

Old Dominion University

ODU Digital Commons

Computer Science Faculty Publications

Computer Science

2021

NPGreat: Assembly of the Human Subtelomere Regions with the Use of Ultralong Nanopore Reads and Linked Reads

Eleni Adam

Desh Ranjan

Harold Riethman

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_fac_pubs



Part of the [Cell Anatomy Commons](#), [Computer Sciences Commons](#), [Genetics Commons](#), and the [Genetic Structures Commons](#)

NPGREAT: Assembly of the human subtelomere regions with the use of ultralong Nanopore reads and Linked-Reads

Eleni Adam (✉ eadam002@odu.edu)

Old Dominion University

Desh Ranjan

Old Dominion University

Harold Riethman



Old Dominion University

Research Article

Keywords: telomeres, segmental duplications, tandem repeats, hybrid assembly, nanopore, Linked-Reads, TELL-Seq, 10x

Posted Date: November 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1080088/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Human subtelomeric DNA regulates the length and stability of adjacent telomeres that are critical for cellular function, and contains many gene/pseudogene families. Large evolutionarily recent segmental duplications and associated structural variation in human subtelomeres has made complete sequencing and assembly of these regions difficult to impossible for many loci, complicating or precluding a wide range of genetic analyses to investigate their function.

Results

We present a hybrid assembly method, NanoPore Guided REgional Assembly Tool (NPGREAT), which combines Linked-Read data with ultralong nanopore reads spanning subtelomeric segmental duplications to potentially overcome these difficulties. Linked-Read sets identified by matches with 1-copy subtelomere sequence adjacent to segmental duplications are assembled and extended into the segmental duplication regions using Regional Extension of Assemblies using Linked-Reads (REXTAL). Telomere-containing ultralong nanopore reads are then used to provide contiguity and correct orientation for matching REXTAL sequence contigs as well as identification/correction of any misassemblies (associated primarily with tandem repeats). While we focus on subtelomeres, the method is generally applicable to assembly of segmental duplications and other complex genome regions. Our method was tested for a subset of representative subtelomeres with ultralong nanopore read coverage in GM12878. 10X Linked-Read datasets with high depth of coverage and a TELL-seq Linked-Read dataset with lower depth of coverage were each combined with the ultralong nanopore reads from the same genome to provide improved assemblies. Tandem repeat regions of the short-read assemblies, which are especially prone to misassembly due to collapse of matching tandemly repeated reads, were readily identified and properly sized by comparison with the nanopore reads.

Conclusion

The NPGREAT method resulted in extension of high-quality assemblies into otherwise inaccessible segmental duplication regions near telomeres, enhancing our ability to accurately assemble human subtelomere DNA. This information will enable improved analyses of the structure, function, and evolution of these key regions.

Background

Telomeres are essential for proper replication and stability of chromosomes. They consist of stretches of (TTAGGG) repeat DNA at the ends of chromosomes with associated proteins and at least one lncRNA, TERRA; their dysfunction can contribute to diseases including cancer through multiple mechanisms ([1], [2], [3], [4], [5]). Subtelomere regions regulate adjacent single-telomere lengths and stabilities of human chromosomes in both telomerase-positive and telomerase-negative contexts ([6], [7], [8], [9]); thus, accurate maps and DNA sequences for human subtelomere regions, along with detailed knowledge of subtelomere variation and long-range

telomere-terminal haplotypes in individuals, are critical for understanding telomere function and its roles in human biology.

Nearly 20 years after completion of the human genome project, human subtelomere regions remain poorly represented in the current human reference sequence (currently HG38) and are typically mis-assembled or entirely absent from short-read whole genome assemblies. The principal obstacles to acquiring complete subtelomeric sequences are the abundance of large, highly similar segmental duplication regions and the very high level of structural variation. We have recently developed a short-read assembly strategy based upon identification of Linked-Reads derived from large source DNA molecules extending from 1-copy DNA into segmental duplications, and their assembly to extend high-quality sequence into the segmental duplication regions [10]. Here, we use Linked-Read datasets from 10X genomic sequencing to show that subtelomeric assemblies can be improved and extended across a group of representative human subtelomeric segmental duplication regions by combining them with telomere-containing ultralong nanopore reads spanning the segmental duplication regions on these subtelomeres. With the recent withdrawal of 10X genomics Linked-Read sequencing technologies from the marketplace, we tested a separate commercially available method for generating Linked-Read datasets (Transposase Enzyme Linked Long-read Sequencing (TELL-Seq; [11])) and show that a TELL-Seq dataset also performs well with NPGREAT, providing an alternative commercial platform for improving assembly of large segmental duplications using this method. These new methodologies may ultimately enable efficient population-based and targeted laboratory studies of the role of subtelomeric sequences in genome function, evolution, and human disease.

Methods

REXTAL [10] identifies reads corresponding to 1-copy DNA adjacent to segmental duplications, then selects Linked-Reads associated with the identified 1-copy reads from barcode information on the identified reads, and assembles all of these reads to extend high-quality assembled sequence from the 1-copy region into the segmental duplication region. A hybrid assembly approach to combine Nanopore reads with REXTAL, called NPGREAT [12] is shown in Fig.1. It consists of five main operations: Orientation, Order and Correction, Connector Segments, Gap Filling and Combination. The final output is a single sequence.

Input:

The initial step, prior to the algorithm, corresponds to the selection of the input data, REXTAL contigs and ultralong nanopore reads. To obtain the input REXTAL contigs, we execute the REXTAL procedure [10] and convert its output, which is initially in the form of scaffolds, to unordered DNA sequence contigs. To detect the telomere-containing nanopore reads, we carry out a telomere-tract motif (TTAGGG)_n screen on the nanopore read sequence database to select all reads containing this motif. A second screen of the same nanopore read database with 1-copy sequences closest to the telomere results in the selection of subtelomeric nanopore reads containing 1-copy/segmental duplication boundaries; typically, a fraction of these reads extends deep into the subtelomeric segmental duplication region. Telomere motif containing nanopore reads also containing 1-copy boundary reads span the entire subtelomeric segmental duplication region and identify the telomere of origin. Telomeric and subtelomeric nanopore reads of length greater or equal to 40,000 bases are used as the input nanopore reads of NPGREAT.

Orientation and Position:

After the input reads and contigs have been selected, the subtelomeric nanopore reads and the REXTAL contigs are oriented. The orientation is anchored by the telomeric nanopore reads, whose orientation is known a priori, given that they always end in the 5'- (TTAGGG)_n -3' telomere tract. Overlapping nanopore reads are aligned and oriented relative to the telomeric nanopore reads and to each other. REXTAL sequence contigs are then aligned, positioned, and oriented relative to the repeat-masked nanopore sequence reads.

Correction:

REXTAL contig alignment with the cognate nanopore read sequence is monitored and alignment discrepancies above a given threshold (typically set at 100 bp) are corrected as described in detail below.

Connector Segments:

Alignment of REXTAL contigs with nanopore reads, yield two possibilities for neighboring REXTAL contigs: their overlap or a gap between them. In the connector segments step, the overlapping REXTAL contigs are merged and nanopore read segments that can bridge gaps between neighboring contigs, are identified and extracted.

Gap Filling:

For each gap between REXTAL contigs, several nanopore segments may be available to bridge it. In order to fill the gap, we select the segment that has the highest average percent identity with the flanking contigs.

Combination:

In the final step, we combine according to their order the REXTAL contigs, the merged REXTAL contigs and the nanopore selected segments that connect as well as extend them. The result is the assembled sequence.

Correction Steps of NPGREAT:

The correction steps of NPGREAT are able to detect misassemblies within REXTAL contigs and correct them using sequence from nanopore reads. In the correction algorithm, NPGREAT scans the local alignments of a REXTAL contig with the corresponding nanopore read. As seen in Fig. 2(A), in the case of a deletion in the REXTAL contig there is a segment in the nanopore read (light blue segment) whose length is much greater than in the REXTAL contig (orange segment); this is identified as a potential need for splitting the REXTAL contig.

To localize the deletion within the REXTAL contig, we align the borders of the non-aligned regions of the REXTAL contig (1 kb on each side) in unmasked mode to identify the exact coordinates of the discrepancy. Then, we split the contig at those coordinates and use the nanopore sequence to correct it.

The threshold of identification can be modulated in the program; a difference in the non-aligned segments above the threshold causes a split in the REXTAL contig and insertion of nanopore-derived sequence to fill the gap. We currently use a 100bp length threshold, allowing insertion/deletions of short length in REXTAL contigs to be identified. In the case of misassemblies caused by Tandem Repeats (Fig. 2(B)) unmasked alignment of the two borders (+1kb to the sides), revealed a specific motif, with the beginning and end portions of the

nanopore sequence matching well with the beginning and end portions of the REXTAL sequence. However, the middle portion of the REXTAL sequence matched equally well with multiple contiguous parts of the cognate nanopore sequence. These observations indicated the presence of a tandem-repeat region not represented correctly by the REXTAL sequence assembly. The tandem repeat (green color) is identified but its length has been compressed in the REXTAL sequence, whereas the nanopore sequence correctly identifies the length of the tandem-repeat region and positions the repeated pattern in the correct location relative to flanking sequence. Thus, in the tandem-repeat regions, we depend on the nanopore sequence to correctly define the total TR length.

In order to correct the REXTAL contigs that contain misassembled tandem-repeat patterns, we remove the entire repeat pattern region from the contig (middle green region) and simultaneously, split it at the defined borders. The two resulting contigs are aligned with the nanopore read and placed on the sides. In between, the tandem-repeat region will be filled by the nanopore sequence. The correction steps enable an assembly where the representation of misassembled repeats collapsed due to short-read assembly errors is accurate.

The most common misassembly is an apparent deletion in the REXTAL assembly caused by collapse of reads in repeats that cannot be bridged by short read assemblies; within this class, the vast majority are associated with Tandem Repeat (TR) arrays in the target sequence, but we have also noted instances of misassemblies within interspersed repeats (Tables S1 and S2, Additional file 1). Three more complex misassemblies in REXTAL were also detected and corrected using NPGREAT; two were associated with conserved gene families within single subtelomeres. In the alpha globin gene cluster at 16p and the zinc finger gene cluster at 19q, misassembled REXTAL contigs were split and corrected using the nanopore sequence (Tables S1 and S2, Additional file 1); additional REXTAL contigs aligning to the nanopore segment could then be accurately positioned between them. A third nanopore-corrected misassembly occurs in the 22q subtelomere in a region annotated as a possible alternative haplotype in HG38; it is not clear at present whether this may have caused the REXTAL misassembly relative to the nanopore coverage.

NPGREAT with TELL-Seq Linked-Reads:

The REXTAL pipeline [10] was designed to work with 10x Linked-Read datasets. To incorporate the TELL-Seq Linked-Reads [11], we modified the REXTAL pipeline, now stated as tREXTAL seen in Fig S1, Additional file 1. The raw TELL-Seq Linked-Read data are acquired, then barcode correction and filtering is performed with the use of the TELL-Read analysis software [11] to obtain the final TELL-Seq 18-base barcoded-reads. In order to convert the barcodes to 16-base 10x compatible ones, we use the ust10x software [13] to uniquely map them to sets of Linked-Reads.

Each of the barcodes is associated with a number of reads; too few reads in a barcode are not constructive in the assembly process and lead to erroneous assumptions whereas, a very high number of reads associated with a barcode is a potential indication that the reads come from different molecules instead of a single molecule. To account for this empirically, we tested Linked-Reads with a range of barcode frequencies for a given target region and examined the output assemblies. The barcode frequency range with a minimum 4 and maximum 2000 reads provided optimal assemblies when used for the rest of the pipeline.

In the next steps, we proceed with the initial REXTAL pipeline, by executing BLAT having as input the selected 16-base barcoded TELL-Seq Linked-Reads and a single-copy hg38 bait of the subtelomeric region in question. From the BLAT-mapped reads, we extract their barcodes and keep only those that have at least 3 mapping reads. From the initially selected 16-base barcoded TELL-Seq Linked-Reads dataset, we pull all the reads that have the BLAT-kept barcodes. We set the pulled Linked-Reads from these bait-identified barcodes as input to the Supernova software [14], using the TELL-Seq converted barcode whitelist file instead of the 10x-supernova software whitelist file. No further modifications of the 10X NPGREAT method were needed in order to incorporate the TELL-Seq technology.

Results

Subtelomeres assembled:

We assembled Linked-Reads and ultralong nanopore reads of the NA12878 human cell line for the subtelomeric regions of 5p, 9p, 10p, 15q, 16p, 18p, 19q, 20p and 22q, (Table 1), utilizing an ultralong Nanopore reads dataset [15], a 10x genomics dataset [14] and a TELL-Seq dataset [11]. This set of subtelomeres sample the known structural variety of human subtelomeres identified by global mapping studies of samples from many human populations [16]. These particular subtelomeres also represent the main types of subtelomeric segmental duplication organizations seen in human populations [16] and are covered by nanopore reads that completely span their subtelomeric segmental duplications in this genome (Table 2), extending from (TTAGGG) n repeats into 1-copy DNA for these specific telomeres.

In subtelomeres 9p, 10p, 16p, 17p, 18p, 19q and 22q there is one segmental duplication region (SD) next to the telomere terminal repeat tract, and the 1-copy region begins on the centromeric side of the subtelomeric segmental duplication. For subtelomeres 5p and 15q there are two SD regions, one next to the telomere and the other flanked by 1-copy DNA. The 20p subtelomere is defined incorrectly by the hg38 human reference by the lack of its large telomere-adjacent SD region in hg38 [16].

Table 1
Coordinates of Subtelomere Regions in hg38.

Subtelomere region	HG38 Reference	Segmental Duplication	1-copy region	10K 1-copy	50K 1-copy	100K 1-copy
5p	10,001 - 677,959	10,001 - 49,495 and 210,596 - 305,378	305,379 - 677,959	49,510 - 59,509	49,510 - 99,509	49,510 - 149,509
9p	10,001 - 403,764	10,001 - 203,763	203,764 - 403,764	205,587 - 215,586	205,587 - 255,586	205,587 - 305,586
10p	10,001 - 588,571	10,001 - 88,570	88,571 - 588,571	88,506 - 98,505	88,506 - 138,505	88,506 - 188,505
15q	101,532,200 - 101,981,189	101,861,029 - 101,981,189 and 101,732,200 - 101,794,395	101,794,396 - 101,861,028 and 101,532,200 - 101,732,199	101,841,838 - 101,851,837	101,801,838 - 101,851,837	101,794,838 - 101,851,837
16p	10,000 - 240,859	10,000 - 40,859	40,860 - 240,859	43,549 - 58,548	43,549 - 93,548	43,549 - 143,548
18p	10,000 - 331,693	10,000 - 131,693	131,694 - 331,693	131,724 - 141,723	131,724 - 181,723	131,724 - 231,723
19q	58,386,558 - 58,607,616	58,586,558 - 58,607,616	58,386,558 - 58,586,557	58,576,001 - 58,586,000	58,536,001 - 58,586,000	58,486,001 - 58,586,000
20p	66,335 - 266,334	N/A	66,335 - 266,334	67,230 - 77,229	67,230 - 117,229	67,230 - 167,229
22q	50,540,514 - 50,808,468	50,740,514 - 50,808,468	50,540,514 - 50,740,513	50,729,932 - 50,739,931	50,689,932 - 50,739,931	50,639,932 - 50,739,931

Table 2
Number of Nanopore Reads in Subtelomere Regions.

Region	Telom NP	Subtel NP	10x Assembly size (bp)	Tell-Seq Assembly size (bp)
5p	2	5	287480	275519
9p	2	8	453049	360172
10p	3	14	449599	410997
15q	2	12	663605	659543
16p	2	5	243871	189308
18p	2	9	462216	394240
19q	5	5	403012	355845
20p	2	9	432027	426683
22q	3	6	376852	280076

Quast Analysis Of Assemblies:

To assess the quality of the NPGREAT assembly, we use the QUAST software [17] and the Icarus genome viewer [18]. QUAST is a tool for the pairwise evaluation and comparison of genome assemblies. We used version 5.0, which uses minimap2 as an aligner to align the assemblies to a reference genome, as specified by the user. We compare each NPGREAT assembly and its nanopore-corrected REXTAL assembly with the corresponding region of the hg38 human reference (Figs. 3-6; Figs S2-S10, Additional file 1). The exact coordinates of the hg38 reference used for each subtelomere can be found in Table 1. In the figures, the first line corresponds to the NPGREAT assembly, the second line to the REXTAL assembly and the third line corresponds to the hg38 reference, designated with gaps at the locations where tandem repeats occur. The telomere is at the left end of the figures representing the p-arms of chromosomes, and the telomere end is on the right end of those figures representing the q-arm.

Possible misassemblies are designated in the Icarus visualization tool with the color gray, red or orange. When the gap or the overlap between the left and right flanking sequences is less than 1 kb but larger than 85 bp, it is designated as a local misassembly and colored gray. If the gap or overlap exceeds 1 kb, then it is considered an extensive misassembly, colored red or orange. For example, if there are local misassemblies before and after an alignment, the contig is colored gray in Icarus. The color however, does not indicate the percent identity of the alignment to the reference, which can be seen in the side table of the software, indicating the analysis of each alignment to the reference.

Telomere-induced “misassembly” Artifact:

In the hg38 human reference, a large portion of the telomere repeat tract sequence 5'- (TTAGGG)_n -3' is missing, due to its clone-based creation. In contrast, thousands of bases of telomere repeat tract sequence are present

in NPGREAT assemblies because the single-molecule nanopore reads include this sequence. As a consequence, when QUASt compares the NPGREAT assembly with the hg38 reference it always generates an artifactual misassembly at the telomere (Figs. 3-6; Figs S2-S10, Additional file 1).

Comparison With The Reference:

In all subtelomere regions, NPGREAT has accomplished essentially complete coverage while maintaining a relatively high percent identity with the reference (Figs. 3-6, Table 3; Figs S2-S10, Additional file 1). The total percent identity of each assembly with the hg38 reference was determined using the percent identities of individual QUASt output alignments. We calculated the weighted percent identity of each assembly as seen in Algorithm 1 (Fig. 7). QUASt generates multiple alignments for every assembly. Due to the inaccurate short alignments in the area of the telomere tract sequence (where the reference is missing), we kept only the alignments after the first occurrence of an alignment whose length is at least 1 kb length. We calculated the weighted average percent identity by using the individual alignment lengths as weights, i.e. multiplying each alignment's percent identity with its weight, then adding all products and finally, dividing them by the sum of the weights (lengths).

As can be seen in Table 3, the NPGREAT assembly nucleotide sequence percent similarity with hg38 varied according to subtelomere from 95 to nearly 99%. The input REXTAL contig % similarities were consistently higher than those of NPGREAT, indicating that the expected lower accuracy nanopore sequence read input (90 to 92% accuracy for the Guppy 2.3.7 base caller we used) of nanopore sequence contributing to NPGREAT (Fig. 1) decreased the overall % identity of the NPGREAT assembly, but NPGREAT provided comprehensive coverage while greatly improving accuracy relative to the nanopore-only coverage. As expected, NPGREAT assembly regions with nearly complete REXTAL coverage had the highest sequence similarity to hg38.

Table 3
Weighted Percent Identity with the HG38 Reference.

Region	NPGREAT 10x	NPGREAT Tell-Seq
5p	96.55	97.34
9p	96.57	96.06
10p	98.35	98.45
15q	97.09	95.51
16p	97.56	97.65
18p	98.59	94.82
19q	96.90	95.95
20p	98.74	97.47
22q	96.94	95.45

Variability In Tandem Repeat Regions:

Tandem repeat regions may vary highly between different alleles, a feature exacerbated in many human subtelomere regions where there are abundant hypervariable tandem repeats known as minisatellites or Variable Number Tandem Repeats [19]. In QUASt comparisons of new assemblies with the human reference sequence (which is a clone-based amalgam of multiple haplotypes) this leads to artifactually called extensive misassemblies where the naturally polymorphic TR tract lengths differ by more than the extensive misassembly threshold (default of 1 kb) between naturally occurring haplotypes being compared. For this reason, we annotate all TR positions in the HG38 reference used for QUASt comparison to pinpoint potentially artifactual misassemblies (Figs. 3 - 6). With the exception of the telomere tract and TR induced artifacts there were no extensive misassemblies called by QUASt in the NPGREAT -hg38 reference sequence comparisons, indicating its high long-range contiguity.

Variability Within The Genome:

As it can be seen in the figures, there are also a number of short local misassemblies of length less than 1Kb that do not occur in Tandem Repeat Regions. These might be attributable to the naturally occurring small polymorphisms between the assembled genome and the region of the reference HG38 it is being compared with; however, it is impossible to distinguish this from true small misassemblies in the new assembly using QUASt comparisons with a reference genome non-identical to the source genome used in the new assembly.

Npgreat With Different Linked-reads Datasets:

The REXTAL pipeline [10] was designed to work with 10x Linked-Read datasets. To incorporate the TELL-Seq Linked-Reads, we modified the REXTAL pipeline, designated as tREXTAL as described in Methods. The output scaffolds of the tREXTAL computational method were processed to create the input contigs for the NPGREAT method, in a manner identical to that described for 10x REXTAL.

We tested the tREXTAL and NPGREAT method for the NA12878 cell-line. A comparison of the NPGREAT with Tell-Seq Linked-Reads against the NPGREAT with 10x Linked-Reads can be seen in Figures 3-6. The NPGREAT with the use of the TELL-Seq dataset successfully assembles the subtelomeric regions into one contig, and addresses the Tandem-Repeat compression which exists in the Linked-Read technology and extends towards the telomere. Therefore, it has the same qualities that were seen previously in the NPGREAT assembly with the 10x data. Both technologies provide similar results, however, the TELL-Seq dataset has lower coverage compared to the 10x Linked-Read dataset, resulting in the higher number of gaps that need to be filled with nanopore sequence. Nonetheless, the NPGREAT method provides similar high-quality output even in this case of lower TELL-seq coverage.

Discussion

We show here that misassemblies in the short-read assemblies of REXTAL are resolved in NPGREAT using Nanopore sequence coverage to guide the correction of REXTAL contigs. Notable errors in REXTAL assemblies

were the length of tandem repeat regions, and more complex misassemblies in the vicinity of high-sequence similarity gene families within single subtelomeres. Two examples of the latter are the regions of 16p and 19q, where the contigs in the QUASt analysis of REXTAL assemblies show that the contigs had to be split, allowing other contigs to be positioned in the resulting gap filled with nanopore sequence (Fig. 5; Fig S6, and Fig S8, Additional file 1).

All NPGREAT assemblies extend to the telomere side beyond the reference and contain thousands of bases of the telomere repeat tract sequence, because the single molecule Nanopore reads include these (TTAGGG)_n sequence tracts that are truncated in the HG38 reference sequence due to its generation from clones unable to propagate the tract. With the use of appropriate bait sequences for REXTAL, the NPGREAT assemblies will extend beyond the QUASt analyzed regions shown here on the centromeric side of the targeted regions, providing a continuous sequence assembly even with low nanopore read coverage. The TELL-Seq Linked-Read dataset provides lower coverage short-read assemblies compared to the 10x genomics Linked-Read dataset, resulting in shorter REXTAL assemblies with a larger number of gaps. Nevertheless, the ultra-long Nanopore reads spanning the entire region enable the creation of a complete NPGREAT assembly and the combination of TELL-seq with Nanopore significantly enhances the accuracy of the sequence.

Recent advances in nanopore sequencing technology provide additional opportunities for improving subtelomere sequence assemblies. An improved nanopore ultralong read dataset with much deeper coverage than that used in this study was used along with high-coverage 10-20 kb reads (from PacBio circular sequencing libraries) to generate a finished-quality telomere to telomere sequence of human chromosome 8 [20]. As with the NPGREAT strategy, a scaffold of ultralong nanopore reads was combined with a second dataset of shorter reads, but in this case shorter reads were 10-20 kb “long reads” and utilized haploid genome-specific single-nucleotide changes that discriminated between highly similar segmental duplications within the haploid genome. A global assembly of CHM13 using the same 10-20 kb long read high-coverage dataset was created for the entire haploid CHM13 genome [21]; in this case a string graph assembly procedure was used to generate the initial assembly and “tangles” caused by repeats and unresolved segmental duplications were resolved by nanopore read coverage. However, this global assembly was unable to resolve all segmental duplication regions even within the CHM13 haploid genome. It is important to note that the assembly methods used in both of these truly impressive landmark papers for CHM13 will not work in diploid genomes ([20]; [21]), and the 10-20 kb long read datasets used in these aforementioned assemblies of CHM13 are much more expensive to generate than Linked-Read datasets. On the other hand, haplotype resolution for subtelomere-sized segmental duplications remains theoretically achievable by combining targeted deep coverage ultralong telomeric nanopore reads with relatively inexpensive Linked-Read libraries of cognate genomes, and we are pursuing this approach with NPGREAT in order to permit analysis of the highly variable subtelomeric regions of large numbers of diploid genomes in human populations.

Conclusions

We present here a hybrid assembly method that combines ultralong nanopore reads with regionally selected Linked-Read assemblies. This NPGREAT method results in extension of high-quality assemblies into otherwise inaccessible segmental duplication regions near telomeres, enhancing our ability to accurately assemble

human subtelomere DNA. This information will enable improved analyses of the structure, function, and evolution of these key genomic regions.

Abbreviations

NPGREAT

NanoPore Guided REgional Assembly Tool

TELL-seq

Transposase Enzyme Linked Long-read sequencing

REXTAL

Regional Extension of Assemblies using Linked-Reads

tREXTAL

Modified REXTAL to incorporate the TELL-seq Linked-Reads

SD regions

Segmental Duplication regions

TR

Tandem Repeat

BLAST

Basic Local Alignment Search Tool

BLAT

BLAST-like alignment tool

QUAST

Quality Assessment Tool for Genome Assemblies

Declarations

Availability of data and materials

The NPGREAT software and the datasets generated/analyzed during the current study are available in the repository: <https://github.com/eleniadam/npgreat> .

The input datasets were derived from the following resources: The ultralong Nanopore reads are available at <https://github.com/nanopore-wgs-consortium/NA12878> [15]. The TELL-Seq Linked-Reads are provided at the NCBI BioProject PRJNA591637 [11]. The 10x Linked-Reads are available at https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2 [14].

Ethics approval and consent to participate

Not applicable

Consent for publication

All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Graduate Summer Award Program from the Graduate School and Office of Research at Old Dominion University, the Dr. Hussein Abdel-Wahab Memorial Scholarship from the Computer Science Department at Old Dominion University and the Old Dominion University Computer Science Department Bioinformatics endowment.

Authors' Contributions

EA developed the software; DR and HR supervised the project; and EA and HR wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Peter Chang and Dr. Z Tom Chen from the Universal Sequencing Technology Corporation for helping in understanding the TELL-seq data and software, Min Dong from the Old Dominion University High Performance Computing Systems group for the support of the cluster infrastructure where the software runs and the Oxford Nanopore Technologies Customer Services for helping in understanding the ONT software.

References

1. Armanios M, Alder JK, Parry EM, Karim B, Strong MA, Greider CW. Short telomeres are sufficient to cause the degenerative defects associated with aging. *The American Journal of Human Genetics*. 2009;85(6):823–32 DOI: 10.1016/j.ajhg.2009.10.028.
2. Armanios M, Blackburn EH. The telomere syndromes. *Nature Reviews Genetics*. 2012;13(10):693–704 DOI: 10.1038/nrg3246.
3. Sahin E, DePinho RA. Linking functional decline of telomeres, mitochondria and stem cells during ageing. *nature*. 2010;464(7288):520–8 DOI: 10.1038/nature08982.
4. Maciejowski J, de Lange T. Telomeres in cancer: tumour suppression and genome instability. *Nature reviews Molecular cell biology*. 2017;18(3):175–86.
5. Sfeir A, De Lange T. Removal of shelterin reveals the telomere end-protection problem. *Science*. 2012;336(6081):593–7 DOI: 10.1126/science.1218498.
6. Baird DM, Rowson J, Wynford-Thomas D, Kipling D. Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nature genetics*. 2003;33(2):203–7 DOI: 10.1038/ng1084.
7. Britt-Compton B, Rowson J, Locke M, Mackenzie I, Kipling D, Baird DM. Structural stability and chromosome-specific telomere length is governed by cis-acting determinants in humans. *Human molecular genetics*. 2006;15(5):725–33.
8. McCaffrey J, Young E, Lassahn K, Sibert J, Pastor S, Riethman H, et al. High-throughput single-molecule telomere characterization. *Genome research*. 2017;27(11):1904–15.
9. Abid HZ, McCaffrey J, Raseley K, Young E, Lassahn K, Varapula D, et al. Single-molecule analysis of subtelomeres and telomeres in Alternative Lengthening of Telomeres (ALT) cells. *BMC genomics*.

2020;21(1):1–17.

10. Islam T, Ranjan D, Zubair M, Young E, Xiao M, Riethman H. Analysis of subtelomeric REXTAL assemblies using QUASt. *IEEE/ACM transactions on Computational Biology and Bioinformatics*. 2021;18(1):365–72 DOI: 10.1109/TCBB.2019.2913845.
11. Chen Z, Pham L, Wu T-C, Mo G, Xia Y, Chang PL, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome research*. 2020;30(6):898–909 DOI: 10.1101/gr.260380.119.
12. Adam E, Islam T, Ranjan D, Riethman H, editors. Nanopore Guided Assembly of Segmental Duplications near Telomeres. 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE); 2019: IEEE; DOI: 10.1109/BIBE.2019.00020.
13. TELL-Seq Software. <https://www.universalsequencing.com/analysis-tool>.
14. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome research*. 2017;27(5):757–67 DOI: 10.1101/gr.214874.116.
15. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*. 2018;36(4):338–45 DOI: 10.1038/nbt.4060.
16. Young E, Abid HZ, Kwok P-Y, Riethman H, Xiao M. Comprehensive analysis of human subtelomeres by whole genome mapping. *PLoS genetics*. 2020;16(1):e1008347 DOI: 10.1371/journal.pgen.1008347.
17. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–5 DOI: 10.1093/bioinformatics/btt086.
18. Mikheenko A, Valin G, Prijibelski A, Saveliev V, Gurevich A. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics*. 2016;32(21):3321–3.
19. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. *Cell*. 2019;176(3):663–75. e19 DOI: 10.1016/j.cell.2018.12.019.
20. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovych MA, Koren S, et al. The structure, function and evolution of a complete human chromosome 8. *Nature*. 2021;593(7857):101–7 DOI: 10.1038/s41586-021-03420-7.
21. Nurk S, Rogaev EI, Eichler EE, Miga KH, Phillippy AM. The complete sequence of a human genome [preprint]. 2021 DOI: 10.1101/2021.05.26.445798.

Figures

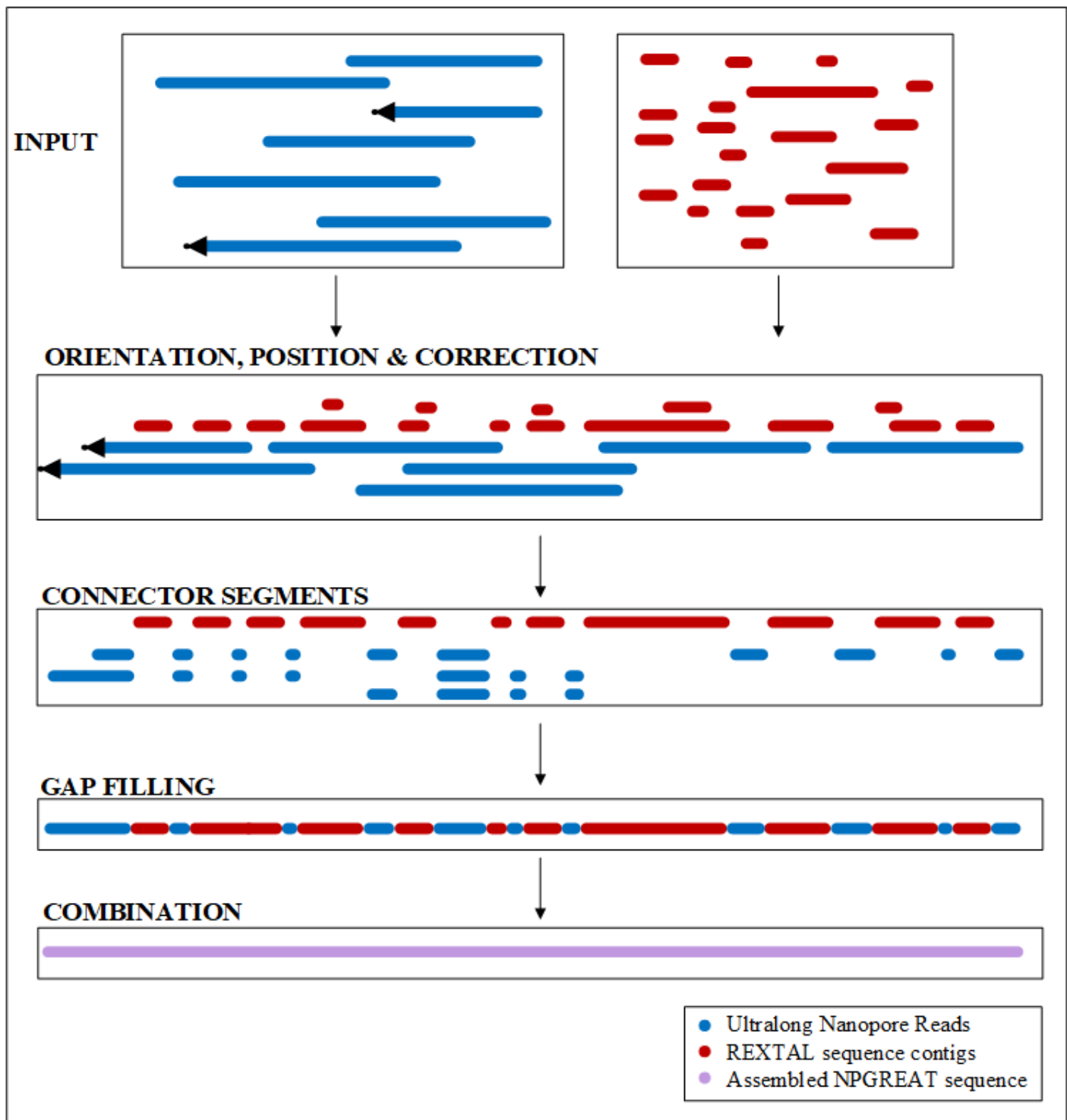


Figure 1

The steps of the NPGREAT method. Telomere-containing ultralong Nanopore reads are shown as blue line segments with black arrows designating terminal ((TTAGGG) n tracts), and are used to anchor the orientation. These along with ultralong Nanopore reads selected from distal 1-copy subtelomere regions are used as scaffolds upon which the Linked-Read assembly (REXTAL) contigs (red line segments) are placed and corrected. The correction of possible misassemblies within the REXTAL contigs is primarily in Tandem Repeat (TR) regions, where the Nanopore reads have a more accurate representation than the short-read Linked-Read assemblies which typically collapse tandem repeats into a short consensus sequence. Properly positioned,

oriented, and corrected REXTAL contigs are merged with nanopore connector segments for the NPGREAT output, a single assembled sequence.

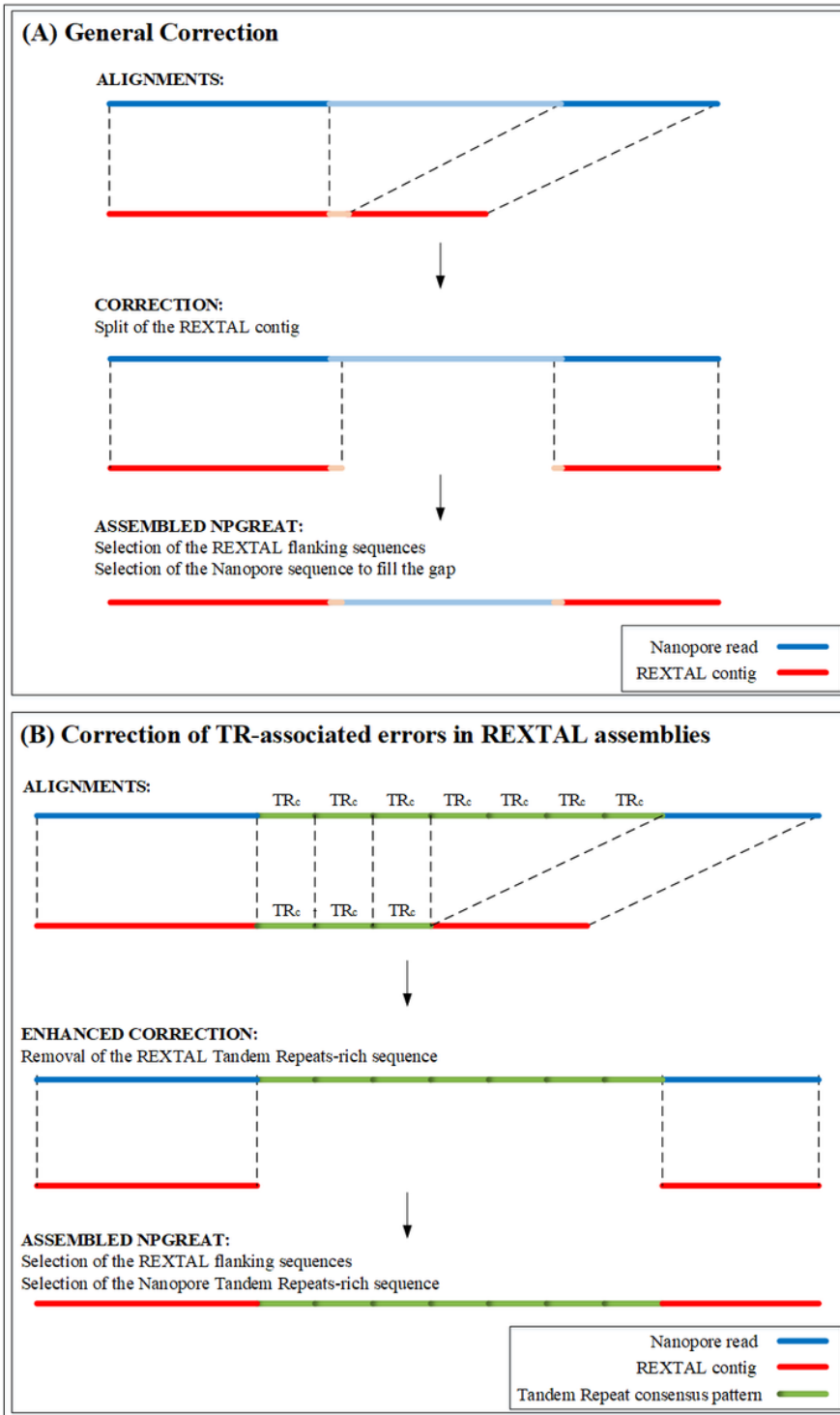
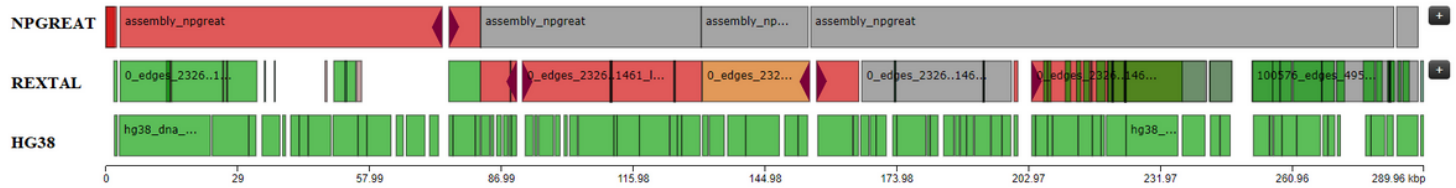


Figure 2

Correction of REXTAL assemblies using Nanopore Reads. (A) The alignments between the nanopore and the REXTAL sequence reveal a segment whose length in the nanopore read (light blue segment) is different than in the REXTAL contig (light red segment). This indicates a potential region where a deletion has occurred in the REXTAL contig. We identify the exact location of the deletion by obtaining the alignments of the deletion boundary regions in question in unmasked mode. Then, we split the REXTAL contig at those coordinates, filling

the missing part with the corresponding nanopore sequence. (B) In most cases where length differences between nanopore and REXTAL assemblies are seen, the initial alignments between the nanopore read and the REXTAL contig reveal a tandem repeat region better represented in the nanopore read. In these cases, the tandem-repeat pattern sequence is removed from the REXTAL contig and the contig is split at those junctions. The properly aligned segments from the REXTAL contig flanking the TR region are joined with the nanopore sequence containing the tandem repeat to form the NPGREAT assembly.

(A) 10x dataset



(B) TellSeq dataset

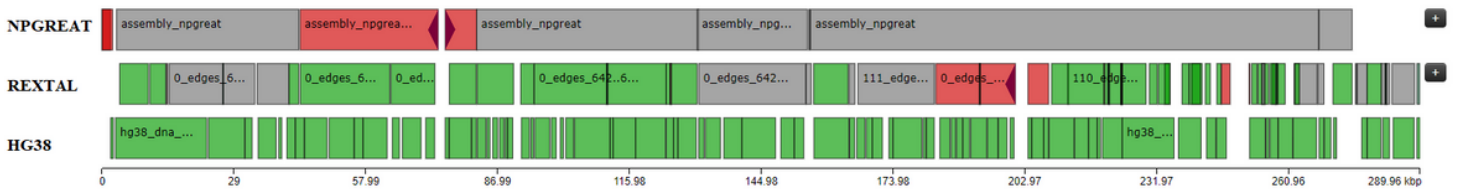
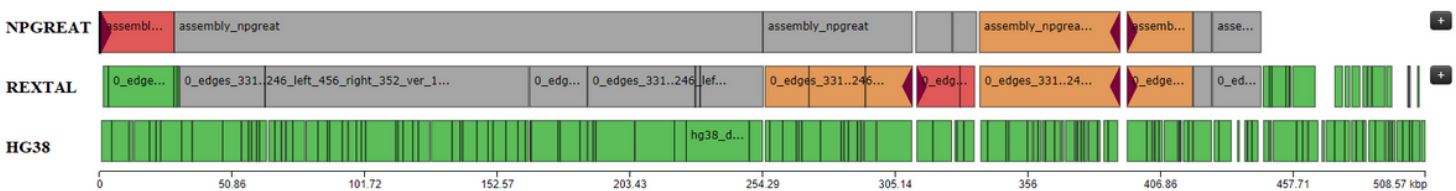


Figure 3

Comparison of the NPGREAT and REXTAL subtelomere assemblies with the HG38 reference using QUAST. In each view there are three parts: (1) the NPGREAT assembly (top part), (2) The REXTAL assembly (middle part), and (3) the hg38 reference genome region to which each of these assemblies are independently compared using QUAST (bottom). The TRs in the hg38 reference are masked to identify their locations, and appear as gaps in the reference sequence. The colors designate QUAST identified misassemblies relative to the reference sequence (misassemblies of length longer than 1 kb are designated with red or orange color, while the gray color signifies gaps or overlaps of length less than 1 kb). (A) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the 10x Linked-Reads datasets as input. (B) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the TELL-Seq Linked-Reads datasets as input.

(A) 10x dataset



(B) TellSeq dataset

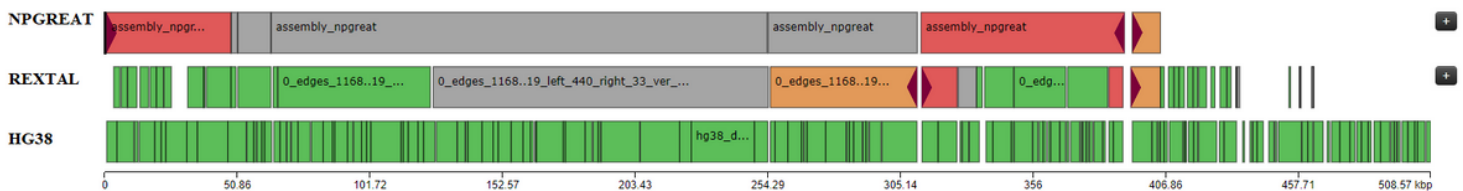


Figure 4

Comparison of the NPGREAT and REXTAL subtelomere assemblies with the HG38 reference using QUASt. In each view there are three parts: (1) the NPGREAT assembly (top part), (2) The REXTAL assembly (middle part), and (3) the hg38 reference genome region to which each of these assemblies are independently compared using QUASt (bottom). The TRs in the hg38 reference are masked to identify their locations, and appear as gaps in the reference sequence. The colors designate QUASt identified misassemblies relative to the reference sequence (misassemblies of length longer than 1 kb are designated with red or orange color, while the gray color signifies gaps or overlaps of length less than 1 kb). (A) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the 10x Linked-Reads datasets as input. (B) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the TELL-Seq Linked-Reads datasets as input.

(A) 10x dataset



(B) TellSeq dataset

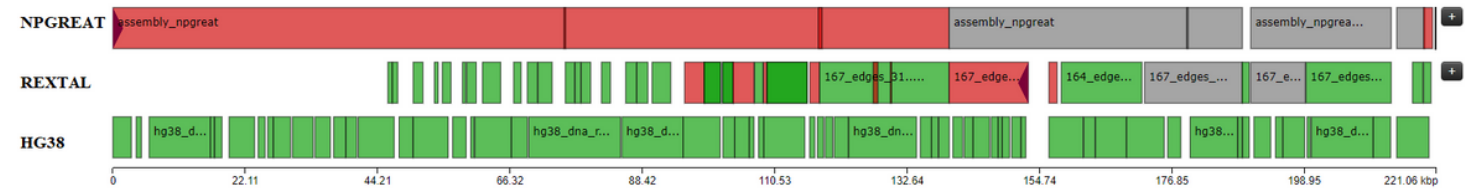
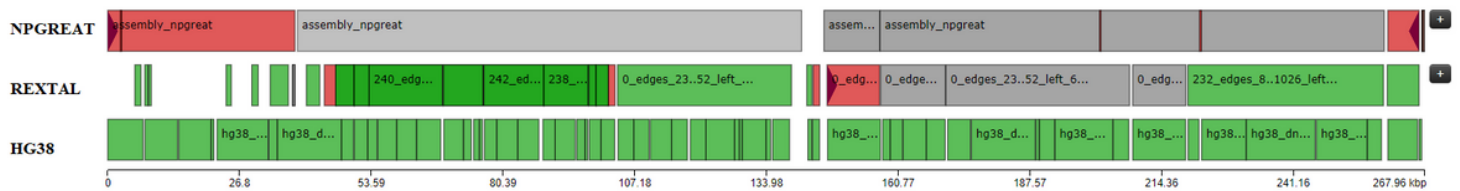


Figure 5

Comparison of the NPGREAT and REXTAL subtelomere assemblies with the HG38 reference using QUASt. In each view there are three parts: (1) the NPGREAT assembly (top part), (2) The REXTAL assembly (middle part), and (3) the hg38 reference genome region to which each of these assemblies are independently compared using QUASt (bottom). The TRs in the hg38 reference are masked to identify their locations, and appear as gaps in the reference sequence. The colors designate QUASt identified misassemblies relative to the reference sequence (misassemblies of length longer than 1 kb are designated with red or orange color, while the gray color signifies gaps or overlaps of length less than 1 kb). (A) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the 10x Linked-Reads datasets as input. (B) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the TELL-Seq Linked-Reads datasets as input.

(A) 10x dataset



(B) TellSeq dataset



Figure 6

Comparison of the NPGREAT and REXTAL subtelomere assemblies with the HG38 reference using QUAST. In each view there are three parts: (1) the NPGREAT assembly (top part), (2) The REXTAL assembly (middle part), and (3) the hg38 reference genome region to which each of these assemblies are independently compared using QUAST (bottom). The TRs in the hg38 reference are masked to identify their locations, and appear as gaps in the reference sequence. The colors designate QUAST identified misassemblies relative to the reference sequence (misassemblies of length longer than 1 kb are designated with red or orange color, while the gray color signifies gaps or overlaps of length less than 1 kb). (A) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the 10x Linked-Reads datasets as input. (B) The NPGREAT and REXTAL assemblies with the use of the Nanopore and the TELL-Seq Linked-Reads datasets as input.

Algorithm 1 WEIGHTED_AVERAGE_IDENTITY (Alignments)

```
1: // Calculation of the weighted average percent identity
2: // of the Assembly's Quast alignments
3:
4: flag ← 0
5:
6: // Search all (sorted) alignments
7: For each align in Alignments do
8:
9:     // Check the threshold
10:    if align.Length < 1000 and flag = 0
11:        Go to next alignment
12:    else
13:        flag ← 1
14:        // Use the length as weight
15:        wIdentity ← align.Identity*align.Length
16:        wsumIdentities ← wsumIdentities + wIdentity
17:
18:        // Sum of all lengths
19:        sumLengths ← sumLengths + align.Length
20:
21: // Calculate the average
22: wAveIdentity ← wsumIdentities/sumLengths
23:
24: return wAveIdentity
```

Figure 7

Algorithm 1 – Weighted Average Percent Identity.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)