RAmbler: *de novo* genome assembly of complex repetitive regions

Sakshar Chakravarty
schak026@ucr.edu
Department of Computer Science and Engineering, University of California Riverside, California, USA

Glennis Logsdon
glogsdon@uw.edu
Department of Genome Sciences, University of Washington Seattle, Washington, USA

Stefano Lonardi*
stelo@cs.ucr.edu
Department of Computer Science and Engineering, University of California Riverside, California, USA

ABSTRACT

Complex repetitive regions (also known as *segmental duplications*) in eukaryotic genomes often contain essential functional and regulatory information. Despite remarkable algorithmic progress in genome assembly in the last twenty years, modern *de novo* assemblers still struggle to accurately reconstruct these highly repetitive regions. When sequenced reads will be long enough to span all repetitive regions, the problem will be solved trivially. However, even the third generation of sequencing technologies on the market cannot yet produce reads that are sufficiently long (and accurate) to span every repetitive region in large eukaryotic genomes.

In this work, we introduce a novel algorithm called RAmbler to resolve complex repetitive regions based on high-quality long reads (i.e., PacBio HiFi). We first identify potentially repetitive regions by mapping the HiFi reads to the draft genome assembly. Regions with sequencing coverage much higher then the average indicate a collapsed repeat in the assembly. Then, (i) we select HiFi reads that map to the repetitive region, including 50 kb upstream and downstream, (ii) we compute the k-mers that are expected to occur only once in the genome (i.e., single copy k-mers, which we call *unikmers*), (iii) we barcode the HiFi reads based on the presence and the location of their uni$k$mers, (iv) we compute an overlap graph solely based on shared barcodes, (v) we reconstruct the sequence of the repetitive region by traversing the overlap graph.

Our statistical analysis of the frequency distribution of k-mers from PacBio HiFi and Oxford Nanopore reads for different choices of k shows that the k-mers occurring in a window of three standard deviations around the mean of the distribution are true uni$k$mers with very high probability, i.e., they occur in the genome only once with high probability.

We present an extensive set of experiments comparing the performance of RAmbler against Hifiasm, HiCANU and Verkko on synthetic HiFi reads generated over a wide range of repeat lengths, number of repeats, heterozygosity rates and depth of sequencing (over 140 data sets). Our experimental results indicate that RAmbler outperforms Hifiasm, HiCANU and Verkko on the large majority of the inputs. We also show that RAmbler can resolve several long tandem repeats in *Arabidopsis thaliana* using real HiFi reads. In order to compare the performance of different assemblers, we introduce a new metric called *assembly-score* that takes into account completeness, contiguity and accuracy in a single number on a scale $[0, 1]$.

In summary, our contributions are as follows. (i) We carry out the first statistical analysis for the method to extract single-copy k-mers for PacBio HiFi and Oxford Nanopore reads; (ii) we introduce a new assembly method that takes advantage of single-copy k-mers to resolve complex repetitive regions; (iii) we compare our method against state-of-the-art assemblers (Hifiasm, HiCANU and Verkko) on more than 140 synthetic data sets; (iv) we show that our method can resolve complex repetitive regions in *A. thaliana* using real HiFi data; (v) we introduce a new quality metric for genome assemblies that incorporates completeness, contiguity and the number of misassemblies. While the initial results are very promising, to establish RAmbler's true ability in resolving complex repeats, it will have to be tested more extensively on real data for other organisms.

The code for RAmbler is available at https://github.com/sakshar/repeat_assembler.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; **Computational genomics**; *Molecular sequence analysis*; • **Theory of computation** → Graph algorithms analysis.

KEYWORDS

Genome assembly, *de novo*, repetitive regions, segmental duplications, sequencing, PacBio HiFi, tandem repeats, k-mers, overlap graph, clustering, connected components

ACM Reference Format:
Sakshar Chakravarty, Glennis Logsdon, and Stefano Lonardi. 2023. RAmbler: *de novo* genome assembly of complex repetitive regions. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3584371.3612971

ACKNOWLEDGMENTS

This project was supported in part by NSF IIS #1814359, NSF CBET #2225878, and NIH 1-R01-AI169543-01 to SL.

*Corresponding author

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09.
https://doi.org/10.1145/3584371.3612971