# Assembly of long, error-pront reads using repeat graphs

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

June 28, 2021

Johannes Hausmann, Luis Kress

# Genome assembly in general

- reconstruct target sequence from the reads

# Genome assembly in general

- reconstruct target sequence from the reads
- different graph structures (De-Bruijn, Overlap-layout, String)

# Genome assembly in general

- reconstruct target sequence from the reads
- different graph structures (De-Bruijn, Overlap-layout, String)
- repeats → assembly fragmentation

# Genome assembly in general

- reconstruct target sequence from the reads
- different graph structures (De-Bruijn, Overlap-layout, String)
- repeats → assembly fragmentation
- error rate long read ↔ short read

# Genome assembly in general

- reconstruct target sequence from the reads
- different graph structures (De-Bruijn, Overlap-layout, String)
- repeats → assembly fragmentation
- error rate long read ↔ short read
- Flye should resolve these repeats correctly

# Disjointigs

- most assemblers spent much time on correct contig assembly

# Disjointigs

- most assemblers spent much time on correct contig assembly
- Flye uses a different approach [1]:

# Disjointigs

- most assemblers spent much time on correct contig assembly
- Flye uses a different approach [1]:
    - we don't care (at least at the initial stage)

# Disjointigs

- most assemblers spent much time on correct contig assembly
- Flye uses a different approach [1]:
  - we don’t care (at least at the initial stage)
  - correct assembly graph

# Disjointigs

- most assemblers spent much time on correct contig assembly
- Flye uses a different approach [1]:
  - we don't care (at least at the initial stage)
  - correct assembly graph
- generate paths from overlapping reads without checking for correct assembly → disjointigs

# Repeat Graph Creation

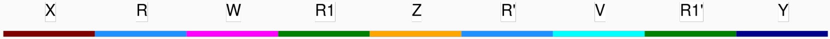


**Figure 1:** Example Genome

# Repeat Graph Creation

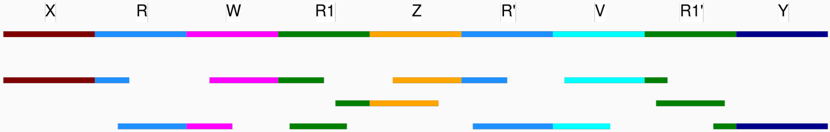


**Figure 2:** Example Genome and Reads

# Repeat Graph Creation

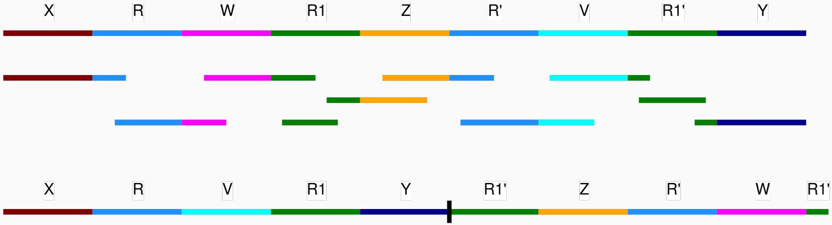


**Figure 3:** Example Genome, Reads and Disjointigs

# Repeat Graph Creation

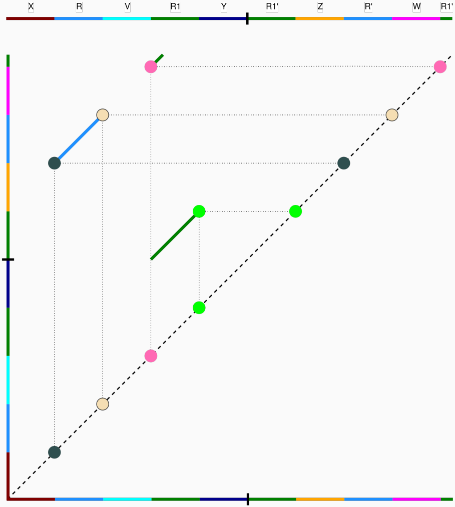


**Figure 4:** Breakpoint Graph

# Repeat Graph Creation

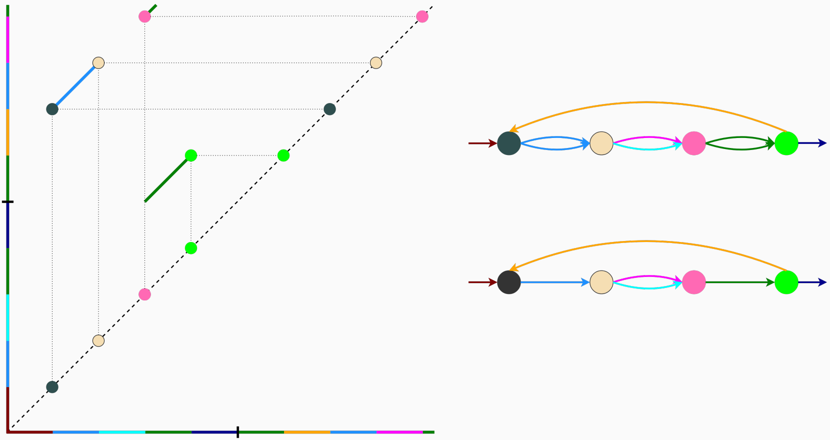


**Figure 5:** Repeat Graph

# Results

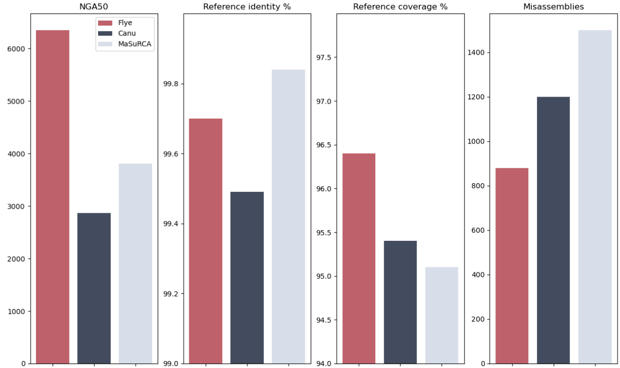


**Figure 6:** Results for HUMAN testset

# References i

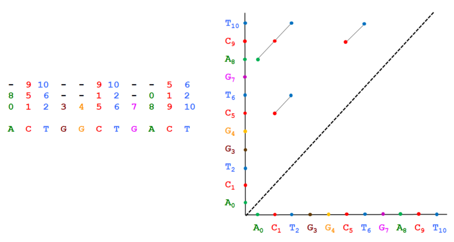
M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner.
**Assembly of long, error-prone reads using repeat graphs.**
*Nature Biotechnology*, 37(5):540–546, May 2019.

# Git (presentation)

**Figure 7:** Link to our git repo

# 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 |

$T_{10}$
$C_9$
$A_8$
$G_7$
$T_6$
$C_5$
$G_4$
$G_3$
$T_2$
$C_1$
$A_0$
$A_0$ $C_1$ $T_2$ $G_3$ $G_4$ $C_5$ $T_6$ $G_7$ $A_8$ $C_9$ $T_{10}$

**Figure 8:** Dot plot creation

# Repeat graphs

- generalization of de bruijn graphs

# Repeat graphs

- generalization of de bruijn graphs
- structure

# Repeat graphs

- generalization of de bruijn graphs
- structure
- creation

# Repeat graphs

- generalization of de bruijn graphs
- structure
- creation
    - from disjointigs = random walk of reads on the repeat graph

# Repeat graphs

- generalization of de bruijn graphs
- structure
- creation
  - from disjointigs = random walk of reads on the repeat graph
  - 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

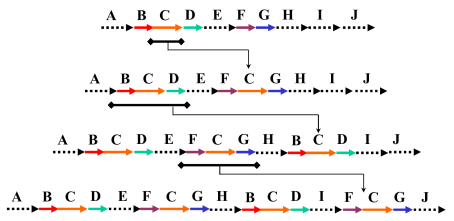

A B C D E F G H I J
A B C D E F C G H I J
A B C D E F C G H B C D I J
A B C D E F C G H B C D I F C G J

**Figure 9:** Segmental Duplications

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

# Segmental duplications

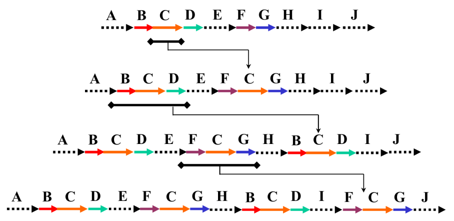

A B C D E F G H I J
A B C D E F C G H I J
A B C D E F C G H B C D I J
A B C D E F C G H B C D I F C G J

**Figure 9:** 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

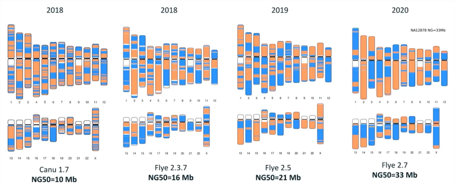


**Figure 10:** Contigity improvements

- colors are contigs

# Contigity improvement

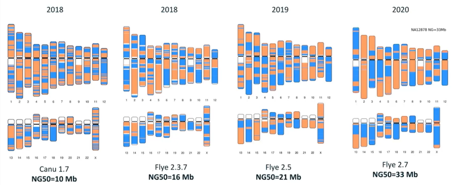


**Figure 10:** Contigity improvements

- colors are contigs

- colors are contigs

# Contigity improvement

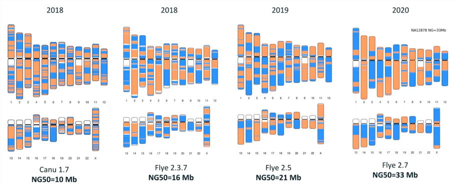


**Figure 11:** Contigity improvements

- colors are contigs

# Contigity improvement

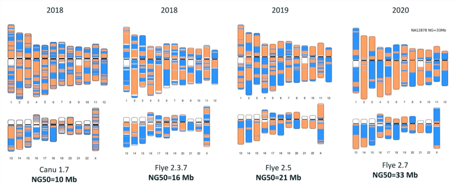


**Figure 11:** Contigity improvements

- colors are contigs
- colors are contigs