Comprehensive resolution of challenging genomic variation with Oxford Nanopore telomere-to-telomere assemblies

Sean McKenzie[1], Phill James[2], Heather Jeffery[2], George Dodd[2], Sissel Juul[1], Philipp Rescheneder[2]

[1]Oxford Nanopore Technologies, Inc. New York, NY
[2]Oxford Nanopore Technologies, plc. Oxford, UK

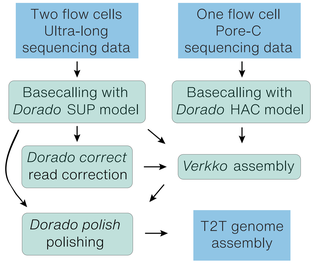
Oxford Nanopore Technologies
Applications

Abstract

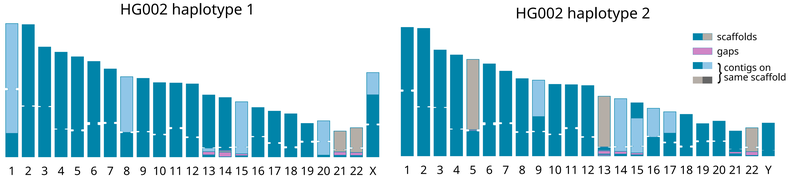
Telomere-to-telomere (T2T) genome assemblies can be generated robustly and efficiently for arbitrary samples using long and ultra-long reads, providing the most complete picture possible of an organism's genetic sequence[1,2]. However, the full utility of these assemblies remains to be demonstrated, especially in comparison to traditional read-alignment approaches for characterizing genomic variation.

Here we apply Oxford Nanopore ultra-long read based T2T assembly to six cell lines in order to demonstrate the power of T2T assembly to resolve potential causes of disease in otherwise un-interrogatable repetitive and structurally divergent regions of the genome. Using the GIAB benchmark sample HG002, we demonstrate that T2T assembly enables comprehensive detection of small nucleotide and structural variants across the genome, including in challenging medically relevant genes (CMRGs). In three cell lines with unresolved balanced chromosomal rearrangements, T2T assembly precisely identified all breakpoints. The approach also fully resolved the duplicated SMN1/SMN2 locus associated with spinal muscular atrophy in two samples with silent carrier indicative marker SNPs, enabling accurate determination of silent carrier status. Together, these analyses highlight the capability of T2T assembly to reveal clinically significant variation inaccessible to traditional methods.

Assembly methods and statistics

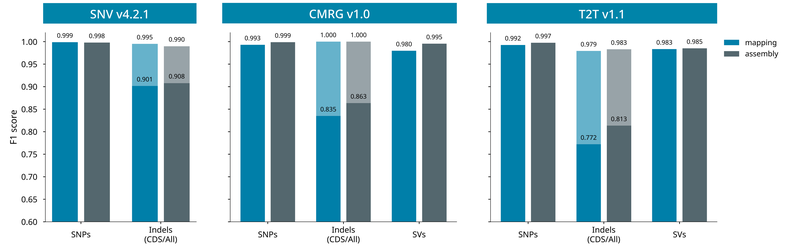


Sample	Assembly size	Scaffold N50	Contig N50	Gapless chromosomes
HG002	6.04 G	146.8 M	135.9 M	30
GM24510	6.09 G	154.4 M	154.4 M	36
GM03786	6.08 G	154.5 M	137.3 M	21
GM21883	6.00 G	145.2 M	135.3 M	24
GM20291	6.04 G	146.2 M	143,1 M	32
GM20359	6.10 G	154.0 M	134.4 M	26



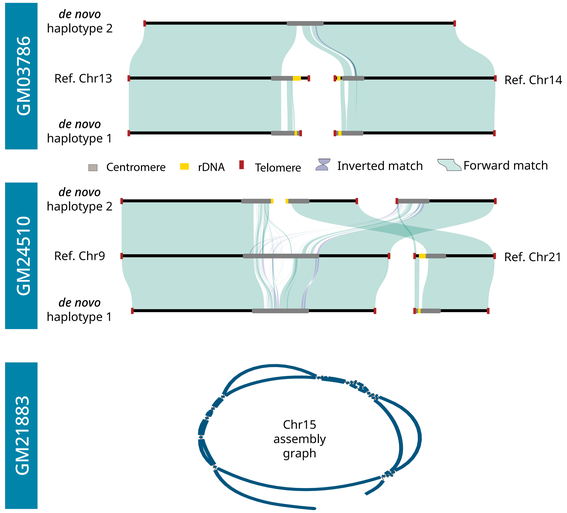
For each sample, cell lines were ordered from Coriell and cultured according to suppliers protocol. Ultra-long and Pore-C sequencing data was generated for all samples following Oxford Nanopore's standard DNA extraction, library preparation, and sequencing protocols[3]. Ultra-long libraries were sequenced over two PromethION Flow Cells per sample, while a single flow cell was used for each Pore-C library. Sequencing data was basecalled with the Dorado[4] basecaller, using the SUP v5.0.0 model for ultra-long data and the HAC v5.0.0 model for the Pore-C data. All data where filtered for a mean Q score of 10 and a minimum read length of 1kb. Ultra-long data was then error-corrected using the Dorado correct tool, and fed into the Verkko assembler[1] as "--hifi" data, along with the uncorrected ultralong data as "--nano" and the Pore-C data as "--porec". Assemblies were subsequently polished with the original ultra-long basecalls and the Dorado polish tool, using the move-table-aware models. Basic assembly quality metrics where calculated using Gfastats[5], and gapless chromosomes were identified by using the SeqTK[6] "cutN" and "telo" functions to find chromosome-sized contigs with telomeres on either side. For plotting, centromeres were annotated by alignment to chm13 centromere sequences, and rDNA arrays were annotated by Minimap2[7] alignment to a representative 45S rDNA sequence. Protein coding genes were identified using Miniprot[8] with RefSeq peptide queries.

HG002 variant calling performance



To investigate general variant calling performance, we sequenced the well characterized HG002 GIAB cell line, which has several well establish small nucleotide and structural variant benchmarks. Assembly-based variant calls were generated with Dipcall[9] and compared to three GIAB variant truthsets[10–12] using the hap.py[13] and truvari[14] tools, along with mapping-based SNPs and SVs from Clair3[15] and LongcallID[16], respectively. Our results show that T2T assembly enables accurate genome-wide calling of small nuclear variants (SNVs) and structural variants (SVs), including in challenging medically relevant genes (CMRGs). On the v4.2.1 small variant benchmark, we see slightly worse SNV accuracies from our assembly methods, however for SVs and SNVs in all other benchmarks, assembly based variant calling outperformed the mapping approach.

Characterization of chromosomal abnormalities



Three samples with complex chromosomal rearrangements were assembled with our T2T protocol: a Robertsonian translocation (GM03786) and a (balanced) reciprocal translocation (GM24510) which had been previously found to be recalcitrant to read-alignment discovery, as well as a ring chromosome (GM21883).

For GM03786, we assembled the whole Robertsonian derivative as a single T2T contig in haplotype 2 (H2), resolving the exact breakpoint. This breakpoint was within a centromeric satellite with low homology to the reference, hence it couldn't be detected with raw read mapping approaches.

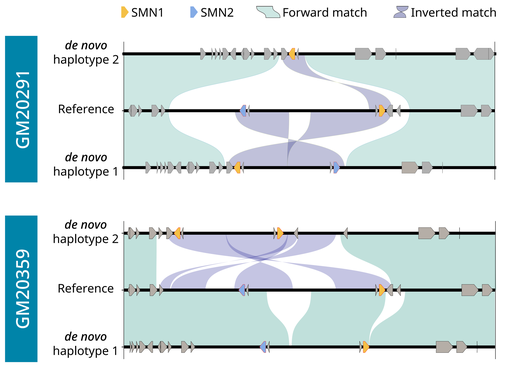
In GM24510 (H2), one derivative assembled T2T and was largely colinear to chr9p, terminating in distal chr21p acrocentric satellites. The second derivative was not scaffolded across the rDNA array by Verkko, but the breakpoint is localized by a chr9 α-satellite → acrocentric p-arm satellite → rDNA transition in a contig largely colinear with chr9q. Association with the H2 chr21q contig is corroborated by Pore-C contacts, further H2 chr21q was the only true acrocentric q arm not scaffolded to a p arm by Verkko.

Verkko lacks a circular contig detection module, however the ring chromosome structure in GM218833 was clear in the assembly graph, and breakpoints can be elucidated by aligning the circularizing edge to the chr15 reference.

Reconstruction of SMA loci in putative silent carriers

Spinal muscular atrophy is a disorder caused by loss-of-function of the SMN1 gene, which sits in a variable segmental duplication with a nearly identical pseudogene (SMN2). Read-alignment approaches can resolve the copy number of both SMN1 and SMN2, however the duplication confounds phasing of the alleles, resulting in missed "silent carriers" where a duplication of SMN1 on one haplotype masks a deletion on the other[17,18]. Several marker SNVs can be used to flag possible silent carriers, but the predictiveness of these markers is limited[17,18,19].

We identified two samples (GM20291 and GM20359) in the Coriell cell repository postive for silent carrier associated SNVs and assembled these with our T2T protocol. The SMA locus was fully resolved in both haplotypes of both samples. GM20291 harbored inversions and large deletions in both haplotypes, with SMN2 being deleted in H2; however, both haplotypes had functional SMN1 copies. GM20359 likewise harbored deletions in both haplotypes and an inversion in H2, however all SMN copies were present, and in fact a gene conversion event appears to have restored function to the SMN2 gene, converting it into a second SMN1 copy. Both samples therefore would have received an incorrect carrier assignment using marker SNVs.



Conclusion

Using Oxford Nanopore ultra-long and Pore-C reads, telomere-to-telomere (T2T) assembly produced haplotype resolved genomes for six cell lines and enabled comprehensive variant discovery in regions that are refractory to read alignment. In HG002 benchmarking, assembly based calling delivered strong genome wide performance, including in challenging medically relevant genes. We observed improved SV accuracy over mapping based approaches, and comparable SNV accuracy with slightly improved metrics for either approach depending on the benchmark. These results indicate that T2T assemblies can serve as a robust method for both small and large variant calling and interpretation.

For chromosomal rearrangements, T2T assemblies localized exact breakpoints that mapping could not resolve: a Robertsonian translocation with a centromeric satellite junction and a reciprocal t(9;21) with junctions from α satellite sequence into acrocentric short arm satellites. T2T assembly also resolved a ring chromosome as a circular assembly graph component. At the SMN1/SMN2 locus, complete, phased reconstructions in two samples positive for silent carrier associated SNVs revealed functional SMN1 copies on all haplotypes, consistent with known limitations for marker based inference. Together, these findings show that ONT T2T assemblies deliver definitive, breakpoint level and haplotype aware interpretation of phenotypically important variation.

References

1. Rautiainen M, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023;41:1474–1482.
2. Nurk S, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
3. nanoporetech.com/document/telomere-to-telomere-sequencing-t2t-on-promethion-sqk-apk114-sqk
4. github.com/nanoporetech/dorado
5. Di Genova A, et al. gfastats: conversion, evaluation and manipulation of genome sequences. Bioinformatics. 2022;38:4214–4216
6. github.com/lh3/seqtk
7. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–3100.
8. Li H. Protein-to-genome alignment with miniprot. Bioinformatics. 2023;39:btad014.
9. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nature Methods. 2018;15:595-597.
10. Wagner J, et al. Benchmarking challenging small variants with linked and long reads. Cell Genomics. 2022; 2(5):100128
11. Wagner J, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol. 2022;40:672-680
12. ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_HG002_DraftBenchmark_defrabbV0.019-20241113
13. github.com/Illumina/hap.py
14. github.com/ACEnglish/truvari
15. Zheng Z, et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nature Computational Science. 2022; 2:797-803
16. github.com/yangao07/longcallD
17. Vitale J, et al. Recommendations for interpreting and reporting silent carrier and disease-modifier variants in SMA. Genes. 2022;13:1657.
18. LabMed Practice Update. Enhanced carrier screening for SMA: detection of silent carriers (rs143838139). Lab Medicine. 2020.
19. Davison D, et al. Duplication Tag SNP g.27134T>G should not be considered diagnostic of SMA carrier status. ACMG poster (2017)