Analysis of Subtelomeric REXTAL Assemblies Using QUAST

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

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

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

Eleanor Young
School of Biomedical Engineering
Drexel University
Philadelphia, PA, USA
eay25@glink.drexel.edu

Ming Xiao
Institute of Molecular Medicine
and Infectious Disease
Drexel University
Philadelphia, PA, USA
mx44@drexel.edu

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

***Abstract*—Genomic regions of high segmental duplication content and/or structural variation have led to gaps and misassemblies in the human reference sequence, and are refractory to assembly from whole-genome short-read datasets. Human subtelomere regions are highly enriched in both segmental duplication content and structural variations, and as a consequence are both impossible to assemble accurately and highly variable from individual to individual. Recently, we developed a pipeline for improved region-specific assembly called Regional Extension of Assemblies Using Linked-Reads (REXTAL) [1]. In this study, we evaluate REXTAL and genome-wide assembly (Supernova; [2]) approaches on 10X Genomics linked-reads data sets partitioned and barcoded using the Gel Bead in Emulsion (GEM) microfluidic method [3]. Our results describe the accuracy and relative performance of these two approaches using the reference-based assessment module of QUAST [4]. We show that REXTAL dramatically outperforms the Supernova whole genome assembler in subtelomeric segmental duplication regions, and results in highly accurate assemblies. Nearly all of the REXTAL "misassemblies" identified using default QUAST parameters simply pinpoint locations of tandem repeat arrays in the reference sequence where the repeat array length differs from that in the cognate REXTAL assembly by $>$ 1000 bp.**

***Index Terms*—regional assembly, quality metric, segmental duplication, subtelomere, tandem repeat, misassembly, genome gap.**

I. Introduction

It is currently impossible to get complete de-novo assembly of segmentally duplicated genome regions using genome-wide short-read datasets. Even using paired-end read approaches with input molecules of various lengths, de novo assembly of human genomes has remained problematic because of abundant interspersed repeats and especially segmental duplication regions which contain $>$ 1 kbp segments of DNA with similar ($>$ 90%) identity. 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 [3]. The bioinformatic pipeline for assembly of these reads called Supernova [2] takes advantage of a 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 bar code. This approach enables efficient de novo assembly of the human genome, with large segments separable into haplotypes [2]. However, it does not solve the problem of segmental duplications such as those found in subtelomeres. To address this problem, we developed a new computational method called REXTAL [1] 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.

In this paper, we do a more extensive analysis of REXTAL and evaluate our regional assemblies with the reference-based alignment tool (QUAST; [4]) on 17 subtelomeric DNA regions. We find dramatically improved coverage of subtelomeric segmental duplication regions in REXTAL vs. whole genome assemblies while maintaining accurate assemblies using REXTAL.

II. Experimental Details

In Subsection A, we present the input data description. Subsection B presents the overview of REXTAL methodology. In subsections C and D we describe QUAST analysis and visualization of our final assemblies.

A. Data

The key input data for REXTAL [1] 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. Human reference genome assembly HG38 is used to select test subtelomere regions for the targeted assemblies. The input data for QUAST [4] is the respective regional assembly generated by REXTAL and the cognate subtelomeric reference sequence from HG38. Details of the reference subtelomeric regions including subtelomeric 1-copy regions as well as subtelomeric segmental duplication regions can be found in Table I.

B. REXTAL Methodology

REXTAL [1] uses linked-read genome sequencing to extend subtelomere assemblies. It differs from the genome-wide assembly method in that we used 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 used RepeatMasker [8] and Tandem Repeats Finder [9] 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. We used BLAT [13] to do the alignment of the masked subtelomeric region with genome-wide reads from GM19440. We initially found reads matching the 1-copy DNA segment (bait DNA segment), then selected all reads for barcodes represented in these initial matching reads in reads selection step (Fig. 1). 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 kbp - 100 kbp segments of flanking DNA. Using barcode read frequency range selection and barcode clustering pattern selection steps (Fig. 1) we selected all reads from a subset of these initial barcodes for assembly [1], enabling the extension of existing assemblies into adjacent segmental duplication and gap regions using the Supernova assembler [2]. Fig. 1 shows the overall REXTAL workflow.

C. QUAST Analysis

QUAST [4] evaluates genome assemblies by computing various metrics from a global alignment of the test assembly with a reference sequence. To measure the quality of the assembly, we ran QUAST with --scaffolds option (keeping other parameters default) using assembled scaffolds generated by REXTAL and using as reference sequence specified subtelomeric regions of HG38 corresponding to our unmasked single-copy bait segments along with their flanking reference DNA segments (including segmental duplication regions).

1) Scaffold: As REXTAL assemblies are scaffolds (rather than contigs) and we ran QUAST with --scaffolds option, this added split versions of assemblies to the comparison (named <assembly_name>_broken). Assemblies are split by continuous fragments of N's of length ≥ 10. Scaffold gap size misassemblies are enabled in this case and we kept default --scaffold-gap-max-size (which is 10 kbp) for setting maximum gap length.

2) Misassembly Detection: QUAST [4] generates a report with the number of misassemblies according to the defined misassembly breakpoint by Plantagora [5]. Misassembly breakpoint is a position in the assembled contigs where the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference, or they overlap by $>$ 1 kbp, or the flanking sequences align on opposite strands or different chromosomes. While running QUAST we kept default threshold of 1 kbp for --extensive-mis-size parameter. Most of the "misassemblies" called in REXTAL generated assemblies relative to reference were due to the gap sizes in a contig slightly exceeding the QUAST default gap limit of 1000 bp.

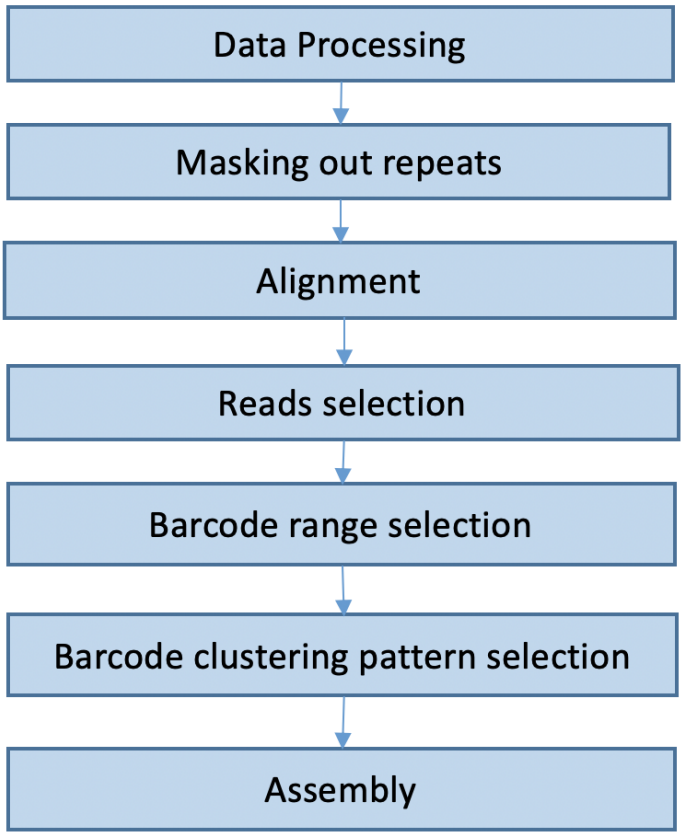

Fig. 1. Overview of REXTAL workflow.

D. Visualization

We used Icarus [6] a genome visualizer for assessment and analysis of genomic assemblies, which is based on QUAST genome quality assessment tool. The contig alignment viewer of Icarus has 2 parts. The top part shows the detailed view of selected region from the bottom part which represents the assembly overview.

1) Broken scaffold: To view only the split versions of assemblies in the Icarus viewer following steps were followed:

Step1: At first we ran the referenced-based QUAST from a command line with --scaffolds and --debug option [12].

TABLE I
CO-ORDINATE OF EXTRACTED SUBTELOMERIC REGION FROM UCSC BROWSER

Region[a]	**Ref**[b]	**Bait**[c]	**SD**[d]	**1-copy**[e]
2p	10,001-700,000	10,001-500,000	N/A	10,001-700,000
5p	10,001-677,959	49,496-210,595 and 305,379-510,000	10,001-49,495 and 210,596-305,378	305,379-677,959
10p	10,001-588,571	88,571-388,571	10,001-88,570	88,571-588,571
16p	10,000-240,859	40,860-140,859	10,000-40,859	40,860-240,859
16q	89,857,010-90,228,345	89,857,010-89,965,857 and 89,968,061-90,057,009	89,965,858-89,968,060 and 90,057,010-90,228,345	89,857,010-89,965,857 and 89,968,061-90,057,009
17p	60,000-341,850	141,851-241850	60,000-141,850	141,851-341,850
17q	83,004,545-83,247,441	83,104,544-83,204,544	83,204,545-83,247,441	83,004,545-83,204,544
18p	10,000-331,693	131,694-231,693	10,000-131,693	131,694-331,693
18q	80,059,053-80,263,285	80159052-80,259,052	80,259,053-80,263,285	80,059,053-80,259,052
19p	10,001-759,447	259,448-559,447	10,001-259,447	259,448-759,447
19q	58,386,558-58,607,616	58486557-58,586,557	58,586,558-58,607,616	58,386,558-58,586,557
20p	66,335-266,334	66,335-166334	N/A	66,335-266,334
20q	64,073,499-64,334,167	64,173,498-64,273,498 and 64,276,019-64,282,623	64,273,499-64,276,018 and 64,282,624-64,334,167	64,073,499-64,273,498 and 64,276,019-64,282,623
21q	46,472,945-46,699,983	46572944-46,672,944	46,672,945-46,699,983	46,472,945-46,672,944
22q	50,540,514-50,808,468	50640513-50,740,513	50,740,514-50,808,468	50,540,514-50,740,513
Xp	222,347-527,305	222,347-320,315 and 327,306-427,306	320,316-327,305	222,347-320,315 and 327,306-527,305
Xq	155,783,780-156,030,894	155,883,778-155,983,778 and 155,987,225-156,000,330	155,983,779-155,987,224 and 156,000,331-156,030,894	155,783,780-155,983,778 and 155,987,225-156,000,330

a. Subtelomeric region.
b. HG38 Co-ordinates of reference subtelomeric region (HG38).
c. HG38 Co-ordinates of 1-copy subtelomeric bait region.
d. HG38 Co-ordinates of subtelomeric segmental duplication region.
e. HG38 Co-ordinates of entire subtelomeric 1-copy region.

Step2: If an output path is not specified manually (we can specify output path of QUAST by using -o option), QUAST generates its output into quast_results/result_<DATE> directory. We chose the <assembly_name>_broken version file under the quast_results/result_<DATE>/quast_corrected_input/ directory [12].

Step3: We reran the referenced-based QUAST from a command line with the same reference but used <assembly_name>_broken instead of the original assembly and did not use --scaffolds option this time.

2) Tandem Repeat Marker: Since there is no special visualization for repeats yet in QUAST [12], we used tandem repeat finder [9] to screen subtelomeric regions of reference DNA segment sequences. We then used this masked reference as input data for QUAST with the same unmasked subtelomeric regions as reference and followed the procedure described in subsubsection II-D1. We used the broken masked reference to locate the positions of tandem repeats in Icarus viewer (Fig. 2 – Fig. 5).

3) Comparative Analysis: To visualize the comparative analysis of REXTAL and Genome-wide method as well as the tandem repeat marker, we ran reference based QUAST with 3 input files i.e. first one is broken masked reference file as tandem repeat marker, then broken REXTAL assembly and 3^{rd} one is broken Genome-wide assembly.

III. RESULTS AND DISCUSSIONS

UCSC browser [7] was used to access HG38 and select subtelomere DNA segments for analysis. We tested REXTAL and

TABLE II
COMPARISON OF QUAST RESULT IN SEGMENTAL DUPLICATION REGION

	REXTAL		**Genome-wide**	
Region[a]	**Genome fraction (%)**	**Misassembly**[b]	**Genome fraction (%)**	**Misassembly**[c]
2p	N/A	N/A	N/A	N/A
5p_1^{st}	88.707	0	57.734	0
5p_2^{nd}	88.038	0	3.034	0
10p	90.103	1	7.092	0
16p	94.18	0	25.493	0
16q_1^{st}	100	0	100	0
16q_2^{nd}	35.423	1	16.549	0
17p	21.52	0	N/A	N/A
17q	85.12	1	0.494	0
18p	69.028	0	9.701	0
18q	82.731	0	60.761	0
19p	28.14	0	2.139	0
19q	92.906	0	1.22	0
20p	N/A	N/A	N/A	N/A
20q_1^{st}	100	0	100	0
20q_2^{nd}	98.462	0	6.955	0
21q	94.941	0	23.806	0
22q	96.087	0	5.055	0
Xp	59.828	0	16.753	0
Xq_1^{st}	100	0	100	0
Xq_2^{nd}	90.211	0	62.174	0

a. Subtelomeric region.
b. "Misassemblies" are tandem repeat arrays in the reference sequence where the repeat array length differs from that in the cognate REXTAL assembly by $>$ 1000 bp.
c. Number of misassembly in genome-wide method.

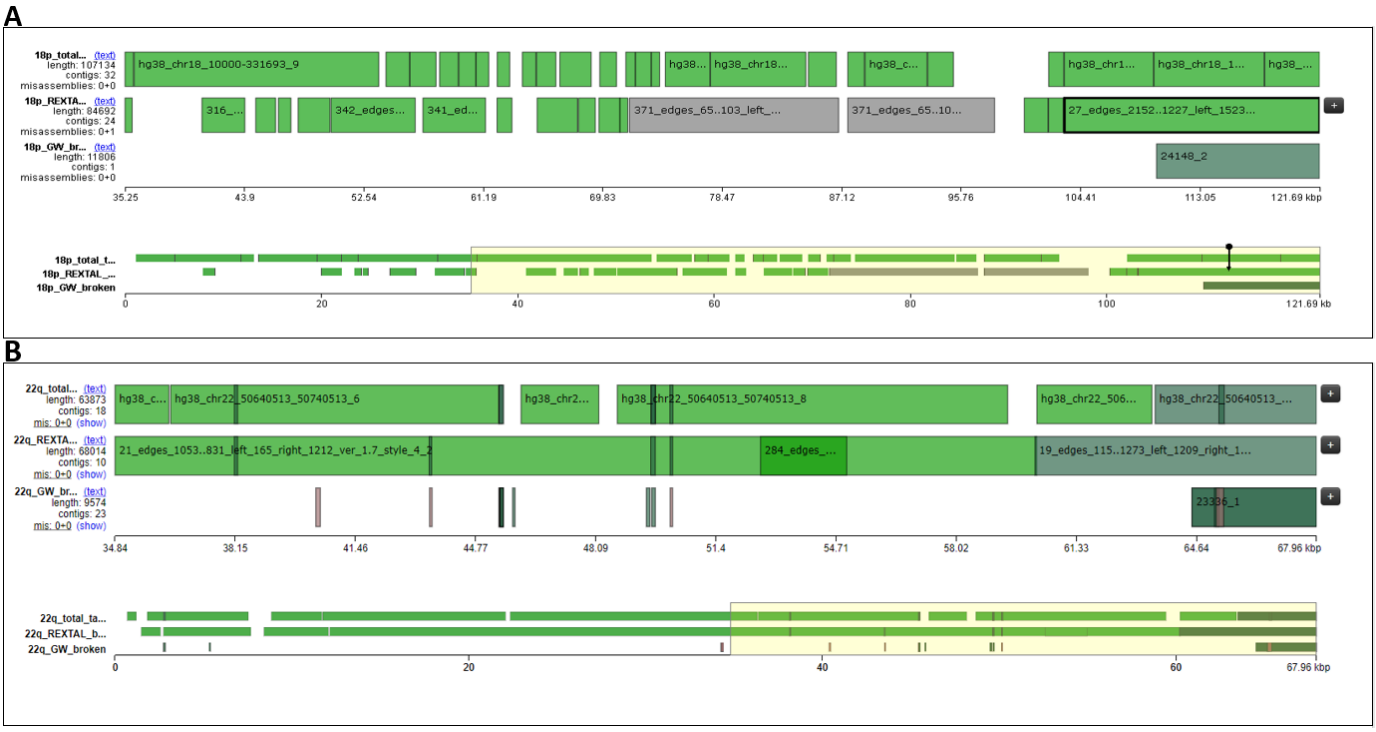

Fig. 2. Contig alignment viewer of Icarus for the segmental duplication region region of 18p and 22q. Each viewer has 2 parts containing 3 rows in each part. The top green bars of the top part represents the reference sequence with white breaks in this track showing the positions and sizes of tandem repeats in the reference. The 2^{nd} and 3^{rd} rows show the REXTAL and the genome-wide assemblies respectively. The bottom part having 3 rows represents the assembly overview with the highlighted yellow box indicating the region expanded in the top 3 rows. **A.** The 2^{nd} and the 3^{rd} row of top part represent the contigs generated by REXTAL and genome-wide method for 18p correspondingly. **B.** The 2^{nd} and the 3^{rd} row of top part represent the contigs generated by REXTAL and genome-wide method for 22q correspondingly. In both **A** and **B** the REXTAL assemblies extend over a large part of the segmental duplication regions.

the QUAST analysis on 17 human subtelomere regions (base pair coordinates listed are from HG38). The 2p subtelomere is a 500 kbp 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 kbp 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 kbp 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). For 16p, 16q, 17p, 17q, 18p, 18q,19q, 20p, 20q, 21q, 22q, Xp, Xq we extracted 100 kbp single copy bait sequences as close as possible to the telomere. Table I shows the details of subtelomeric region.

For a fair comparison of REXTAL with genome-wide assembly method, we extracted all contigs in the genome-wide assembly that overlap (including potential extensions into flanking DNA) with the 1-copy bait sequences using SAMtools [11].

A. QUAST report on genome fraction in segmental duplication region

As segmental duplication regions contain segments of DNA with near-identical duplicated subtelomere sequences, these regions are hard to assemble de novo with whole genome reads. We can extend REXTAL into subtelomere segmental duplication regions. To measure the quality of REXTAL vs. genome-wide assemblies in segmental duplication regions, we ran reference based QUAST for these regions (Table II). 2p and 20p do not have segmental duplication regions. 5p, 16q, 20q, and Xq have multiple segmental duplication regions. For 17p, genome-wide method could not extend the assembly up to segmental duplication region.

It is easy to observe that the % of genome fractions obtained by REXTAL (2^{nd} Column of Table II) are significantly better than the % of genome fractions obtained by genome-wide method (4^{th} Column of Table II) for all loci that have been tested.

Fig. 2A and Fig. 2B show the visualization of QUAST analysis of 18p and 22q in segmental duplication regions.

B. QUAST report on misassembly in segmental duplication region

Generally, QUAST report contains a classification of misassembly events (using Plantagoras [5] definition) into three groups: relocations, translocations, and inversions (subsubsection II-C2).

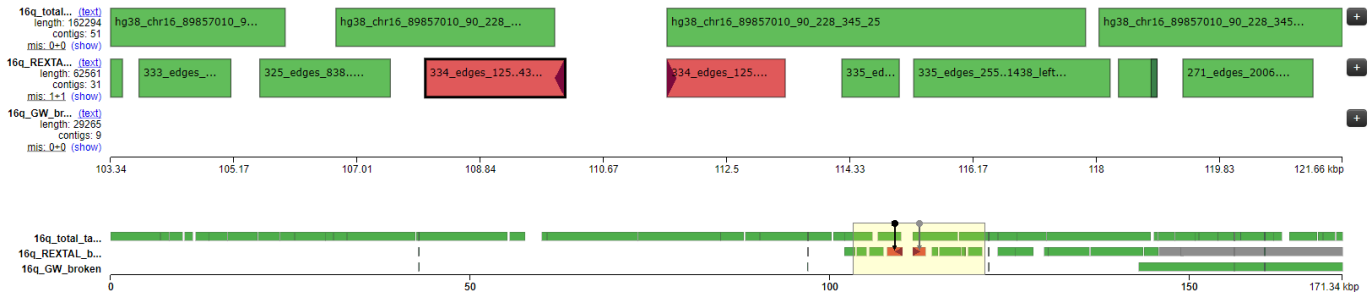

Fig. 3. Contig alignment viewer of Icarus for the segmental duplication region of 16q_2^{nd}.The top green bars of the top part represents the reference white breaks in this track showing the positions and sizes of tandem repeats in the reference. The 2^{nd} row represents the contigs generated by REXTAL and two red blocks represents the misassembled contig with gap 1512 bp. 3^{rd} row is supposed to be the contigs generated by genome-wide method for segmental duplication region of 16q_2^{nd} and this row shows nothing here because genome-wide method could not extend the assembly up to this point. The bottom three rows represent the assembly overview with the highlighted yellow box indicating the region expanded in the top 3 rows. Note that the "misassembled contig" is in fact a gap in the contig corresponding exactly to a tandem repeat. It is called a QUAST misassembly only because it exceeds the 1000 bp default when aligned to the reference sequence.

TABLE III
COMPARISON OF QUAST RESULT IN THE BAIT SEGMENT INTO ADJACENT 1-COPY REGION

	REXTAL		Genome-wide	
Region[a]	**Genome fraction (%)**	**Misassembly**[b]	**Genome fraction (%)**	**Misassembly**[c]
2p	96.226	1	96.839	0
5p_1^{st}	93.471	0	94.818	0
5p_2^{nd}	91.68	1	93.789	1
10p	97.966	1	98.142	0
16p	75.873	0	55.733	0
16q_1^{st}	96.759	3	96.146	0
16q_2^{nd}	97.719	1	96.606	0
17p	68.839	5	39.774	0
17q	77.839	1	45.235	0
18p	93.448	0	99.71	0
18q	87.391	0	98.668	1
19p	87.173	2	85.141	0
19q	81.459	1	55.077	0
20p	86.243	1	100.00	0
20q_1^{st}	71.204	0	53.714	0
20q_2^{nd}	100.00	0	100.00	0
21q	73.621	0	99.974	0
22q	70.74	0	96.547	1
Xp_1^{st}	72.381	1	26.064	1
Xp_2^{nd}	72.103	0	13.322	0
Xq_1^{st}	87.678	0	90.088	0
Xq_2^{nd}	100.00	0	100.00	0

a. Subtelomeric region.
b. "Misassemblies" are tandem repeat arrays in the reference sequence where the repeat array length differs from that in the cognate REXTAL assembly by $>$ 1000 bp.
c. Number of misassembly in genome-wide method.

The number of misassemblies obtained in segmental duplication region by REXTAL and genome-wide method are shown correspondingly in 3^{rd} and 5^{th} Column of Table II. QUAST generated one misassembly for 10p, 16q_2^{nd} (2^{nd} segmental duplication region of 16q), and 17q and all these three misassemblies happened because of relocation according to QUAST report. Fig. 3 shows the misassembled contig (two red blocks) for segmental duplication region of 16q_2^{nd}. The cause of the misassembly was relocation with inconsistency = 1512 bp. As the top green bars represent tandem repeat marker and the gap between green top bars represent the tandem repeat region, the misassembly happened in tandem repeat region. These misassembled blocks are in one contig. Genome-wide method could not extend the assembly up to this point. To run the QUAST we used default value of parameter --extensive-mis-size and that is 1000. If we set the parameter of --extensive-mis-size with higher value, we would not find these misassemblies.

TABLE IV
COMPARISON OF QUAST RESULT IN THE BAIT SEGMENT INTO ADJACENT DNA INCLUDING SEGMENTAL DUPLICATION REGION

	REXTAL		Genome-wide	
Region[a]	**Genome fraction (%)**	**Misassembly**[b]	**Genome fraction (%)**	**Misassembly**[c]
2p	75.535	1	71.337	0
5p	75.448	1	70.945	1
10p	72.933	2	53.707	0
16p	78.2	0	51.56	0
16q	68.623	2	59.263	0
17p	54.829	5	28.38	0
17q	78.813	2	37.142	0
18p	84.151	0	65.66	0
18q	87.138	0	96.764	1
19p	51.108	3	35.316	1
19q	78.18	0	47.087	0
20p	86.208	1	99.999	0
20q	77.442	0	46.099	0
21q	75.861	0	90.902	0
22q	77.156	0	73.266	1
Xp	70.569	1	16.982	0
Xq	88.709	0	88.709	0

a. Subtelomeric region.
b. "Misassemblies" are tandem repeat arrays in the reference sequence where the repeat array length differs from that in the cognate REXTAL assembly by $>$ 1000 bp.
c. Number of misassembly in genome-wide method.

Fig. 4. Contig alignment viewer of Icarus for 1-copy region of 19q and 17q. Each viewer has 2 parts containing 3 rows in each part. The top green bars of the top part represents the reference with white breaks in this track showing the positions and sizes of tandem repeats in the reference. The 2^{nd} and 3^{rd} rows show the REXTAL and the genome-wide assemblies respectively. The bottom part having 3 rows represents the assembly overview with the highlighted yellow box indicating the region expanded in the top 3 rows. **A.** The 2^{nd} row (including the expansion (yellow area)) represents the contigs generated by REXTAL and the 3^{rd} row represents the contigs generated by genome-wide method for 19q. The expanded version of 2^{nd} row shows that a misassembled contig has seven blocks, among them two blocks (red blocks) are misassembled because of relocation with inconsistency = 1115 bp. This misassembled contig is located entirely within another higher-quality contig (1 green block in 2^{nd} row). **B.** The 2^{nd} row represents the contigs generated by REXTAL and the 3^{rd} row represents the contigs generated by genome-wide method for 17q. The misassembled contig has four blocks (in assembly overview image there is a light yellow rectangle representing the selected region and four down arrows (↓) represent four blocks in one contig.). Among them two blocks (red blocks) are misassembled with inconsistency = 1168 bp. These two misassembled blocks are in one contig in REXTAL assembly but two different contigs in genome-wide assembly. In the selected region of genome-wide method has seven different assembled contigs whereas REXTAL has one contig with four blocks with gaps. Note that the "misassembled contig" is in fact a gap in the contig corresponding exactly to a tandem repeat. It is called a QUAST misassembly only because it exceeds the 1000 bp default when aligned to the reference sequence.

C. QUAST report on the bait segment into adjacent 1-copy region

We extracted subtelomeric region containing 1-copy and 1-copy bait region as reference from UCSC genome browser to compare the extending assemblies of the bait segment into adjacent 1-copy region for REXTAL and genome-wide method. To show the quality of REXTAL vs. genome-wide assembly, we ran QUAST with both assemblies (Table III).

Fig. 4A and 4B show the contig alignment viewer of Icarus in the bait segment into adjacent 1-copy region for 19q and 17q correspondingly.

For 19q QUAST reports 1 misassembly in 1-copy region for REXTAL (3^{rd} Column of Table III). Fig. 4A shows that the misassembled contig corresponds to a small contig matching less accurately (identity 96% – 98%) among seven small blocks to the reference than a larger, more complete and closely matching contig (99.92% identity) that completely encompasses the smaller contig. Among the seven blocks two blocks (red blocks) are misassembled because of relocation with inconsistency value 1115 bp. However, this misassembled contig is located entirely within the other higher-quality (99.92% identity) contig (1 green block in 2^{nd} row in Fig. 4A). To avoid this situation in our prior work we proposed a metric called Lengthwise Assembled Fraction (LAF) [1] for quality measurement of the regional assemblies. Before measuring the quality, we extracted reference sequences from HG38 and then aligned them with corresponding assembled scaffolds using BLAST [10], requiring $\geq$ 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 [1]. The LAF metric avoids the secondary more weakly matching assemblies like that shown above.

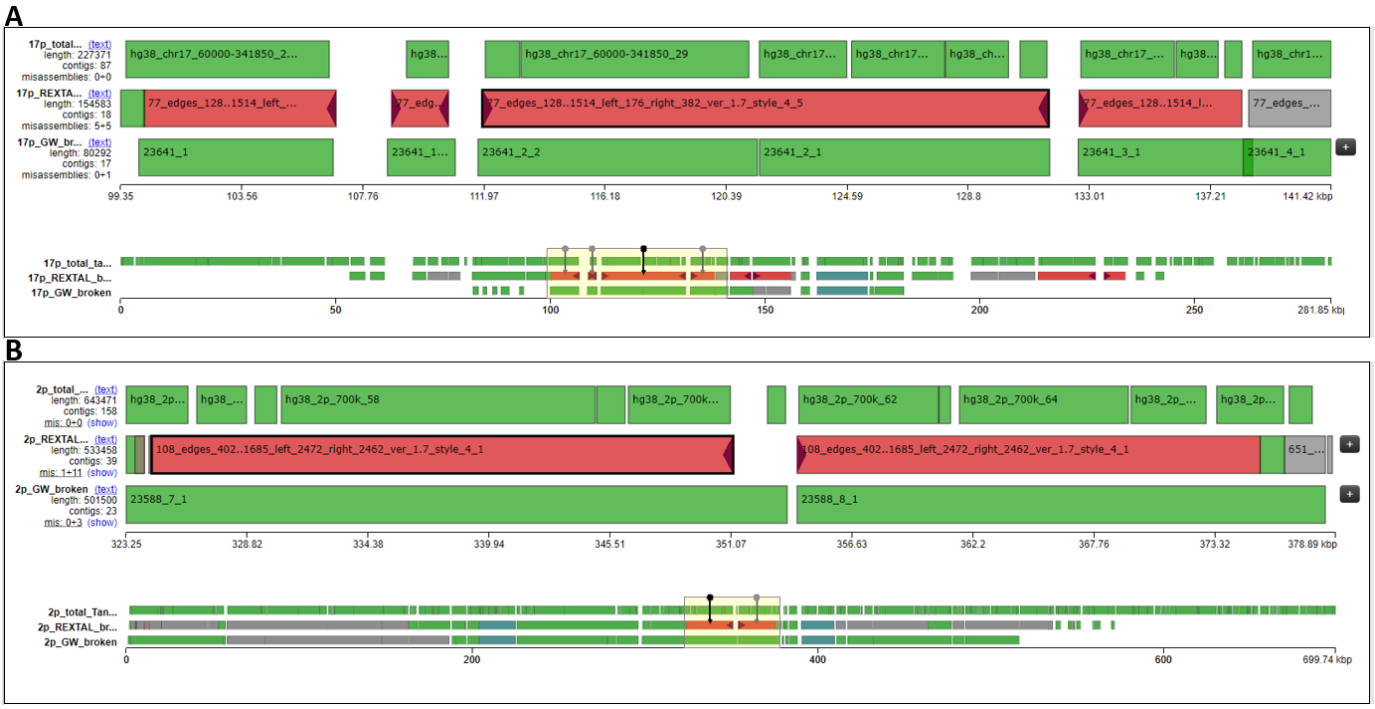

Fig. 5. Contig alignment viewer of Icarus for the bait segment extension into adjacent DNA including 1-copy and segmental duplication region of 17p and 2p. Each viewer has 2 parts containing 3 rows in each part. The top green bars of the top part represents the reference with white breaks in this track showing the positions and sizes of tandem repeats in the reference. The 2^{nd} and 3^{rd} rows show the REXTAL and the genome-wide assemblies respectively. The bottom part having 3 rows represents the assembly overview with the highlighted yellow box indicating the region expanded in the top 3 rows. **A.** The 2^{nd} row represents the contigs generated by REXTAL and the 3^{rd} row represents the contigs generated by genome-wide method for 17p. There are four red blocks in a contig that is misassembled because of relocation with inconsistency value 1920 bp, 1172 bp, and 1055 bp. The genome-wide assembly calls the single REXTAL assembly five separate unrelated contigs. **B.** The 2^{nd} row represents the contigs generated by REXTAL and the 3^{rd} row represents the contigs generated by genome-wide method for 2p. The two red block represents the misassembly because of 2935 bp gap between two blocks within a contig. Note that the "misassembled contig" is in fact a gap in the contig corresponding exactly to a tandem repeat. It is called a QUAST misassembly only because it exceeds the 1000 bp default when aligned to the reference sequence.

In Fig. 4B, for 17q the misassembled contig has four blocks (in assembly overview image there is a light yellow rectangle representing the selected region and four down arrows (↓) represent four blocks in one contig). Among them two blocks (red blocks) are misassembled because of relocation with inconsistency value 1168 bp. These two misassembled blocks are in one contig in REXTAL assembly but two different contigs in genome-wide assembly. Overall the selected region of genome-wide method has seven different assembled contigs in the genome-wide assembly whereas REXTAL has one contig with four blocks with gaps. The gaps all correspond to tandem repeat regions where REXTAL was able to assemble across the tandem repeat region putting gaps in a contig rather than creating separate unrelated contigs. We can avoid these misassembly calls by setting the parameter of --extensive-mis-size with slightly higher value during running the QUAST.

D. QUAST report on the bait segment into adjacent DNA including segmental duplication region

We extracted subtelomeric region containing 1-copy, 1-copy bait and segmental duplication region as reference from UCSC genome browser to compare the extending assemblies of the bait segment into adjacent DNA for REXTAL and genome-wide method.

To show the quality of REXTAL vs. genome-wide assembly, we ran reference based QUAST for these regions and compared results in Table IV.

1) Analysis of genome fraction and misassmblies: REXTAL has better % of genome fraction than whole genome assembly except for 18q, 20p, and 21q (2^{nd} Column of Table IV). In subsection III-A we showed that 18q and 21q have noticeably good extension in segmental duplication region (2^{nd} Column of Table II). 20p is all single copy region and the genome-wide method gave a better genome fraction here than the REXTAL.

Fig. 5A and Fig. 5B show the contig alignment viewer of Icarus in the bait segment into adjacent DNA including segmental duplication region for 17p and 2p correspondingly.

For 17p, QUAST generates total 5 misassemblies (3^{rd} Column of Table IV) on the bait segment into adjacent DNA region. Fig. 5A shows that there are four red blocks in a contig that were misassembled because of relocation with inconsistency value 1920 bp, 1172 bp, and 1055 bp, where the genome-wide method has five separate unrelated contigs

instead of these misassemblies.

In Fig. 5B for 2p similar case happened in tandem repeat region where misassembly happened because of the gap (inconsistency = 2935 bp) between two blocks within a contig. Genome-wide method considered these two blocks as two separate contigs.

Both for 17p and 2p (Fig. 5), it is noticeable that the "misassembled contig" is in fact a gap in the contig corresponding exactly to a tandem repeat. It is called a QUAST misassembly only because it exceeds the 1000 bp default when aligned to the reference sequence. We can avoid these errant misassembly calls by setting the parameter of --extensive-mis-size with higher value during running QUAST.

IV. Conclusion

We successfully used REXTAL [1] on 17 subtelomeric bait regions and extended the assembly of single-copy diploid DNA into adjacent DNA including inaccessible subtelomere segmental duplication regions. We evaluated REXTAL and genome-wide assemblies using the reference-based assessment module of QUAST and showed that REXTAL dramatically outperformed the Supernova whole genome assembler in subtelomeric segmental duplication regions, and produced highly accurate assemblies. In future experiments, we will combine REXTAL and Nanopore single-read datasets to achieve complete long-range assemblies throughout all human subtelomere regions.

Acknowledgments

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] Islam, T. et al.,"REXTAL: Regional Extension of Assemblies Using Linked-Reads", International Symposium on Bioinformatics Research and Applications, pp. 63–78, 2018.

[2] Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB, "Direct determination of diploid genome sequences", Genome research, 27, pp. 757–767, 2017.

[3] Zheng, G. X.-L.-P. et al., "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing", Nature biotechnology, 34, pp. 303–311, 2016.

[4] Gurevich A, Saveliev V, Vyahhi N, Tesler G, "QUAST: quality assessment tool for genome assemblies," Bioinformatics, 29, pp. 1072–1075, 2013.

[5] Barthelson R, et al., "Plantagora: modeling whole genome sequencing and assembly of plant genomes," PLoS One, 6:e28436, 2011.

[6] Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A, "Icarus: visualizer for de novo assembly evaluation", Bioinformatics, 32, pp. 3321–3323, 2016.

[7] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D., "The human genome browser at UCSC", Genome research, 12, 996–1006, 2002.

[8] Smit, A. F. (1996) RepeatMasker Open-3.0 http://www.repeatmasker.org/.

[9] Benson, G., "Tandem repeats finder: a program to analyze DNA sequences", Nucleic acids research, 27, 573, 1999.

[10] Altschul SF, Madden TL, Schffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic acids research, 25, 3389–3402, 1997.

[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, "The sequence alignment/map format and SAMtools", Bioinformatics, 25, 2078–2079, 2009.

[12] Alexey Gurevich, email = "alexeigurevich@gmail.com", Affiliation = Research Scientist at Center for Algorithmic Biotechnology, Saint Petersburg State University.

[13] Kent, W. J., "BLAT the BLAST-like alignment tool", Genome research, 12, 656–664, 2002.

Tunazzina Islam received her M.Sc. degree in Computer Science from Old Dominion University (ODU), Norfolk, VA, USA in August 2018 and B.Sc. degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2013. Tunaz started pursuing Ph.D. degree in Computer Science at Purdue University, West Lafayette, IN, USA in January 2019.

Desh Ranjan received his undergraduate degree in computer science from Indian Institute of Technology, Kanpur (INDIA) (1987) and a Masters' and Ph.D. in Computer Science, with a minor in Mathematics, from Cornell University in Ithaca, New York in 1990 and 1992 respectively. After finishing his Ph.D., he spent one year (1992-93) as a post-doctoral fellow at the Max Planck Institute for Computer Science in Saarbrucken, Germany. He was a member of the faculty of the Department of Computer Science at New Mexico State University for 16 years and served as the chair of the department from 2004 to 2009. He was the founding director of the NSF funded Center for Bioinformatics and Computational Biology at NMSU. He is currently the Bioinformatics Endowed Professor in the Department of Computer Science at Old Dominion University in Norfolk, Virginia. Desh's research interests include Algorithms, Bioinformatics and Computational Complexity.

Mohammad Zubair has more than twenty-five years of research experience in the area of experimental computer science and engineering both at the university as well as at the Industry. He is a professor of computer science at Old Dominion University. His primary interest is in the area of application of high-performance computing to bioinformatics, scientific computing and big data analytics. Mohammad Zubair has experience working with industry, his major industrial assignment was at IBM T.J. Watson Research Center, where his research focus was in high-performance computing and some of his work was integrated into IBM products: Engineering Scientific Subroutine Library (ESSL), and Parallel ESSL. He has been successful in obtaining funds to support his research work from NSF, DTIC, DARPA, Jefferson Laboratory, NASA, Los Alamos, AFRL, NRL, JTASC, Sun Microsystems, and IBM Corporation.

Eleanor Young is a PhD Candidate at Drexel University in the department of Biomedical Engineering. She received a BE in Chemical Engineering from the University of Dayton. Her research interests include human genomics and bioinformatics.

Ming Xiao received his BA degree from Huazhong University of Science and Technology, and his PhD degree from Baylor University. Dr. Xiao's career in genomics has bridged both industry and academia. He has developed many single molecule DNA analysis methods. He is currently an Associate Professor in the School of Biomedical Engineering, Science and Health Systems, Drexel University. His current research is focused on genomic technology development and human genome variation.

Harold Riethman received his BS and MS from the University of Cincinnati, and his PhD from the University of Missouri. Following postdoctoral fellowships in Genetics at Washington University in St. Louis, he began his independent career at The Wistar Institute in Philadelphia, with adjunct appointments in Genetics and Biology at the University of Pennsylvania. His laboratory played a key role in the sequencing and analysis of subtelomeres and telomeres, complex DNA regions near the tips of chromosomes, as part of the human genome project. Dr. Riethman joined Old Dominion University in 2015 and is currently Chair of the School of Medical Diagnostic and Translational Sciences. His current research is focused on the structure, molecular genetics and biology of human telomeres as they relate to cancer, aging, and stem cell biology.