REXTAL: Regional Extension of Assemblies Using Linked-Reads

Interdisciplinary Research Team from Dept. of Computer Science & School of Medical Diagnostic & Translational Sciences

Graduate Student: Tunazzina Islam

Professors: Desh Ranjan, Mohammad Zubair, Harold Riethman

Objective

- It is currently impossible to get complete de novo assembly of segmentally duplicated genome regions using genome-wide short-read datasets.
- Using REXTAL, it is possible to extend assembly of single-copy diploid DNA into adjacent, otherwise inaccessible subtelomere segmental duplication regions and other subtelomeric gap regions.

Background

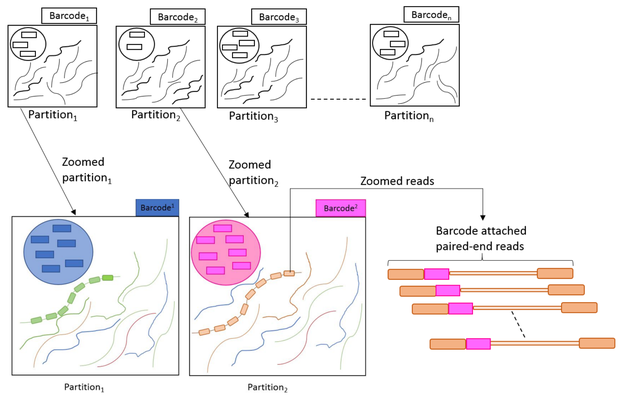
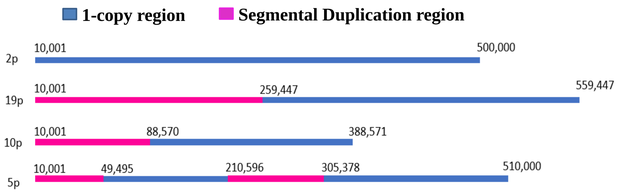


Figure 1. Conceptual description of GEM microfluidic method. Circle (blue, magenta) represents gel beads. Each bead contains many copies of a 16-base barcode (Rectangles inside the circle) unique to that bead. Each partition gets one gel bead. The 10 curve lines inside the large square (represents partition) represent molecules of length approximately 50 kb – 100 kb. The green and orange ovals represent short reads of length 150 bases which are obtained from these molecules (curve lines) with the barcode for the partition attached at the beginning of the first read in a pair. Sets of these read pairs having same barcodes attached to them are called linked-reads.

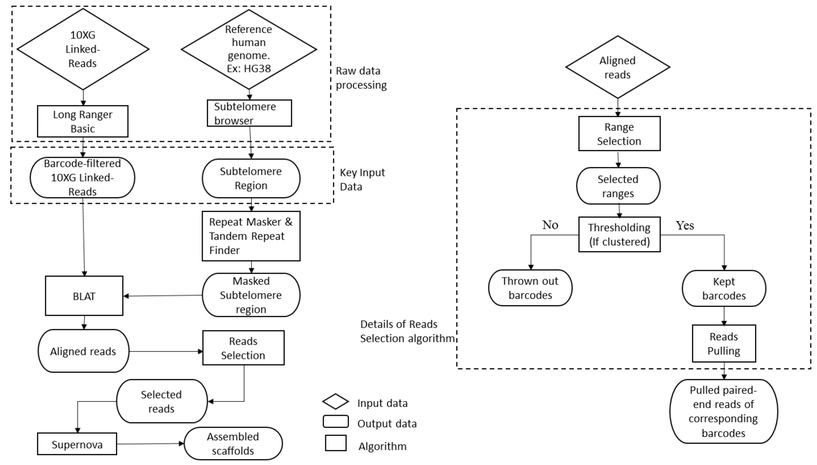
Data and Test set

- 1.49 billion 10X Genomics linked-reads in paired-end format from individual human genomes.
- Human reference genome HG38 was used to select test subtelomere regions for the targeted assemblies.
- Four different chromosomes with different representative patterns.



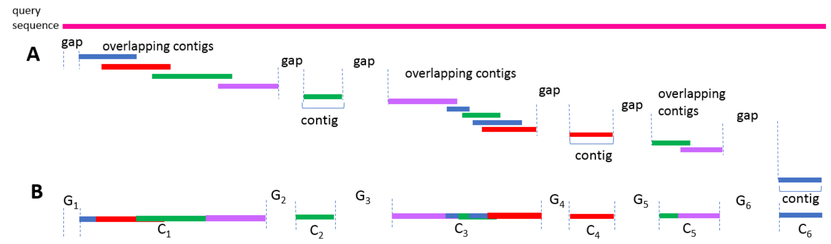
Method

REXTAL Workflow



Quality Measurement

- Align unmasked flanking reference DNA segments with assembled scaffold generated by REXTAL using NCBI BLAST with high identity matches (≥ 98%).
- Lengthwise Assembled Fraction (LAF) is defined as the fraction of a targeted reference sequence that is accurately assembled by the regional sequence assembly.
- To measure the accuracy in 1-copy region, we calculated the LAF with regular intervals
- For extended 1-copy and segmental duplication region, we calculated the LAF with regular intervals only from the edge of the bait segment into the extended region.
- To decide the cut-off point, we stopped where we noticed a sharp drop of the LAF.



Results

Table 1: Quality comparison for 1-copy region

Chromosomal region	Interval size[a]	LAF[b]	LAF[c]	Chromosomal region	Interval size[a]	LAF[b]	LAF[c]
19p	50kb	0.90	0.91	5p (1st 1-copy)	30kb	0.97	0.98
	100kb	0.91	0.91		60kb	0.94	0.90
	150kb	0.89	0.87		90kb	0.94	0.91
	200kb	0.88	0.86		120kb	0.94	0.92
	250kb	0.88	0.86		150kb	0.95	0.93
	300kb	0.89	0.87				
10p	50kb	0.99	0.99	5p (2nd 1-copy)	30kb	0.99	0.99
	100kb	0.99	0.99		60kb	0.96	0.96
	150kb	0.99	0.99		90kb	0.93	0.96
	200kb	0.99	0.99		120kb	0.92	0.95
	250kb	0.98	0.97		150kb	0.93	0.95
	300kb	0.97	0.68		180kb	0.93	0.94
					210kb	0.93	0.93
2p	200kb	0.99	0.98				
	300kb	0.98	0.98				
	400kb	0.97	0.97				
	500kb	0.97	0.97				

a: Starting position of 1st gap after the given interval size.
b: LAF for REXTAL. For 2p the range is 10-60 and for 19p, 10, 5p 1-copy the range is 3-70.
c: LAF for genome-wide assembly method.

➢ REXTAL achieves better LAF compare to Genome-wide assembly method in different interval sizes.

Table 2: Quality comparison for extended 1-copy region

Chromosomal region	(EL, LAF)[a]	(EL, LAF)[b]	Chromosomal region	(EL, LAF)[a]	(EL, LAF)[b]
2p	(33798, 0.99)	(16954, 1.00)	10p	(52022, 0.93)	(12437, 1.00)
19p	(43666, 0.93)	(6738, 0.99)	5p (2nd 1-copy)	(42326, 0.98)	(22485, 0.97)

a: Extension length (in bases) and LAF for REXTAL. For 19p, 10p, and 5p 1-copy region the range is 3-70.
b: Extension length (in bases) and LAF for genome-wide assembly method.

➢ Results obtained by REXTAL for extended 1-copy region are significantly better than the genome-wide method.

Table 3: Quality comparison for segmental duplication region

Chromosomal region	SD_L[a]	(EL, LAF)[b]	(EL, LAF)[c]
19p	249446	(67099, 0.98)	(5549, 1.00)
10p	78569	(40089, 0.98)	(4606, 1.00)
5p (1st 1-copy extends to 1st SD)	39495	(36477, 0.98)	(23129, 0.99)
5p (1st 1-copy extends to 2nd SD)	94782	(51860, 0.96)	(65, 1.00)
5p (2nd 1-copy extends to 2nd SD)	94782	(43090, 0.92)	(1307, 1.00)

a: Length of SD (segmental duplication) region (in bases) of corresponding chromosomal region.
b: Extension length (in bases) and LAF for REXTAL. For 19p, 10p, and 5p 1-copy region the range is 3-70.
c: Extension length (in bases) and LAF for genome-wide assembly method.

➢ For segmental duplication region the results obtained by REXTAL are notably superior to the genome-wide method.
➢ In particular, extensions from the 5p 1st 1-copy and the 2nd 1-copy region together (94950 bp) cover the entire 2nd segmental duplication region .

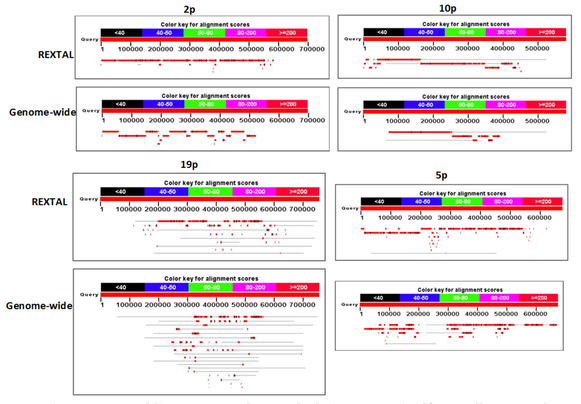


Figure 2. Assemblies corresponding to the bait segment itself as well as extending assemblies of the bait segment into adjacent DNA.

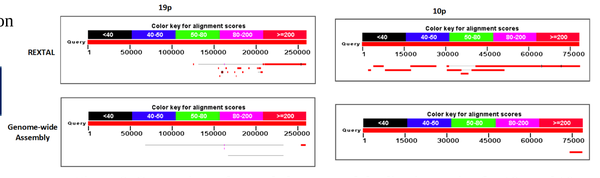


Figure 3. Comparison of extended segmental duplication region for 19p and 10p.

Acknowledgements

I would like to thank Eleanor Young and Ming Xiao from Drexel University, Philadelphia, PA for providing data. This work is partially supported by NIH R21CA177395 (HR and MX), and Modeling and Simulation Scholarship (to TI) from Old Dominion University.

References

1. Zheng, G. X.-L.-P. et al., (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature biotechnology, 34, 303-311.
2. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. (2017). Direct determination of diploid genome sequences. Genome research, 27, 757-767.
3. Alkan C, Sajjadian S, Eichler EE (2011). Limitations of next-generation genome sequence assembly.Nature methods, 8, 61.
4. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002). The human genome browser at UCSC. Genome research, 12, 996-1006.
5. Smit, A. F. (1996). 2010 RepeatMasker Open-3.0. http://www.repeatmasker.org/.
6. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research, 27, 573.
7. Kent, W. J. (2002). BLAThe BLAST-like alignment tool. Genome research, 12, 656{664.
8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25, 3389-3402.
9. Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 1072-1075.
10. Li H., Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760.
11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25, 2078-2079.