TELL-Seq™ Application Note:
Targeted Phasing of Cas9-Excised Human DNA

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

October 2021
Document # 100029, version 1.0

Introduction

Every personal genome contains maternal and paternal copies of each chromosome, and both copies are homologous in autosomal chromosomes except for positions containing heterozygous variants. For over a decade, Illumina®-powered next-generation sequencing (NGS) has paved the path for massive growth in annotating heterozygous variants. However, many clinical and research applications require resolving multi-variant sets by chromosomal copy (haplotypes)—a process known as phasing. Illumina NGS relies on short sequencing reads, which does not allow phasing beyond a few hundred bases using standard DNA library preparation kits. Long sequencing reads—offered by companies such as Pacific Biosciences® and Oxford Nanopore Technologies®—are better suited for phasing. However, long-read sequencing is not as cost-effective or accurate as short-read sequencing for variant calling, which is a critical step in the phasing process.

To add phasing capabilities to Illumina instruments, Universal Sequencing Technology has developed the Transposase Enzyme-Linked Long-read-Sequencing (TELL-Seq™) Library Prep kit (Chen et al., 2020). TELL-Seq is a linked-read methodology that requires as little as 0.1 ng of input DNA, processed in a single tube in less than three hours. This application note demonstrates successful phasing of approximately 190 kb segments of DNA containing the *BRCA1* or *BRCA2* genes—associated with an increased risk for breast and ovarian cancers. These two DNA segments were excised and isolated without PCR amplification from human cells using CRISPR/Cas9 technology and the HLS-CATCH™ system (Sage Sciences®). After TELL-Seq processing of CATCH'ed DNA and Illumina sequencing, phasing was conducted using a reference genome or by *de novo* assembly—i.e., reference-free (**Figure 1**).

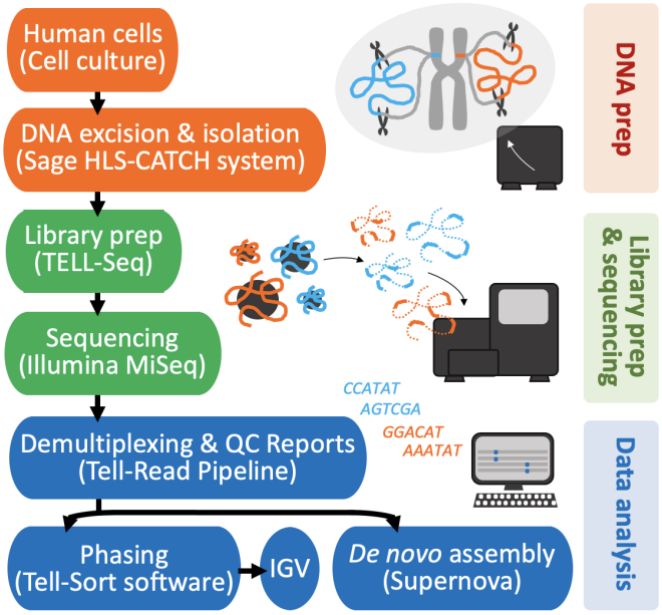

Figure 1. Experimental workflow

Methods

DNA preparation

Genome in a Bottle (GIAB) Personal Genome Project cells (B lymphocytes) for GM24385 (HG002), GM24149 (HG003), and GM24143 (HG004) were used to excise high molecular weight (HMW) DNA segments containing the *BRCA1* and *BRCA2* genes. For each sample, one million (M) cells were lysed and processed with the HLS-CATCH system (Sage Sciences). HMW DNA was excised using 2uM Cas9 and a pair of guide RNAs in each case. After 4 min/80 V electrophoretic injection, each targeted fragment was size separated and eluted by pulse-field electrophoresis. This approach yielded 220,000-400,000 copies per targeted locus with an enrichment of 200-400-fold over genomic DNA (measured by quantitative PCR). Total DNA of the two target fractions ranged between 3 and 7 ng.

Library preparation and sequencing

A small amount of CATCH'ed DNA (0.1 ng) was processed using the TELL-Seq WGS Library Prep kit with 2M TELL beads and sequenced on a MiSeq™ instrument (Illumina): 8.5M and 8.8M/8.2M of 2×150 and 2x146 bp paired-end reads for HG003 and HG002/HG004, respectively.

Data analysis and visualization

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

Contact information: technicalsupport@universalsequencing.com

software (Universal Sequencing Technology). Variant calling and phasing were performed using Tell-Sort (Universal Sequencing Technology), which incorporates BWA-MEM for read alignment, GATK-v4.0.3.0 for variant calling (Broad Institute), and HapCUT2 for phasing (github/Vibansal). The Integrative Genomics Viewer, IGV-v2.11.1 (Broad Institute) was used for data visualization, assisted with WhatsHap-v1.1 (github/whatshap). *De novo* assembly was conducted with the assembler Supernova-v2.2.1 (10x Genomics®).

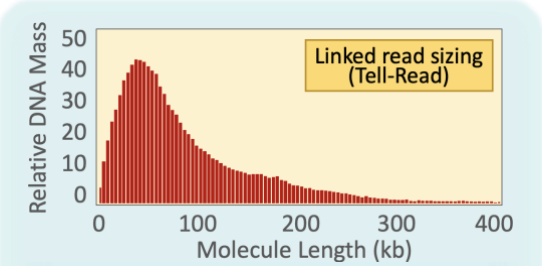

Figure 2. Linked-read molecule length

Results

Highly effective short-read linkage

GIAB materials are often used to benchmark variant calling and phasing methods. To evaluate the variant calling and phasing capabilities of the TELL-Seq method, we excised *BRCA1*- and *BRCA2*-containing fragments of genomic DNA from GIAB HG003 cells: 198.4 and 187.5 kb, respectively (chr17:43,018,501-43,216,880 and chr13:32,258,275-32,445,810). Excised DNA was processed with the TELL-Seq WGS Library Prep kit and sequenced on a MiSeq instrument (median insert size, 114-115 bp; mean sequencing depth, 106.4-110.7x). Up to 3.7% reads mapped to the targeted loci (139-fold enrichment over genomic DNA); and, on average, five short reads could be assigned to a single linked-read molecule. Linked reads could be mapped along regions larger than 20 and 100 kb in the human genome in approximately 60% and 14% of the targeted molecules, respectively, supporting a highly effective read linkage (**Figure 2**).

Successful phasing

Using GRCh38 as reference genome, we annotated 308 and 71 heterozygous single-nucleotide variants (hetSNVs) in the *BRCA1*- and *BRCA2*-locus containing fragments, respectively. Tell-Sort was able to phase 379 hetSNVs (100% phasing efficacy), which enabled us to resolve a long phase block—a continuous set of phased variants—that almost entirely covered each of the two targeted loci: 98.7% and 90.9% (or 195.9 and 170.5 kb) of the *BRCA1* and *BRCA2* loci, respectively (compare phase block and coverage extension in **Figure 3**).

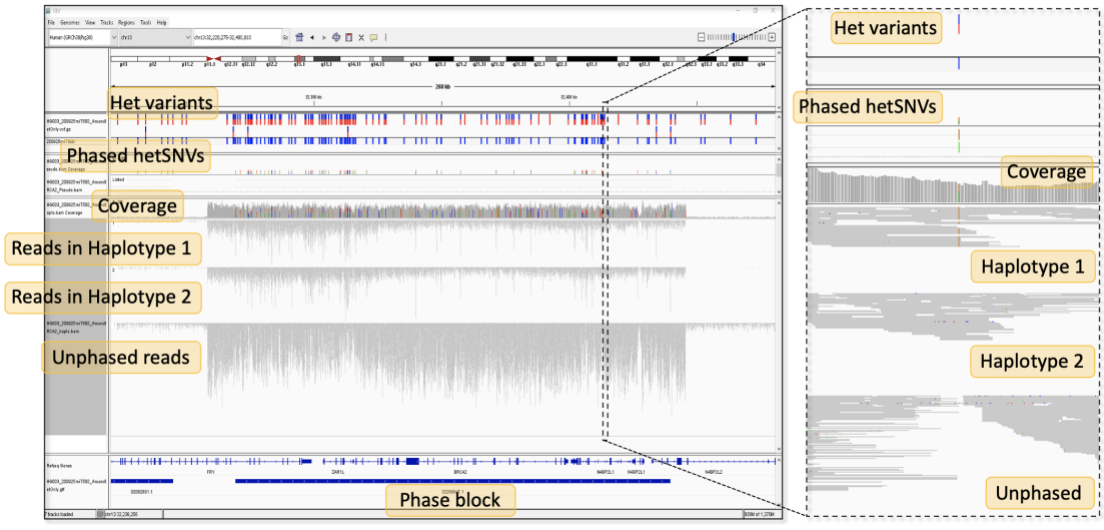

Figure 3. IGV visualization of linked-read data along the *BRCA2* locus (HG003 cells). Right image depicts a zoomed-in view of a phased hetSNV.

Validated haplotype predictions

Imputation of phase blocks and haplotypes can be validated using a parent-offspring set (family trio). We leveraged that HG003 (father) is a member of

Contact information: technicalsupport@universalsequencing.com

the GIAB Ashkenazim Trio (father-mother-son) to validate TELL-Seq-based results. As with HG003 cells, we excised *BRCA1*- and *BRCA2*-containing genomic segments from HG002 (son) and HG004 (mother) cells and then processed the target-enriched DNA with the TELL-Seq WGS Library Prep kit. Using GRCh38 as reference, we annotated 10/10 and 87/116 (son/mother) hetSNVs in the *BRCA1*- and *BRCA2*-containing fragments, respectively. Tell-Sort successfully phased 87/116 (son/mother) hetSNVs in the *BRCA2* locus (a 100% phasing efficacy), which enabled us to predict a long phase block in both family members (**Figure 4**, bottom track). Moreover, haplotype comparisons enabled us to infer the origin of each of the son's chromosomal copies. In **Figure 4**, 'S' refers to the maternal and paternal haplotypes passed on to the son by the mother and the father; and 'M' and 'F' refers to the inferred transmission of a chromosomal copy from the mother and the father, respectively, to the son (in each family member, the grey lines connect hetSNVs and homozygous variants inferred in the same haplotype).

Robust phasing capabilities

In the Ashkenazim Trio, the son has the most complete phasing annotation (Wagner et al., 2020), which we used to further validate the phasing capabilities of the TELL-Seq method. In the *BRCA2* locus—the fragment with the highest number of hetSNVs (see above)—TELL-Seq was able to correctly recall the phasing of 59 hetSNVs (previously phased by GIAB). TELL-Seq was also able to phase 21 additional hetSNVs cataloged as unphased by GIAB. We were able to validate the phasing of these 21 hetSNVs using the mother and father phasing information. Together, these analyses suggest TELL-Seq as a robust technology for phasing applications.

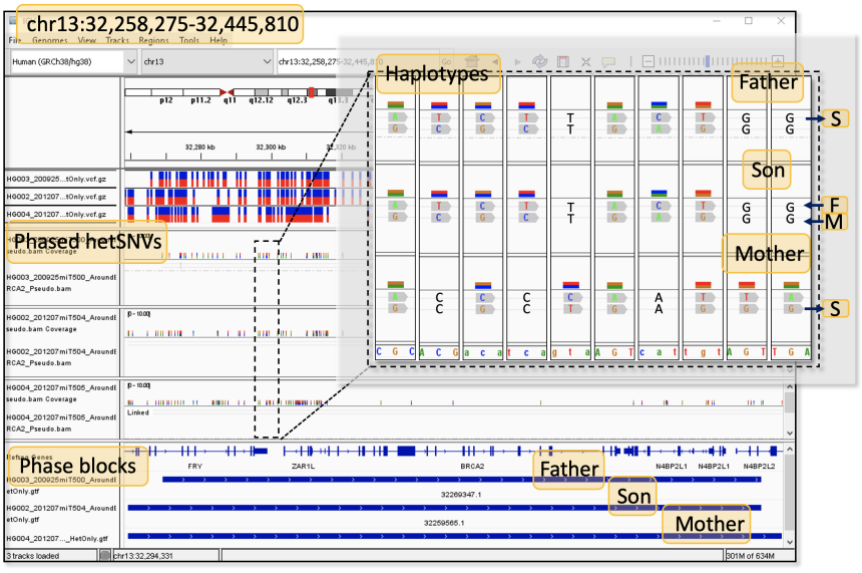

Figure 4. IGV visualization of linked-read data (*BRCA2* locus, family trio). The zoomed-in view shows inferred haplotypes connected by grey lines.

Accurate haplotype-resolved *de novo* assembly

Linked-read technology enables *de novo* assembly, which can be used as an alternative method to resolve haplotypes when a reference genome is lacking, or when a large structural variant (SV) is present. None of these two scenarios fits with the case of the *BRCA1/2* loci in the GIAB Ashkenazim Trio. Still, we sought to test *de novo* assembly as a strategy to resolve haplotypes using TELL-Seq data. In support of the validity of this option, Supernova assembled long haplotypes in both loci in the three family members: 108-183 and 147 kb assemblies ('scaffolds') in the *BRCA1* and *BRCA2* loci, respectively. The validity of the assemblies could be demonstrated by the detection of the son's scaffolds in the father and the mother. This result confirmed *de novo* assembly as an accurate haplotype-reconstruction method for TELL-Seq data.

Summary

Advances in NGS technologies have transformed the field of human molecular genetics over the last decade, including the annotation of millions of new human genetic variants. Unfortunately, this massive volume of newly captured genetic diversity has not been yet resolved at the level of haplotypes, limiting its impact in research and clinical settings (Tewhey et al., 2011). This application note shows

Contact information: technicalsupport@universalsequencing.com

that the TELL-Seq WGS Library Prep kit can add haplotype information to the power and accuracy of Illumina instruments. TELL-Seq is a fast and accurate assay that requires only picogram amounts of input DNA providing outstanding phasing performance (Chen et al., 2020).

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

Cas9 HLS-CATCH system (Sage Sciences)

IGV viewer (Broad Institute)

Supernova software (10x Genomics)

Additional TELL-seq applications: microbial (Illumina)

Link to data (hyperlinks)

NCBI/bioproject: PRJNA771708

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

Acknowledgements

T. Christian Boles, Sage Science, Inc., Beverly, MA

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)

Tewhey et al. The importance of phase information for human genomics. *Nat Rev Genet.* 12:215–223 (2011)

Wagner et al. Benchmarking challenging small variants with linked and long reads. *BioRxiv*, Dec 05, 2020

Contact information: technicalsupport@universalsequencing.com