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 4, 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 algorithm and before any of the filters are applied to that read.

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.

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.

2 The TLX File

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`*.*

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.

Columns

The Primary Fields

- **Qname** - Query name, the character string assigned to this query sequence by the sequencer
- **Rname** - Reference name, the name of the reference sequence which the capture alignment mapped to
- **Junction** - the coordinate of the junction end of the capture alignment
- **Strand** - the strandedness of the capture alignment

Bait Alignment

- **B_Rname** -
- **B_Rstart** -
- **B_Rend** -
- **B_Strand** -
- **B_Qstart** -
- **B_Qend** -

Capture Alignment

- **Rstart** -
- **Rend** -
- **Qstart** -
- **Qend** -

Note that Rname and Strand were included as part of the primary fields.

The Rest

- **Qlen** -
- **J_Seq** -
- **Seq** -

3 Column Descriptions

Qname

Rname

Junction

Strand

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