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

December 17, 2013

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

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
- **Breaksite** - breaksite cassette sequence, if applicable
- **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 the columns of the metadata file is different for each library type.

3.1 Endogenous Breaksite Libraries

The `Chr` is the name of the chromosome that contains the breaksite (*e.g.* `chr15`). The `Strand` is the orientation in which your primer aligns to this chromosome, either "`+`" or "`-`". We will annotate the area between the first base pair of the primer to the cut site, and will follow the convention of keeping the end coordinate greater than the start coordinate regardless of orientation. For endogenous breaksite libraries, the cut site end does not need to be exact, however the primer end does. For libraries in the `+` orientation the `Start` column denotes the primer end and for `-` orientation libraries, it is the `End` column.

3.2 Non-endogenous Breaksite Libraries