An Optimized Single-Tube Linked-Read Method Empowers Short-Read Sequencers to Phase 2 to 200 Kilobase DNA targets

Ivan Garcia-Bassets[1], Veronika Mikhaylova[1], Madison Rzepka[1], Tetsuya Kawamura[1]Yu Xia[1], Peter L. Chang[1], Richard Zhou[4], Long Pham[1], Naisarg Modi[1], Likun Yao[2], Adrian Perez-Agustin[3], Sara Pagans[3], Christian Boles[4], Ming Lei[5], Yong Wang[5], Zhoutao Chen[1]

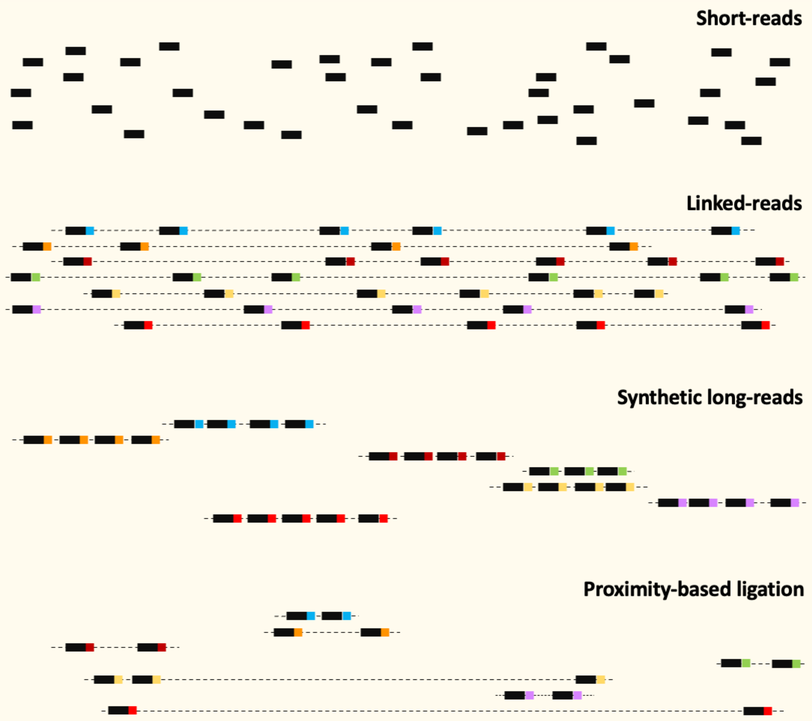
[1] Universal Sequencing Technology Corp., Carlsbad, CA 92011, USA; [2] Cellular & Molecular Medicine Department, University of California, San Diego, La Jolla, CA 92093 USA; [3] Medical Sciences Department, School of Medicine, University of Girona, Girona, Spain; [4] Sage Sciences Inc., Beverly, MA 01915, USA; [5] Universal Sequencing Technology Corp., Canton, MA 02021, USA

AGBT-2023
Poster #504

A Brief Intro to Linked Reads

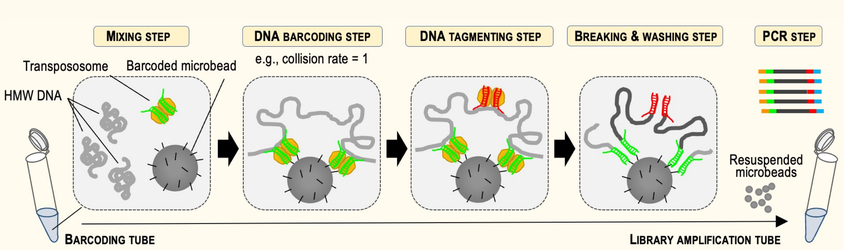
What are linked reads?

Linked reads are a type of short reads tagged with a unique molecular identifier (**barcode**) that informs about the compartmentalized fragmentation of a **high-molecular-weight** (HMW) DNA molecule. On a high-throughput scale, millions of HMW DNA molecules can be virtually or physically compartmentalized and individually fragmented for independent tagging (**co-barcoding**) of the derived subfragments. After sequencing, barcodes are used to 'link' short-reads (subfragments) and reconstruct the original HMW DNA molecules (***compare short-reads and linked reads; below***). Linked reads are sequenced with a short-read sequencer.



What is TELL-Seq?

Generation of linked-read libraries can be cheap, but generation of linked-read reagents is complex and costly; it also requires substantial NGS technical experience. To free researchers from this hassle, **Universal Sequencing Technology Corp.** (UST) has developed a commercial library preparation kit – TELL-Seq™, that provides easy-&-ready-to-use linked-read reagents. **TELL-Seq** (Transposase Enzyme Linked Long-read Sequencing) is a **single-tube-based linked-read method** that leverages a dense solution to enable efficient DNA partitioning and co-barcoding in the open space of **a PCR tube**. Linked-read libraries can be generated in **only three hours** with as little as **0.1-5 ng** of input material, without the need for special instrumentation (**microfluidics-free, plate-free**). TELL-Seq relies on the surface of **micron-sized beads** to capture ~1-8 DNA molecules per microbead and enable **millions of independent co-barcoding reactions** in the open space of a PCR tube (**Workflow shown below).** TELL-Seq libraries are then sequenced in a short-read sequencer, which provide **higher accuracy, higher throughput, and low cost**.



Abstract - **In the human genome, heterozygous sites are genomic positions with a different nucleotide or length option (allele) inherited from each parent. Often, neighboring heterozygous sites are separated by distances longer than the length of a short sequencing read. Hence using a short-read platform to resolve these differences by parental copy (known as phasing) requires a library preparation method that incorporates long-range information into short reads. Linked-read library preparation methods allow tagging short reads that contain genomic sites separated up to 200 kb in the genome with the same unique molecular identifier (barcode). Linked reads enable the phasing of a whole genome. However, phasing a whole genome is not cost-effective for clinical or research applications interested in a single or a few loci (a targeted approach). Here, we report conditions to phase one or a few targets with linked-read method that does not require special instrumentation, single-tube TELL-Seq (microfluidics-free). We validate this method with picogram amounts of clinically relevant targets and a variety of purity levels, heterozygosity densities, and lengths. We isolated the targets according to size using CRISPR/Cas9-mediated excision coupled with pulse-field electrophoresis for 180-200 kb sizes, CRISPR/Cas9-mediated protection from exonuclease digestion for 20-40 kb sizes, and PCR for 2-13 kb sizes. We also report an analysis pipeline that outputs phased TELL-Seq data on the Integrative View Genome (IGV). Together, our analyses demonstrate that single-tube TELL-Seq enables high-quality phasing of one or a few 2-200 kb targets with the low cost and high accuracy of a short-read sequencer.**

Phasing 40-200 kb Targets

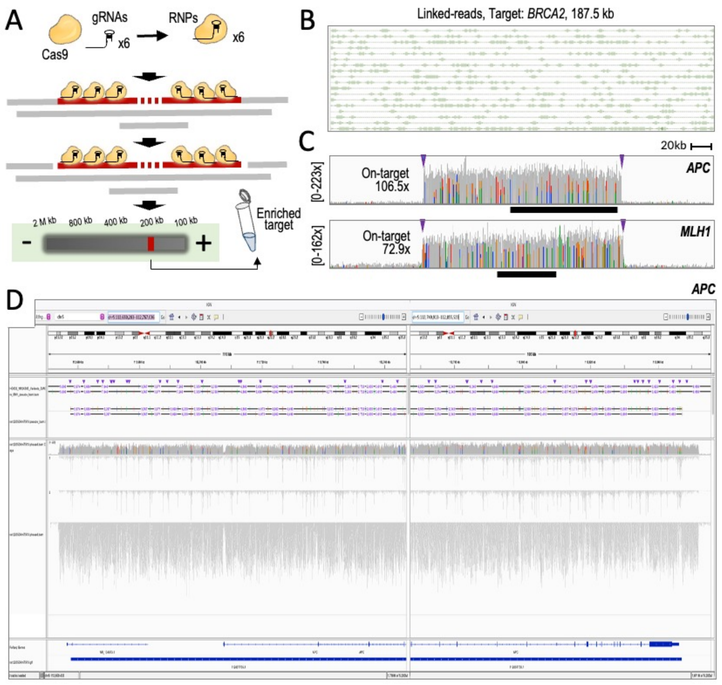


Figure 1. Phasing of 180-200 kb targets enriched with a CRISPR-Cas9/size selection-based method (HLS2). A. Experimental workflow. **B.** Representative example of linked reads (186 kb region). Boxes represent short reads. **C.** Representative read profiles (coverage, ranges included in panels) recovered from TELL-Seq libraries showing robust on-target recovery compared to background. Targets: 200 kb *APC*- and *MLH1*-containing loci. Arrowheads represent gRNA binding sites. Black bars represent the *APC* and *MLH1* genes. **D.** Screenshot from IGV portal showing TELL-Seq results. Tracks (from top to bottom): phased reference/haplotypes (GIAB); phased target/haplotypes (*APC* locus; numbers indicate large distances between phased sites; coverage (target); haplotype 1 (target); haplotype 2 (target); unphased reads (target); gene annotations; phase block (blue bar)

Phasing 20-40 kb Targets

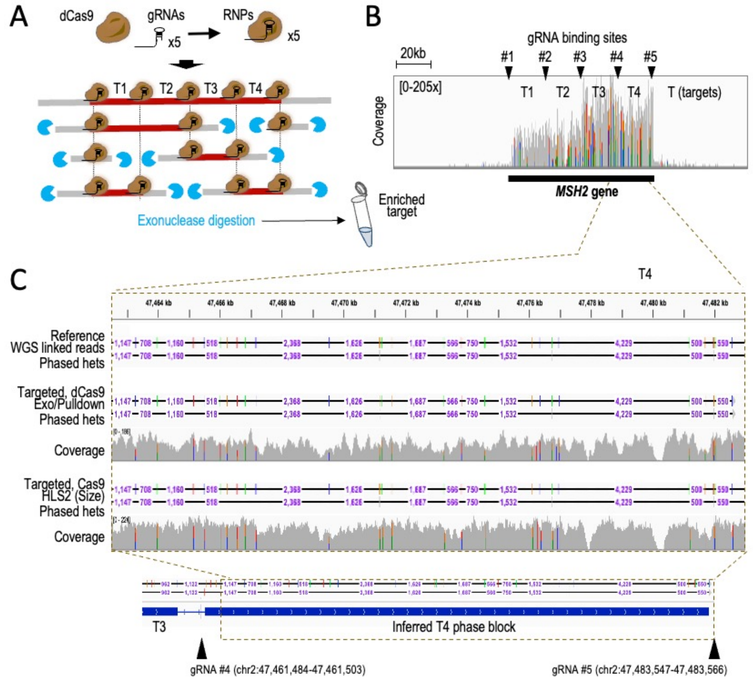


Figure 2. TELL-Seq-mediated phasing of four adjacent 20 kb targets enriched with a modified CaBagE protocol that leverages dCas9 to protect both sides of the targeted regions. A. Experimental workflow. Targets can be enriched as individual loci (~20 kb) or as longer ~40, 60, or 80 kb fragments, depending on the HMW properties of input genomic DNA. Protected DNA (target) is represented in dark red; unprotected DNA (non-target) is represented in grey. **B.** Coverage across the *MSH2* locus. Arrowheads indicate gRNA binding sites (#1-#5). **C.** Screenshot from IGV Portal showing phased data and coverage from Targeted TELL-Seq using dCas9/Exo or HLS2/Cas9 experiments as DNA enrichment methods (WGS linked-read phased data shown on top as a reference, HG002). Bottom tracks show phased data and inferred phase blocks in the Targeted TELL-Seq using dCas9/Exo experiment.

Phasing 2-20 kb Targets

A

	PIK3CA amplicons							*SCN10A* amplicons						
Sample characteristics	C#1	C#2	C#3	C#4	C#5	C#6	C#7	C#8	C#9	C#10	C#11	C#12	C#13	C#14
Priming method for cDNA (R6 vs dT)	R6	R6	R6	R6	dT	dT	dT	--	--	--	--	--	--	--
DNA source (genomic, gen.)	--	--	--	--	--	--	--	gDNA	gDNA	gDNA	gDNA	gDNA	gDNA	gDNA
PCR product size (kb)	3.4	3.4	1.8	1.8	3.4	3.4	3.4	13	2.7	2.7	3.9	3.9	4.1	4.1
Target input amount (pg)	20	20	20	5	20	5	0.4	20	20	20	20	20	20	20
Target fraction over total DNA	1/4	--	1/4	1/14	1/5	1/20	1/250	1/4	1/4	--	1/4	--	1/4	--
E.coli gDNA amount (pg)	60	--	60	65	--	--	--	60	60	--	60	--	60	--
Lambda phage gDNA amount (pg)	--	--	--	--	80	95	99.6	--	--	--	--	--	--	--
Other human fragments (pg)	--	60	--	--	--	--	--	--	--	60	--	60	--	60
Data analysis														
Technical replicates	6	6	3	3	3	2	2	6	3	6	3	6	3	6
Average on-target coverage (x)	213	257	3,574	834	423	73	58	480	307	728	301	397	58.8	124
Average on-target, dedup reads (number)	6.0k	7.2k	5.5k	12.8k	12.4k	2.1k	1.6k	57.8k	6.8k	15.9k	9.9k	12.6k	2.1k	4.2k
Expected hets per test	3	3	3	3	3	3	3	23	4	4	6	6	8	8
Tests w/all expected hets recalled correctly	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Tests w/all expected hets phased correctly	100%	100%	100%	100%	66%	100%	100%	100%	100%	100%	100%	100%	100%	100%

B

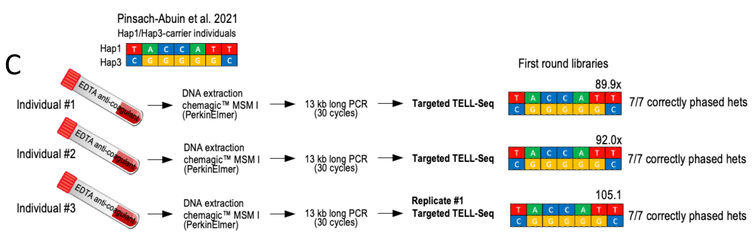


Figure 3. Phasing of *SCN10A* and *PIK3CA* PCR products (sizes: 1.8-13 kb). A. Summary table of genotyping and phasing results using *SCN10A* and *PIK3CA* amplicons as input from NA12878 genomic DNA (n = 33 tests) or breast cancer HCC202 cDNA (n = 25 tests). In red, highlighted the number of technical replicates, and genotyping and phasing accuracies. R6 (random hexamers) and dT (oligo-dT) for cDNA priming, as indicated. **B.** (*Top haplotypes*) A representative example is shown of n = 23 phased positions (100% genotype and phasing accuracy). Clinically relevant alleles (*); minor alleles are indicated (_). TACCATT corresponds to Hap1 (bottom copy) and CGGGGGC corresponds to Hap3 (top copy). (*Screenshot from IGV portal*) Representative examples of *SCN10A* phasing (NA12878). Fragment sizes: 3.9 kb (tracks 1-3), 4.1 kb (tracks 4-6), 2.7 kb (tracks 7-9) in triplicate (n = 3 replicates each), and 13 kb (tracks, 10-12; n = 6 replicates). Screen capture shows phased data on IGV. **C**. Targeted TELL-Seq with three 13 kb amplicons generated from peripheral blood-extracted genomic DNA from three known Hap1/Hap3 carrier individuals.

CONCLUSIONS: TELL-Seq enables high-quality phasing of one or a panel of 2-200 kb targets without the need for special instrumentation (a single-tube, microfluidics-free, plate-free method) and the low cost and high accuracy of a short-read sequencer. TELL-Seq is commercially available.

REFERENCES

Chen, Z. *et al.* Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.* 30, 898–909 (2020) → **TELL-Seq**

Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16, 1215 (1988) → **HMW genomic DNA extraction**

Pinsach-Abuin, M. L. *et al.* Analysis of Brugada syndrome loci reveals that fine-mapping clustered GWAS hits enhances the annotation of disease-relevant variants. *Cell Rep. Med.* 2, 100250 (2021) → **Patient samples**

Shin, G. *et al.* Targeted short read sequencing and assembly of re-arrangements and candidate gene loci provide megabase diplotypes. *Nucleic Acids Res.* 47, e115 (2019) → **HLS2 system**

Wallace, A. D. *et al.* CaBagE: A Cas9-based Background Elimination strategy for targeted, long-read DNA sequencing. *PloS One* 16, e0241253 (2021) → **CaBagE method**