GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Assembly of long, error-prone reads using repeat graphs

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

June 30, 2021

Johannes Hausmann, Luis Kress

Background

- Assembly: reconstruct target sequence from the reads

Background

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

Background

- Assembly: reconstruct target sequence from the reads
- Different assemblers, different graph structures (De-Bruijn, Overlap-layout, String)
- Repeats $\rightarrow$ assembly fragmentation

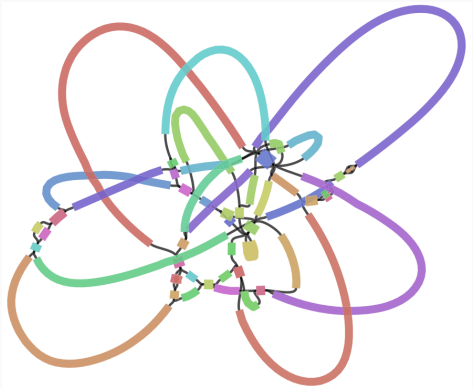


Figure 1: Tangled assembly graph[1]

Background

- Assembly: reconstruct target sequence from the reads
- Different assemblers, different graph structures (De-Bruijn, Overlap-layout, String)
- Repeats $\rightarrow$ assembly fragmentation

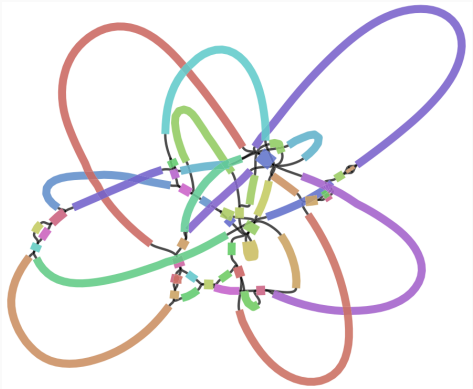


Figure 1: Tangled assembly graph[1]

- Error rate long read $\leftrightarrow$ short read

Background

- Assembly: reconstruct target sequence from the reads
- Different assemblers, different graph structures (De-Bruijn, Overlap-layout, String)
- Repeats $\rightarrow$ assembly fragmentation

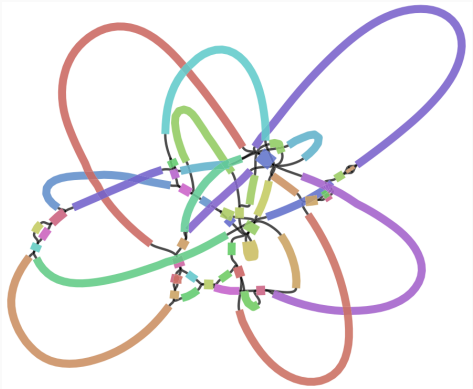


Figure 1: Tangled assembly graph[1]

- Error rate long read $\leftrightarrow$ short read
- Flye $\rightarrow$ resolve these repeats correctly, create contiguous assemblies

Disjointigs

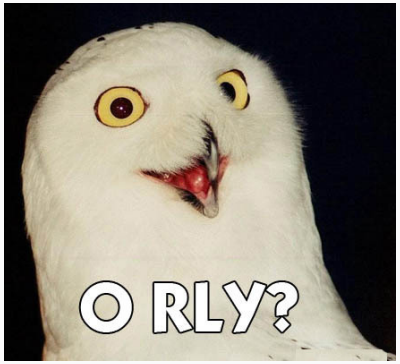
- Most assemblers spend much time on correct contig assembly

Disjointigs

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

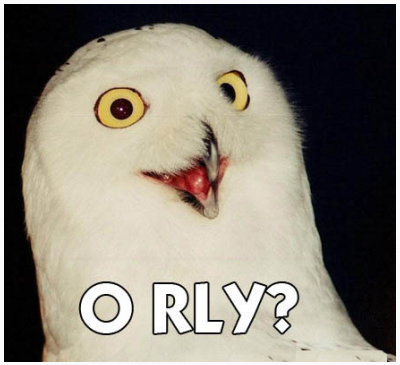
Disjointigs

- Most assemblers spend much time on correct contig assembly
- Flye uses a different approach [1]:
  - we don't care



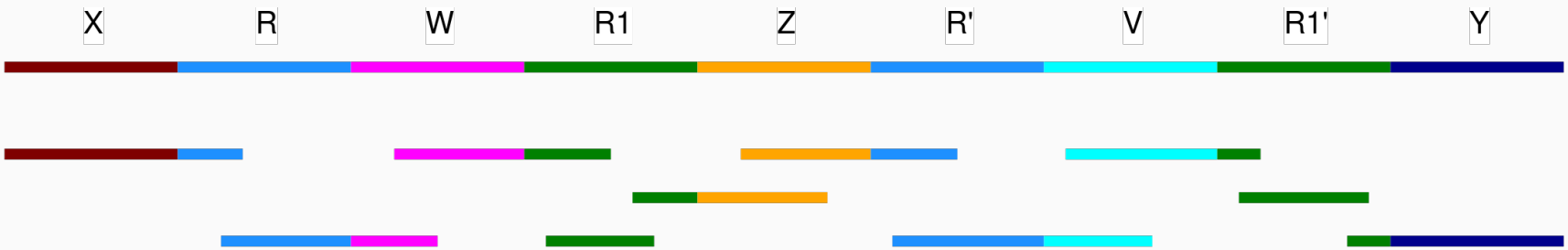
Disjointigs

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

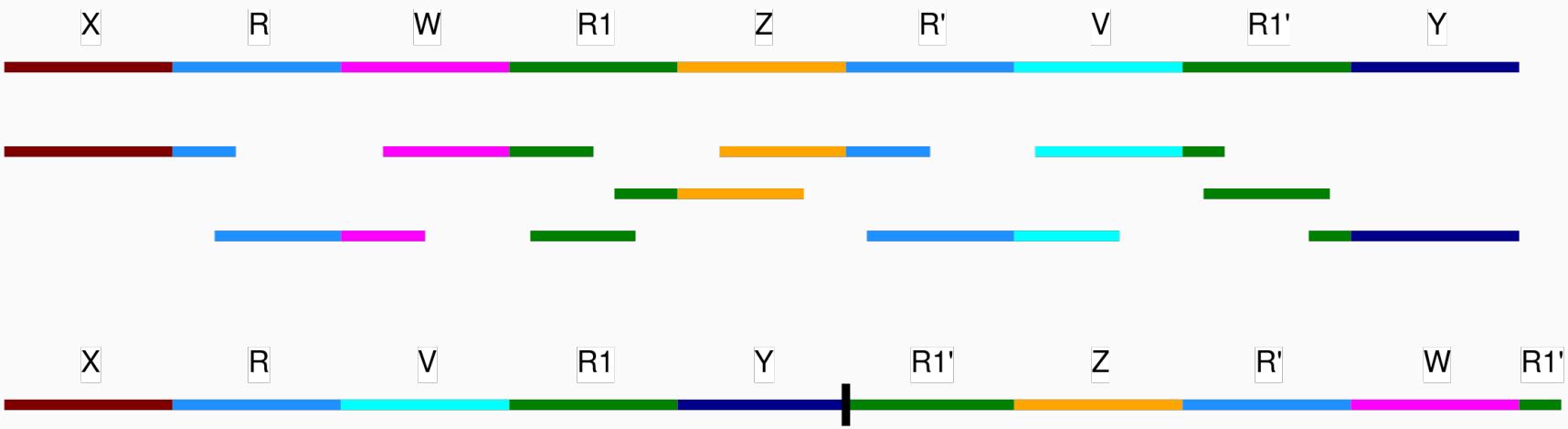


- Generate paths from overlapping reads without checking for correct repeat resolution $\rightarrow$ Disjointigs

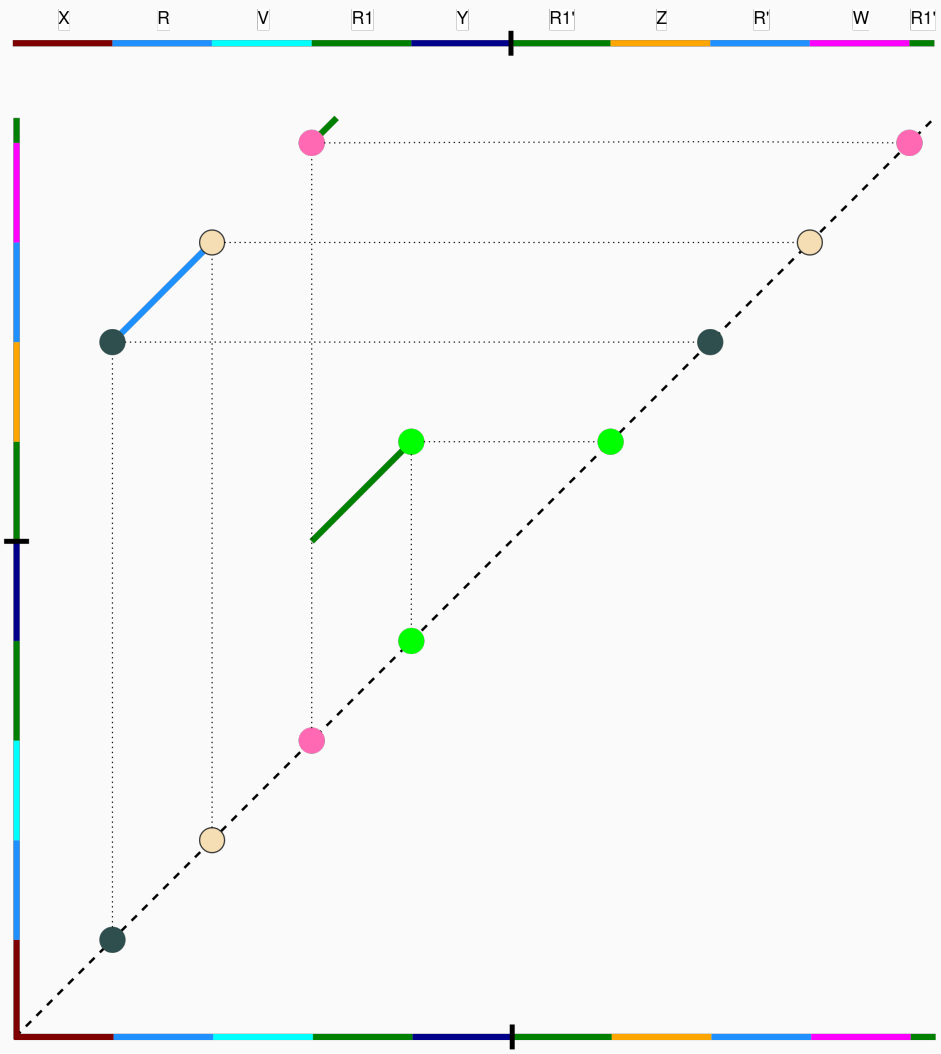
Repeat Graph Creation



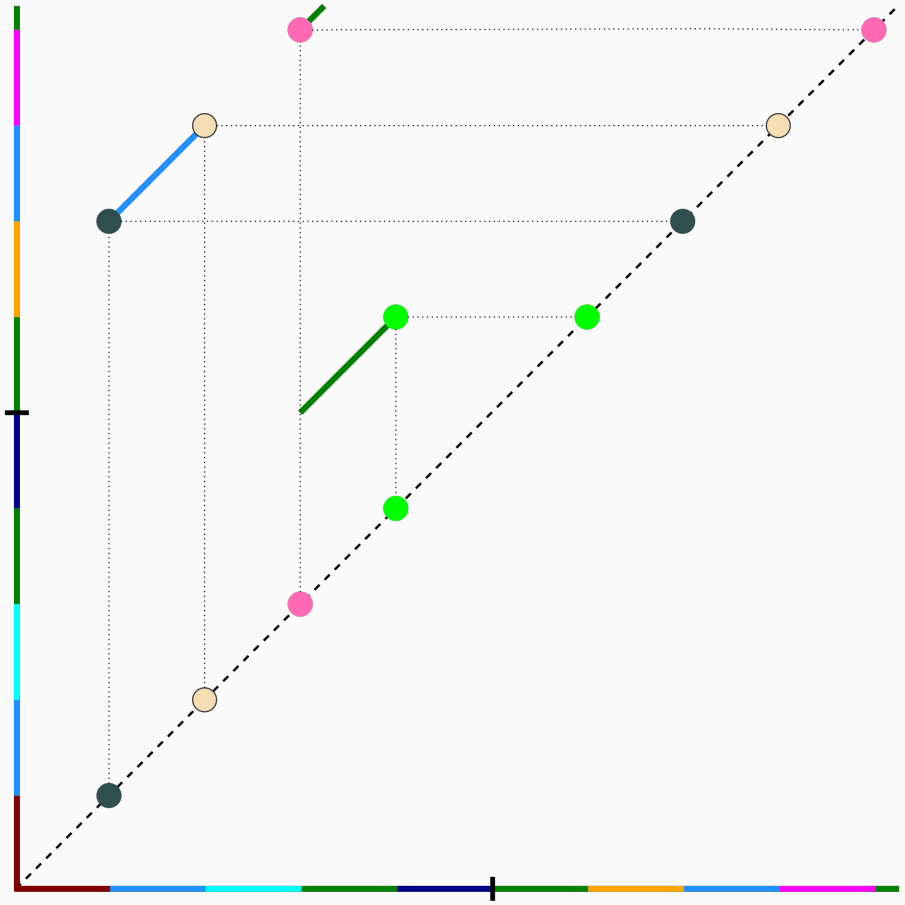
Repeat Graph Creation

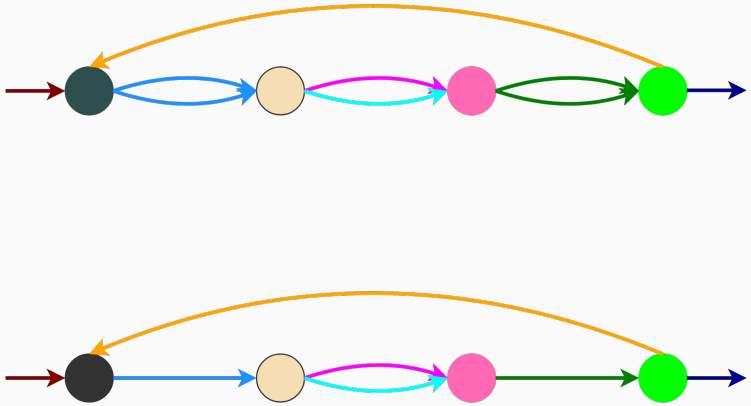


Repeat Graph Creation



Repeat Graph Creation





Results

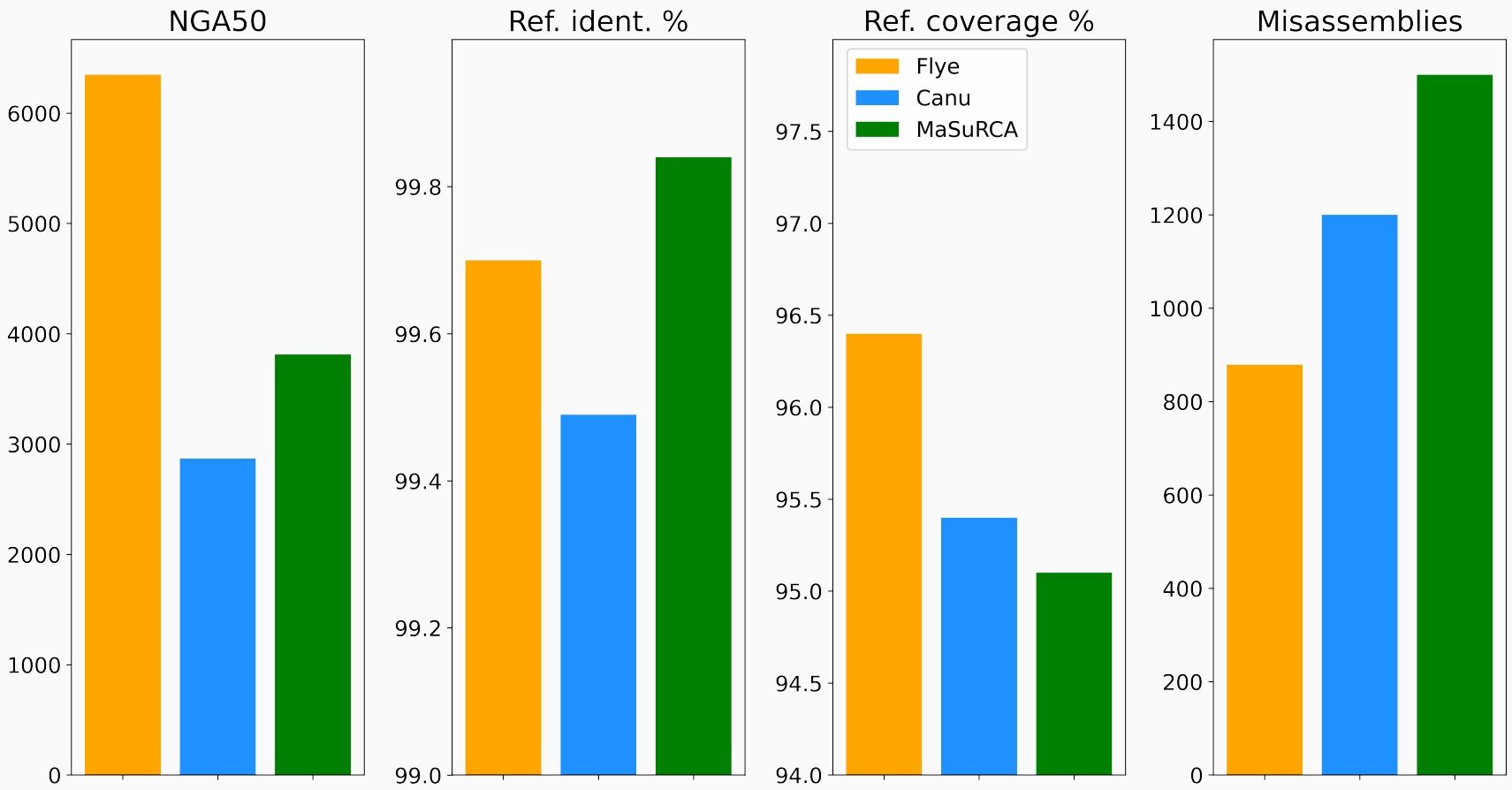


Figure 2: Results for HUMAN testset

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.


P. A. Pevzner, P. A. Pevzner, H. Tang, and G. Tesler.
De novo repeat classification and fragment assembly.
Genome Research, 14(9):1786–1796, Sept. 2004.


B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl.
Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.
PloS One, 9(11):e112963, 2014.

References ii

Git (presentation and poster)

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Assembly of long, error-prone reads using repeat graphs

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

June 30, 2021

Johannes Hausmann, Luis Kress

Appendix

Dot plot creation

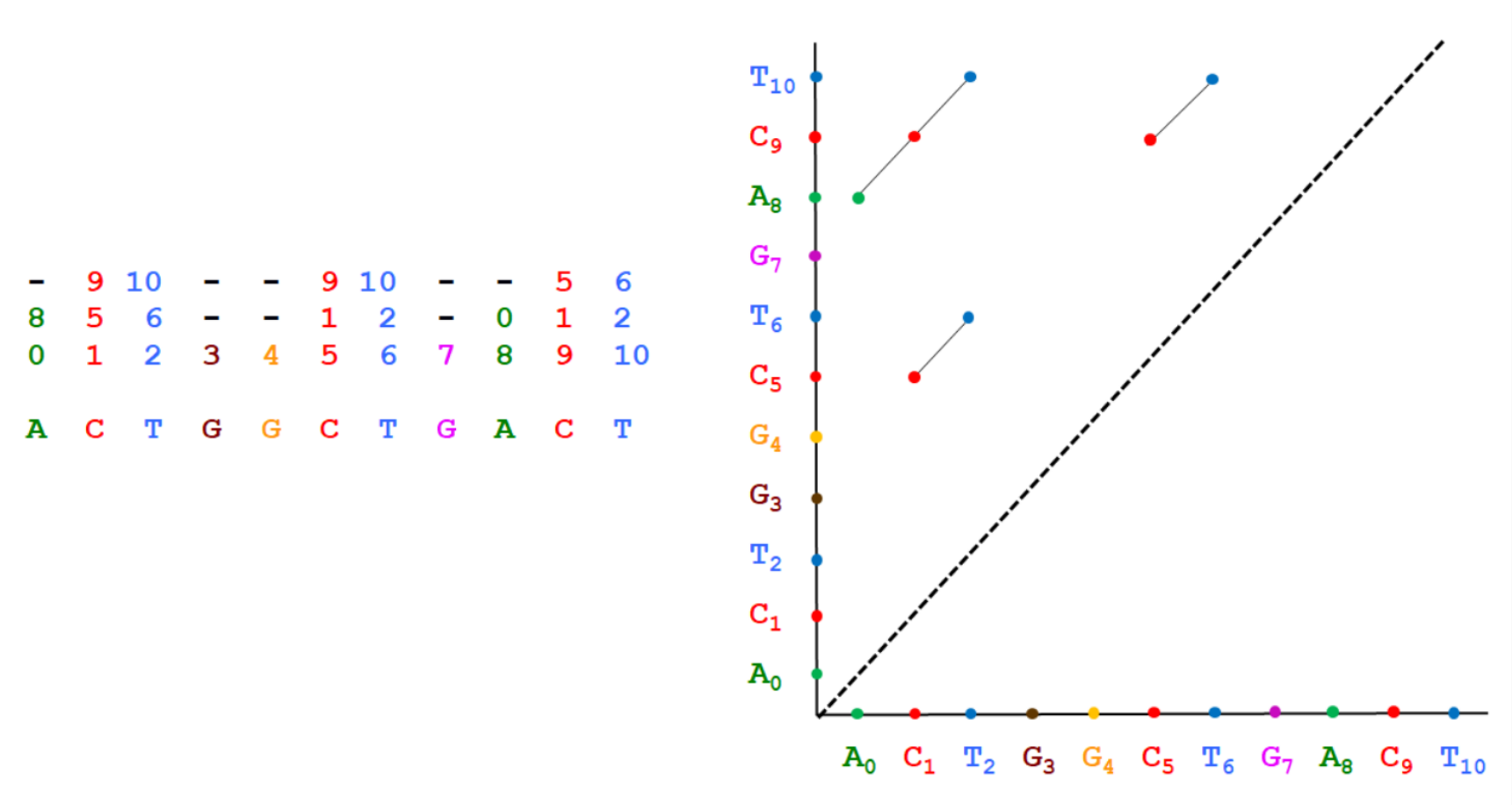


Figure 3: Dot plot creation

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
- "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." [2]
- de Bruijn graphs need correct bases
- otherwise tangled graph

Segmental duplications

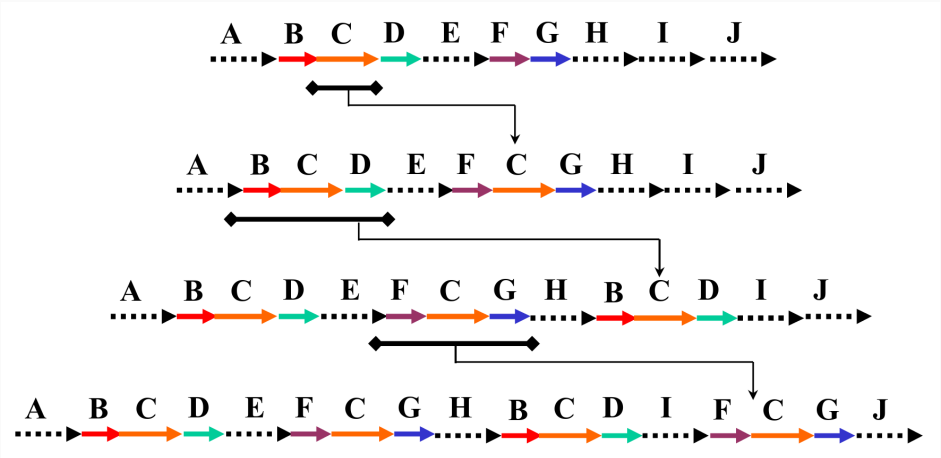


Figure 4: Segmental Duplications

- 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

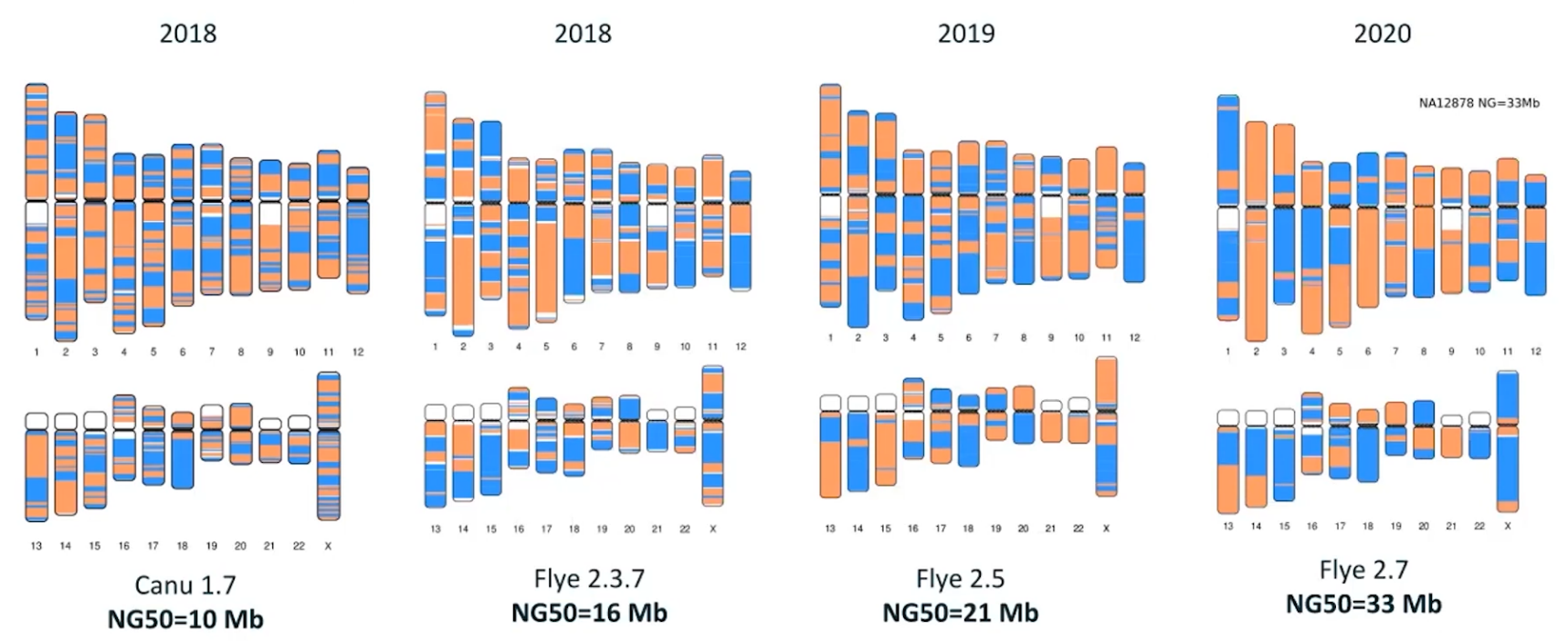


Figure 5: Contigity improvements

- colors are contigs $\rightarrow$ change in color means fragmentation