Technical Note:

A Brief Introduction to Linked-Read Technology

UNIVERSAL SEQUENCING
innovation for all

What are linked reads?

Linked reads are short reads tagged with a unique molecular identifier (barcode) that informs about the compartmentalized fragmentation of a long DNA molecule. On a high-throughput scale, millions of high-molecular-weight (HMW) DNA molecules are virtually or physically compartmentalized at the same time and individually fragmented for the independent co-barcoding of the subfragments. After sequencing, barcodes are used to 'link' subfragments and reconstruct the original HMW DNA molecules (**Figure 1**). This technology can use sequencing reads generated from a short-read sequencer, such as Illumina®'s platforms, to generate long-range information up to 200-300 kb.

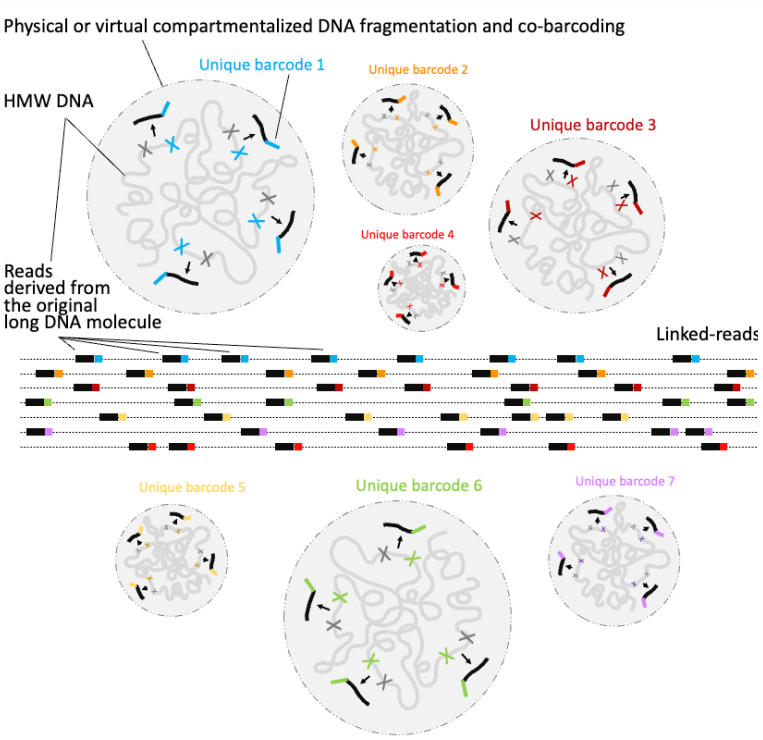

Figure 1. Physical or virtual DNA fragmentation and co-barcoding

Categories of linked-read methods

There are three categories of linked-read methods based on the level of physical or virtual DNA compartmentalization: droplet-based, microwell-based, and single-tube-based (virtual compartmentalization or partition-free). Droplet-based methods leverage an emulsification oil to generate micron-sized droplets to compartmentalize DNA into pools of ~1-20 DNA molecules per droplet (**Figure 2**, left). Microwell-based methods rely on ultra-high dilutions in 384-well plates to compartmentalize DNA into pools of a few thousand molecules per microwell (**Figure 2**, middle). Single-tube-based methods rely 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 (**Figure 2**, right).

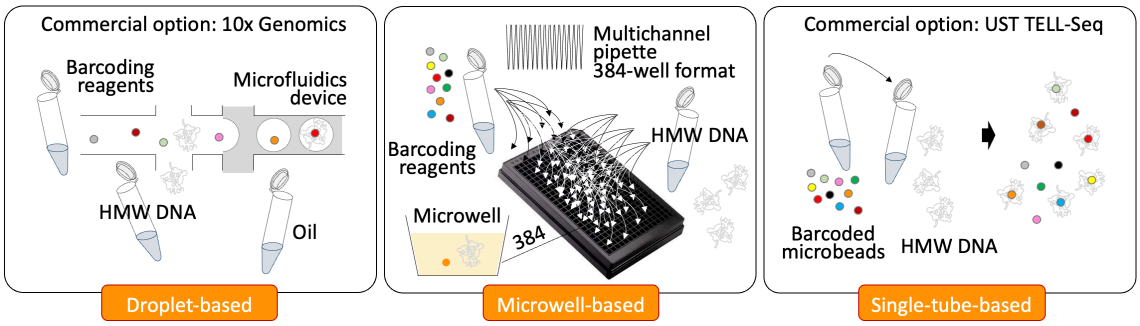

Figure 2. Three categories of linked read methods based on the level of DNA compartmentalization

Contact information: technicalsupport@universalsequencing.com

What are the commercially available linked-read methods?

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 (**Figure 2**, right). TELL-Seq is the only linked-read method that is commercially available worldwide. A previously commercialized droplet-based linked-read method developed by 10x Genomics® that required expensive instrumentation was discontinued in 2020 (**Figure 2**, left).

What is TELL-Seq?

TELL-Seq (Transposase Enzyme Linked Long-read Sequencing) is a single-tube-based 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 and without the need for special laboratory instrumentation (workflow shown in **Figure 3**). TELL-Seq libraries are then sequenced in short-read instruments (e.g., Illumina), which provide higher accuracy (for base calling), higher throughput, and lower cost than long-read options.

What are synthetic long reads?

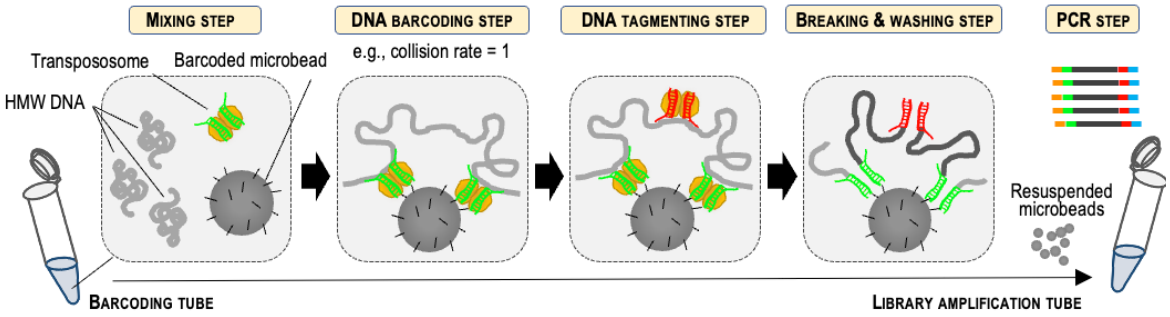

Figure 3. Overview of TELL-Seq library preparation workflow

Synthetic long reads represent a second category of DNA co-barcoding methods. The principle behind synthetic long reads and linked reads is similar: short reads tagged with the same barcode that informs about their shared origin from a long DNA molecule. The key difference is the extent of long-range information that can be captured into short reads. Synthetic long reads enable co-barcoding within only 6-10 kb, whereas linked reads enable co-barcoding up to 200-300kb (**Figure 4**, synthetic long reads). For best results, synthetic long reads require deeper sequencing. Some short-read sequencer manufacturers commercialize synthetic long-read methods.

Are there non-co-barcoding-based alternatives to capture long-range information?

Yes. Proximity-based ligation methods (such as Hi-C or Omni-C™) capture long-range information before DNA extraction, as opposed to co-barcoding methods, which capture long-range information after DNA extraction. Proximity-based ligation methods rely on chromatin-mediated contacts to ligate otherwise distal regions along the genome and require using fixatives to stabilize the contacts. Generally, proximity-based ligation data is sparse and can be vulnerable to genetic variation that alter chromatin organization (**Figure 4**, proximity-based ligation). It is preferred as a scaffolding method in combination with other methods rather than as a standalone approach.

Contact information: technicalsupport@universalsequencing.com

What are (continuous) long reads?

Continuous long reads developed by Pacific Biosciences® (PacBio) and Oxford Nanopore Technology® (ONT) are, on average, 20 and 100 kb long sequencing reads, respectively (**Figure 4**, continuous long reads). Despite their undoubted value, long-read technologies remain less cost-effective and show much lower accuracy when read-length is greater than 20 kb, and lower genotyping fidelity and throughput than more popular, short reads, which are used by linked-read methods (**Figure 4**, linked reads). For these and other reasons, capturing long-range information with short reads remains a go-to option for many applications, typically through a DNA co-barcoding method.

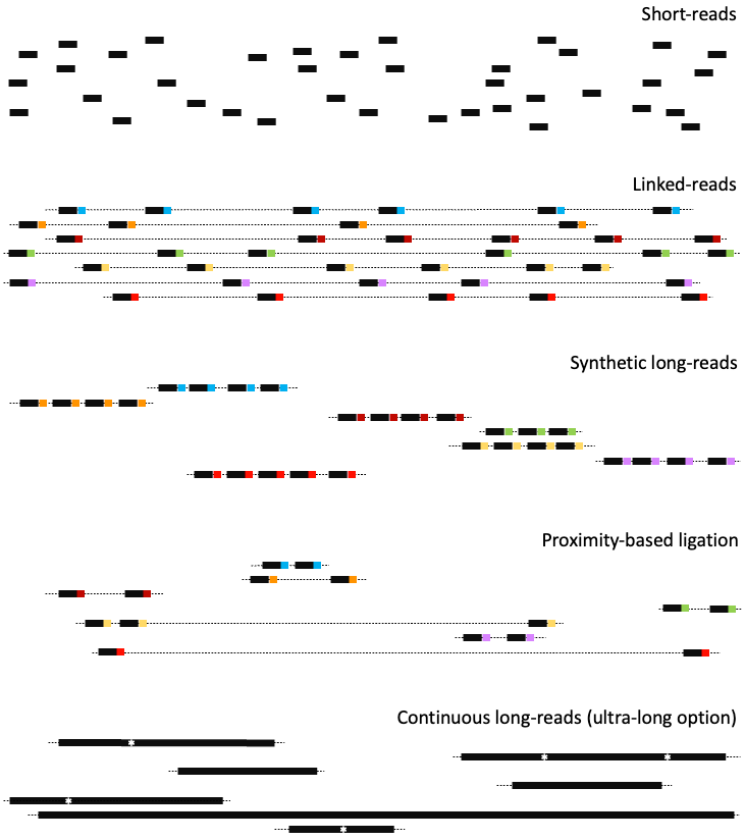

Figure 4. Types of NGS reads

Popular linked-read applications

- *De novo* assembly (reference-free) of single genomes (animal, plant, invertebrate, or microbial) and metagenomes (microbial communities) to annotate new genomes
- *De novo* assembly of metagenomes (microbial communities and environmental/habitat samples) for taxonomy and abundance estimation purposes
- Reconstructing large structural variations (SVs) and long repeats in the human genome
- Haplotype-resolved analysis of whole human genomes
- Targeted haplotyping of relevant loci (e.g., *BRCA2*, *PMS2*, *PIK3CA*, *MLH1*, *MSH2*, MHC region, etc.), which is critical for interpreting genomic data in the clinic and for understanding the association of genetic variation and gene expression (eQTL or expression Quantitative Trait Loci analyses)
- Targeted phasing of rare variants (defined as allele frequency lower than 0.1%)
- Genome-wide meiotic recombination analyses in gametes
- Whole-genome risk prediction of common disease in preimplantation embryos

Finally, we note that researchers often combine different library preparation methods and sequencing technologies for best results (for example, a combination of linked-read, long-read, and Hi-C data). We provide references below that illustrate this comprehensive strategy.

Contact information: technicalsupport@universalsequencing.com

Additional information

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) PMID: 32540955

Mikhaylova et al. Targeted phasing of discrete 2-200 kilobase DNA fragments with a single-tube linked-read method and a short-read platform. *BioRxiv*, http://doi.org/ (2023)

Some examples of linked-read applications:

Kumar et al., Whole-genome risk prediction of common diseases in human preimplantation embryos. *Nature Med*. 28(3):513-516 (2022) PMID: 35314819

Sun et al., Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nature Commun*. 10(1):4310 (2019) PMID: 31541084

Chin et al., A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nature Commun*. 11(1):4794 (2020) PMID: 32963235

Xia et al., Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acis Res*. 46(4):e19 (2018) PMID: 29186506

Wohlers et al., An integrated personal and population-based Egyptian genome reference. *Nature Commun*. 11(1):4719 (2020) PMID: 32948767

Jarvis et al., Semi-automated assembly of high-quality diploid human reference genomes. *Nature*. 611:519-531 (2022) PMID: 36261518

Rhie et al., Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 592:737-746 (2021) PMID: 33911273

Yang et al., Evolutionary and biomedical insights from a marmoset diploid genome assembly. *Nature*. 594:227-233 (2021) PMID: 33910227

Zhang et al., A comprehensive investigation of metagenome assembly by linked-read sequencing. *Microbiome*. 8(1):156 (2020) PMID: 33176883

Zlitni et al. Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale *Genome Med*. 12(1):50 (2020) PMID: 32471482

Marks et al., Resolving the full spectrum of human genome variation using linked reads. *Genome Res*. 29(4):635-645 (2019) PMID: 30894395

Ling et al. Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature*. 557:424-428 (2018) PMID: 29743678

Seo et al., De novo assembly and phasing of a Korean human genome. *Nature*. 538:243-247 (2016) PMID: 27706134

Dreau et al., Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nature Commun*. 10(1):4309 (2019) PMID: 31541091

Zheng et al., Haplotyping germline and cancer genomes with high-throughout linked-reads sequencing. *Nature Biotech*. 34:303-311 (2016) PMID: 26829319

Contact information: technicalsupport@universalsequencing.com