GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Assembly of long, error-prone reads using repeat graphs

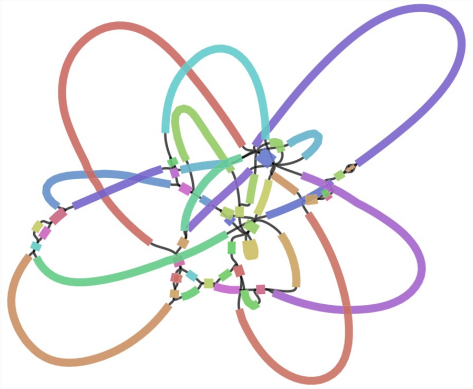
Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel A. Pevzner

July 5, 2021

Presented by: Johannes Hausmann, Luis Kress

Background

- Assembly: reconstruct target sequence from the reads
- Different assemblers, different graph structures
- Repeats $\rightarrow$ assembly fragmentation



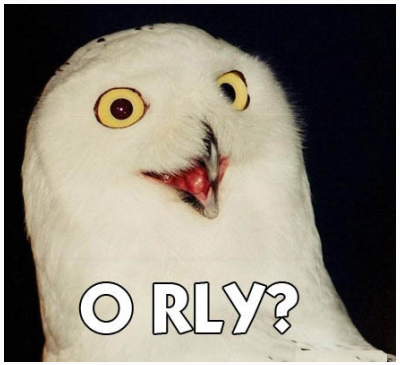
Tangled assembly graph of *E.coli* [2]

- Small differences between repeat copies $\rightarrow$ hard to resolve with error-prone reads

Disjointigs

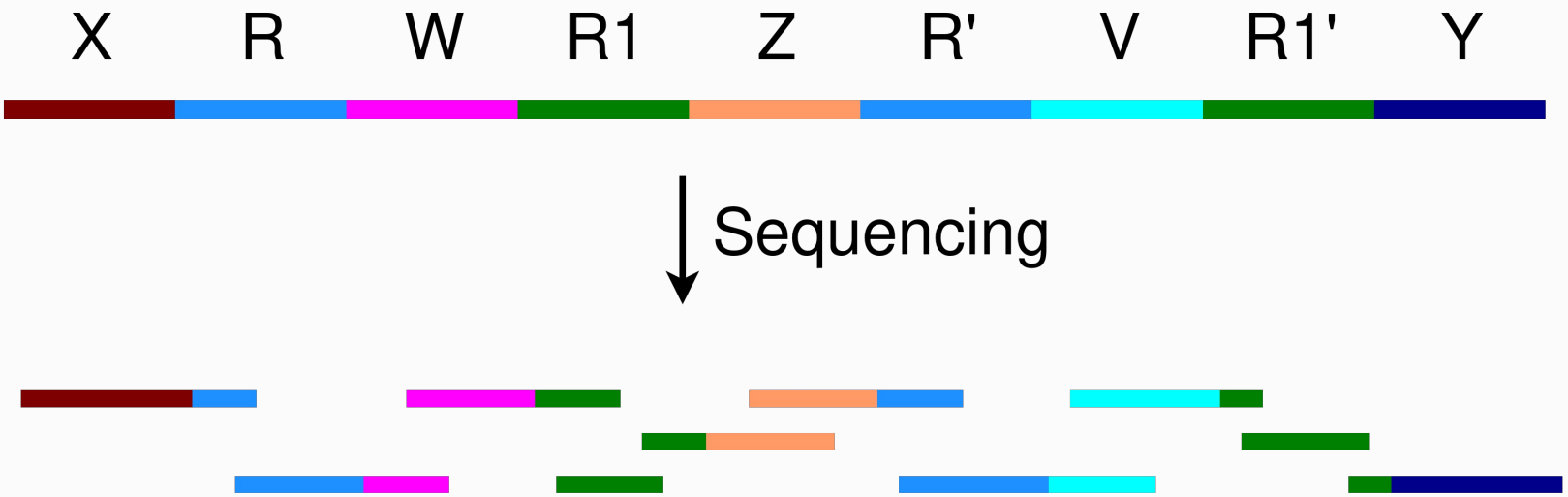
- Most assemblers spend much time on correct contig assembly
- Flye uses a different approach:
  - we don't care about the correct contig assembly

    $\rightarrow$ at least at the initial stage

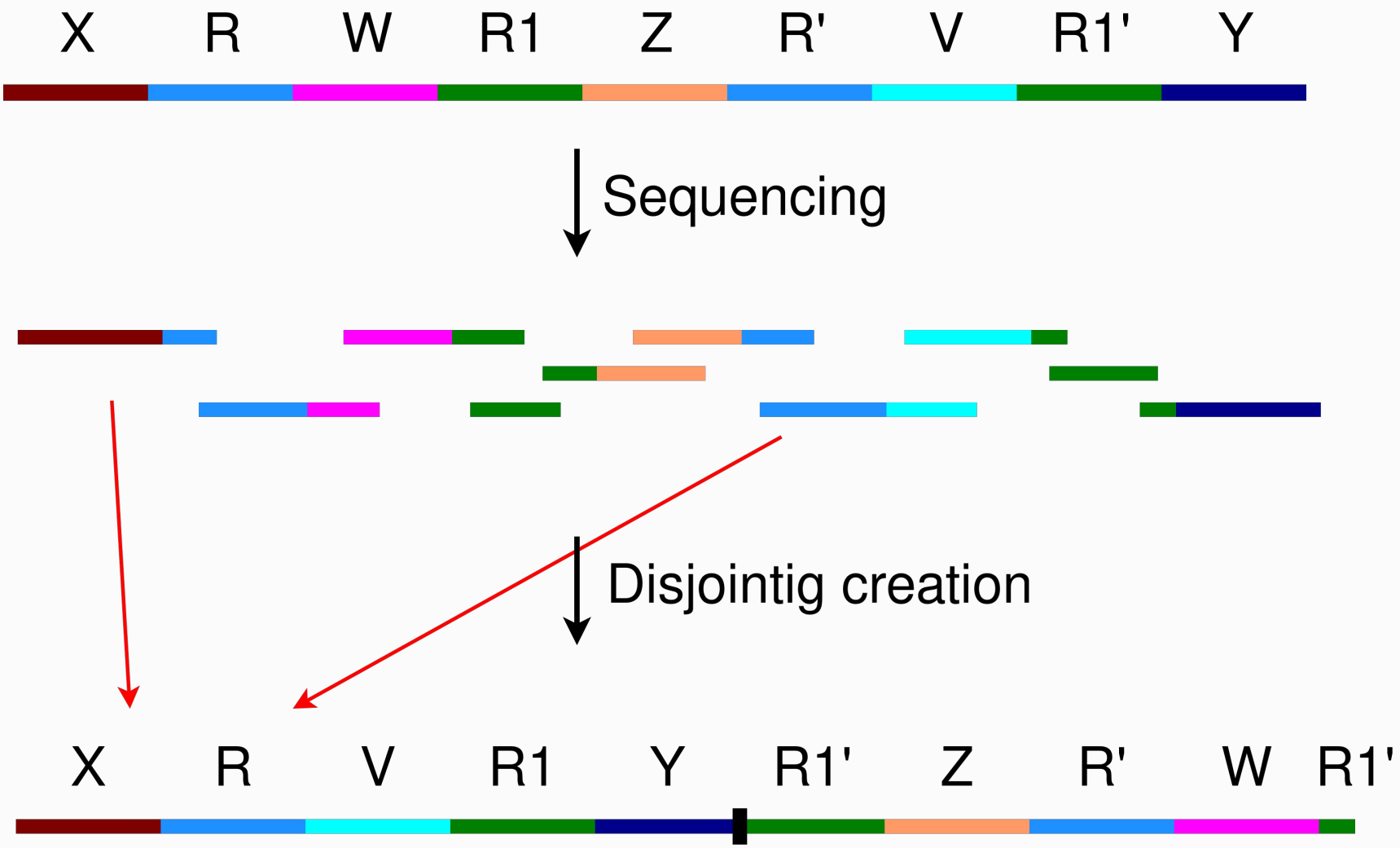


- Generate arbitary paths from overlapping reads $\rightarrow$ Disjointigs

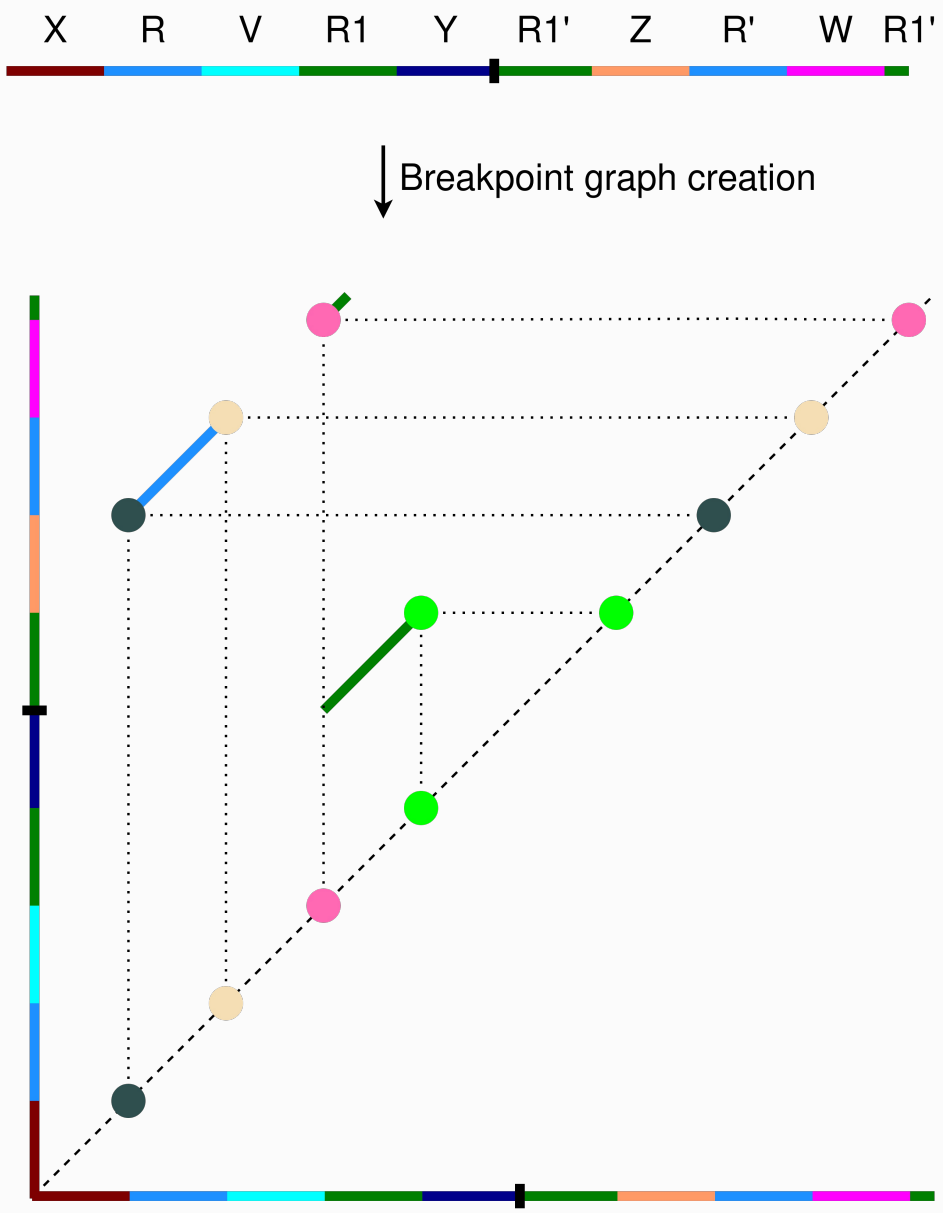
Repeat Graph Creation



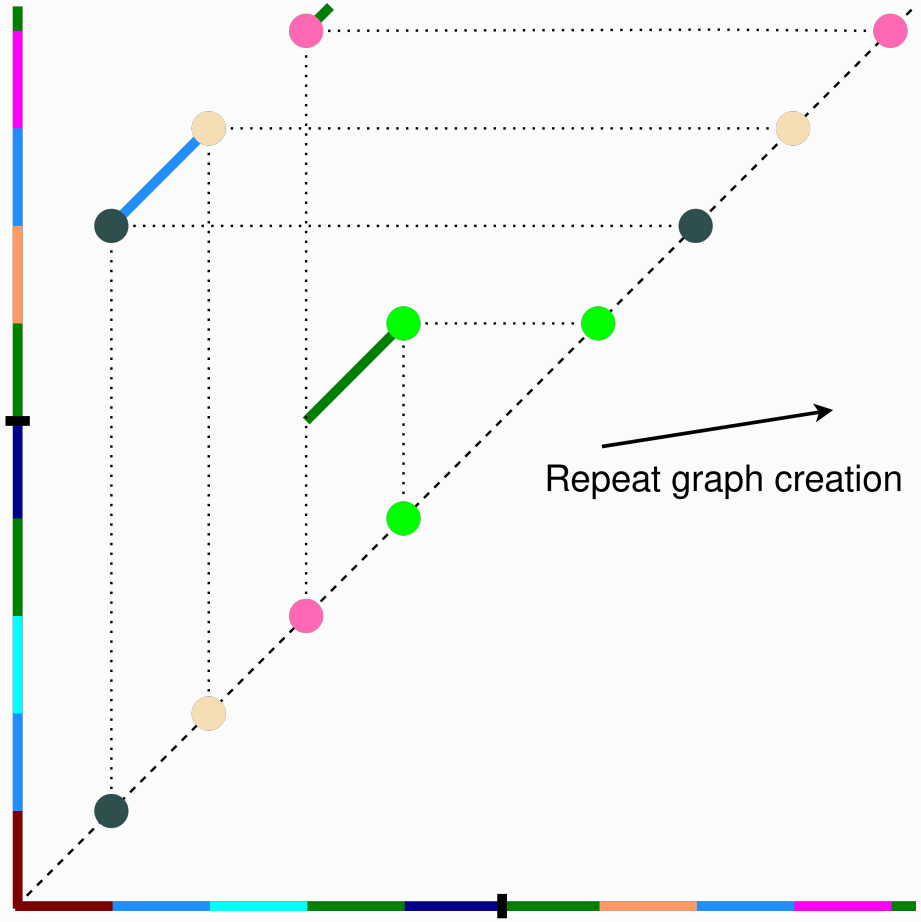
Repeat Graph Creation



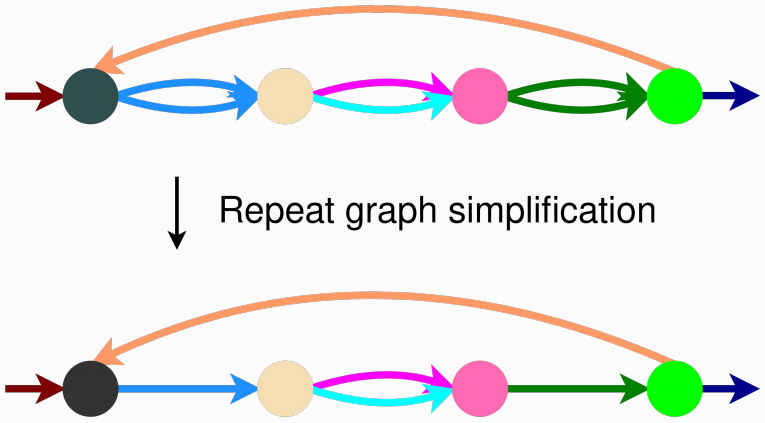
Repeat Graph Creation



Repeat Graph Creation



Repeat graph creation



Repeat graph simplification

Results

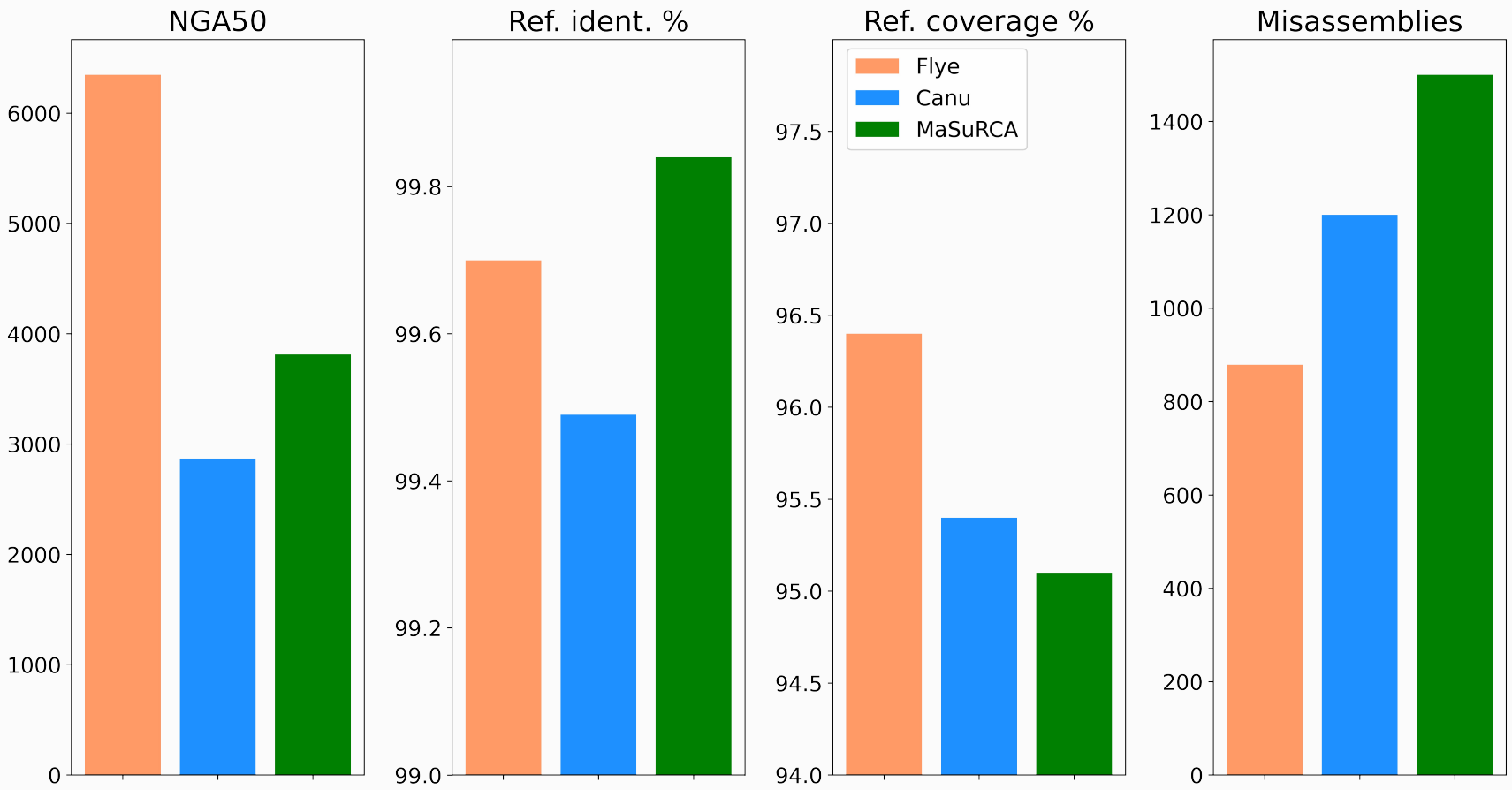


Figure 1: Results for HUMAN testset

Git (presentation and poster)

https://github.com/LKress/ASA

References i


A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler.
QUAST: quality assessment tool for genome assemblies.
Bioinformatics (Oxford, England), 29(8):1072–1075, Apr. 2013.


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.


Y. Lin, S. Nurk, and P. A. Pevzner.
What is the difference between the breakpoint graph and the de Bruijn graph?
BMC genomics, 15 Suppl 6:S6, 2014.

References ii

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 iii


A. V. Zimin, G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke.
The MaSuRCA genome assembler.
Bioinformatics, 29(21):2669–2677, Nov. 2013.