TLX File Format Specification

The Output File of the Translocation Pipeline

Robin M. Meyers
Laboratory of Fredrick W. Alt,
Boston Children's Hopsital, Harvard Medical School
robin.meyers@childrens.harvard.edu

December 9, 2013

Abstract

The format of the TLX file, which is the primary output file of the Translocation Pipeline, is defined and discussed in depth. Suggestions are also given for interpretting the data included and further manipulating it.

1 Introduction

The TLX file is the primary output file of the Translocation Pipeline. It consists of TLX records, which is the fundamental data type of the pipeline's output. Each TLX record represents a translocation junction, which is defined, for our purposes, as two adjacent alignments on a single query sequence that each map to distinct sites on the reference genome. The TLX records for each read are computed in the pipeline after the application of the optimal query coverage (OQC) algorithm and before any of the filters are applied to that read.

1.1 File Naming Convention

There are in fact a few TLX files that are output as part of the Translocation Pipeline. They will all be named like `LIBNAME[_DESC].tlx` where LIBNAME is the name of the library and DESC is an optional descriptor about the tlx records in that file. Currently, the tlx records that passed all filters during the pipeline has no descriptor and tlx records that were filtered will exist in either the `LIBNAME_unjoined.tlx` or `LIBRARYNAME_filtered.tlx` file. See the current Translocation Pipeline Documentation for more details.

1.2 TLX Records

The TLX record represents a single translocation junction. It is composed of two adjacent alignments on the query sequence read. Every query may have zero, one, or more TLX records associated with it. The two alignments will be termed the bait alignment and the capture alignment. Assuming that the library was sequenced from the forward primer, the bait alignment occurs on the 5' end of the junction. That is, it occurs first when reading from left to right. The capture alignment is at the 3' end of the junction, occuring second in the sequence. The bait alignment will typically also be the alignment to the breaksite. However, on a query with multiple junctions, the capture alignment of the first TLX record will also be the bait alignment in the second TLX record.

1.3 Single-End and Paired-End Reads

The processing of paired-end (PE) reads requires some extra computational work compared to single-end (SE) reads and it is important to understand this difference when drawing interpretations from TLX records. The query coordinates of TLX records from an SE library refer to the position of the alignment on each read. However, for PE libraries, this is not necesarily the case. The pipeline attempts to merge PE reads during the execution of the OQC algorithm and construction of TLX records. In order to be merged, the optimal coverage set (OCS) must include at least one concordant alignment that links the two ends. See the Bowtie2 manual for the definition of a concordant alignment. The pipeline will then stitch together the two ends if the concordant alignment overlaps on the reference and will fill in the likely sequence if a gap exists on the reference between the two ends of the concordant alignment. The concordant alignment can be in either a bait alignment or capture alignment; it makes no difference. The end result of this merge is that the query coordinates reported in the TLX record are not actually the coordinates of the original PE query sequences, but instead refer to the reconstructed sequence of the entire DNA fragment reported in the
texttttSeq field. For queries that were not merged, the query coordinates simply refer to the first read of the pair.

2 The TLX File

2.1 Format

The TLX file is a tab-delimited plain text file that includes a header row, followed by a row for each TLX record. Each row should be separated by the UNIX newline character. However, researchers will most frequently edit the TLX file using Microsoft Excel which, upon saving regular text files, writes the carriage-return character between rows in place of the newline character. *Therefore, any downstream processing modules must first replace these carriage-return characters, usually denoted* `\r`*, with the newline character,* `\n`.

2.2 Header

The header row consists of the names of the columns each separated by a tab. There is no strict ordering of the columns and the column names are not case-sensitive. *Future development of the pipeline, and downstream processing modules, must allow for flexibility with respect to these features.* There is however a recommended ordering of the columns, see below, and a recommended capitalization convention - namely that the first letter of any string of alphanumeric characters is uppercase and the rest are lowercase. For example, "`B_Qstart`" would be the recommended capitalization for that column name, since an underscore is not considered alphanumeric.

2.3 Columns

- **Qname** - Query Name, the name of the query sequence
- **Rname** - Reference Name, the reference sequence name of the capture alignment
- **Junction** - the coordinate of the junction end of the capture alignment
- **Strand** - the strandedness of the capture alignment
- **B_Rname** - Bait Reference Name, the reference sequence name of the bait alignment
- **B_Rstart** - Bait Reference Start, the reference start coordinate of the bait alignment
- **B_Rend** - Bait Reference End, the reference end coordinate of the bait alignment
- **B_Strand** - Bait Strand, the strandedness of the bait alignment
- **B_Qstart** - Bait Query Start, the query start coordinate of the bait alignment
- **B_Qend** - Bait Query End, the query end coordinate of the bait alignment
- **Rstart** - Reference Start, the reference start coordinate of the capture alignment
- **Rend** - Reference End, the reference end coordinate of the capture alignment
- **Qstart** - Query Start, the query start coordinate of the capture alignment
- **Qend** - Query End, the query end coordinate of the capture alignment
- **Qlen** - Query Length, the computed length of the query sequence
- **J_Seq** - Junction Sequence, a short sequence from the reference genome encompassing the junction end of the capture alignment
- **Seq** - Sequence, the computed query sequence
- **Flags** - extra information about the junction; e.g. filter, gene name, etc.

3 Column Descriptions

3.1 The Primary Fields

These are the most important fields, and therefore it is recommended that they occur first in the TLX File. They contain the vital information about the capture alignment.

Qname

The Query Name is a unique character string assigned to each read by the sequencer. For paired end reads, both ends have the same `Qname`. Although there can only be one junction per `Qname` in the final TLX File, there may be subsequent junctions with the same `Qname` in the filtered TLX file which may be of some interest to the researcher.

Rname

The Reference Name is the name of the reference sequence that the capture alignment was mapped to. Generally speaking, this will be the chromosome of the translocation.

Junction

The Junction is the reference coordinate of the capture alignment that is at the junction end of the alignment, that is, adjacent to the bait alignment. For capture alignments that map to the "`+`" strand, `Junction` will be the same as `Rstart`, and for ones that map to the "`-`" strand, `Junction` will be the same as `Rend`.

Strand

The Strand is the orientation of the capture alignment. This column will be reported as `1` for an alignment to the "`+`" strand of the reference and `-1` for the "`-`" strand. Any other values in this field are not allowed.

3.2 Bait Alignment

Researchers have become very interested in what is now termed the bait alignment, and the distributions across their libraries. The bait alignment is generally the alignment of the query to the breaksite, starting at the forward primer and progressing towards the cut site, but strictly speaking it is the lefthand alignment of the junction. This means that for sequential junctions, the bait alignment can map anywhere in the reference.

B_Rname

B_Rstart

B_Rend

B_Strand

B_Qstart

B_Qend

3.3 Capture Alignment

The capture alignment is the true translocation alignment, the "many" in the translocation sequencing one-to-many strategy. Much about this alignment is already reported as part of the primary fields, which leaves only the start and end coordinates left to report.

Rstart

The Reference Start

Rend

The Reference End

Qstart

The Query Start

Qend

The Query End

3.4 Other Fields

Qlen

The Query Length

J_Seq

The Junction Sequence

Seq

The Sequence

4 Glossary of Terms

alignment the mapping of a query sequence to a reference sequence. Key features include the name, coordinates, and strand of the reference sequence that the query has mapped to, as well as the portion of the query sequence that has aligned.

bait the first of the two alignments that make up a junction.

breaksite same as bait alignment

junction what a junction is

capture

query

reference

split junction

TLX record tlx record represents a junction