APPLICATION NOTE

Ultra-long range phasing with linked-read sequencing technology

UST TELL-Seq using Illumina NGS platforms enables genome-scale haplotype phasing

Linked-read sequencing powered by

UNIVERSAL SEQUENCING innovation for all

illumina®

Introduction

Advances in next-generation sequencing (NGS) technologies have enabled comprehensive human whole-genome sequencing (WGS). Historically, WGS generated a single consensus sequence without distinguishing between variants on homologous chromosomes. Phased sequencing, or genome phasing, provides haplotype information about a given genome, distinguishing between alleles on maternal and paternal chromosomes.[1] By identifying haplotype information, phased sequencing can inform studies of complex traits, which are often influenced by interactions among multiple genes and alleles.[2] For example, it can provide valuable information for genetic disease research, as phased sequencing can help researchers to analyze complex heterozygote variants (eg, structural variants), measure allele-specific expression, identify variant linkages, and more.

Universal Sequencing Technology Corporation offers Transposase Enzyme Linked Long-read Sequencing (TELL-Seq), a simple, scalable library prep solution. TELL-Seq uses linked-read sequencing to apply the advantages of short-read Illumina NGS for generating highly accurate and cost-effective long-range sequencing information (Figure 1).[3] Highly accurate long reads are important for human whole-genome phased sequencing. This application note presents results from an investigation into the effects of gDNA extraction and input mass on TELL-Seq phasing metrics and demonstrates the exceptional performance of TELL-Seq as part of a comprehensive workflow for human phased sequencing using Illumina NGS systems (Figure 2).

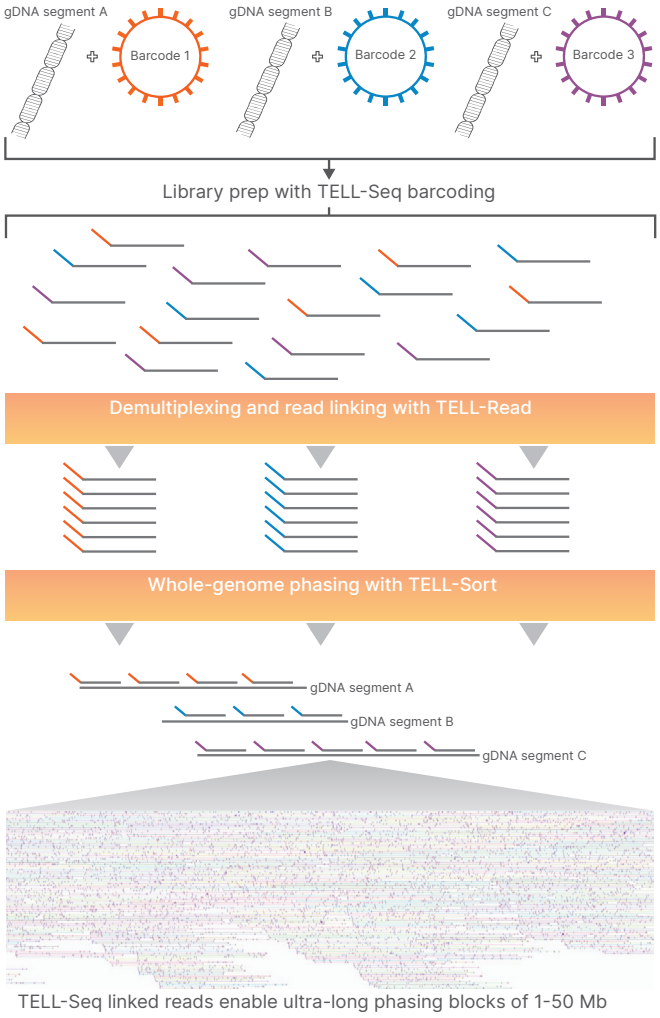

Figure 1: TELL-Seq barcoding during library preparation combined with Illumina NGS is capable of generating highly accurate, ultra-long phasing blocks 1-50 Mb in size.

Methods

Sample preparation

A trio of cell lines were obtained from the Coriell Institute for Medical Research (Table 1) for this evaluation. Two methods were evaluated for genomic DNA (gDNA) extraction: the MagAttract HMW DNA Kit (QIAGEN, Catalog no. 67563) and a salting out method used by Universal Sequencing Technologies based on a previously described protocol from 10x Genomics.[4] One DNA aliquot isolated using the MagAttract Kit was further purified using the BluePippin High-Pass PlusDNA size-selection using High-Pass Plus < 20 kb Removal Size Selection Kit (Sage Science, Catalog no. BPLUS10).

Library preparation

Libraries were prepared from either 3 ng or 5 ng of input gDNA using the TELL-Seq WGS Library Prep Kit (Universal Sequencing Technologies, Catalog no. 100001) following the manufacturer's protocol.

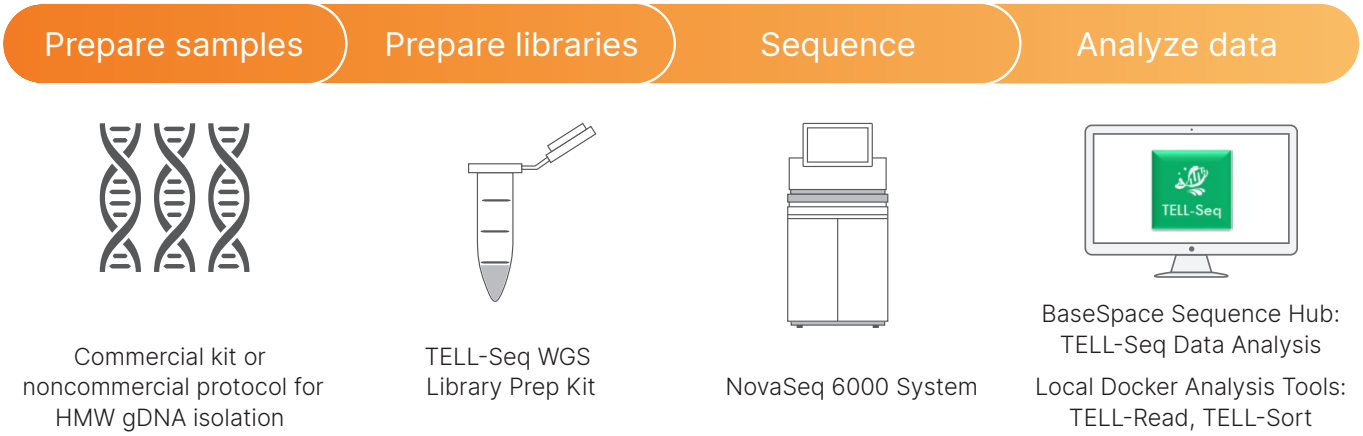

Figure 2: TELL-Seq phased sequencing workflow—Linked-read sequencing is an integrated, DNA-to-data workflow that includes TELL-Seq library preparation, sequencing on the NovaSeq 6000 System, and data analysis with the TELL-Seq BaseSpace App or TELL-Read and TELL-Sort Local Docker analysis tools.

Table 1: Samples for phased sequencing analysis

Coriell ID	NIST ID	Data source	Gender	Race	Ethnicity	Relationship to proband
GM24385	HG002	PGP	Male	White	Ashkenazim Jewish	Son (proband)
GM24149	HG003	PGP	Male	White	Ashkenazim Jewish	Father
GM24143	HG004	PGP	Female	White	Ashkenazim Jewish	Mother

NIST, National Institute of Standards and Technology; PGP, Personal Genome Project.

Sequencing

Libraries were sequenced on a NovaSeq™ 6000 System with a run configuration of 146 + 18 + 8 + 146 bp. Libraries can also be sequenced on any compatible Illumina sequencing system.

Data analysis

Sequencing data was streamed directly from the instrument into the cloud ecosystem for analysis using the BaseSpace™ TELL-Read + TELL-Sort Data Analysis App for linked-read analysis and phasing.

Results

DNA extraction and size selection

A challenge of TELL-Seq linked-read technology is that gDNA extraction methods can over-fragment DNA, resulting in smaller DNA fragments that preferentially occupy barcoded TELL-beads and effectively reduce N50 phased block sizes. Therefore, it can be advantageous to reduce or remove smaller gDNA fragments (< 20 kb) prior to TELL-Seq library preparation. To evaluate this and maximize phasing performance with the TELL-Seq WGS Library Prep Kit, high-molecular weight (HMW) gDNA from GM24385 (HG002, proband) was isolated using each of the described gDNA extraction and size selection strategies. The resulting fragment sizes were

visualized using pulse field gel electrophoresis (PFGE), showing that the salting out method yields longer gDNA fragments, as indicated by a lower intensity smear < 10 kb and higher intensity main band near the 200 kb ladder marker, when compared to the MagAttract HWM gDNA approach (Figure 3). Additionally, results demonstrate successful removal of small target fragments < 20 kb from the initial MagAttract HMW gDNA prep following high-pass DNA size < 20 kb removal on the BluePippin Instrument (Sage Science).

Primary WGS metrics and coverage

Following library preparation and sequencing, each sample was subsampled to ~40× unique genome coverage before analysis. More than 97% of reads from all sample conditions tested were mapped to the GRCh38 reference. As expected, conditions with higher gDNA inputs resulted in lower duplication rates than lower inputs (Table 2). Relatively good coverage uniformity was observed from regions between 20% and 65% GC content, while coverage uniformity declined for very high AT-rich regions and for high GC-rich regions (data not shown). Variant calling without incorporating any barcoding information resulted in 93%-98% recall rate and 93%-97% precision rate on SNVs compared to the Genome In A Bottle (GIAB) HG002 high-confidence benchmark v4.2.1 truthset (data not shown). Together, these results indicate that the salt out protocol with 5 ng input was the highest performing condition.

TELL-Seq WGS barcoded metrics

Barcoding analysis revealed linked long reads that were successfully mapped along large regions in the human genome across all sample conditions tested. There were 7-9M effective barcodes identified across all sample conditions and approximately 15-30M unique molecules recovered (Table 3). Greater proportions of DNA molecules larger than 20 kb and 100 kb were captured using the salt out protocol and lower gDNA inputs (Table 3). Mean molecule lengths were consistently higher using the salt out protocol, compared to other methods (Table 3, Figure 4). Sequencing coverage per molecule by TELL-Seq linked long-reads averaged from ~10% to 16% (data not shown).

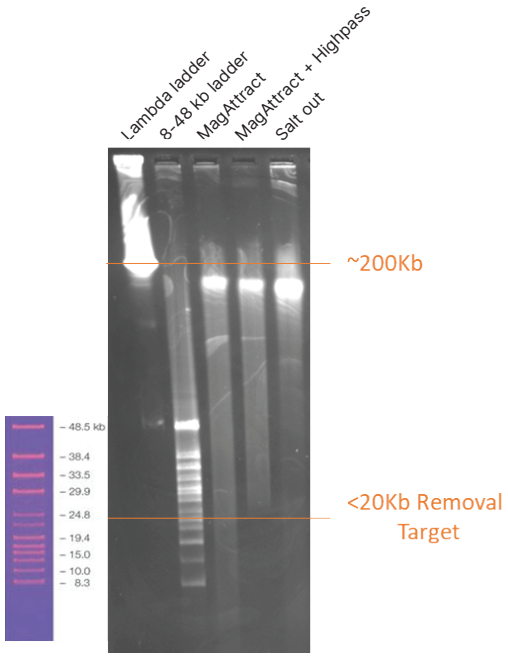

Figure 3: Extracted gDNA size selection—Sizing of extracted gDNA from the proband was determined by PFGE.

Table 2: WGS mapping metrics

gDNA isolation method	MagAttract	MagAttract + Highpass	Salt out	Salt out
gDNA input mass	3 ng	5 ng	3 ng	5 ng
Total reads	1.7B	1.2B	1.3B	1.2B
% ≥ Q30 bases	97.52%	91.65%	91.71%	92.18%
Mapped reads	97.62%	97.87%	97.96%	98.14%
Duplicate reads	68.08%	55.20%	60.05%	54.27%
Median insert size	194 bp	187 bp	206 bp	187 bp
Unique genome coverage	38×	41×	40×	41×

Table 3: Barcoding metrics

gDNA isolation method	MagAttract	MagAttract + Highpass	Salt out	Salt out
gDNA input mass	3 ng	5 ng	3 ng	5 ng
Barcodes detected	7,529,942	9,098,428	8,608,068	9,241,689
Molecules detected	15,909,526	30,657,519	20,223,171	26,096,693
DNA in molecules > 20 kb	89.70%	90.00%	93.40%	93.20%
DNA in molecules > 100 kb	31.60%	17.90%	34.80%	27.00%
Mean molecule length	36,590	36,844	45,056	43,981
Mean molecule per barcode	2.1	3.4	2.3	2.8
N50 reads per molecule	42	25	41	35

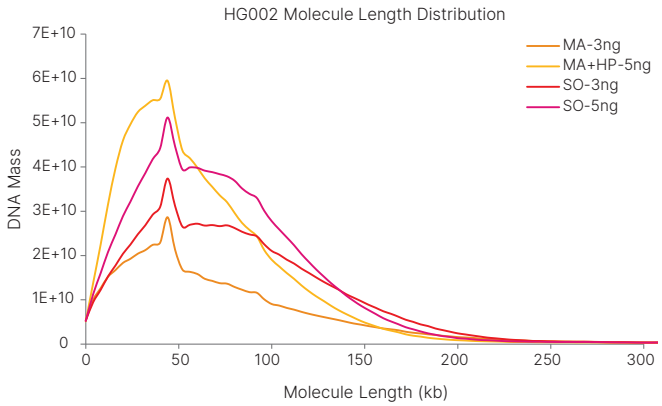

Figure 4: Distribution of molecule lengths—Molecule length distribution analysis showed high proportion of larger molecules across conditions.

Haplotype phasing human genomes

TELL-Sort software requires heterozygous single nucleotide variants (SNVs) to be present within a phased block to perform variant phasing (Figure 5). The TELL-Sort phasing tool, which utilizes the HapCUT2 application, was used for analysis.[5] Phasing of over 98.8% of heterozygous SNVs in each sample was achieved with the TELL-Sort tool, with low switch error rates and high N50 phasing block sizes (Table 4). The salt out protocol with 3 ng DNA input yielded the highest N50 phased block size (8.1 Mb), longest phased block size (28 Mb), second highest phased variant fraction (99.5%), and second lowest switch and mismatch error rates (0.063% and 0.094%, respectively). The lowest performing method was the MagAttract kit with 3 ng DNA input (Table 4). This difference is likely attributed to the salt out protocol yielding a lower proportion of DNA fragments below 10 kb and a higher proportion of long gDNA fragments that approached 200 kb. Additional sub-sampling analysis demonstrated a strong positive correlation between N50 phased block size and unique genome coverage (Figure 6). This finding may be useful for researchers looking to increase phased block sizes, as it may be achieved by sequencing TELL-Seq libraries at higher depths than the recommended levels of 30-40× unique genome coverage.

Figure 5: SNV phasing with TELL-Sort—TELL-Sort software requires heterozygous SNVs (green) to be present in a phased block to perform variant phasing.

Table 4: Phasing metrics

gDNA isolation method	MagAttract	MagAttract + Highpass	Salt out	Salt out
gDNA input mass	3 ng	5 ng	3 ng	5 ng
Switch rate	0.135%	0.124%	0.063%	0.052%
Mismatch rate	0.243%	0.170%	0.094%	0.072%
Phased count	2,442,655	2,653,451	2,524,207	2,532,128
Phased count (%)	98.8%	99.1%	99.5%	99.6%
N50 phased block	4,170,165	7,281,053	8,071,353,	5,968,988
Number of SNVs in longest phased block	293,731	379,791	397,949	336,565
Longest phased block length	19,425,298	25,123,081	28,431,382	20,056,668

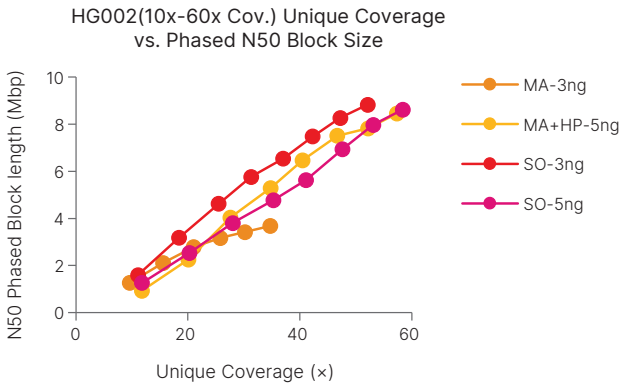

Figure 6: N50 phased block size and unique genome coverage—A positive correlation was observed between N50 phased block size and unique genome coverage across DNA isolation methods and input amounts.

Ashkenazim trio phasing analysis

Based on initial results using the proband (HG002) to benchmark different sample processing conditions (gDNA extraction, size selection, and input mass to TELL-Seq library preparation), the salt out protocol with 5 ng gDNA input mass was determined to be the best. Samples from the remaining members of the Ashkenazim trio (HG003 and HG004) were likewise prepared for TELL-Seq phasing for complete phased trio analysis.

Across the trio of samples exceptionally high long-range phasing metrics were observed (Table 5), including high fraction of SNVs phased (99.3% to 99.6%), large phased blocks (~6 Mb to ~19 Mb), and low switch (0.14% to 0.05%) and mismatch (0.09% to 0.07%) error rates. HG003 yielded the highest phased block sizes with N50 phased blocks at 18.7 Mb and longest phased block size at 46.7 Mb (Table 5). This is likely due to the larger mean molecule length observed with HG003 (51,389 kb), as compared to HG004 (47,939 kb) and HG002 (43,981 kb). Visualization of phased blocks for the trio across the genome clearly shows the exceptional performance of TELL-Seq (Figure 7).

Consistently lower switch rates were seen in HG002 (0.05%), compared to HG003 (0.20%) and HG004 (0.14%). This finding is likely a direct result of the availability of a higher fidelity GIAB phased truthset available for comparison for HG002 (HG002_GRCh38_1_22_v4.2.1 benchmark_phased_ MHCassembly_StrandSeqANDTrio.vcf.gz), which is currently unavailable for HG003 or HG004.

Table 5: Ashkenazim trio phasing metrics

Phasing metric	HG004	HG003	HG002
Relationship to proband	Mother	Father	Proband
Switch rate (%)	0.139	0.210	0.052
Mismatch rate (%)	0.079	0.087	0.072
Phased count	2,502,129	2,510,719	2,532,128
Phased count (%)	99.3	99.6	99.6
N50 phased block (bp)	7,226,432	18,717,182	5,968,988
No. of SNVs in the longest pahsed block	354,781	584,354	336,565
Longest phased block length (bp)	32,370,477	46,733,675	20,056,668

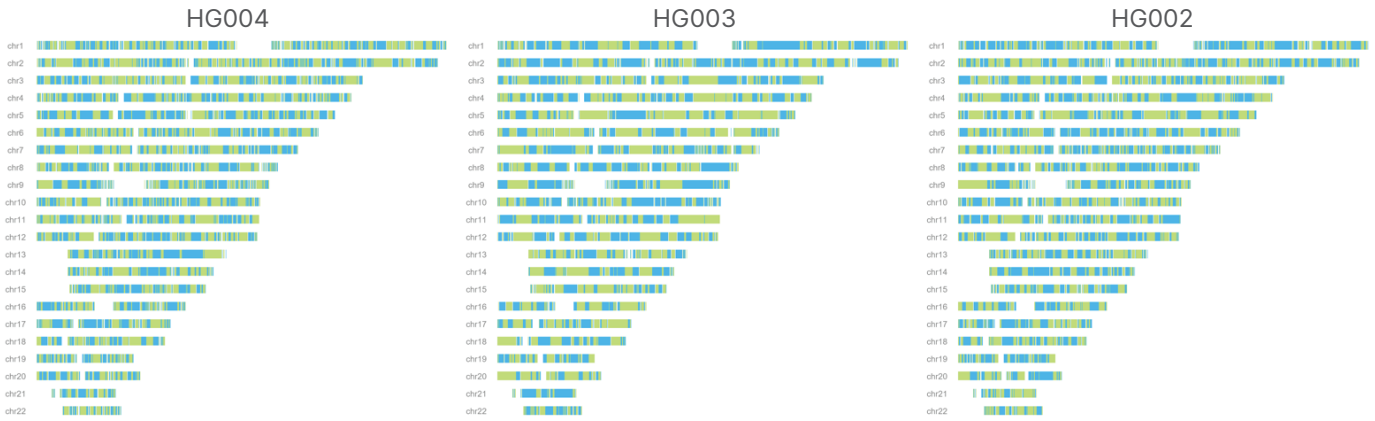

Figure 7: Genome-wide phasing blocks—TELL-Seq reads were converted to a 10x Genomics file format for analysis and visualization using Long Ranger software. This view shows the distribution of ultra-long phased blocks in alternating colors of blue and green across the genome (excluding X and Y chromosomes). Of note, larger phased blocks were seen in HG003 (N50 phased block length ~19Mb) compared to HG004 (N50 phased block length ~7Mb) and HG002 (N50 phased block length ~6Mb.

Summary

Historically, acquiring genomic phasing information has been a challenging, costly process that required specialized long-read sequencing technology. TELL-Seq technology enables Illumina NGS systems to generate highly accurate data while reducing costs, turnaround time, and DNA input requirements. This application note demonstrates the exceptional performance of TELL-Seq library preparation combined with Illumina NGS for generating ultra-long phasing blocks, providing an accessible solution for researchers to perform genome phasing studies using existing instrumentation.

Learn more

Phased sequencing,
illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing/phased-sequencing.html

TELL-Seq technology,
universalsequencing.com/technology

References

1. Browning SR, Browning BL. Haplotype phasing: existing methods and new develoments. *Nat Rev Genet*. 2011; 12(10):703-714. doi: 10.1038/nrg3054.
2. Porubsky D, Ebert P, Audano PA, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol*. 2021;39(3):302-308. doi: 10.1038/s41587-020-0719-5.
3. Chen Z, Pham L, Wu TC, 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*. 2020;30(6):898-909. doi: 10.1101/gr.260380.119.
4. 10X Genomics. 2017. Salting Out Method for DNA Extraction from Cells. Accessed December 1, 2021.
5. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*. 2017;27(5):801-812. doi:10.1101/gr.213462.116.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-00581 v1.0