SpLitteR: Diploid genome assembly using linked TELL-Seq reads and assembly graphs

Ivan Tolstoganov[1,2*], Pavel A. Pevzner[3], Anton Korobeynikov[1,4]

[1] Center for Algorithmic Biotechnology, Saint Petersburg State University, Saint Petersburg, 199004, Russia
[2] Department of Mathematics, Science for Life Laboratory, Stockholm University, 106 91, Stockholm, Sweden.
[3] Department of Computer Science and Engineering, University of California, San Diego, San Diego CA, USA
[4] Department of Statistical Modelling, Saint Petersburg State University, Saint Petersburg, 198504, Russia
*To whom correspondence should be addressed. Email: ivan.tolstoganov@math.su.se

Abstract

Recent advances in long-read sequencing technologies enabled accurate *de novo* assembly of large genomes and metagenomes. However, even long and accurate high-fidelity (HiFi) reads do not resolve repeats that are longer than the read lengths. This limitation significantly affects the contiguity of diploid genome assemblies since two haplomes share many long identical regions. To generate telomere-to-telomere assemblies of diploid genomes, biologists use additional experimental technologies, such as linked reads or ultralong Oxford Nanopore reads. In particular, the barcoded linked-reads, generated using an inexpensive TELL-Seq technology, provide an attractive way to bridge unresolved repeats in phased assemblies of diploid genomes.

Here, we present a SpLitteR tool for haplotype phasing and scaffolding in an assembly graph using barcoded linked-reads. We benchmark SpLitteR on assembly graphs produced by various long-read assemblers and TELL-Seq reads and show that it improves upon the state-of-the-art linked-read scaffolders in the accuracy and contiguity metrics.

Introduction

Recently developed barcoding linked-read technologies, such as stLFR (McElwain *et al.*, 2017), TELL-Seq (Chen *et al.*, 2020), and LoopSeq (Callahan *et al.*, 2021), are based on co-barcoding of short reads from the same long DNA fragment. They start with the distribution of long DNA fragments over a set of containers marked by a unique barcode. Afterward, long fragments within the containers are sheared into shorter fragments and sequenced. The resulting library consists of short reads marked by the barcode corresponding to the set of long fragments.

Various tools, such as Athena (Bishara *et al.*, 2018), cloudSPAdes (Bishara *et al.*, 2018; Tolstoganov *et al.*, 2019), Supernova (Weisenfeld *et al.*, 2017), and TuringAssembler (Chen *et al.*, 2020), were developed to generate *de novo* genome assembly from linked-reads alone. However, even though linked short reads result in more contiguous assemblies than assemblies based on non-linked short reads, all these tools generate highly fragmented assemblies of large genomes and metagenomes. For large genomes and metagenomes, long high-fidelity (HiFi) reads proved to be useful in generating highly-accurate and contiguous assemblies (Kolmogorov *et al.*, 2019; Bankevich *et al.*, 2022; Nurk *et al.*, 2020; Rautiainen *et al.*, 2022; Cheng *et al.*, 2021; Kolmogorov *et al.*, 2020). However, even though HiFi reads enabled the first complete assembly of the human genome by the Telomere-to-Telomere (T2T) consortium (Nurk *et al.*, 2022), HiFi assemblies do not resolve some long repeats and thus are often scaffolded using supplementary technologies, such as Hi-C reads, Oxford Nanopore (ONT) ultralong reads, and Strand-seq reads (Nurk *et al.*, 2022). Scaffolding methods based on inexpensive linked-read sequencing can be used as a viable alternative to other supplementary technologies by combining the low cost of short reads and the long-range information encoded by linked reads originating from the same barcoded fragment.

Although the state-of-the-art linked-read scaffolders, such as Architect (Kuleshov *et al.*, 2016), ARKS (Coombe *et al.*, 2018), SLR-superscaffolder (Guo *et al.*, 2021), and Physlr (Afshinfard *et al.*, 2022), improve the contiguity of HiFi assemblies, they do not take advantage of the assembly graph and thus ignore important connectivity information. In addition, these tools are not applicable to diploid assemblies, limiting their applicability to large genomes and complex metagenomes with many similar strains.

We present the SpLitteR tool that uses linked reads to improve the contiguity of HiFi assemblies. In contrast to existing linked-reads scaffolders, it utilizes the assembly graph and supports diploid assemblies. Given a linked-read library and a HiFi assembly graph in the GFA format, SpLitteR resolves repeats in the assembly graph using linked-reads and generates a simplified (more contiguous) assembly graph with corresponding scaffolds. SpLitteR is implemented in C++ as a part of the freely available SPAdes package and is available at https://cab.spbu.ru/software/splitter.

Methods

SpLitteR is a tool for resolving repeats in the input assembly graph using Tell-Seq data. First, the barcoded TELL-Seq reads are mapped to the edges of the assembly graph and a set of barcodes from uniquely mapped reads is stored for each edge. A vertex in a graph is classified as *branching* if both its in-degree and out-degree exceed 1 (each branching vertex in the contracted graph represents a *repeat vertex*). In a nutshell, the repeat resolution procedure identifies branching vertices in the preprocessed assembly graph (pre-processing details are described in Supplementary Sections Data preparation and Assembly graph transformation). Given an incoming edge *e* into a vertex *v*, SpLitteR attempts to find a *matching* outgoing edge *next*(*e*) from this vertex that immediately follows *e* in the (unknown) genomic traversal of the assembly graph. To find *next(e)*, it analyzes all linked reads that map to both the in-edge *e* and

all out-edges from *v* (see Supplementary Sections Aligning barcoded reads and Repeat resolution for details). SpLitteR outputs the results of the repeat resolution procedure both as the set of scaffolds and as the simplified assembly graph, where the resolved vertices are split and each matched pair of in-edges and out-edges is merged into a single edge. The repeat resolution procedure has both diploid and metagenomic modes.

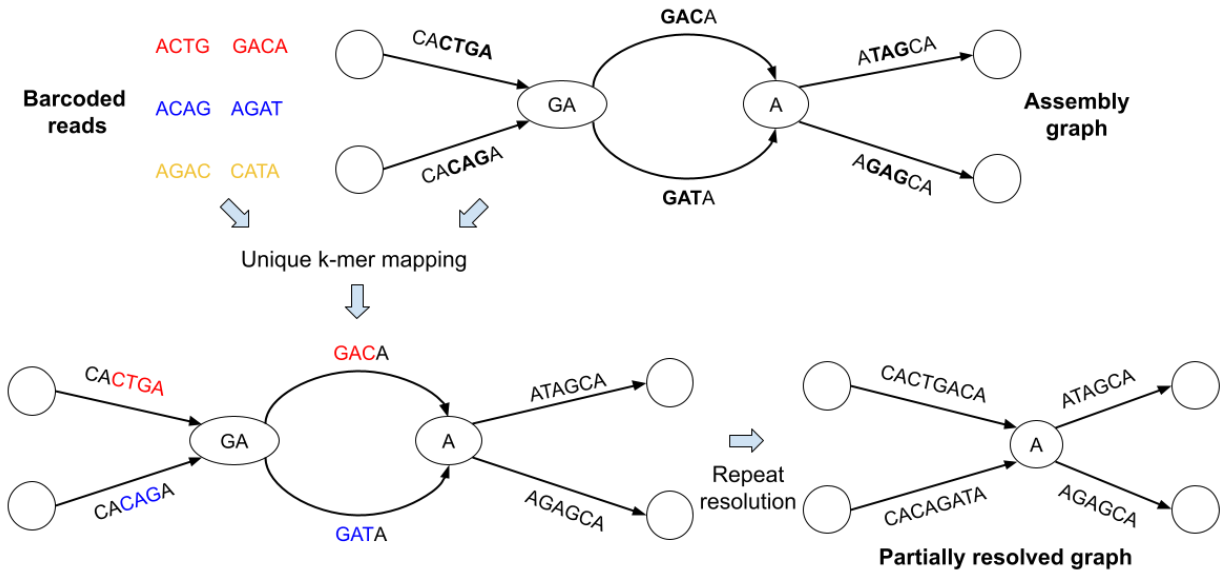

Figure 1. Outline of the SpLitteR workflow. Reads with the same barcode are represented by the same color. Red and blue reads are mapped to the assembly graph based on their *k*-mers (*k*=3 for this toy example). Only unique 3-mers which occur only once in the assembly graph edges are used for mapping. Since yellow reads do not contain a unique 3-mer. SpLitteR first contracts the input multiplex de Bruijn graph by contracting short edges, maps the barcoded reads to uncontracted edges, and resolves the vertices in the multiplex de Bruijn graph by assigning in-edges to their matching out-edges based on the barcode information. Variants within the contracted subgraphs remain unresolved.

Results

SpLitteR (0.1) was benchmarked against ARKS 1.2.4 (Coombe *et al.*, 2018), and SLR-superscaffolder 0.9.1 (Guo *et al.*, 2021) on a human TELL-Seq dataset. The HUMAN dataset contains two TELL-Seq libraries and a HiFi read-set from a diploid human genome (Nurk *et al.*, 2022). We used LJA (Bankevich *et al.*, 2022) to generate the assembly graph (*multiplex de Bruijn graph*) from the HiFi reads in the HUMAN dataset and further scaffolded it using ARKS, SLR-superscaffolder, and SpLitteR. We used QUAST-LG (Mikheenko *et al.*, 2018) to compute various metrics of the resulting assembly (NGA50 values, the largest alignment, etc.) with the homopolymer-compressed T2T HG002 assembly as the reference (Rautiainen *et al.*, 2022). For the HUMAN dataset, the NGA50 values are 301, 303, 233, and 453 kb for LJA,

ARKS, SLR-superscaffolder, and SpLitter, respectively. The largest alignment is 21.5 Mb for LJA, SLR-superscaffolder, and SpLitteR, and 35 Mb for ARKS. The results of the benchmark and the SpLitteR repeat resolution procedure are explained in more detail in the Repeat resolution section.

Discussion

We presented SpLitteR, a tool for scaffolding and haplotype phasing in assembly graphs using linked-reads. We demonstrated that SpLitteR significantly increases the assembly contiguity compared to previously developed HiFi assemblers and linked-read scaffolders. We thus argue that linked-read technologies have the potential to become an inexpensive supplementary approach for generating contiguous assemblies of large genomes from the initial HiFi assemblies, in line with ONT and Hi-C reads, which were used by the T2T consortium to assemble the first complete human genome sequences (Nurk *et al.*, 2022). Due to the assembly graph simplification procedure that yields longer contig sequences, SpLitteR can be integrated as a preprocessing step in the assembly pipeline with other tools that employ supplementary sequencing technologies, such as Hi-C (Cheng *et al.*, 2021) and Strand-seq (Porubsky *et al.*, 2021) scaffolders.

Acknowledgments

The research was carried out in part by computational resources provided by the Resource Center “Computer Center of SPbU.”

Funding

IT and AK are grateful to Saint Petersburg State University for the overall support of this work. IT and AK were supported by the Russian Science Foundation (grant 19-14-00172).

Conflict of Interest: none declared.

References

Afshinfard,A. *et al.* (2022) Physlr: Next-Generation Physical Maps. *DNA*, **2**, 116–130.
Bankevich,A. *et al.* (2022) Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.*, **40**, 1075–1081.
Bishara,A. *et al.* (2018) High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.*
Callahan,B.J. *et al.* (2021) Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome*, **9**, 130.
Cheng,H. *et al.* (2021) Haplotype-resolved de novo assembly using phased assembly graphs

with hifiasm. *Nat. Methods*, **18**, 170–175.
Chen,Z. *et al.* (2020) Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.*, **30**, 898–909.
Coombe,L. *et al.* (2018) ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics*, **19**, 234.
Guo,L. *et al.* (2021) SLR-superscaffolder: a de novo scaffolding tool for synthetic long reads using a top-to-bottom scheme. *BMC Bioinformatics*, **22**, 158.
Kolmogorov,M. *et al.* (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.
Kolmogorov,M. *et al.* (2020) metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods*, **17**, 1103–1110.
Kuleshov,V. *et al.* (2016) Genome assembly from synthetic long read clouds. *Bioinformatics*, **32**, i216–i224.
McElwain,M.A. *et al.* (2017) Long Fragment Read (LFR) Technology: Cost-Effective, High-Quality Genome-Wide Molecular Haplotyping. *Methods Mol. Biol.*, **1551**, 191–205.
Mikheenko,A. *et al.* (2018) Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, **34**, i142–i150.
Nurk,S. *et al.* (2020) HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.*, **30**, 1291–1305.
Nurk,S. *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.
Porubsky,D. *et al.* (2021) Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.*, **39**, 302–308.
Rautiainen,M. *et al.* (2022) Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv*, https://doi.org/10.1101/2022.06.24.497523.
Tolstoganov,I. *et al.* (2019) cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs. *Bioinformatics*, **35**, i61–i70.
Weisenfeld,N.I. *et al.* (2017) Direct determination of diploid genome sequences. *Genome Res.*, **27**, 757–767.