UNIVERSAL
SEQUENCING
innovation for all

TELL-Seq^{TM} Data Analysis Roadmap User Guide

Table of Contents

1. Introduction

This document describes data analysis roadmaps for DNA sequencing data generated by the TELL-Seq WGS Library Prep Kit.

The TELL-Seq WGS library prep kit uses an innovative Transposase Enzyme Linked Long-read Sequencing (TELL-Seq$^{\text{TM}}$) technology to prepare a paired-end library to generate barcode linked reads from an Illumina sequencing system. Linked reads can then be processed and analyzed by for genome wide variant calling, haplotype phasing, structural variation detection, metagenomic studies and *de novo* sequencing assembly, etc. TELL-Seq sequencing data can be processed by analytic software pipelines, such as UST's Tellysis accompanied by the library prep kit itself, or, by other commonly available linked read or read cloud analysis pipelines, such as, Longranger, Supernova, Loupe, etc. Figure 1 illustrates analysis options for TELL-Seq linked reads.

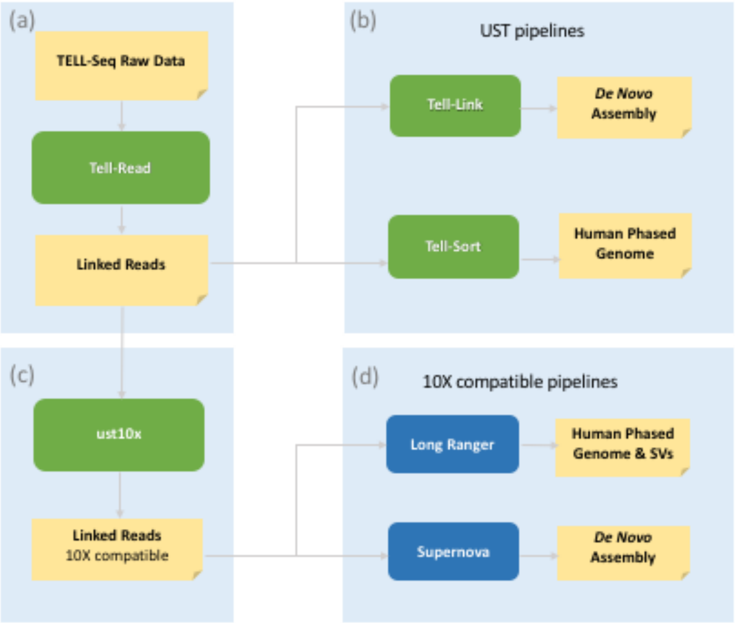

Figure 1: TELL-Seq Linked Reads analysis worflows. **(a)** TELL-Seq raw data generated by Illumina sequencers is processed by Tell-Read pipeline to produce linkd read data in fastq format. **(b)** The fastq files genreated are the input files for both UST phasing pipeline Tell-Sort and UST assembly pipeline Tell-Link. **(c)** Alternatively, the TELL-Seq linked reads can be converted to 10X compatible format to be processed by pipelines that may be already used by users' current workflows. **(d)** 10X compatible pipelines

2. Processing TELL-Seq Linked Reads by UST Pipelines

UST pipeline software comes in the form of three main pipelines.

- **Tell-Read**

  a set of pipeline processes that takes as input the sequencing output from an NGS sequencing instrument and generates linked-read FASTQ data, as well as QC reports.

- **Tell-Sort**

  a set of pipeline processes that takes as input the linked-read data from Tellread result and performs variant calling, phasing and SV.

- **Tell-Link**

  de novo assembly pipeline processes that builds barcode-aware assembly graph, assembles contigs and performs scaffolding.

For detailed descriptions and usage of UST pipelines, please refer to TELL-Seq Software User Guide documents.

3. Processing TELL-Seq Linked Reads by 10X Compatible Pipelines

Users currently running linked read analysis workflows using 10X compatible software pipelines can also process TELL-Seq linked reads after converting data to 10X data format.

This section describes detailed steps and tools to convert TELL-Seq data to 10X data format. Examples will be given on how to run Longranger and Supernova on TELL-Seq linked reads.

3.1. Data Conversion

TELL-Seq and 10X have different barcode designs. TELL-Seq barcode is 18 bases long while 10X barcode is 16 bases long. The TELL-Seq barcode is also sequenced as part of the I1 index and must be prepended to R1 reads for Longranger and Supernova analyses.

The Linux binary executable `ust10x` is the tool for the data conversion, provided as a separate downloadable file.

The command line to convert TELL-Seq sequencing reads to 10X format takes following form,

```
$ ust10x \
     -sz <size> \
     -i1 <path/to/i1.fastq.gz> \
     -r1 <path/to/r1.fastq.gz> \
     -r2 <path/to/r2.fastq.gz> \
     -wl <path/to/whitelist>
```

The command line options are explained in the table below.

-sz	parameter to determine maximum barcodes to retain. See below for more details.
-i1	gzipped index 1 fastq file
-r1	gzipped read 1 fastq file
-r2	gzipped read 2 fastq file
-wl	barcode whitelist. See Section “Barcode Whitelist” for details.

This should generate two unzipped fastq files R1_sl.fastq.gz.4tenx.fasq and R2_sl.fastq.gz.4tenx.fastq in the local directory.

Compress these two files,

```
$ gzip *_sl.fastq.gz.4tenx.fastq
```

Both Longranger and Supernova are expecting files to conform with certain file name pattern. So the file names will need to be changed to following format.

```
$ mv R1_sl.fastq.gz.4tenx.fastq.gz <runname>_S1_L001_R1_001.fastq.gz
$ mv R2_sl.fastq.gz.4tenx.fastq.gz <runname>_S1_L001_R2_001.fastq.gz
```

3.2. Barcode Whitelist

TELL-Seq barcoding capacity is over 2 billion which is way over the number of barcodes routinely used. 10X platform provides a standard whitelist of 4 million barcodes for sequencing runs to choose from. Longranger and Supernova also use this whitelist as a reference for its data analysis pipeline process.

Included in this conversion package is a whitelist containing more than 24 million barcode sequences. For routine TELL-Seq runs this is more than enough currently. To convert TELL-Seq data to 10X format to run Longranger and/or Supernova, replace existing 10X barcode whitelist file `4M-with-alts-february-2016.txt` with the one supplied by this package.

3.3. Downside Barcodes If Necessary

Currently 10X's genome visualization tool Loupe seems to have a limitation on the number of unique barcodes it can support (16M), therefore, if the actual number of unique barcodes used in the TELL-Seq runs exceeds this maximum unique barcode number, the conversion tool allows user options to remove the barcodes that have fewer associated reads.

For example, if the user wishes to reduce the total number of barcodes to 16M, the command line option is `-sz 16000000.` The conversion tool will first remove all barcodes with only single read. If the remaining number of unique barcodes still exceeds the limit, it will then remove barcodes with 2 reads. This procedure will continue until the remaining number of unique barcodes is fewer than the limit. If `-sz` option is not set, the default behavior is all barcodes will be retained.

3.4. Run Longranger for Phasing and SV Detections Applications

To run Longranger on converted TELL-Seq data,

1. Locate 10X original barcode whitelist in the Longranger installation path, `longranger-2.2.2/longranger-cs/2.2.2/tenkit/lib/python/tenkit/barcodes/4M-with-alts-february-2016.txt`

2. Back up the original whitelist and replace it with the new whitelist

```
$ cd longranger-2.2.2/longranger-cs/2.2.2/tenkit/lib/python/tenkit/barcodes
$ mv 4M-with-alts-february-2016.txt 4M-with-alts-february-2016.txt.bk
$ cp /path/to/4M-with-alts-february-2016.txt .
```

3. Run Longranger as usual, for example,

```
$ longranger-2.2.2/longranger wgs \
                       --id=mytest \
                       --fastq=/path/to/fastq/directory \
                       --reference=/path/to/reference/directory \
                       --sex=m \
                       --vcmode=freebayes
```

3.5. Run Supernova for Genome Assembly application

To run Supernova on converted TELL-Seq data,

1. Locate 10X original barcode whitelist in the Supernova installation path, `supernova-2.1.1/supernova-cs/2.1.1/tenkit/lib/python/tenkit/barcodes/4M-with-alts-february-2016.txt`

2. Back up the original whitelist and replace it with the new whitelist

```
$ cd supernova-2.1.1/supernova-cs/2.1.1/tenkit/lib/python/tenkit/barcodes
$ mv 4M-with-alts-february-2016.txt 4M-with-alts-february-2016.txt.bk
$ cp /path/to/4M-with-alts-february-2016.txt .
```

3. Run Supernova as usual, for example,

```
$ supernova-2.1.1/supernova run \
                     --id=myrun \
                     --fastq=/path/to/fastq/directory \
                     --maxreads='all' \
                     --accept-extreme-coverage
```