REXTAL: Regional Extension of Assemblies Using Linked-Reads

Tunazzina Islam[1(✉)], Desh Ranjan[1], Eleanor Young[2], Ming Xiao[2,3], Mohammad Zubair[1], and Harold Riethman[4]

[1] Department of Computer Science, Old Dominion University, Norfolk, VA, USA
{tislam,dranjan,zubair}@cs.odu.edu
[2] School of Biomedical Engineering, Drexel University, Philadelphia, PA, USA
eay25@glink.drexel.edu
[3] Institute of Molecular Medicine and Infectious Disease, School of Medicine, Drexel University, Philadelphia, USA
mx44@drexel.edu
[4] School of Medical Diagnostic and Translational Sciences, Old Dominion University, Norfolk, VA, USA
hriethma@odu.edu

Abstract. It is currently impossible to get complete de novo assembly of segmentally duplicated genome regions using genome-wide short-read datasets. Here, we devise a new computational method called Regional Extension of Assemblies Using Linked-Reads (REXTAL) for improved region-specific assembly of segmental duplication-containing DNA, leveraging genomic short-read datasets generated from large DNA molecules partitioned and barcoded using the Gel Bead in Emulsion (GEM) microfluidic method [1]. We show that 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. Moreover, REXTAL is computationally more efficient for the directed assembly of such regions from multiple genomes (e.g., for the comparison of structural variation) than genome-wide assembly approaches.

Keywords: 10X sequencing · Linked-read sequencing · Subtelomere Assembly · Segmental duplication · Structural variation · Genome gaps

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. However, even with the use of paired-end read approaches using input molecules of various lengths, de novo assembly of human genomes has remained problematic because

F. Zhang et al. (Eds.): ISBRA 2018, LNBI 10847, pp. 63–78, 2018.
https://doi.org/10.1007/978-3-319-94968-0_6

of abundant interspersed repeats. 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. This approach enables efficient de novo assembly of much of the human genome, with large segments separable into haplotypes [2]. 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 designed to leverage linked-reads from genomic GEM datasets 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 Fig. 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.

Supernova assembly [2] takes advantage of linked reads to separate haplotypes over long distances, and these separated haplotypes are represented as megabubbles in the assembly. The chain of megabubbles generates scaffolds [2]. Supernova uses the barcode information after initial whole-genome assembly for bridging long gaps. It finds all the reads of corresponding barcodes that are present in sequence adjacent to the left and right sides of the assembly gap. Then it assembles this set of reads and tries to fill the gap [2]. We refer to this method as genome-wide assembly method. As in all genome-wide assemblies, reads from evolutionarily recent segmental duplications such as those near subtelomeres are collapsed into artifactual DNA segment assemblies; these assembly artifacts are typically either located at a single genomic locus or excluded entirely from the initially assembled genome [3]. 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. We initially find reads matching the 1-copy DNA segment (bait DNA segment) based upon the reference human genome (HG38), then select all reads for barcodes represented in these initial matching reads. This set of reads should represent a very limited subset of all genomic reads, and approximately 10% of the barcode-selected reads should be derived specifically from the selected 1-copy DNA and 50 kb-100 kb segments of

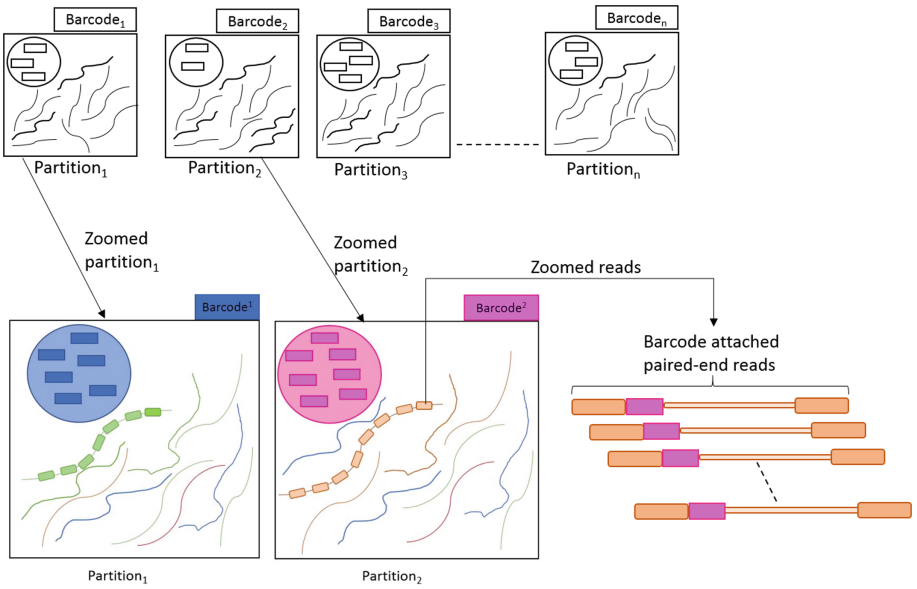

Fig. 1. Conceptual description of GEM microfluidic method. Circle (blue, magenta) represents gel beads. Each bead contains many copies of a 16-base barcode (Rectangles inside the circle) unique to that bead. Each partition gets one gel bead. The 10 curve lines inside the large square (represents partition)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

flanking DNA. We show here that this is indeed the case, enabling the extension of existing assemblies into adjacent segmental duplication and gap regions.

While the primary motivation of our work is to improve the assembly of subtelomeric gap regions and extend the assembly to inaccessible subtelomere segmental duplication regions of genomes of human individuals from their 10X genomic data, 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

In Subsect. 2.1, we present the input data description. Subsection 2.2 presents processing of raw data to get our key input data. In Subsects. 2.3, 2.4 and 2.5 we show our assembly pipeline step by step. Subsection 2.6 shows further analysis after assembly.

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 150 bp. The Supernova whole genome assembly using these data had an overall coverage of 103 and a Supernova N50 scaffold of 19.1 Mb. The loupe file shows a mean depth coverage of 67.4. Human reference genome assembly HG38 was used to select test subtelomere regions for the targeted assemblies.

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. The Long Ranger basic pipe-line performs basic read and barcode processing including read trimming, barcode error correction, barcode whitelisting, and attaching barcodes to reads. We used the UCSC browser [4] to access HG38 and selected subtelomere DNA segments for analysis.

2.3 Alignment of Subtelomeric Region with Linked-Reads

Masking Out Repeats. We used RepeatMasker [5] and Tandem Repeats Finder [6] to screen bait DNA segment sequences for interspersed repeats, low complexity DNA sequences, and tandem repeats in order to minimize the possibility of false-positive contaminant read identification in the initial selection of reads matching specified 1-copy DNA segments.

Alignment Using BLAT. We used BLAT (BLAST-like alignment tool) [7] with default parameter to do the alignment of the masked subtelomeric region with genome-wide reads from GM19440.

Reads Selection. The output of BLAT gives reads which have a good match with a given subtelomeric bait region. However, it is possible that many reads that would have originated within this given subtelomeric region could have been missed because of masking out repeat regions done previously. More importantly, we were especially interested in capturing reads from the large source DNA molecules extending from the flanks of the targeted 1-copy bait segment. 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 on 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 in order 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 in order 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 [2] was used. It can generate assembled scaffolds in four styles named: raw, megabubbles, pseudohap, and pseudohap2. We used pseudohap2 style here. An overview of our assembly strategy is shown in Fig. 2.

2.6 Alignment of Assembled Scaffolds with Reference

To measure the quality of the assembly, we aligned specified subtelomeric regions of the HG38 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 [8], requiring high identity matches ($\geq$98%) for retention of each local alignment. The resulting output hit table of these local alignments lists the sequence identifier, the start and stop points for each local stretch of sequence similarity, and the percent identity of the match. From this information one can map high-similarity alignments

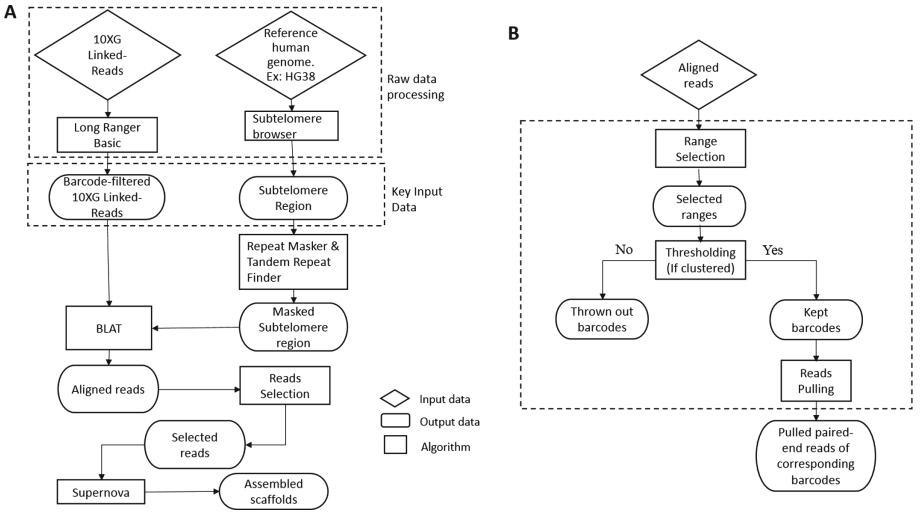

Fig. 2. A: Flowchart of REXTAL. B: Details of Reads Selection algorithm is shown inside dotted box.

of our regional assembly (prepared using barcode-selected linked reads) 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 Fig. 5). 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. Details of LAF calculation are presented in 3.4.

3 Results and Discussions

We tested our read selection and regional assembly strategy (Fig. 2) on four human subtelomere regions with representative patterns of sequence organization (base pair coordinates listed are from HG38; Fig. 3). The 2p subtelomere is a 500 kb sized segment of 1-copy DNA (10,001 to 500,000); 19p subtelomere has a very large segmental duplication region next to the telomere (10,001–259,447) followed by a 300 Kb-sized 1-copy region (259,448–559,447), 10p has a smaller segmental duplication region near the telomere (10,001–88,570) followed by a 300 kb 1-copy region (88,571–388,571); 5p has multiple segmental duplication regions (10,001–49,495 and 210,596–305,378) separated and flanked by two 1-copy regions (49,496–210,595 and 305,379–510,000).

We processed the raw input data from GM19440 as described in Subsect. 2.2. Table 1 presents some characteristics of the output obtained after processing

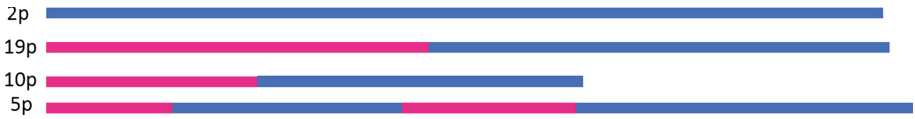

Fig. 3. Four different chromosomes with different characteristics. The blue rectangle represents single copy region and the magenta rectangle represents segmental duplication region. (Color figure online)

the raw data with Long Ranger Basic software. Interspersed repeats and tandem repeats from the 1-copy regions of these subtelomeres were masked and used as bait segments to select matching reads from the GM19440 linked-read dataset using BLAT. Barcodes for matching reads were identified and characterized according to occurrence frequency and clustering within the bait DNA segments.

Table 1. Some characteristics of the obtained data.

Number of reads	$1.49 * 10^9$	Number of reads without barcode	$9.8 * 10^7$
Number of paired-end reads	$0.75 * 10^9$	Barcode whitelist	0.933959
Number of barcoded reads	$1.39 * 10^9$	Barcode diversity	743369.62

3.1 Barcode Range and Clustering Analysis

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 (Fig. 4) as well as extending assemblies of the bait segment into adjacent DNA (Fig. 6). In all cases, a secondary filter was applied requiring that barcodes used for reads selection contained only reads mapping to a single 100 kb segment of the bait DNA (cluster) as anticipated from linked-read library preparation (Table 2). Initial experiments with 2p focused on selection of reads from barcode ranges that produced high-quality assemblies of the 500 kb bait segment, and follow-up work with all four subtelomeres fine-tuned these parameters to optimize both high-quality assembly of bait segments as well as maximal extension into adjacent segmental duplication regions and single-copy regions. Table 2 shows the selected number of barcodes and number of reads after thresholding for 2p and 19p 1-copy region for our chosen ranges.

3.2 Generation of Assembled Scaffolds

After pulling out reads according to our selected range and clustering parameters, we used Supernova assembler for the assembly of the collected paired-end reads. We analyzed assembled scaffolds in pseudohap2 style and calculated the length of each assembled scaffolds.

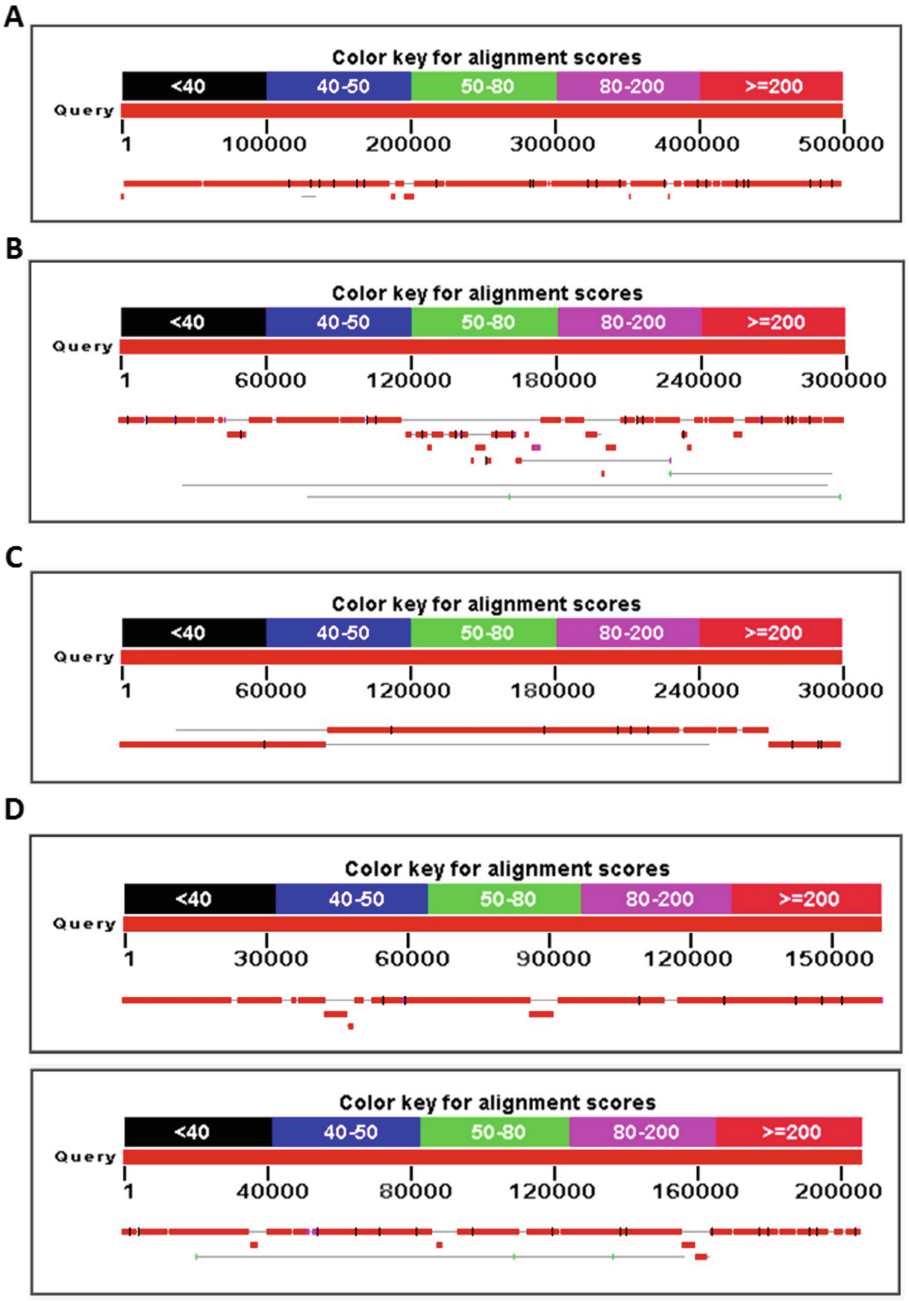

Fig. 4. A: Alignment of 2p 500 kb as query with assembled scaffolds of 2p for range 10–60 as subject in BLAST. B: Alignment of 19p 1-copy 300 kb as query with assembled scaffolds of 19p 1-copy for range 3–70 as subject in BLAST. C: Alignment of 10p 1-copy 300 kb as query with assembled scaffolds of 10p 1-copy for range 3–70 as subject in BLAST. D: Alignments of two 1-copy regions of 5p as query with assembled scaffolds of 5p 1-copy regions for range 3–70 as subject in BLAST.

Table 2. Results after range selection and clustering step

chr[a]	freq[b]	bc[c]	bc[d]	read[e]	chr[a]	freq[b]	bc[c]	bc[d]	read[e]
2p	10–50	1639	1223	2074096	19p	3–70	1493	1378	2482446
	10–60	1726	1281	2177142		5–70	1142	1026	1870206
	10–70	1807	1330	2265538		10–70	770	662	1265324

a: Chromosomal region.
b: Barcode frequency ranges.
c: Number of selected barcode after range selection.
d: Number of selected barcode after clustering.
e: Number of collected reads of corresponding barcodes.

3.3 Alignment of the Scaffolds with Reference

We aligned the 2p 1-copy region, 19p 1-copy region, 10p 1-copy region and 5p 1-copy regions as the query with corresponding generated assembled scaffolds of 2p, 19p, 10p and 5p as the subject using BLAST with default parameters and retaining only local alignments with ≥98% identity. Figure 4 shows a graphical representation (using the NCBI BLAST output visualization tool) of these BLAST alignments with near-optimal barcode frequencies for retention of linked-reads prior to assembly. While the respective assemblies cover most of each of the 1-copy bait regions, the extent of coverage as well as the number of scaffolds contributing substantially to coverage vary according to subtelomere. We, therefore, developed a more quantitative metric for assembly coverage in order to better quantify the assembly quality and compare them with the assemblies generated de novo from the whole-genome dataset using Supernova.

3.4 Assembly Quality Measurement

Standard assembly quality measurements (QUAST [9]) 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. As mentioned previously, LAF measures the fraction of a targeted reference sequence that is accurately assembled by the regional sequence assembly.

Quality in Single Copy Region. We extracted reference sequences of 2p, 19p, 10p, and 5p from HG38 and then aligned them with corresponding assembled scaffolds using BLAST, requiring ≥98% of identity for retention of each local alignment. This generates positions of each local alignment including query start positions and query end positions. The starting positions of the query were sorted in increasing order. Local alignments were merged by (1) deleting local alignments located entirely within other higher-quality alignments; and (2) Local

alignments with partial overlap, the overlap regions were merged by selecting the alignment with equivalent or higher % identity in the overlap region. The regions of the query sequence not aligned with sequences in the assembly scaffold are designated as gaps.

For LAF calculation, we considered a number of subsequences of the assembly. More precisely we considered subsequences of the assembly whose end points are start and end positions of n contigs (Fig. 5).

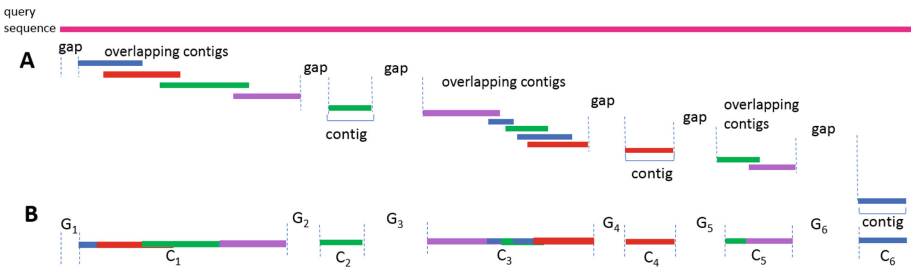

Fig. 5. Top magenta rectangle represents the query sequence. A: Partially overlapped local alignment regions and gaps in coverage of the query sequence. B: Considering partially overlapped local alignment regions as sequence contigs and each sequence contig region (C) is followed by one sequence gap (G). Dotted blue lines represent starting position and ending position of gap. (Color figure online)

We present an algorithm (Algorithm 1) to compute the LAF of given contig and gap lengths. The input to the algorithm are two arrays C and G each of size n. C[i] is the length of i^{th} contig and G[i] is the length of gap before the i^{th} contig. The algorithm computes LAF and outputs an array S of size $2n$. 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. 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 1^{st} gap after 200 kb, 300 kb, 400 kb, and

Algorithm 1. CALCULATE LAF (C, G)

1: **construct** C', $C' \leftarrow [C_1, C_1 + C_2, ..., (C_1 + C_2 + ... + C_n)]$
2: **construct** G', $G' \leftarrow [G_1, G_1 + G_2, ..., (G_1 + G_2 + ... + G_n)]$
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

500 kb respectively. For range 10–60 of 2p subtelomeric region we achieve good LAF (Table 3).

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 1^{st}t gap after 50 kb, 100 kb, 150 kb, 200 kb, 250 kb, and 300 kb respectively. We achieve good LAF for range 3–70 of 19p 1-copy (Table 3). We fixed the range 3–70 for 10p and 5p. Table 3 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. 1^{st} segmental duplication region is 10,001–49,495 bp, 1^{st} 1-copy region is 49,496–210,595 bp, 2^{nd} segmental duplication region is 210,596-305,378 bp, and 2^{nd} 1-copy region is 305,379–677,959 bp. We applied our assembly pipeline both for 1^{st} 1-copy and 2^{nd} 1-copy (305,379–510,000 bp) region. Because of the length variation of 1-copy region we chose different set of intervals for 1^{st} 1-copy and 2^{nd} 1-copy. We calculated the LAF from coordinate 1 of the reference query sequence up to the starting positions of 1^{st} gap after 30 kb, 60 kb, 90 kb, 120 kb, and 150 kb respectively for 1^{st} 1-copy region and for the 2^{nd} 1-copy region we chose the intervals from coordinate 1 of the reference query sequence to the starting position of 1^{st} gap after 30 kb, 60 kb, 90 kb, 120 kb, 150 kb, 180 kb, and 210 kb (Table 3).

Quality in Extended Region. We can extend our assembly of single-copy diploid DNA into adjacent and other subtelomeric gap regions. To see the extension of our assembly to extended single copy region, we extracted the reference 2p (10,001–700,000 bp) with length 700 kb, 19p (259,448–759,447 bp) with length 500 kb, 10p (88,571–588,571 bp) with length 500 kb, and 5p 2^{nd} 1-copy (305,379–677,959 bp) with length 372,580 bp from HG38. Following BLAST analysis using the extended reference sequence as the query and the assembled scaffolds as subject, we used Algorithm 1 to measure the LAF only for the extended region i.e. >500k for 2p, >300k for 19p and 10p 1-copy, >204,621 bp for 5p 2^{nd} 1-copy.

We calculated the LAF with regular intervals only from the edge of the bait segment into the extended region. We took the intervals as from the end of the bait segment to the starting positions of 1^{st} gap after 10 kb, 20 kb, 30 kb, 40 kb, and 50 kb respectively. To decide the cut-off point for the extended region, we checked all LAFs of the extended region and we stopped where we noticed a sharp drop of the LAF. The reason for this sharp drop is after this contig there is a big gap and after that, there is no significant length of assembled contig to increase the LAF (Table 4).

Quality in Segmental Duplication Region. As segmental duplication region contains segments of DNA with near-identical duplicated subtelomere sequence, this region is hard to assemble de novo with whole genome reads. We can extend our assembly into subtelomere segmental duplication regions. Following BLAST analysis using the HG38 reference subtelomere assembly including the segmental duplication region along with the adjacent bait region, we used Algorithm 1, to

Table 3. 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	50 kb	0.9	0.91	5p (1^{st} 1-copy)	30 kb	0.97	0.98
	100 kb	0.91	0.91		60 kb	0.94	0.9
	150 kb	0.89	0.87		90 kb	0.94	0.91
	200 kb	0.88	0.86		120 kb	0.94	0.92
	250 kb	0.88	0.86		150 kb	0.95	0.93
	300 kb	0.89	0.87				
10p	50 kb	0.99	0.99	5p (2^{nd} 1-copy)	30 kb	0.99	0.99
	100 kb	0.99	0.99		60 kb	0.96	0.96
	150 kb	0.99	0.99		90 kb	0.93	0.96
	200 kb	0.99	0.99		120 kb	0.92	0.95
	250 kb	0.98	0.97		150 kb	0.93	0.95
	300 kb	0.97	0.68		180 kb	0.93	0.94
					210 kb	0.93	0.93
2p	200 kb	0.99	0.98				
	300 kb	0.98	0.98				
	400 kb	0.97	0.97				
	500 kb	0.97	0.97				

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

measure the LAF only for the segmental duplication region of 19p, 10p, and 5p and then chose the cut-off point. Table 5 shows the analysis of segmental duplication region with extension length as well as LAF.

3.5 Comparison with Genome-Wide Assembly

For a fair comparison with genome-wide assembly method, we need to extract all contigs in the genome-wide assembly that overlap (including potential extensions into flanking DNA) with the reference sequence. To do so we use BWA index [10] of the reference genome (hg38). We have the genome-wide assembly of our input data using Supernova. For alignment using BWA-MEM [10], we aligned the genome-wide assembled reads against the indexed reference genome and generated a .sam file. Using SAMtools [11] we converted the .sam file into a .bam file, sort, and index the results. We extracted specific region of specific chromosomes (here 2p, 19p, 10p, and 5p) from that indexed results using SAMtools and aligned them with the same reference queries used for analysis of the barcode-selected read assemblies using BLAST with $\geq$98% of identity (see Fig. 6B, D, F and H).

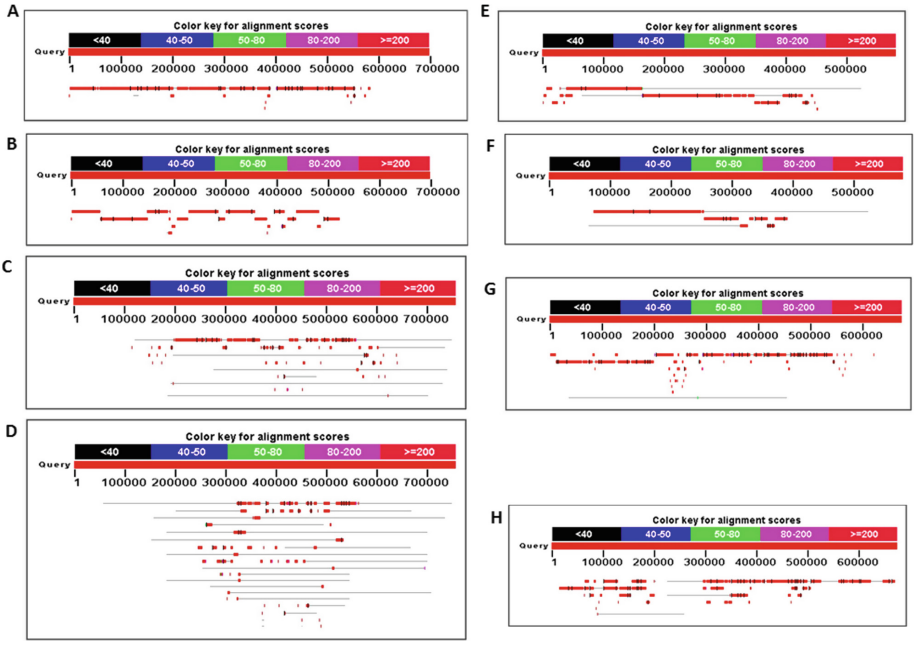

Fig. 6. A: Alignment of 2p with assembled scaffolds of 2p for range 10–60 of REXTAL. B: Alignment of 2p as query with assembled scaffolds of 2p extracted from genome-wide assembly. C: Alignment of 19p with assembled scaffolds of 19p 1-copy for range 3–70 of REXTAL. D: Alignment of 19p with assembled 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.

Comparison in Single Copy Region. To measure the quality of subtelomeric region assembly of extracted 2p, 19p, 10p, and 5p 1-copy region from the genome-wide assembly, we followed the same steps that mentioned previously for REXTAL to measure quality in the 1-copy region (see 3.4). We calculated the LAF with regular intervals using Algorithm 1. Table 3 shows the comparison of LAF between REXTAL and genome-wide assembly method. For 2p and 5p 2^{nd} 1-copy we get similar LAF with genome-wide method (Table 3). We get better LAF using REXTAL for 19p, 10p 1-copy, and 5p 1^{st} 1-copy than genome-wide method (Table 3).

Comparison in Extended Region. To show the extension of single copy region in genome-wide assembly method, we followed the same steps that we discussed for REXTAL to measure quality in the extended 1-copy region. We calculated the LAF using Algorithm 1. Then we decided the cut-off point. We compared our result for the extended 1-copy region with the genome-wide method

in Table 4. It is easy to observe that the results obtained by REXTAL are significantly better than the genome-wide method for these four loci.

Table 4. 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 (2^{nd} 1-copy)	(42326, 0.98)	(22485, 0.97)

a: 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.

Comparison in Segmental Duplication Region. We used Algorithm 1 to calculate the LAF for segmental duplication region that we got from genome-wide assembly method and compared the extension achieved by REXTAL with the extension achieved by genome-wide assembly method. Table 5 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. In particular, extensions from the 5p 1^{st} 1-copy and the 2^{nd} 1-copy region together (94,950 bp) cover the entire 2^{nd} segmental duplication region (Table 5).

Table 5. 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 (1^{st} 1-copy extends to 1^{st} SD)	39495	(36477, 0.98)	(23129, 0.99)
5p (1^{st} 1-copy extends to 2^{nd} SD)	94782	(51860, 0.96)	(65, 1.00)
5p (2^{nd} 1-copy extends to 2^{nd} SD)	94782	(43090, 0.92)	(1307, 1.00)

a: Length of segmental duplication region (in bases) of corresponding chromosomal region.
b: Extension length (in bases) and LAF of 19p, 10p, and 5p for REXTAL.
c: Extension length (in bases) and LAF for genome-wide assembly method.

Figure 7 shows the comparison of extended segmental duplication region for 19p and 10p using REXTAL and genome-wide assembly method.

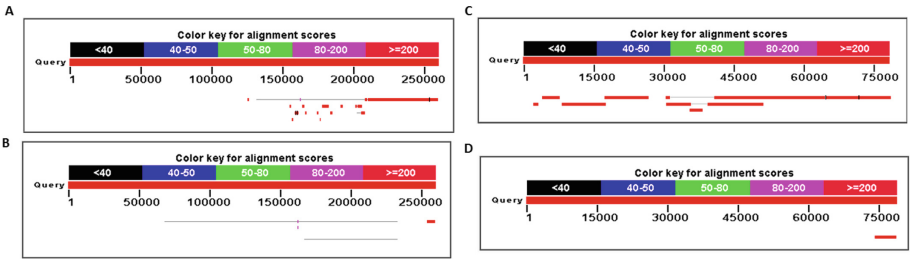

Fig. 7. A: Alignment of 19p segmental duplication region with assembled scaffolds of 19p 1-copy for range 3–70 of REXTAL. B: Alignment of 19p segmental duplication region with assembled scaffolds of 19p 1-copy region extracted from genome-wide assembly. C: Alignment of 10p segmental duplication region with assembled scaffolds of 10p 1-copy for range 3–70 of REXTAL. D: Alignment of 10p segmental duplication region with assembled scaffolds of 10p 1-copy region extracted from genome-wide assembly.

3.6 Efficiency Considerations

For targeted region-specific assemblies from multiple individuals for which 10X datasets are available, REXTAL is faster and more accurate than genome-wide assembly method. For genome-wide assembly, we need to assemble the whole genome of the individuals first and then extract the assembled portion of the specific region. To do whole genome assembly using Supernova takes approximately 36–48 h [2]. Before extraction of the specific region, we need to align the genome-wide assembled reads against the indexed reference genome (hg38) using BWA-MEM. This takes approximately 25 h. Then we can extract the specific region of the specific chromosome. For multiple individuals, although we want to do region-specific assembly, the genome-wide assembly method assembles the whole genome for each individual first and does the alignment – these two steps are time-consuming.

In contrast, REXTAL extracts the reads relevant for assembling the targeted region from the 10X dataset by aligning the targeted region with a 1-copy segment of the reference genome (hg38) using BLAT. This step takes 2–5 h. Reads selection step mentioned in 3^{rd} paragraph of Subsect. 2.3 and Barcode frequency range and clustering pattern selection step described in Subsect. 2.4 together take around 2–3 h. Assembly of the subset of selected reads using Supernova takes approximately 5–15 min. So in total, the region-specific assembly using REXTAL takes approximately 4–8 h. In case of targeted region-specific assembly for multiple individuals, our method REXTAL is approximately 9 times faster than genome-wide assembly method. The configuration of the machine where we ran REXTAL is CPU: 32 cores (2, 16 core processors → Intel(R) Xeon(R) CPU E5-2683 v4/Broadwell @ 2.10 GHz), Memory: 128 GB RAM, Network: FDR IB (56 Gbps fabric).

4 Conclusion

We successfully used a new computational method called Regional Extension of Assemblies Using Linked-Reads (REXTAL) for improved region-specific assembly of segmental duplication-containing DNA, leveraging genomic short-read datasets generated from large DNA molecules partitioned and barcoded using the Gel Bead in Emulsion (GEM) microfluidic method [1]. 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.

Acknowledgement. The work in this paper is supported in part 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.: Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnol. **34**, 303–311 (2016)
2. Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B.: Direct determination of diploid genome sequences. Genome Res. **27**, 757–767 (2017)
3. Alkan, C., Sajjadian, S., Eichler, E.E.: Limitations of next-generation genome sequence assembly. Nature Methods **8**, 61 (2011)
4. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome Res. **12**, 996–1006 (2002)
5. Smit, A.F.: 2010 RepeatMasker Open-3.0 (1996). http://www.repeatmasker.org/
6. Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**, 573 (1999)
7. Kent, W.J.: BLAT the BLAST-like alignment tool. Genome Res. **12**, 656–664 (2002)
8. Altschul, S.F., Madden, T.L., Schffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
9. Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G.: QUAST: quality assessment tool for genome assemblies. Bioinformatics **29**, 1072–1075 (2013)
10. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754–1760 (2009)
11. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The sequence alignment/map format and SAMtools. Bioinformatics **25**, 2078–2079 (2009). 1000 Genome Project Data Processing Subgroup