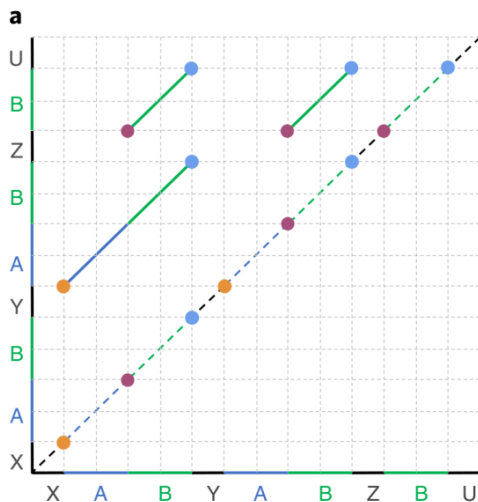
Repeat graph creation



Repeat graph creation

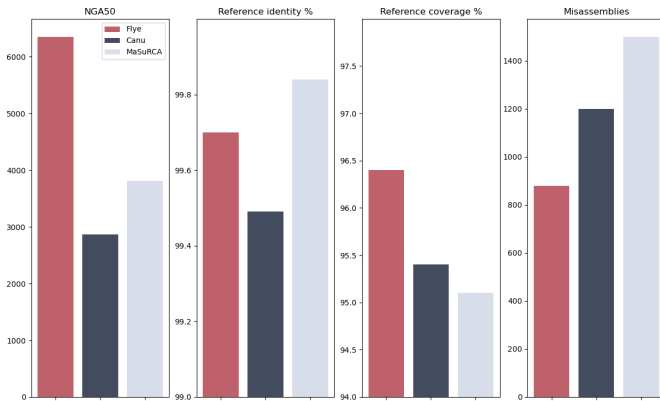


Repeat graph creation



Repeat resolution

Results

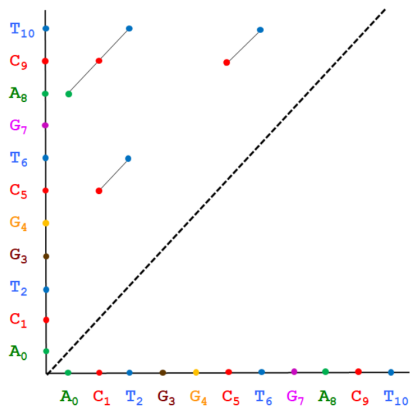


References

Appendix

Dot plot creation

-	9	10	-	-	9	10	-	-	5	6
8	5	6	-	-	1	2	-	0	1	2
0	1	2	3	4	5	6	7	8	9	10
A	C	T	G	G	C	T	G	A	C	T



Repeat graphs

► generalization of de bruijn graphs

>- from disjointigs = random walk of reads on the repeat

>- means the repeat graph hasn't to be known

Repeat graphs

- ▶ generalization of de bruijn graphs
- ▶ structure

>- from disjointigs = random walk of reads on the repeat
>- means the repeat graph hasn't to be known

Repeat graphs

- ▶ generalization of de bruijn graphs
- ▶ structure
- ▶ creation

>- from disjointigs = random walk of reads on the repeat

>- means the repeat graph hasn't to be known

Difference repeat graph de Bruijn graph

- ▶ A-Bruijn graph (alignments) generalizes the de Bruijn graph

Difference repeat graph de Bruijn graph

- ▶ A-Bruijn graph (alignments) generalizes the de Bruijn graph
- ▶ We thus argue that the time has come to explain that the breakpoint graphs and the de Bruijn graphs are two identical data structures (if one ignores a cosmetic difference between them) as they both represent specific instances of a general notion of the A-Bruijn graph introduced in [13]. The A-Bruijn graphs are based on representing genomes as sets of labeled paths and further gluing identically labeled edges (breakpoint graphs) or vertices (de Bruijn graphs) in the resulting paths.

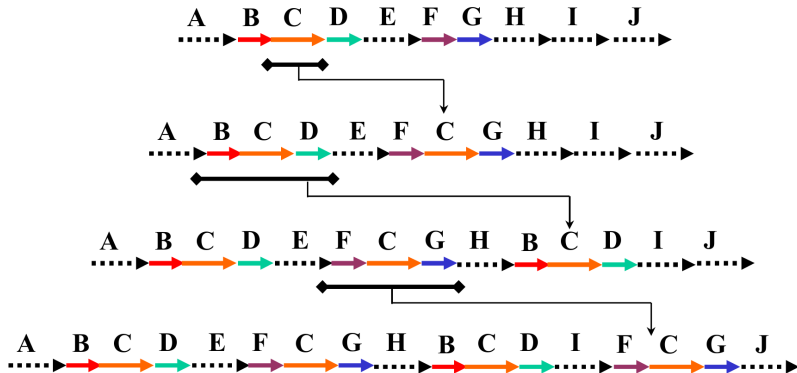
Difference repeat graph de Bruijn graph

- ▶ A-Bruijn graph (alignments) generalizes the de Bruijn graph
- ▶ We thus argue that the time has come to explain that the breakpoint graphs and the de Bruijn graphs are two identical data structures (if one ignores a cosmetic difference between them) as they both represent specific instances of a general notion of the A-Bruijn graph introduced in [13]. The A-Bruijn graphs are based on representing genomes as sets of labeled paths and further gluing identically labeled edges (breakpoint graphs) or vertices (de Bruijn graphs) in the resulting paths.
- ▶ de Bruijn graphs need correct bases

Difference repeat graph de Bruijn graph

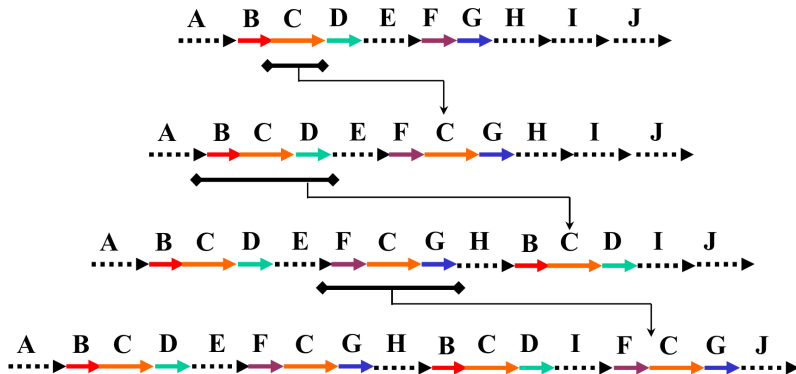
- ▶ A-Bruijn graph (alignments) generalizes the de Bruijn graph
- ▶ We thus argue that the time has come to explain that the breakpoint graphs and the de Bruijn graphs are two identical data structures (if one ignores a cosmetic difference between them) as they both represent specific instances of a general notion of the A-Bruijn graph introduced in [13]. The A-Bruijn graphs are based on representing genomes as sets of labeled paths and further gluing identically labeled edges (breakpoint graphs) or vertices (de Bruijn graphs) in the resulting paths.
- ▶ de Bruijn graphs need correct bases
- ▶ otherwise tangled graph

Segmental duplications



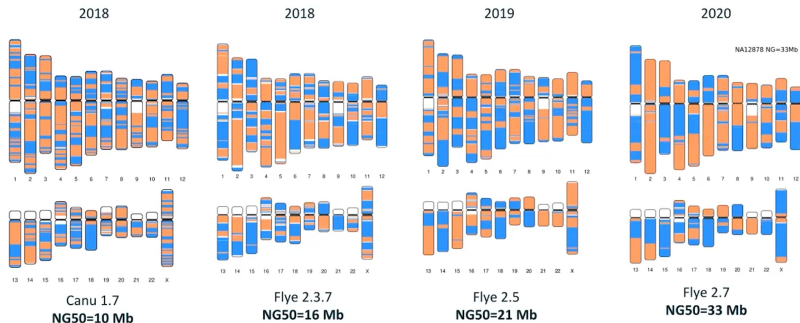
- ▶ Segmental duplications are duplicated blocks of genomic DNA typically ranging in size from 1-200 kb (IHGSC 2001)

Segmental duplications



- ▶ Segmental duplications are duplicated blocks of genomic DNA typically ranging in size from 1-200 kb (IHGSC 2001)
- ▶ They often contain sequence features such as high-copy repeats and gene sequences with intron-exon structure.

Contigity improvement



- colors are contigs
- color changes -> fragmented