TELL-Seq™ Application Note:

Benchmarking Algorithms for Structural Variation Detection

For Research Use Only.
Not for use in diagnostic procedures.

March 2022
Document # 100033, version 2.0

Introduction

The TELL-Seq WGS library prep kit uses an innovative Transposase Enzyme-Linked Long-read-Sequencing (TELL-Seq™) technology to generate a paired-end (PE) library that captures long-range information in short sequencing reads—known as linked reads (Chen et al., 2020). Linked reads can be sequenced on an Illumina® instrument and enable haplotype phasing, *de novo* genome assembly, and identification of small and large structural variations.

Structural variants (SVs) generally refer to sequence alterations larger than 50 bp that can be categorized into insertions, deletions, duplications, inversions, and translocations. More complex scenarios involving combinations of these categories can also exist. Compared with single-nucleotide variants (SNVs) and small insertions and deletions (i.e., shorter than 50 bp), large SVs are difficult to detect using short-read technology (Mahmoud et al., 2019; Sudmant et al., 2015). Short reads generate a high false-negative rate in SV detection. In contrast, long-read technology has higher accuracy for SV detection, although coverage needs to be high enough to detect with confidence the sequences spanning the structural variation and their breakpoints. This is often cost-ineffective.

No single computational algorithm can accurately and sensitively detect all types and all sizes of SVs (Kosugi et al., 2019). It is reasonable, therefore, to use multiple methods to process data and then analyze the results in a merged view. This application note compares five SV calling methods for the detection of large deletions and complex SVs in genomic DNA from GM12878 cells (NA12878) using TELL-Seq data. NA12878 is a popular reference for the characterization of SV methods.

Methods

Library preparation

Genomic DNA from immortalized human lymphocyte cells GM12878 (Coriell Institute) was used to benchmark SV detection using TELL-Seq technology (NA12878). We extracted NA12878 DNA using an improved salting-out protocol (Chen et al., 2020; Miller et al., 1988). Five nanograms of DNA were processed with the TELL-Seq WGS Library Prep kit in a 0.2 mL PCR tube and approximately 8M TELL beads.

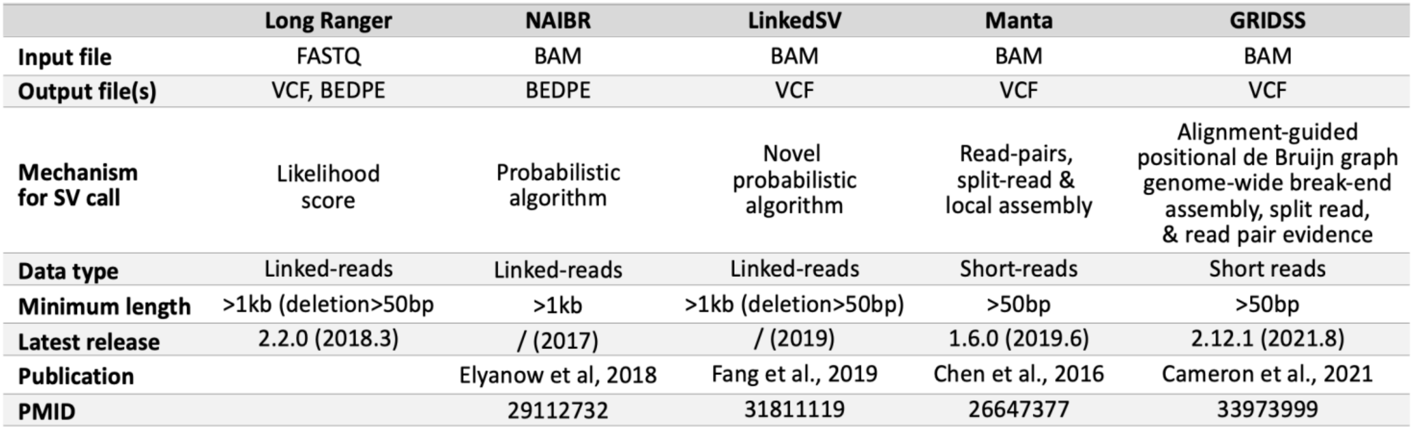
Sequencing

The TELL-Seq library was sequenced on the NovaSeq™ 6000 sequencer (Illumina) using the S1 flow cell, the TELL-Seq Illumina Sequencing Primer kit, and the 2 × 146 bp PE sequencing protocol, generating approximately 1 billion reads.

Data analysis

For sample demultiplexing and QC reporting, sequencing outputs were processed with Tell-Read

Table 1. SV callers



	Long Ranger	NAIBR	LinkedSV	Manta	GRIDSS
Input file	FASTQ	BAM	BAM	BAM	BAM
Output file(s)	VCF, BEDPE	BEDPE	VCF	VCF	VCF
Mechanism for SV call	Likelihood score	Probabilistic algorithm	Novel probabilistic algorithm	Read-pairs, split-read & local assembly	Alignment-guided positional de Bruijn graph genome-wide break-end assembly, split read, & read pair evidence
Data type	Linked-reads	Linked-reads	Linked-reads	Short-reads	Short reads
Minimum length	>1kb (deletion>50bp	>1kb	>1kb (deletion>50bp)	>50bp	>50bp
Latest release	2.2.0 (2018.3)	/ (2017)	/ (2019)	1.6.0 (2019.6)	2.12.1 (2021.8)
Publication		Elyanow et al, 2018	Fang et al., 2019	Chen et al., 2016	Cameron et al., 2021
PMID		29112732	31811119	26647377	33973999

software (Universal Sequencing Technology), identifying 7.7M valid barcodes. In support of efficient linkage, more than 93% and 33% linked reads spanned over genomic regions larger than 20kb and 100kb, respectively. Next, we used Tell-Sort (Universal Sequencing Technology) and the GRCh38 assembly as reference genome for mapping and data processing purposes, generating a phased_sorted.bam file. Bam files can be used as an input for common SV callers. We tested five SV callers: NAIBR, LinkedSV, Manta™, GRIDSS, and Long Ranger™ (**Table 1**). Novel Adjacency Identification with Barcoded Reads (NAIBR; github/raphael-group/NAIBR), LinkedSV (github/WGLab/LinkedSV), and Long Ranger (10x Genomics) were developed to analyze linked-read data. Manta (Illumina; github/Illumina/Manta) and the Genomic Rearrangement Identification Software Suite (GRIDSS; github/PapenfussLab/3rids) were developed to analyze short-read data. We used the Tell-Sort-generated phased_sorted.bam file as input for NAIBR, LinkedSV, Manta, and GRIDSS. For Long Ranger, however, we needed to convert TELL-Seq reads first into a Long Ranger data format. We developed the ust10x tool (Universal Sequencing Technology) to perform this conversion. In the conversion process, 18bp TELL-Seq barcodes are mapped to 16bp sequences that conform to 10X Genomics barcode format. In addition, TELL-Seq barcodes, sequenced as the I1 index, are prepended to R1 reads. Converted R1 and R2 reads are used together as input for Long Ranger. Detailed steps for TELL-Seq-Long Ranger data conversion can be found in the TELL-Seq Data Analysis Roadmap User Guide. We used default parameters when applying the five SV-calling methods. Hardware details and operative system: Linux® Ubuntu 20.04.2 LTS, 60-core Intel® (CPU), 500GB (RAM), and 12TB hard drive.

Data visualization

The Tell-Sort-generated phased_sorted.bam file can be uploaded into the Integrative Genomics Viewer, IGV developed by the Broad Institute. With IGV features that support linked-read data, haplotype-resolved information can be displayed, e.g., homozygous and heterozygous SVs.

Results

Sequencing summary

After TELL-Seq processing of NA12878, more than 97% of the reads successfully mapped to the reference genome (**Table 2**). After removing duplications, the mean depth coverage was 39.8x (**Table 2**).

Table 2. Sequencing results

	NA12878
Total sequencing reads	2,034,862,326
Barcodes	7,702,045
Sequencing condition	2x146bp PE
Cluster read number	1,065,878,321
Mapped reads	97.05%
Duplicate reads	43.29%
Mean depth coverage (total)	76.6x
Mean depth coverage (w/o duplicates)	39.8x
Mean DNA/TELL bead	197,545
DNA in molecules mapping >20kb	93.3%
DNA in molecules mapping >100kb	33.8%
N50 reads per molecule	44

Comparing run time and peak memory consumption among SV callers

Run time for SV calling using the five SV callers (Manta, NAIBR, GRIDSS, LinkedSV, and Long Ranger) is shown in

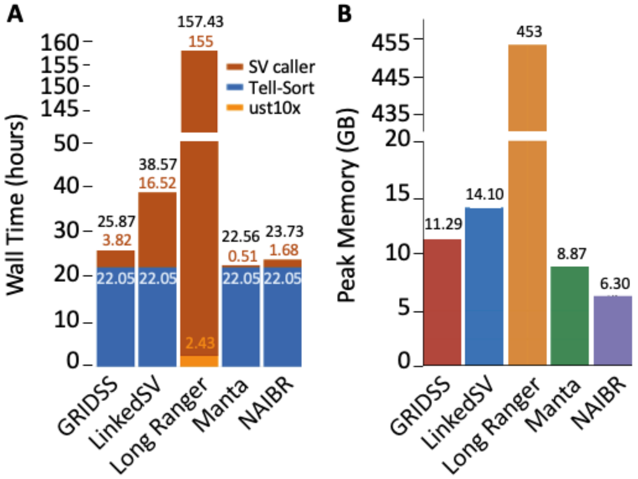


Figure 1. Run time (in A) and peak memory consumption (in B) by SV caller

Table 3. Benchmarking SV callers for the detection of 10 known deletions in NA12878

Benchmark			Deletion detection by SV caller				
Chr.	**hg38 coordinates**	**Size (kb)**	**Long Ranger**	**NAIBR**	**LinkedSV**	**Manta**	**GRIDSS**
Chr8	39,374,555-39,529,710	155.12	155.12	NA	155.20	155.15	155.15
Chr3	162,794,346-162,908,547	114.20	114.20	114.43	114.32	114.20	114.20
Chr1	189,735,379-189,814,229	78.85	79.12	79.08	78.84	78.85	78.85
Chr6	78,257,477-78.326,702	78.85	69.28	69.51	69.27	69.22	69.26
Chr5	105,096,412-105,167,972	71.56	71.56	71.84	71.77	NA	71.56
Chr3	65,203,325-65,228,324	25.00	25.88	26.11	25.87	25.88	25.88
Chr4	115,245,844-115,255,843	10.00	NA	10.55	NA	NA	NA
Chr7	110,541,943-110,547,942	6.00	NA	7.05	6.68	NA	6.47
Chr16	62,511,096-62,516,095	5.00	6.33	6.57	6.33	6.34	6.33
Chr4	186,172,846-186,176,845	4.00	NA	NA	NA	NA	4.46
		Total deletions detected:	7	8	8	6	6 + 3

High confident calls / Low confident calls

Figure 1. Manta required the shortest time to complete the analysis, 22.87 hours. NAIBR, GRIDSS, and LinkedSV required 38.57, 22.56, and 23.73 hours, respectively (**Figure 1A**). Meanwhile, Longer Ranger required the longest time, 157.43 hours. We note that Tell-Sort required 22.05 hours to process fastq files and generate the bam files prior to SV calling—which is applicable to the Manta, NAIBR, GRIDSS, and LinkedSV workflows. In terms of RAM usage, NAIBR, LinkedSV, Manta, and GRIDSS required similar amounts of memory consumption, but much less than Long Ranger (**Figure 1B**). Long Ranger's singularity in memory requirement was due, again, to the need to map fastq files as part of the Long Ranger workflow (a prior step for the other four SV callers). Lastly, we note that using the minimum required memory according to the official 10x Genomics website—96GB RAM—to run Long Ranger will take even longer time to complete the SV calling process.

Resolving known deletions

Through PCR and linked-read methods, 10 large deletions (4-155kb) have previously been annotated in NA12878 (Zhang et al., 2017; Zheng et al., 2016; Wang et al., 2019). We tested the accuracy of the five SV callers to resolve these 10 deletions using TELL-Seq data (**Table 3**). The three linked-read-based callers (Long Ranger, NAIBR, and LinkedSV) had a slightly better recall performance than the two short-read-based callers, correctly calling 7-8 over 6, respectively, out of the annotated 10 deletions. This observation was expected as linked-reads have barcode information that can more accurately reflect SV status. Notably, only GRIDSS was able to identify the smallest deletion (4kb), although it was a low-quality SV call (chr4:186,172,846-186,176,845). At present, we are still investigating the quality score threshold for each software to obtain the best results. We expect to give an update in a later application note version.

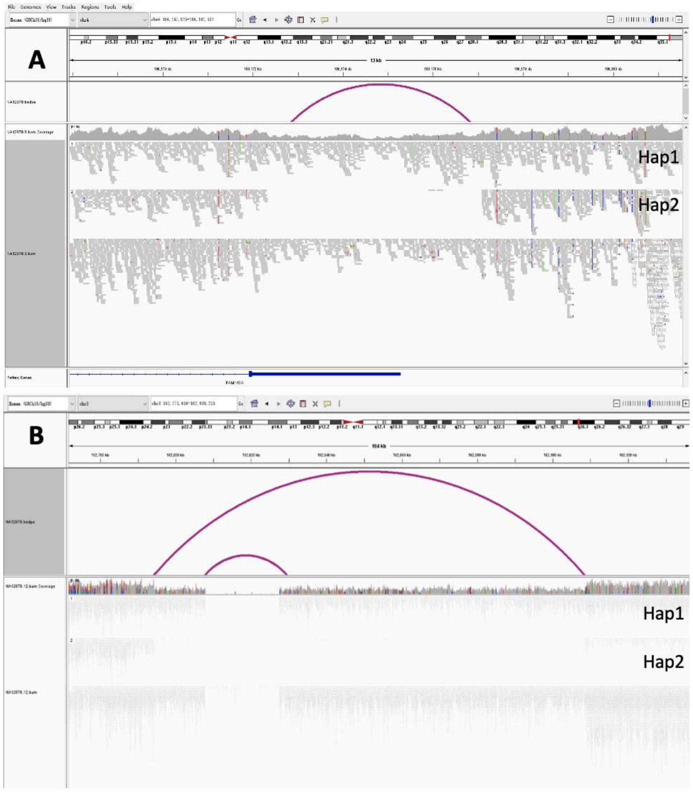


Figure 2. IGV visualization of a 4kb heterozygous deletion in chr4 (in A) and a 19kb homozygous deletion overlapping a 114kb heterozygous deletion in chr3 (in B) in NA12878. Tracks of phased haplotypes are indicated (Hap1 and Hap2).

We note that direct visualization of phased linked reads using the IGV portal confirms the presence of the 4kb heterozygous deletion at chr4:186,172,846-186,176,845 (**Figure 2A**, note ends connected by the colored arch). In the past, Zhang et al. (2016) also failed to detect this deletion using Manta. Likewise, Wang et al. (2019) also were unable to identify this deletion and the two other smallest deletions in NA12878, which the authors attributed to low sequencing coverage in their case.

Visual inspection of IGV tracks also shows that TELL-Seq reads can capture complex SVs, e.g., chr3:162,512,134-162,626,335. This genomic location has a large 114kb heterozygous deletion that overlaps with a 19kb homozygous deletion (**Figure 2B**, indicated by the large and small colored arches, respectively). Notably, the analysis of these overlapping deletions has been inconclusive in the past. Zheng et al. (2016) cataloged it only as a 114kb heterozygous deletion, while Wang et al. (2019) considered it a 19kb homozygous deletion. Our analyses, therefore, further demonstrate the robustness of TELL-Seq to identify complex SVs.

Summary

SV analysis in human cells is notoriously challenging and often requires the integration of multiple computational analyses. Traditional short reads cannot fully cover large SVs; and SVs are often complex, which requires a robust linked-read method that can resolve overlapping SVs. Although long-read sequencing technology can address these challenges, it is still impractical on a whole genome-wide scale due to sequencing cost and/or the need for relatively large amounts of input DNA. Additionally, long-read platforms have data quality issues either inherent to the state of the technology or that require a functional trade-off between maximum read length and data quality. Using 5ng of DNA on high-quality short-read platforms solves these problems at least to a certain extent, and it has the resolution to identify complex SVs.

Learn more (hyperlinks)

TELL-Seq technology

TELL-Seq technology video

TELL-Seq guides

TELL-seq software guides: Tell-Read, Tell-Sort, and IGV visualization of TELL-Seq data

IGV viewer (Broad Institute)

Additional TELL-seq applications: microbial (Illumina)

Link to raw data (hyperlinks)

www.ncbi.nlm.nih.gov/bioproject/PRJNA591637

Ordering information (hyperlinks)

Reagent boxes	
TELL-Seq WGS Library Reagent Box 1	#100001
TELL-Seq WGS Library Reagent Box 2	#100002
Primer boxes	
TELL-Seq Library Index Primer Kit	#100003*
TELL-Seq Illumina Sequencing Primer Kit	#100004

* 100009 and 100010 contain additional indexes

TELL-Seq Safety Data Sheets

References (hyperlinks)

Chen et al. Ultra-low input single tube linked-read library method enables short-read second-generation sequencing systems to generate highly accurate and economical long-range sequencing information routinely. *Genome Res*. 30:898-909 (2020)

Kosigi et al., Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*., 20:117 (2019)

Mahmoud et al. Structural variant calling: the long and the short of it. *Genome Biol*. 20:246 (2019)

Miller et al. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 16:1215 (1988)

Sudmant et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75-81 (2015)

Wang et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res*. 29:798–808 (2019)

Zhang et al. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. Biotechnol*. 35:852–857 (2017)

Zheng et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34:303-311 (2016)

For research use only. Not for use in diagnostic procedures.