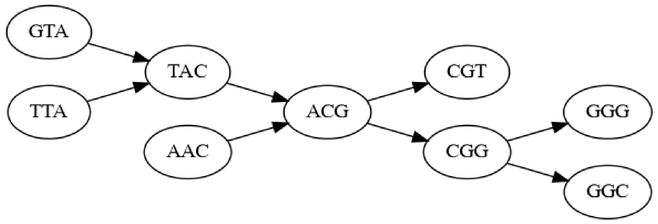
RResolver: short read repeat resolution with sliding window

Vladimir Nikolić [1] Justin Chu [1] Johnathan Wong [1] Lauren Coombe [1] Ka Ming Nip [1] René Warren [1] Inanç Birol [1]

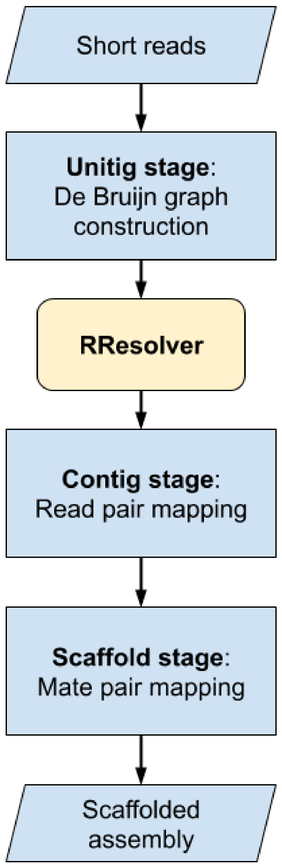
[1]Canada's Michael Smith Genome Sciences Centre at BC Cancer, Vancouver, BC V5Z 4S6, Canada

Background

- *De novo* short read assemblers commonly use de Bruijn graphs (DBG) where node sequences are of same length, specified with parameter k.



- K parameter selection is a trade-off between connectivity and contiguity, but is inherently a crude approach that only works on average. Lesser covered regions need lower k value to preserve connectivity while well-covered regions benefit with better contiguity from high k value.



- RResolver is a pluggable component into a short read assembler. The figure to the left shows important stages of a short read assembler — in this case ABySS, and where RResolver fits.

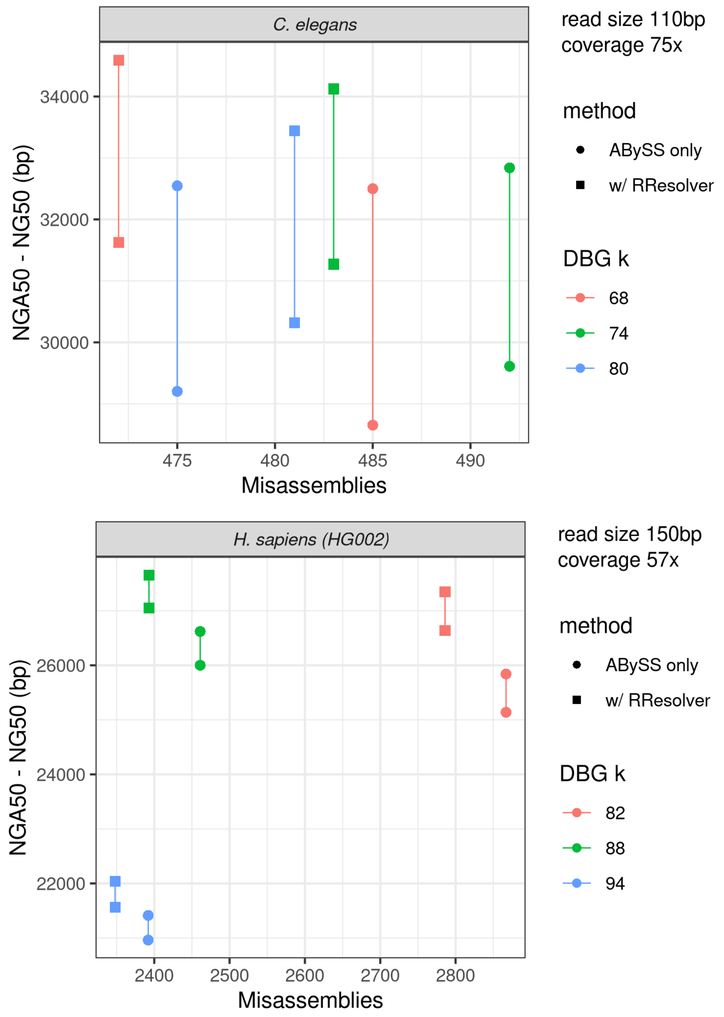
- RResolver improves upon the DBG by resolving repeats in well-covered regions using a k value larger than the one used to construct DBG, increasing contiguity.

- Code can be found at:

github.com/bcgsc/abyss/tree/master/RResolver

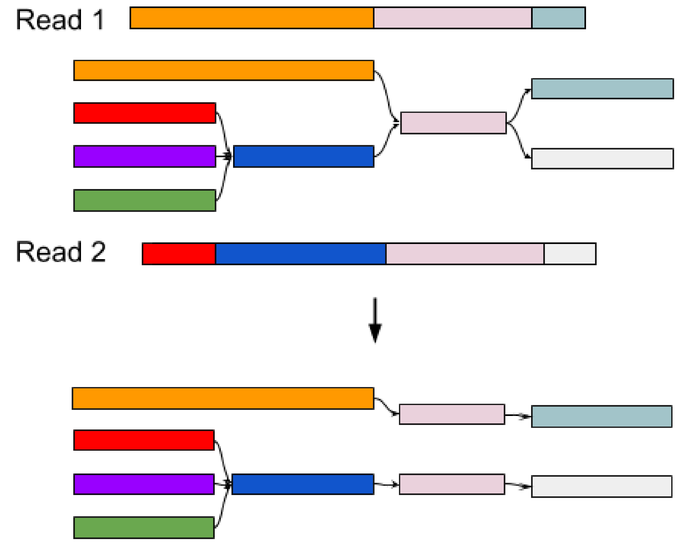
Results

The following plots show assembly quality improvements at scaffold stage.

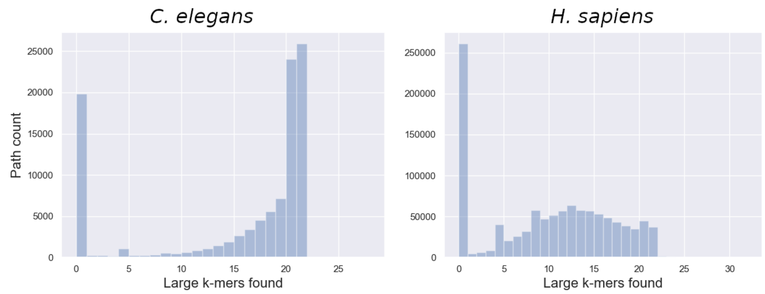


Algorithm

- A large k value is dynamically determined close to read size.
- A Bloom filter is constructed from short reads, storing large k-mers.
- A window of large k size is slided along all possible paths of a repeat, in an attempt to find large k-mers.



- Paths that find no large k-mers along them are considered unsupported and removed.
- To deal with Bloom filter false positives, a number of large k-mers need to be found along a path.



Conclusions

- Right choice for k parameter varies depending on the local coverage.
- Revisiting well-covered regions of the graph with a larger k value improves assemblies.
- Using Bloom filters to reduce memory usage is viable.

Funding

Genome BritishColumbia
GenomeCanada
National Institutes of Health
NSERC CRSNG