Metadata File Specification

Creating the Annotation File for Libraries in the Translocation Pipeline

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

January 9, 2014

Abstract

Guidelines for creating a metadata file are presented. The metadata file is important for delivering details about the translocation library and how it should be analyzed to the pipeline software. A record in the metadata file is created for each library.

1 Introduction

The metadata file was designed to deliver important details about the translocation library to the pipeline software so that it can be analyzed correctly. As little extraneous information is included as possible, while still accounting for the diversity of library types that the lab generates. Entries are created for each library and teh following guidelines should be followed as strictly possible to ensure consistency between libraries.

2 The Metadata File

2.1 Format

The metadata file is a tab-delimited plain text file that includes a header row, followed by a row for each library. Each row should be separated by the UNIX newline character. However, researchers will most frequently create/edit the metadata 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 processing 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.*

2.3 Columns

- **Library** - name of the library, typically the researcher's initials followed by a three digit, zero-padded number (*e.g.* `AB001`)
- **Sequencing** - name of the sequencing run, typically "Alt" followed by a three digit, zero-padded number(*e.g.* `Alt001`)
- **Researcher** - the researcher's name
- **Assembly** - the genome build used, at time of writing must be either *mm9* or *hg19*
- **Chr** - name of the breaksite chromosome
- **Start** - start coordinate of the breaksite
- **End** - end coordinate of the breaksite
- **Strand** - strandedness of the breaksite, *i.e.* the primer orientation
- **Breakseq** - breaksite cassette sequence, if applicable
- **Breaksite** - specific breaksite coordinate with respect to `Breakseq`
- **MID** - multiplex identifier sequence
- **Primer** - primer sequence
- **Adapter** - adapter sequence
- **Cutter** - frequent cutter sequence, if applicable
- **Description** - a description of the library

3 Annotating the Breaksite

There are two major types of translocation libraries that the Alt Lab currently produces, characterized by either an endogenous breaksite or non-endogenous breaksite (*i.e.* an I-SceI cassette). The pipeline was engineered to handle these two cases as similarly as possible, and is described fully in the pipeline documentation. The meaning of some columns in the metadata file is different for each library type.

3.1 Endogenous Breaksite Libraries

For endogenous breaksite libraries, the breaksite is fully characterized by the priming site and cutting site. Thus, the the metadata file is used to annotate these two ends of the breaksite and the region in between, using only the columns `Chr`, `Start`, `End`, and `Strand`. The `Chr` is the name of the chromosome that contains the breaksite (*e.g.* `chr15`). The `Strand` is the orientation in which the primer aligns to this chromosome, either "`+`" or "`-`". We will annotate the region between the priming end of the breaksite and the cutting end of the breaksite, and will follow the convention of requiring the end coordinate greater than the start coordinate regardless of orientation. Therefore, for libraries in the `+` orientation, the `Start` column denotes the primer end and the `End` coordinate denotes the cutting end, and vice versa for `-` orientation libraries. The specific coordinates used will be the first coordinate of the start of the region and the first coordinate after the end of the region. This means, for `+` orientation libraries, the `Start` will be the first coordinate of the primer and the `End` will be the first coordinate after the cut site. For `-` orientation libraries, the `Start` will be the first coordinate of the cutting site and the `End` will be the first coordinate after the priming site.

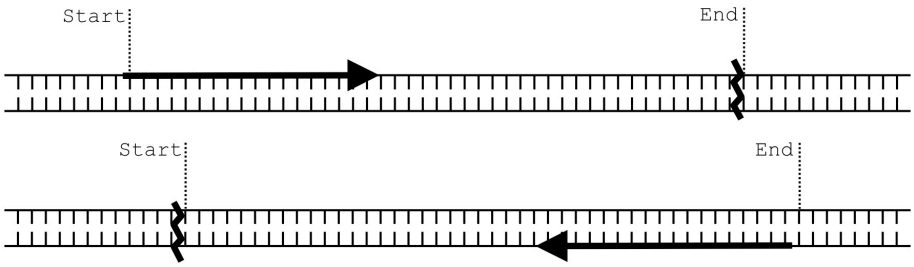

Figure 1. Endogenous breaksite strategies are shown in the `+` strand orientation **(Top)** and `-` orientation **(Bottom)**. `Start` and `End` coordinates are marked. Diagrams are oriented so that chromosomal coordinates increase from left to right.

3.2 Non-Endogenous Breaksite Libraries (Breaksite Cassette)

The annotation is marginally more complicated for non-endogenous breaksite libraries than for endogenous breaksites because the pipeline needs a little extra information about where the breaksite cassette has been inserted into the genome as well as what that sequence is. The `Breakseq` column contains the sequence of the entire cassette written from priming end to cutting end. The `Chr`, `Start`, `End`, and `Strand` columns together fully annotate the cassette position and direction. The `Strand` is the orientation of the cassette, either `+` or

-, and is determined by the orientation of the primer. *Note: regardless of the orientation, the breaksite cassette sequence must be given from priming end to cutting end.* Again, we will follow the convention of requiring the end coordinate to be greater than the start coordinate. To determine the `Start` column, find the last basepair coordinate of endogenous locus before the start of the cassette and add 1 to this number. The `End` column denotes the first basepair coordinate of endogenous locus after the end of the cassette.

While in endogenous breaksite libraries the priming site is given explicitly, this is not so in non-endogenous breaksite libraries because the priming site can be determined using the primer sequence and the breaksite sequence. However, that still leaves one last piece of information to be accounted for in the non-endogenous breaksite libraries and that is the specific location of the cutting site. This coordinate is with respect to the sequence entered in `Breakseq` and is located in the `Breaksite` column.

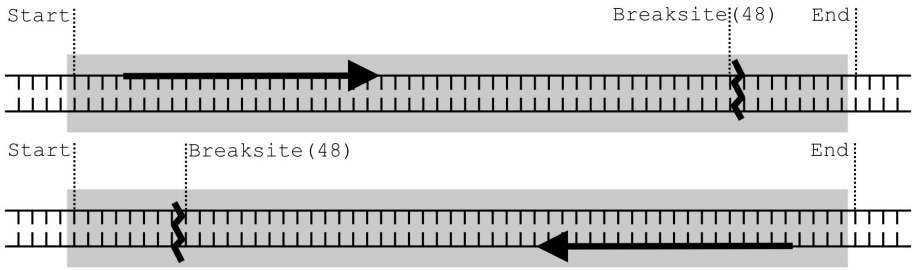

Figure 2. Non-endogenous breaksite strategies are shown in the `+` strand orientation **(Top)** and `-` orientation **(Bottom)**. `Start` and `End` coordinates are marked. Diagrams are oriented so that chromosomal coordinates increase from left to right. The shaded area represents the non-endogenous cassette.