Phased NGS Library Generation via Tethered Synaptic Complexes

Joseph C. Mellor, PhD (joe.mellor@seqwell.com), James H. Smith, PhD and Jack T. Leonard, PhD (jack.leonard@seqwell.com)

seqWell Inc., 376 Hale Street, Beverly MA 01915 USA

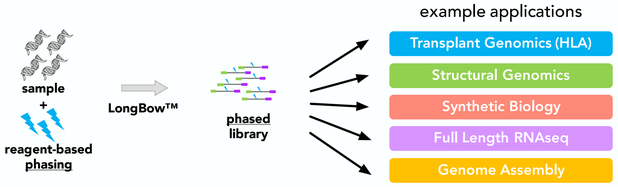
AGBT 2017
Poster #303

Introduction

The ability to efficiently resolve sequence phase information of long DNA fragments is of high potential impact for a large number of research and clinical sequencing applications. Example applications that stand to benefit from improved phased sequencing technology include phased genome sequencing and assembly, metagenomic sequencing, transcript isoform characterization and quantitation, and sequencing of long polymorphic loci.

Here, we report the development of a novel technology for obtaining phased sequence read information from long DNA fragments using an engineered molecular approach we term Tethered Synaptic Complexes™ (TSCs). TSCs are a single-tube, multivalent reagent for producing coordinated transposition events on multiple tandem sites that occupy proximal cis positions on individual target DNA fragments. By exploiting the biophysical proximity effect of transposition with TSCs, we show that sequenceable short-read (Illumina®) libraries can be efficiently and quickly derived while maintaining accurate phasing information of sequences present in samples containing known mixture of long fragments and known allelic mixtures. We further characterize the performance of TSCs in terms of the number and spacing of cis transposition contacts, with discussion of application and relevance to more complexed phasing and assembly problems such as complex heterozygotes, long polymorphic loci (e.g HLA) and transcript isoform characterization.

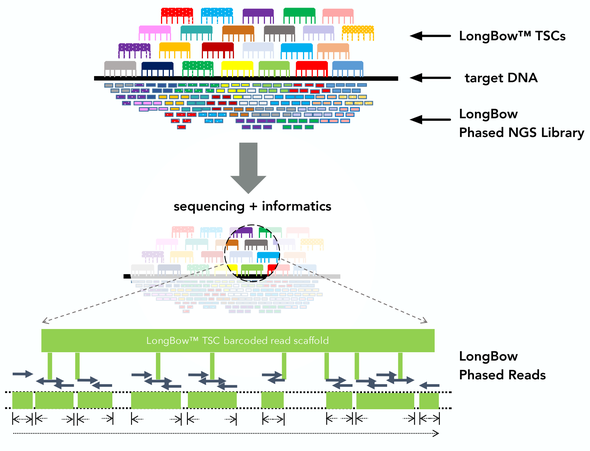
LongBow™: Library Prep 2.0



Simple, Scalable Long-Range Phasing Information

LongBow™ TSC480 library prep reagent contains 480 molecular scaffolds in a single tube, each individual scaffold molecule carrying hundreds of identical barcoded adapters (**Fig 1**). The LongBow TSC480 reagent inserts linked adapter molecules containing the same barcode into discrete regions of individual target DNA molecules.

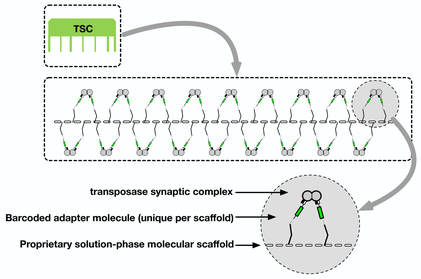
Fig 1. Schematic Illustration of LongBow™ TSC Phased Read Library Preparation and Sequencing,



LongBow™ TSC Technology

LongBow - An Engineered Molecular Reagent

Fig 2. LongBow™ TSC Functional Structure



A synaptic complex (SC) is a nucleoprotein structure of dimerized transposase proteins in which each transposase is bound to a transposase binding site on DNA in a reactive configuration. When multiple synaptic complexes are tethered to a single LongBow scaffold molecule, we refer to that as a tethered synaptic complex (TSC) (**Fig 2**). A key feature of LongBow TSC reagent is that it inserts many barcoded TSCs from a single scaffold molecule into a single target DNA molecule (multiple proximal cis transposition events). The overall length of the soluble scaffold, as well as the number and spacing between points for tethering SCs can be easily modified and precisely controlled when manufacturing the proprietary LongBow reagent.

Because of its reagent-based nature, the standard single-tube LongBow workflow follows a simple protocol with library prep only requiring two hours of hands-on time. After tagging the target DNA with LongBow TSC480 reagent, the library fragments are filled-in and amplified by PCR. Amplified libraries are then purified using magnetic beads and quantified by qPCR before loading the NextSeq® or MiSeq® DNA sequencing instruments. LongBow libraries can be sequenced using the standard Illumina paired read chemistry (e.g, 2 x 150bp, dual index). Standard Illumina dual indexing allows for sample-level indexing and multiplexing in addition to read phasing information provided by TSC indexing.

Single-Tube Phased Library Generation

Fig 3. LongBow™ TSC Library Prep Workflow Overview

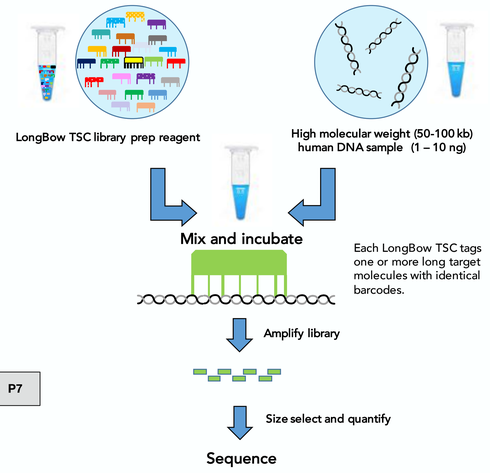


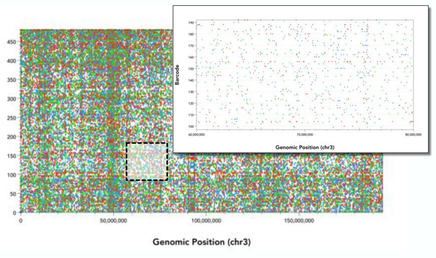
Fig 4. LongBow™ TSC Library Barcoding Scheme

Results

LongBow library prep and sequencing of NA12878

LongBow TSCs were used to prepare an Illumina-compatible library NA12878 DNA, and linked reads were identified by deconvolution after sequencing. Linked reads were identified by sampling large numbers of mapped reads and statistically deriving linked read sets based on their mapping quality and proximity to reads with the same barcode. The distance between linked transposition events on the NA12878 reference genome ranged from approximately 50 - 50,000 bp with an average distance of 20 kb.

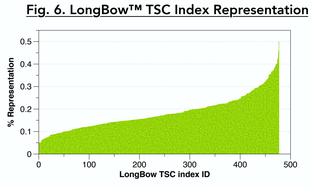
Fig. 5 LongBow™ TSC Data from human chromosome 3 (NA12878, Genome in a Bottle)



Uniformity of Multiplex Tethered Transposition

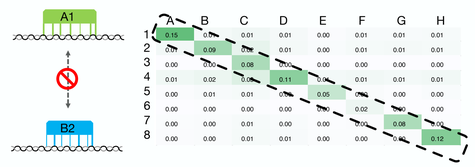
The ability of LongBow TSCs to produce resolvable phased read sets depends heavily on the relatively uniform formulation and efficient recovery of each of many different barcoded sequencing fragments from hundreds of uniquely barcoded TSCs. As shown in **Fig 6** (right), we measured the recovery of reads produced by the LongBow TSC480 reagent and assessed index representation, finding that 99% of the barcoded read counts fall within a 10-fold range.

Fig. 6. LongBow™ TSC Index Representation



Fidelity of Phased Library Generation

Fig. 7. LongBow™ TSC Phased Read Fidelity



We analyzed phased reads originating from a controlled mixture of eight (8) barcoded LongBow TSCs, each TSC having a specific combination of two unique adapters. Only specific pairs of barcodes (outlined box on diagonal) are expected to be observed. **Fig. 7** (left) shows that >95% of the signal from each TSC comes from proximal transposition events.

LongBow™ TSCs Produce Phased Reads from Long DNA Molecules

Sequence reads originating from a single LongBow TSC molecule produce high accuracy phasing events that can be hundreds to thousands of basepairs apart. Higher order phase blocks were iteratively assembled from phased read sets with different barcodes and shared haplotype content. **Fig 8** and **Fig 9** show the proximity of LongBow reads versus control (unphased) reads, and the distance of inferred phasing events on a human gDNA sample.

Fig. 8. Proximity of LongBow Reads vs Control Reads

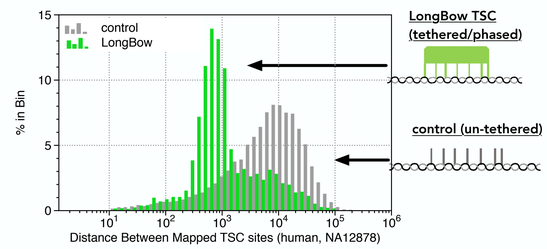
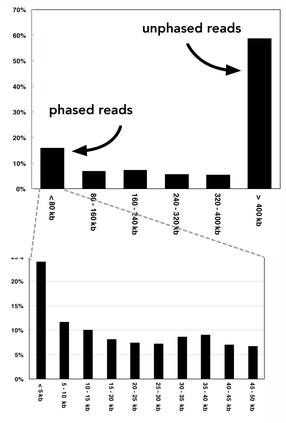


Fig. 9. Distance between LongBow TSC sites on human chromosome 21 (NA12878)



Conclusions

A number of different approaches to phased sequencing exist, including direct long raw sequencing read technologies (e.g. PacBio® and MinION®), haploid-level dilution and parallel library preparation (e.g., Moleculo®, Complete Genomics LFR and CPT-seq), microfluidic droplet-based technologies (e.g., 10X®), as well as more traditional fosmid and BAC clone sequencing. Many current technologies for phased sequence generation, while accurate, suffer from relatively high capital and consumable cost and/or difficult workflows, all of which are barriers to wider adoption and scaling.

- **LongBow™ is a breakthrough single-tube library prep platform that generates long linked reads on short read DNA sequencing instruments.**
- **Easy, single tube workflow for library prep (no boxes required!)**
- **Fully-compatible with on-board Illumina sequencing reagents and libraries**

Given the single tube format and solubility of LongBow reagent, we anticipate that the workflow will be easily portable to automated liquid handling instruments and microfluidic devices alike. A clear advantage of LongBow is the efficiency of phasing. Since long linked reads from a single-barcoded scaffold can form read contigs that extend over 1,000 basepairs, the depth of coverage required for phasing is less than those platforms that assemble much larger numbers of short reads into phased contigs.

We anticipate that the resolving power of 480 LongBow barcoded TSCs will be adequate for phasing most diploid genomes. The LongBow platform extends the read length of Illumina sequencers, consequently, LongBow will improve any sequencing application that benefits from longer phased reads. Phasing diploid genomes is the first LongBow application that seqWell is investigating with research partners, but the LongBow reagent platform also holds promise for de novo sequencing, metagenomics, genotyping, and diagnostics research applications.

Availability

seqWell's Early Access Program for users to test LongBow™ TSC technology will initiate in late Q1. For more information please send email to info@seqwell.com.