REXTAL: Regional Extension of Assemblies Using Linked-Reads

Tunazzina Islam
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
tislam@cs.odu.edu

Desh Ranjan
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
dranjan@cs.odu.edu

Mohammad Zubair
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
zubair@cs.odu.edu

Harold Riethman
School of Medical Diagnostic & Translational Sciences
Old Dominion University
Norfolk, VA 23529
hriethma@odu.edu

1. INTRODUCTION

Massively parallel short-read DNA sequencing has dramatically reduced the cost and increased the throughput of DNA sequence acquisition; it is now cheap and straightforward to do a variety of whole-genome analyses by comparing datasets of newly sequenced genomes with the human reference sequence. A recently developed approach pioneered by 10X Genomics generates short-read datasets from large genomic DNA molecules first partitioned and barcoded using the "Gel Bead in Emulsion" (GEM) microfluidic method [1]. The bioinformatic pipeline for assembly of these reads ("Supernova"; [2]) takes advantage of the very large number of sets of "linked reads". Each set of linked reads is comprised of low-read coverage of a small number of large genomic DNA molecules (roughly 10) and is associated with a unique barcode. However, even with these new methods, evolutionarily recent segmentally duplicated DNA such as that found in subtelomere regions remain inaccessible to de novo assembly due to the long stretches of highly similar (> 95% identity) DNA. The problem for subtelomere DNA analysis is amplified by the relative lack of high-quality reference assemblies and abundance of structural variation in these regions. To address this problem and attempt to better assemble human subtelomere regions, we have developed a computational approach called Regional Extension of Assemblies Using Linked-Reads (REXTAL) to extend de novo assemblies from subtelomeric 1- copy DNA regions into adjacent segmentally duplicated and gap regions of human subtelomeres.

Conceptually, what the "Gel Bead in Emulsion" (GEM) [1] microfluidic method enables us to do is illustrated in Figure 1. There are approximately one million partitions, each with a unique barcode. Each partition receives approximately 10 molecules of length approximately 50 kb–100 kb. Short reads of length 150 bases are obtained from these molecules with the barcode for the partition attached at the beginning of the first read in a pair [2]. Sets of these read pairs having same barcodes attached to them are called linked-reads.

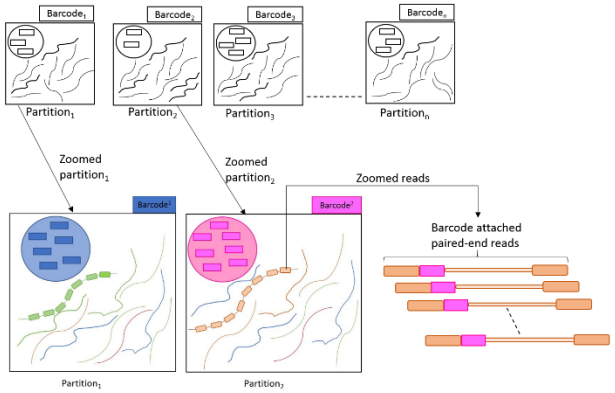

Figure 1. Conceptual description of GEM microfluidic method. Circle (blue, magenta) represents gel beads. Each bead contains many copies of a 16-base barcode (Rectangles inside the circle) unique to that bead. Each partition gets one gel bead. The 10 curve lines inside the large square (represents partition) 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).

Supernova uses the barcode information after initial whole-genome assembly for bridging long gaps. We refer to this method as "genome-wide assembly method". REXTAL differs from the genome-wide assembly method in that we use the barcode information for selection of reads from anticipated segmental duplication or gap regions adjacent to a specified 1-copy DNA segment before doing the assembly. REXTAL can be applied more generally for enriching region-specific linked reads and improving the assembly of any specified 1-copy genome region of an individual from any species for which a reference genome exists. For targeted region-specific assemblies from many individuals for which 10X datasets are available (e.g., analysis of structural variation at specific loci), REXTAL is faster and more accurate than genome-wide assembly method. In this scenario, for genome-wide assembly, we need to assemble the whole genome of the individuals and then extract the assembled portion of the specific region. But in our case, we first extract the specific region from the 10X dataset by aligning with a 1-copy segment of the reference genome, and then use our bioinformatic pipeline to do the assembly.

2. METHOD

2.1 Data

The key input data is 10X Genomics linked-reads from individual human genomes, in our case from the genome of a publically available cell line GM19440. Our dataset has approximately 1.49 billion 10X Genomics linked-reads in paired-end format, with each read about 150bp.

2.2 Data Processing

We processed the raw 10X Genomics data using Long Ranger Basic software developed by 10X Genomics (and freely available to any researcher) to generate barcode-filtered 10XG linked-reads. We used the UCSC browser [3] to access HG38 and selected subtelomere DNA segments for analysis.

2.3 Alignment of Subtelomeric Region with Linked-Reads

We used RepeatMasker [4] and Tandem Repeats Finder [5] to screen bait DNA segment sequences for interspersed repeats, low complexity DNA sequences, and tandem repeats to minimize the possibility of false-positive contaminant read identification in the initial selection of reads matching. We used BLAT (BLAST-like alignment tool) [6] with default parameter to do alignment of masked subtelomeric region with genome-wide reads from GM19440. We therefore initially collected all reads that shared a barcode with any read matching the 1-copy segment.

2.4 Barcode Frequency Range and Clustering Pattern Selection

We further reduced this subset of selected reads based on the frequency of occurrence and the clustering pattern of reads from each barcode identified as matching within the specified 1-copy segment. We estimated that each barcode should have approximately 800 reads based upon the following calculation: we assumed there are 1 million partitions in the genome with each partition containing 10 molecules of 50 kb each [2]. With the length of each read 150 bp and 0.25X coverage of each single molecule in the partition, we should have approximately (0.25 x 500000 bp)/ 150 = 833 reads with each barcode. For each barcode, approximately 1/10 of these reads (about 80) should originate from a single locus, and since about 50% of the bait locus (the specified 1-copy region used for BLAT) is masked, about 40 reads/partition should be matched if the entire 50 kb is within the bait locus. If the source DNA molecule partially

overlaps the bait locus and extends into the adjacent region, then this number would be smaller and dependent on the extent of the overlap. So, a key challenge was to identify the range of matching reads for each barcode that would minimize inclusion of false positive barcodes while maximizing inclusion of true positive barcodes that would permit extension of the assembly into adjacent DNA. Histogram analysis to check the frequency of the occurrence of each barcode revealed vast over-representation of barcodes with one or two reads, so we required a minimum of three reads per barcode to include that barcode for read selection. In addition, we required all matching reads from a single barcode to originate within less than the estimated maximum input molecule size of 100 kb within a given bait region to qualify for inclusion. We then empirically tested a variety of barcode frequency ranges meeting both of the above requirements for final read selection, using the ability of the selected reads to assemble the original bait region and extend into flanking DNA as the metric for optimization as described below.

2.5 Assembly of subset of reads

To get the assembly of the selected paired-end barcode reads Supernova was used. We used pseudohap2 style here. An overview of our assembly strategy is shown in Figure 2.

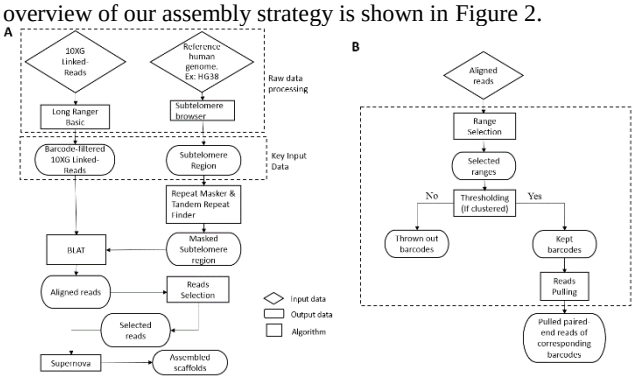

Figure 2. A: Flowchart. B: Details of Reads Selection algorithm is shown inside dotted box.

2.6 Alignment of assembled scaffolds with reference

To measure the quality of the assembly, we aligned specified subtelomeric regions of the reference sequence corresponding to our unmasked single-copy bait segments along with their flanking reference DNA segments as query with our generated assembled scaffolds as subject using NCBI BLAST [7], requiring high identity matches (≥ 98%) for retention of each local alignment. We mapped high-similarity alignments of REXTAL across the query reference sequence and, by merging the high-quality local alignments, evaluate assembly coverage relative to regions of the reference sequence using a parameter we define as the Lengthwise Assembled Fraction (LAF; see Figure 3). Intuitively, LAF is defined as the fraction of a targeted reference sequence that is accurately assembled by the regional sequence assembly. Regions of the reference query sequence with highest LAF have the best coverage of assembled sequence, and the limit of assembly extension regions corresponding to flanking reference sequence can be ascertained by a sudden decrease in LAF.

In Figure 4 we present an algorithm to compute the LAF of given contig and gap lengths. The input to the algorithm are two arrays C (contig length) and G (gap length) each of size *n*. The values in this S array correspond to LAF for *2n* different subsequences of the assembled sequence, all starting at the reference start position and ending at the end of each contig and gap.

3. RESULTS AND DISCUSSIONS

We tested our read selection and regional assembly strategy on four human subtelomere regions with representative patterns of sequence organization (base pair coordinates listed are from HG38). The 2p subtelomere is a 500kb sized segment of 1-copy DNA (10001 to 500,000); 19p subtelomere has a very large segmental duplication region next to the telomere (10001-259447) followed by a 300kb-sized 1-copy region (259448-559447), 10p has a smaller segmental duplication region near the telomere (10001-88570) followed by a 300kb 1-copy region (88571-388571); 5p has multiple segmental duplication regions as well as multiple single copy regions. 1st segmental duplication region is 10,001-49,495bp, 1st 1-copy region is 49,496-210,595bp, 2nd segmental duplication region is 210,596-305,378bp, and 2nd 1-copy region is 305,379-677,959bp. We tested a wide variety of Barcode ranges empirically for their ability to select read sets capable of generating high-quality regional assemblies corresponding to the bait segment itself as well as extending assemblies of the bait segment into adjacent DNA. For 2p we chose barcode range 10-60 and for 19p, 10p and 5p we chose barcode range 3-70.

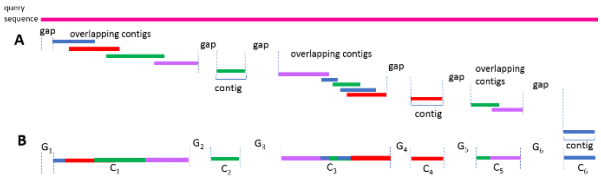

Figure 3. Top magenta rectangle represents the query sequence. A: Partially overlapped local alignment regions and gaps in coverage of the query sequence. B: Considering partially overlapped local alignment regions as sequence contigs and each sequence contig region (C) is followed by one sequence gap (G). Dotted blue lines represent starting position and ending position of gap.

Algorithm 1: CALCULATE_LAF (C, G)

1: **Construct** C', $C' \leftarrow [C_1, C_1 + C_2, \ldots . + (C_1 + C_2 + \cdots + C_n)]$
2: **Construct** G', $G' \leftarrow [G_1, G_1 + G_2, \ldots . + (G_1 + G_2 + \cdots + G_n)]$
3: $S[1] \leftarrow 0$
4: $S[2] \leftarrow C'[1]/\,(C'[1] + G'[1])$
5: ***for*** $i = 1$ ***to*** $n - 1$ ***do***
6: $\quad S[2i + 1] \leftarrow C'[i]/\,(C'[i] + G'[i + 1])$
7: $\quad S[2i + 2] \leftarrow C'[i + 1]/\,(C'[i + 1] + G'[i + 1])$
8: ***return*** S

Figure 4. Algorithm to calculate LAF.

3.1 Assembly Quality Measurement

Standard assembly quality measurements ("QUAST" [8]) are not suitable to our case as we are doing region specific assemblies rather than genome-wide assemblies. We are focused on coverage and accuracy of our assembly over the targeted region and have developed a metric called Length-wise Assembled Fraction (LAF) for quality measurement of our regional assemblies.

To see the accuracy of REXTAL in subtelomeric region, we calculated the LAF with regular intervals. For example: for all ranges of 2p, we took the intervals as the distance from coordinate 1 of the reference query sequence to the starting positions of the 1st gap after 200kb, 300kb, 400kb, and 500kb respectively. For range 10-60 of 2p subtelomeric region we achieve good LAF (Table 1).

For the calculation of LAF, for all ranges of 19p 1-copy, we calculated the LAF from coordinate 1 of the reference query sequence up to the starting positions of 1st gap after 50kb, 100kb, 150kb, 200kb, 250kb, and 300kb respectively. We achieve good LAF for range 3-70 of 19p 1-copy (Table 1). We fixed the range 3-70 for 10p and 5p. Table 1 shows the LAF of 10p 1-copy with same intervals taken for 19p 1-copy.

The 5p has multiple segmental duplication regions as well as multiple single copy regions. Because of the length variation of 1-copy region we chose different set of intervals for 1st 1-copy and 2nd 1-copy. We calculated the LAF from coordinate 1 of the

reference query sequence up to the starting positions of 1st gap after 30kb, 60kb, 90kb, 120kb, and 150kb respectively for 1st 1-copy region and for the 2nd 1-copy region we chose the intervals from coordinate 1 of the reference query sequence to the starting position of 1st gap after 30kb, 60kb, 90kb, 120kb, 150kb, 180kb, and 210kb (Table 1).

Table 1: Quality comparison for 1-copy region

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

We calculated the LAF with regular intervals only from the edge of the bait segment into the extended 1-copy region and segmental duplication region. We took the intervals as from the end of the bait segment to the starting positions of 1st gap after regular intervals. To decide the cut-off point, we checked all LAFs of extended region and segmental duplication region and we stopped where we noticed sharp drop of the LAF. The reason for this sharp drop is after this contig there is big gap and after that there is no significant length of assembled contig to increase the LAF. Table 2 and Table 3 show the corresponding analysis of extended 1-copy region and segmental duplication region with extension length as well as LAF.

3.2 Comparison with Genome-Wide Assembly

For fair comparison with genome-wide assembly method, we extracted all contigs in the genome-wide assembly that overlap (including potential extensions into flanking DNA) with the reference sequence using BWA [9] and SAMtools [10]. We compared our results for the extended 1-copy region with the genome-wide method in Table 2. It is easy to observe that the results obtained by REXTAL are significantly better than the genome-wide method for these four loci.

Table 2: Quality comparison for extended 1-copy region

Chr[a]	(EL, LAF)[b]	(EL, LAF)[c]	Chr[a]	(EL, LAF)[b]	(EL,LAF)[c]
2p	(33798,0.99)	(16954,1.00)	10p	(52022, 0.93)	(12437,1.00)
19p	(43666,0.93)	(6738, 0.99)	5p	(42326, 0.98)	(22485,0.97)

a: Chromosomal region.
b: Extension length (in bases) and LAF for REXTAL.
c: Extension length (in bases) and LAF for genome-wide assembly method.

Table 3 shows the comparison of REXTAL result for the segmental duplication region with the genome-wide method. Once again note that for segmental duplication region the results obtained by REXTAL are notably superior to the genome-wide method for all loci that have been tested. Extensions from the 5p 1st 1-copy and the 2nd 1-copy region together (94950bp) cover the entire 2nd segmental duplication region (Table 3).

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.
c: Extension length (in bases) and LAF for genome-wide assembly method.

Figure 5 shows the quality and comparison of 1-copy region, extended region and extended segmental duplication region for our tested chromosomes using REXTAL and genome-wide assembly method.

4. CONCLUSION

We showed that using REXTAL, it is possible to extend assembly of single-copy diploid DNA into adjacent, otherwise inaccessible subtelomere segmental duplication regions. In future experiments, using larger source DNA molecules for barcode sequencing approaches could further extend assemblies into and through segmental duplications, and optical maps of large single molecules extending from the 1-copy regions through segmental duplications and gaps could be used to optimally guide and validate these assemblies.

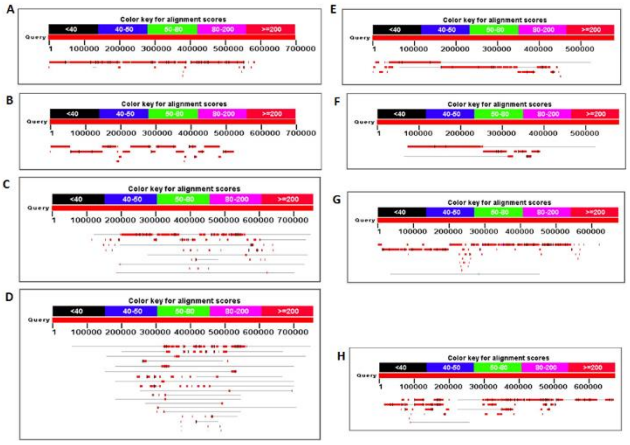

Figure 5. A: Alignment of 2p with assembled scaffolds of 2p for range 10-60 of REXTAL. B: Alignment of 2p as query with assembled scaffolds of 2p extracted from genome-wide assembly. C: Alignment of 19p with assembled scaffolds of 19p 1-copy for range 3-70 of REXTAL. D: Alignment of 19p with assembled scaffolds of 19p 1-copy region extracted from genome-wide assembly. E: Alignment of 10p with assembled scaffolds of 10p 1-copy for range 3-70 of REXTAL. F: Alignment of 10p with assembled scaffolds of 10p 1-copy region extracted from genome-wide assembly. G: Alignment of 5p with assembled scaffolds of 5p 1-copy regions for range 3-70 of REXTAL. H: Alignment of 5p with assembled scaffolds of 5p 1-copy regions extracted from genome-wide assembly.

6. ACKNOWLEDGEMENT

The work in this paper is supported in part by NIH R21CA177395 (HR), and Modeling and Simulation Scholarship (to TI) of Old Dominion University.

7. 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. 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.
4. Smit, A. F. (1996). 2010 RepeatMasker Open-3.0. URL: http://www. repeatmasker. org.
5. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research, 27, 573.
6. Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. Genome research, 12, 656—664.
7. 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.
8. Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 1072--1075.
9. Li H., Durbin R. (2009). Fast and accurate short read alignment with Burrows--Wheeler transform. Bioinformatics, 25, 1754--1760.
10. 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.