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

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

1. Introduction

- **Linked-Reads** data and tag DNA molecules with **barcodes** before short-read sequencing
- Combine the **high-quality** of the short-reads and a **long-range** information

- Multiple sequencing technologies:
  - 10x Genomics (2016) [1]
  - stLFR (2019) [2]
  - Haplotagging (2020) [3]
  - TELL-Seq (2020) [4]
- Useful in a broad range of applications: **assembly**, **phasing**, **scaffolding** and **SV calling**
- **No tool** dedicated to Linked-Reads / barcodes management was previously **available**

2. Contribution

We propose **LRez**, a **toolkit** and **C++ API** that allows to **manage barcodes** from Linked-Reads data

- **Various functionalities** (indexing, querying, comparison, ...)
- Allows to process **BAM** and **FASTQ / gzipped FASTQ** files
- **Compatible** with **all** currently available **Linked-Reads technologies**
- **Easily usable** in any external tool or pipeline to **improve performances**

3. Functionalities

Toolkit command	Description	API module
compare	Compute the number of common barcodes between pairs of regions or between pairs of contig ends	BarcodesComparison
extract	Extract the barcodes from a given region of a BAM file	BarcodesExtraction
index bam	Index the BAM offsets or genomic positions of the barcodes contained in a BAM file	IndexManagementBam
index fastq	Index by barcode the offsets of the sequences contained in a FASTQ or gzipped FASTQ file	IndexManagementFastq
query bam	Query the index to retrieve alignments in a BAM file given a barcode or list of barcodes	AlignmentsRetrieval
query fastq	Query the index to retrieve sequences in a FASTQ or gzipped FASTQ file given a barcode or list of barcodes	ReadsRetrieval

4. Methods

- **Index** is represented as a **map**
  - Keys: barcodes
  - Values: list of occurrences positions

```
ACGTAGCTGTAGTTAG:   0,3512,5340,...,576948
TTAGTTACGATTGAGG:   440,6598,9549,...,657483
...
GGCCTAAAGCGATTCG:   842,4560,8756,...,458765
```

- **Query** the index and **browse** the BAM / FASTQ file to **retrieve reads or alignments**

5. Indexing performances

- **Datasets** from **all** Linked-Reads sequencing **technologies** and various species
  - *E. coli*
  - *H. sapiens*
  - *H. erato*

Dataset	BAM size (GB)	# Barcodes	Runtime	RAM (MB)	Disk (MB)
TELL-Seq (*E. coli*)	1	634,133	1 min	293	340
10x Genomics (*H. sapiens*)	61	609,058	52 min	9,320	15,062
Haplotagging (*H. erato*)	70	36,645,651	1 h 09 min	10,751	10,125
stLFR (*H. sapiens*)	206	38,779,362	3 h 06 min	26,769	34,256

6. Results: querying with LRez vs. samtools

- **Querying** experiments on the **previous datasets** from *E. coli*
- Comparison against a **naive methods** based on **samtools**
- Reported statistics from a **thousand queries**

Dataset	Overall runtime			Runtime per query	
	Samtools	LRez (index + query)	LRez (query)	Samtools	LRez
TELL-Seq (*E. coli*)	8 h 18 min	1 min	11 sec	30 sec	4 ms
10x Genomics (*H. sapiens*)	11.8 days	1 h 02 min	10 min	17 min	290 ms
Haplotagging (*H. erato*)	7.6 days	1 h 14 min	5 min	11 min	11 ms
stLFR (*H. sapiens*)	41.6 days	3 h 14 min	8 min	1 h	15 ms

- **LRez** is much **faster**, even when counting indexing time
- Runtime per query varies according to the number of reads / alignments
- **LRez** reaches a runtime of **15 ms per query** on a **206 GB BAM file**

7. Conclusion

- **Novel** and open-source **toolkit** and **C++ API** for processing Linked-Reads barcodes
- **Compatible** with **all** Linked-Reads **technologies**
- **Various functionalities** (indexing, querying, comparison, extraction)
- **Easily usable** in **external projects** (used in a SV calling tool and a gap-filling pipeline)
- **Efficient:** Time-saving and scaling barcode processing

[1] Brock Medsker et al. Haplotyping germline and cancer genomes using high- throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311, 2016.

[2] Ou Wang et al. Efficient and unique co-barcoding of second-generation sequencing reads from long dna molecules enabling cost effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Research*, 2019.

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

[4] Zhoutao Chen et al. Ultra-low input single tube linked-read library method enables short-read second-generation sequencing systems to generate highly accurate and economical long-range sequencing information routinely. *Genome Research*, 2020.

github.com/morispi/LRez

Contact: pierre.morisse@inria.fr