Accurate full-length gene diplotyping using ultralow input linked-read library and short reads

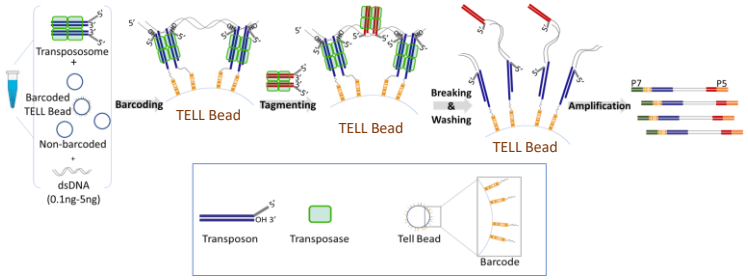
Tom Chen[1], Peter Chang[1], Chris Boles[2], Long Pham[1], Yu Xia[1], Yong Wang[1]

1. Universal Sequencing Technology, Carlsbad, CA 92011 *2. Sage Science, Inc., Beverly, MA 01915* *email: tchen@universalsequencing.com*

Abstract

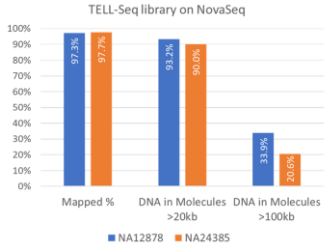
The human genome is composed of two haplotypes. Long-range sequencing information is required for diplotyping by phasing haplotypes properly and identifying structural variations efficiently. **TELL-Seq**™ library technology[1], which enables a low-cost, high-accuracy and high-throughput short-read sequencer to generate over 100 kb long-range sequencing information with as little as 0.1 ng input material, is an excellent method for diplotyping application. With TELL-Seq linked reads, we phased NA12878 and NA24385 human genomes with N50 phased block size up to 15Mb long. We further demonstrated a diplotype sequencing method for targeting genomic regions of interest greater than 0.1 megabases. We used CRISPR-Cas9 digestion and electrophoresis size selection (Sage HLS-CATCH[2]) to isolate intact DNA targets for a 200 kb *BRCA1* locus and a 200 kb *BRCA2* locus from live cells. Without any intermediate amplification step, 0.1 ng of target-enriched samples from an Ashkenazi trio were directly used for preparation of a TELL-Seq library and sequenced on a MiSeq system. With TELL-Seq reads, we successfully established the complete diplotypes of *BRCA1* and *BRCA2* loci for the family by reference-based phasing method. In addition, we were also able to construct the complete diplotypes of targeted genes by de novo assembly without any references.

Single-tube 3-hour Linked-read Library Prep

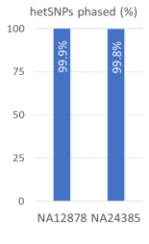


TELL-Seq Library Workflow is Simple & Scalable

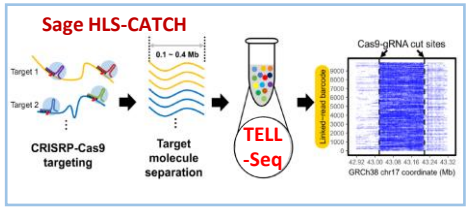
Whole Genome Phasing

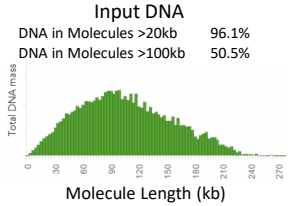


TELL-seq Human Phasing	2×146 PE	
	NA12878	NA24385
Coverage Depth (Unique)	38×	46×
Longest Phased Block	67.5 Mb	59.2 Mb
N50 Phased Block	14.4 Mb	13.4 Mb
Switch Error Rate	0.04%	0.08%

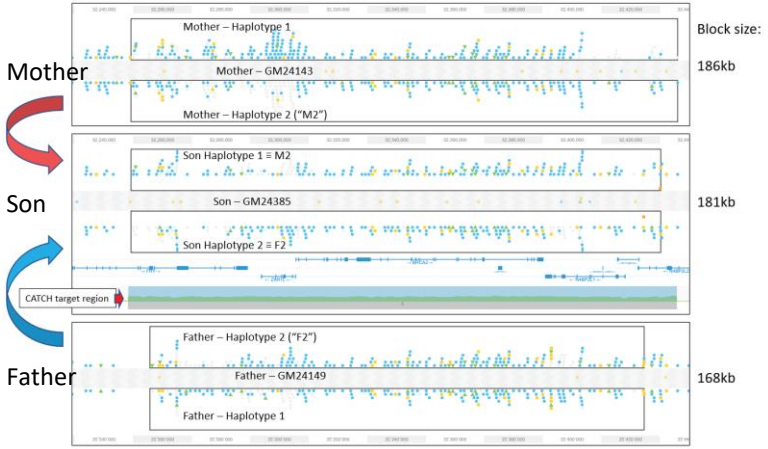


Targeted Full-Length Gene Diplotyping





Diplotype Phased *BRCA2* Gene of the Ashkenazi Trio



BRCA1 & *BRCA2* Diplotype Variants Among the Trio

BRCA1 ⇨

BRCA2 ⇩

Sample / Relationship	Haplotype (1 & 2)	Genome Coordinate (GRCh38) 43,020,890	43,048,520	43,099,914	43,129,533	43,132,602	43,153,380	43,183,966	43,187,470	43,188,035	43,199,383	43,202,523
GM24143 / Mother	M1	G	T	G	G	C	C	C	G	C	T	T
	M2	A	C	G	G	C	T	C	G	C	A	A
GM24385 / Son	S1 = M2	A	C	G	G	C	T	C	G	C	A	A
	S2 = F2	A	T	A	A	T	C	A	A	G	T	T
GM24149 / Father	F2	A	T	A	A	T	C	A	A	G	T	T
	F1	G	C	G	G	C	T	C	A	C	A	A

Sample / Relationship	Haplotype (1 & 2)	Genome Coordinate (GRCh38) 32,274,573	32,283,544	32,291,422	32,296,745	32,296,952	32,298,850	32,299,707	32,302,159	32,303,184	32,304,467	32,305,909	32,310,367	32,310,516	32,313,393	32,313,397	32,313,428	32,313,925	32,314,336	32,315,831	32,316,435	32,340,099	32,402,221	32,402,471	32,405,001	32,405,510	32,406,619	32,406,662	32,407,005	32,407,142	32,409,113	32,410,060	32,417,289	32,424,225	32,424,473	32,424,754	32,425,814
GM24143 / Mother	M1	A	C	T	A	A	C	T	G	A	A	G	C	T	C	C	G	C	T	A	A	C	A	G	C	G	A	G	T	A	A	G	G	A	T	G	G
	M2	G	T	T	G	G	G	G	A	G	T	A	G	G	T	T	A	A	C	A	A	T	A	G	C	G	A	G	T	A	A	G	A	G	G	T	C
GM24385 / Son	S1 = M2	G	T	T	G	G	G	G	A	G	T	A	G	G	T	T	A	A	C	A	A	T	A	G	C	G	A	G	T	A	A	G	A	G	G	T	C
	S2 = F2	G	T	C	A	A	C	T	G	A	A	G	C	T	C	C	G	C	T	G	G	C	G	A	T	T	G	T	C	T	G	T	A	G	G	T	C
GM24149 / Father	F2	G	T	C	A	A	C	T	G	A	A	G	C	T	C	C	G	C	T	G	G	C	G	A	T	T	G	T	C	T	G	T	A	G	G	T	C
	F1	A	C	T	G	G	G	G	A	G	T	A	G	G	T	T	A	A	C	A	A	T	A	G	C	G	A	G	T	A	A	G	A	G	G	T	C

Conclusion

TELL-Seq is a simple and scalable linked-read library for haplotype phasing and de novo sequencing. In a PCR tube, under a standard NGS laboratory setting and without the need for any expensive protocol-specific instrument, TELL-Seq Library Prep Kit will generate an Illumina sequencing library in 3 hours from as low as 0.1ng DNA input. Combined with HLS-CATCH (Sage Science) target enrichment, TELL-Seq successfully de novo assembled 200kb long diplotypes of *BRCA1* and *BRCA2* loci. With the picogram input requirement for library preparation and highly accurate short-read sequencing platforms, targeted full-length gene diplotyping by linked reads gets closer to become a routine application, especially with further improvement on enrichment technologies for long contiguous targets.

References

1. Chen, Z et al. 2020. *Genome Research* **30**: 898-909.
2. Shin, G et al. 2019. *Nucleic Acids Research* **47**: e115.

UNIVERSAL SEQUENCING
innovation for all