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

Ivan Tolstoganov[1*], Zhoutao Chen[2], Pavel A. Pevzner[3], Anton Korobeynikov[4,5]

[1] Department of Mathematics, Science for Life Laboratory, Stockholm University, 106 91, Stockholm, Sweden
[2] Universal Sequencing Technology Corporation, Carlsbad, California 92011, USA
[3] Department of Computer Science and Engineering, University of California, San Diego, San Diego CA, USA
[4] Center for Algorithmic Biotechnology, Saint Petersburg State University, Saint Petersburg, 199004, Russia
[5] 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 and contiguous *de novo* assemblies 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 negatively affects the contiguity of diploid human genome assemblies since two haplomes share many long identical regions. To generate the telomere-to-telomere assemblies of diploid genomes, biologists now construct their HiFi-based phased assemblies and use additional experimental technologies to transform these phased assemblies into more contiguous diploid assemblies. 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 show how TELL-Seq reads facilitate phasing and scaffolding in these graphs. This benchmarking demonstrates that SpLitteR improves upon the state-of-the-art linked-read scaffolders in the accuracy and contiguity metrics.

Introduction

The recently developed 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 linked-reads, and short reads marked by the barcode corresponding to the set of long fragments.

Various tools, such as Athena (Bishara *et al.*, 2018), cloudSPAdes (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-reads result in more contiguous assemblies than assemblies based on non-linked short reads, all these tools generate rather 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 (Nurk *et al.*, 2020; Shafin *et al.*, 2020; Kolmogorov *et al.*, 2020; Cheng *et al.*, 2021; Bankevich *et al.*, 2022; Rautiainen *et al.*, 2023). Still, 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-reads represent a viable alternative to other supplementary technologies since they combine 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), and SLR-superscaffolder (Guo *et al.*, 2021) improve the contiguity of HiFi assemblies, they do not take advantage of the assembly graph and thus ignore the important connectivity information encoded by this graph. In addition, these tools are not applicable to diploid assemblies and complex metagenomes with many similar strains.

We present the SpLitteR tool that uses linked-reads to improve the contiguity of phased HiFi assemblies. In contrast to existing linked-reads scaffolders, it utilizes the assembly graph and was developed with diploid assemblies in mind. 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 assembly graph using Tell-Seq data. We assume that the genome defines an (unknown) *genomic traversal* of the assembly graph. Given an incoming edge e into a vertex v, we define a *follow-up edge* $next(e)$ as the edge that immediately follows e in this traversal. A vertex in a graph is classified as *branching* if both its in-degree and out-degree exceed 1 (each branching vertex in the graph represents a genomic repeat).

Figure 1 illustrates the SpLitteR workflow. First, SpLitteR maps the barcoded TELL-Seq reads to the edges of the assembly graph, identifies the uniquely mapped reads, and stores their barcodes for each edge (see Supplementary Section Aligning barcoded reads for details). Given an incoming edge *e* into a branching vertex *v*, SpLitteR attempts to find a follow-up outgoing edge *next*(*e*) by analyzing all linked reads that map to both the in-edge *e* and all out-edges from *v* (see Supplementary Section Repeat resolution). A vertex is classified as *resolved* if SpLitteR finds a follow-up edge for each incoming edge into this vertex. SpLitteR further simplifies the assembly graph by *splitting* the resolved vertices in such a way that each matched pair of an in-edge and an out-edge is merged into a single edge. Finally, it outputs the results of the repeat resolution procedure both as the set of scaffolds and as the simplified assembly graph. The repeat resolution procedure has both diploid and metagenomic modes.

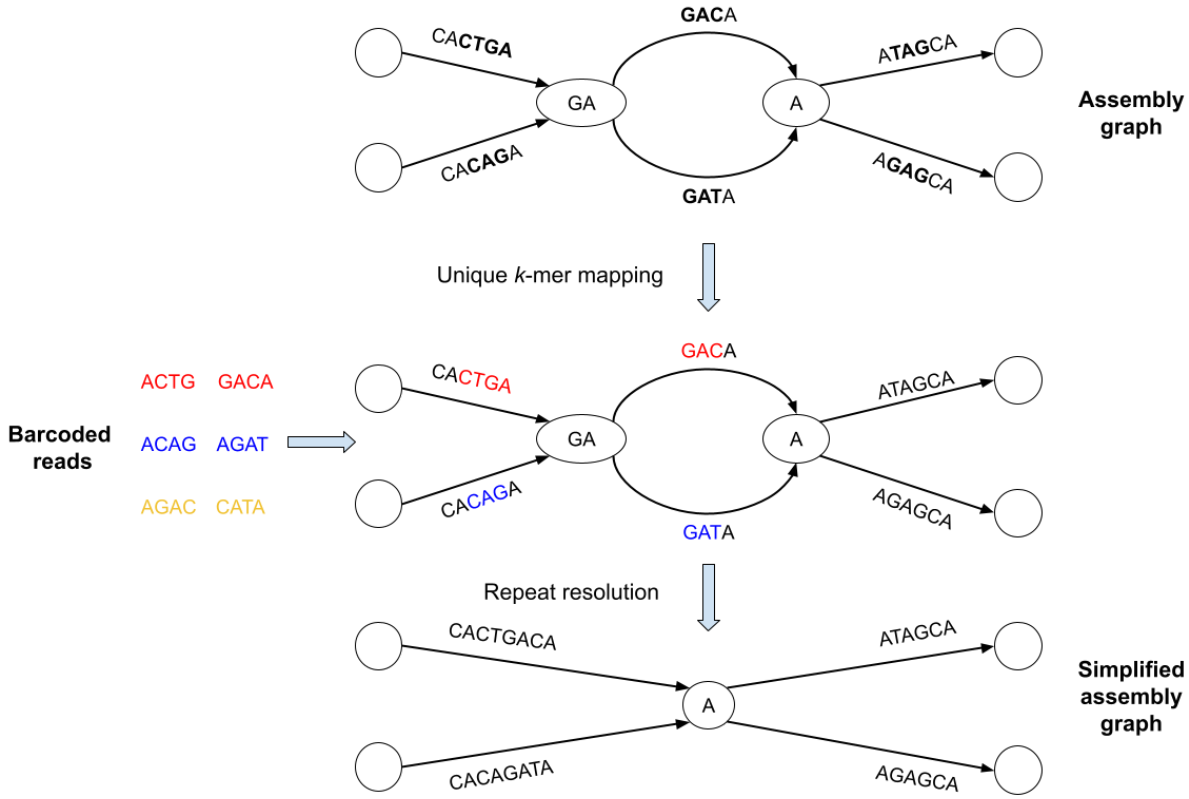

Figure 1. Brief summary of the SpLitteR workflow. In this toy example, the assembly graph is represented as the *multiplex de Bruijn graph* (Bankevich *et al.*, 2022) where vertices are labeled by *k*-mers of varying sizes. Reads with the same barcode are represented by the same color (each barcode contains only two reads) Reads are mapped to the assembly graph based on their *unique k*-mers, i.e., *k*-mers which occur only once in the edges of the assembly graph (*k*=3 for this toy example). Since yellow reads do not contain unique 3-mers, they remain unmapped. SpLitteR resolves vertices in the multiplex de Bruijn graph by assigning in-edges to their follow-up out-edges based on the barcode information.

Results

We benchmarked SpLitteR on a *HUMAN* dataset (Chen *et al.*, 2020) from a diploid human HG002 genome that was recently assembled from HiFi reads (Rautiainen *et al.*, 2023). The HUMAN dataset includes a TELL-Seq library which contains ~994 million barcoded TELL-Seq reads and a HiFi read-set from HG002. Since both TELL-Seq (Chen *et al.*, 2020) and HiFi technologies (Wenger *et al.*, 2019) emerged only three years ago, there are currently very few datasets that include both HiFi and TELL-Seq reads. We thus generated additional TELL-Seq datasets described below.

HUMAN+ dataset includes two additional TELL-Seq libraries which contain an additional ~4,585 million barcoded TELL-Seq reads.

The SHEEP dataset includes a TELL-Seq library containing ~1004 million barcoded reads and a HiFi library from a sheep fecal metagenome. Supplementary Table 1 provides additional information about these datasets, such as approximate fragment length. The Data Preparation section specifies the details of the TELL-Seq library preparation.

SpLitteR (version 0.1) was benchmarked against ARKS 1.2.4 (Coombe *et al.*, 2018), and SLR-superscaffolder 0.9.1 (Guo *et al.*, 2021) on the HUMAN, HUMAN+, and SHEEP datasets. We used LJA (Bankevich *et al.*, 2022) to generate the assembly graph (multiplex de Bruijn graph) from HiFi reads in the HUMAN and HUMAN+ datasets, and metaFlye (v.2.9) (Kolmogorov *et al.*, 2020) to generate the assembly graph for the SHEEP dataset. Assemblies for both datasets were further scaffolded using SpLitteR, ARKS, and SLR-superscaffolder. We used QUAST-LG (Mikheenko *et al.*, 2018) to compute various metrics of the resulting assemblies (NGA50 values, the largest alignment, etc.) with the homopolymer-compressed T2T HG002 assembly as the reference (Rautiainen *et al.*, 2023) for the HUMAN and HUMAN+ datasets.

For the HUMAN dataset, the NGA50 values are 301, 303, 233, and 461 kb for LJA (input graph), ARKS, SLR-superscaffolder, and SpLitteR, respectively. For the HUMAN+ dataset, the LJA assembly scaffolded with SpLitteR resulted in a 479 kb NGA50 value.

For the SHEEP dataset, SLR-superscaffolder and SpLitteR scaffolding did not result in any increase in contiguity compared to the initial metaFlye assembly, while ARKS result in a minor increase, as shown in the Supplementary section SHEEP dataset benchmark. Since ARKS and SLR-superscaffolder have very high RAM requirements, we only report SpLitteR results on the high-coverage HUMAN+ dataset. Benchmarking of the SpLitteR repeat resolution procedure is described in detail in the Supplementary sections HUMAN dataset benchmark and Repeat resolution. The SpLitteR results for the HUMAN dataset were additionally validated using trio-binning (Koren *et al.*, 2018) as shown in the Trio-binning validation section. The Supplementary section Coverage effects on the repeat resolution describes how the increase in TELL-Seq coverage improves the SpLitteR assembly quality.

Discussion

We presented a SpLitteR tool for scaffolding and haplotype phasing in assembly graphs using linked-reads. Our benchmarking demonstrated that it significantly increases the assembly contiguity compared to the previously developed HiFi assemblers and linked-read scaffolders. We thus argue that linked-reads have the potential to become an inexpensive supplementary technology for generating more 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 genomes (Nurk *et al.*, 2022; Rautiainen *et al.*, 2023). Since the assembly graph simplification procedure in SpLitteR yields longer contigs as compared to the initial HiFi-based assembly, 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).

Acknowledgments

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

Data availability

The sequencing reads for the HUMAN dataset are available in the NCBI BioProject database under accession number SRX7264481. The remaining reads for the HUMAN+ and SHEEP datasets generated in this study have been submitted to the NCBI BioProject database under accession number PRJNA956112. Baseline LJA assembly and trio binning results for the HUMAN+ dataset are available at https://figshare.com/articles/dataset/HG002/21678842. Baseline metaFlye assembly for the SHEEP dataset is available at https://figshare.com/articles/dataset/SHEEP/22864043.

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.
Antipov,D. *et al.* (2022) LJATrio development branch. *GitHub*.
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.* (2022) Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.*, **40**, 1332–1335.
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.
Koren,S. *et al.* (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*
Kuleshov,V. *et al.* (2016) Genome assembly from synthetic long read clouds. *Bioinformatics*, **32**, i216–i224.
Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
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.* (2023) Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.*
Shafin,K. *et al.* (2020) Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.*, **38**, 1044–1053.
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.
Wenger,A.M. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, **37**, 1155–1162.