LRez: C++ API and toolkit for analyzing and managing Linked-Reads data

Pierre Morisse, Claire Lemaitre, Fabrice Legeai

LRez: C++ API and toolkit for analyzing and managing Linked-Reads data

Pierre Morisse[1], Claire Lemaitre[1] and Fabrice Legeai[1,2]
[1] Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France
[2] IGEPP, INRAE, Institut Agro, Univ Rennes, 35000, Rennes, France

Corresponding author: pierre.morisse@inria.fr

Abstract

Linked-Reads technologies, such as 10x Genomics, Haplotagging, stLFR and TELL-Seq, partition and tag high-molecular-weight DNA molecules with a barcode using a microfluidic device prior to classical short-read sequencing. This way, Linked-Reads manage to combine the high-quality of the short reads and a long-range information which can be inferred by identifying distant reads belonging to the same DNA molecule with the help of the barcodes. This technology can thus efficiently be employed in various applications, such as structural variant calling, but also genome assembly, phasing and scaffoling. To benefit from Linked-Reads data, most methods first map the reads against a reference genome, and then rely on the analysis of the barcode contents of genomic regions, often requiring to fetch all reads or alignments with a given barcode.

However, despite the fact that various tools and libraries are available for processing BAM files, to the best of our knowledge, no such tool currently exists for managing Linked-Reads barcodes, and allowing features such as indexing, querying, and comparisons of barcode contents. LRez aims to address this issue, by providing a complete and easy to use API and suite of tools which are directly compatible with various Linked-Reads sequencing technologies.

LRez provides various functionalities such as extracting, indexing and querying Linked-Reads barcodes, in BAM, FASTQ, and gzipped FASTQ files (Table 1). The API is compiled as a shared library, helping its integration to external projects. Moreover, all functionalities are implemented in a thread-safe fashion.

Our experiments show that, on a 70 GB Haplotagging BAM file from *Heliconius erato* [1], index construction took an hour, and resulted in an index occupying 11 GB of RAM. Using this index, querying time per barcode reached an average of 11 ms. In comparison, using a naive approach without a barcode-based index, querying time per barcode reached an hour.

LRez is available on GitHub at https://github.com/morispi/LRez and as a bioconda module. Additionally, its features are already used in the SV calling tool LEVIATHAN (https://github.com/morispi/LEVIATHAN) and in the gap-filling pipeline MTG-Link (https://github.com/anne-gcd/MTG-Link).

Command	Description
compare	Compute the number of common barcodes between pairs of regions or between pairs of contig ends
extract	Extract the barcodes from a given region of a BAM file
index bam	Index the BAM offsets or genomic positions of the barcodes contained in a BAM file
index fastq	Index by barcode the offsets of the sequences contained in a FASTQ or gzipped FASTQ file
query bam	Query the index to retrieve alignments in a BAM file given a barcode or list of barcodes
query fastq	Query the index to retrieve sequences in a FASTQ / gzipped FASTQ file given a barcode or list of barcodes

Tab. 1. LRez features.

Acknowledgements

This project has received funding from the French ANR ANR-18-CE02-0019 Supergene grant.

References

[1] Joana I. Meier et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *bioRxiv*, pages 1–27, 2020.