illumina®

Sequencing Sample Sheet Format Specifications

Formatting rules and software-based differences in formatting for Illumina Sample Sheets.

Introduction

The Sample Sheet is a file format used by Illumina for storing biological sample information and metadata associated with a given experiment. This file is used widely across the Illumina informatics ecosystem as an input to many pieces of software, such as bcl2fastq and BaseSpace® Sequence Hub. The Sample Sheet for sequence data uses American Standard Code for Information Interchange (ASCII) character encoding. This plain-text, comma-delimited format supports multiple sections with metadata describing experimental setup, demultiplexing settings, or analysis options. This document describes the Sample Sheet format and delineates how the file is used by Illumina software.

Formatting

Character Encoding

Valid Sample Sheet files are encoded in unicode transformation format, 8 bit (UTF-8) without byte order mark (BOM). A specific list of characters is permitted in the file (Table 1). Some sections of the Sample Sheet are more stringent than others, only permitting a subset of the legal characters listed.

Table 1: Permitted Characters in Sample Sheet File

ASCII Decimal Code	Description
10	LF (NL line feed, new line)
13	CR (carriage return)
32	Space
33-46	Special Characters ! " # $ % & ' ( ) * + , - . /
48-57	Numbers 0 1 2 3 4 5 6 7 8 9
58-64	Special Characters : ; < = > ? @
65-90	Upper-Case Alphabet A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
91-96	Special Characters [\ ] ^ _ `
97-122	Lower-Case Alphabet a b c d e f g h i j k l m n o p q r s t u v w x y z
123-126	Special Characters { | } ~

Delimiters

Fields

Commas (ASCII code 44) are used as a delimiter to separate adjacent fields on the same line. If a field contains a comma, then the field needs to be wrapped in double quotes (ASCII code 34).

Example:

Library type

is represented in the Sample Sheet as:

"Library type"

If a field contains double quotes, then the entire field needs to be wrapped in double quotes and the embedded double quotes need to be escaped using a pair of adjacent double quotes.

Example:

This is in "quotes", as well as commas

is represented in the Sample Sheet as:

"This is in ""quotes"", as well as commas"

There is no global minimum or maximum number of comma delimiters required to for each line. Use as many comma delimiters as necessary to separate the number of fields required per line in a given section. For instance, if a section of the sample sheet requires exactly two fields per line, the minimum number of comma delimiters for that line is one. However, the end of the line can be padded with as many commas as desired, which are ignored.

Lines

Windows-style and Linux-style line endings are both permitted in the Sample Sheet format. It is recommended that the line ending style is consistent throughout the entire Sample Sheet.

Using Windows-style line endings, lines are separated using a carriage return (ASCII code 13), followed by a line feed (ASCII code 10). This is often represented in programming languages as "\r\n".

Using Linux-style line endings, lines are separated using only a line feed. This is often represented in programming languages as "\n".

Empty lines or lines that consist entirely of commas and/or whitespace characters are valid, but ignored.

Sections

The Sample Sheet format is divided into multiple sections, which are denoted by a line starting with a section label. The section label starts with an open square bracket (ASCII code 91) containing a string of text for the section name, and ending with a closing square bracket (ASCII code 93).

Example "Header" section label:

```
[Header]
```

Section labels are case-sensitive. No characters after the section label are permitted other than commas and a line ending. All lines following a section label are considered part of a section until the next section label is encountered.

Standard Sections

The following are standard Sample Sheet sections used by Illumina software. Sections may be in any order, except that the sample sheet must begin with the Header section and end with the Data section. Depending on the software using the Sample Sheet file, some sections may be optional. However, the Header and Data sections are always required.

Header

The Header section is required, and must be located on the first line of the Sample Sheet file. The Header contains informational fields describing the context around which a sequencing run or analysis was performed (eg, date, workflow, library prep kit, chemistry, etc.).

Header records are represented as a series of key-value pairs. As such, each line requires exactly two fields. The first field in each line is the "key," which names the piece of metadata being recorded. Each key in the Header section must be unique. The second field in each line is the "value," which is the actual piece of metadata being recorded. Values do not necessarily need to be unique.

Example of a legal "Header" section containing records describing "Date" and "Investigator":

```
[Header]
Date,2007-01-26
Workflow,GenerateFASTQ
Investigator,John Smith
```

Example of a legal "Header" section (with padded commas):

```
[Header],,,,,,
Date,2007-01-26,,,,,
Workflow,GenerateFASTQ,,,,,
Investigator,John Smith,,,,,
```

Settings

Settings is an optional section used to control various parameters for some Illumina software, such as bcl2fastq and BaseSpace Sequence Hub. If no Settings section is present in the Sample Sheet, the software will assume the default values for all options. Some software will require different options available to configure in the Settings Section. For more detailed information on the available options, please refer to the respective user guides.

Settings records are represented as a series of key-value pairs. As such, each line requires exactly two fields. The first field in each line is the "key," which is the name of the option being configured. Each option in the Settings section can only be used once. The second field in each line is the "value," which is the actual piece of metadata being recorded. Values do not necessarily need to be unique.

Example bcl2fastq settings:

```
[Settings]
Adapter,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGT
```

Reads

The Reads section describes the number of sequencing cycles used for read 1 and read 2. Reads records are represented as a single positive integer per read. As such, each line requires exactly one field. For single-end sequencing, there should be exactly one record. For paired-end sequencing, there should be exactly two records with the first representing read 1 and the second representing read 2. Index reads are not included in this section.

Example for a single-end 1 × 151 sequencing run:

```
[Reads]
151
```

Example for a paired-end 2 × 151 sequencing run:

```
[Reads]
151
151
```

The Reads section is only a required section when using a sample sheet file to set up a sequencing run through the MiSeq® Control Software.

Manifests

Manifests is a section required by some secondary analysis workflows for targeted resequencing to specify a manifest file that contains the targets of interest.

Settings records are represented as a series of key-filename pairs. As such, each line requires exactly two fields. The first field in each line is the "key" which is an arbitrary string (usually a single letter). The second field in each line is the "filename," which is either the name or full path of the Manifest file including the file extension. The Manifest file must reside in the same run folder as the SampleSheet.csv file if a full path to the Manifest is not provided. Each key and filename in the Manifests section should be unique.

The Data section should also contain a Manifest column to assign a Manifest to each sample. Each sample can be assigned exactly one Manifest, but different samples may be assigned different Manifests.

Example with two TruSeq® Amplicon manifest files:

[Manifests]
A,TruSeqAmpliconManifest-1.txt
B,TruSeqAmpliconManifest-2.txt

[Data]
Sample_ID,Sample_Name,I7_Index_ID,index,I5_Index_ID,index2,Manifest,GenomeFolder
A10001,Sample_A,A701,ATCACGAC,A501,TGAACCTT,A,Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFasta
A10002,Sample_B,A702,ACAGTGGT,A501,TGAACCTT,A,Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFasta
A10003,Sample_C,A703,CAGATCCA,A501,TGAACCTT,B,Bos_taurus\Ensembl\UMD3.1\Sequence\WholeGenomeFasta
A10004,Sample_D,A704,ACAAACGG,A501,TGAACCTT,B,Bos_taurus\Ensembl\UMD3.1\Sequence\WholeGenomeFasta

Data

The Data section is required and must be located at the end of the Sample Sheet file. The Data section is a table and captures all sample-specific metadata.

The line immediately following the Data section label is a header line. The Data header line lists the names of each of the table columns in this section, separated by commas. No specific ordering of the column names is required and they are not case-sensitive. Each column name may only appear in data header line once.

At a minimum, the one column that is universally required is Sample_ID, which provides a unique string identifier for each sample. However, different softwares will require different columns to be present in the Data section. For more detailed information on the columns required by different software platforms, refer to the respective user guides.

User-defined columns are allowed and will be ignored by Illumina software, provided they do not conflict with an existing column name that the software uses.

Immediately following the Data header line are the records for each sample. Each record spans exactly one line and must contain as many comma-separated fields as there are column names in the Data header line. Empty fields are permitted in cases where the column is optional. The field for the Sample_ID column has special character restrictions as only alphanumeric (ASCII codes 48-57, 65-90, and 97-122), dash (ASCII code 45), and underscore (ASCII code 95) are permitted. The Sample_ID length is limited to 100 characters maximum. Other fields may have their own character and length restrictions, depending on the software being used.

Example of typical Data section to be used with bcl2fastq:

[Data]
Sample_ID,Sample_Name,I7_Index_ID,index,I5_INdex_ID,index2
A10001,Sample_A,D701,AATACTCG,D501,TATAGCCT
A10002,Sample_B,D702,TCCGGAGA,D501,TATAGCCT
A10003,Sample_C,D703,CGCTCATT,D501,TATAGCCT
A10004,Sample_D,D704,GAGATTCC,D501,TATAGCCT

User-Defined Sections

Additional user-defined sections may be added to the Sample Sheet, as long as they adhere to the Sample Sheet formatting rules described above. They also cannot be named the same as sections defined by Illumina, and must be placed between the Header section and the Data section. These sections will be ignored by Illumina software, but can be used to store additional information for record keeping purposes or for processing in third-party software. Illumina may add to the standard Sample Sheet sections in the future without regard for naming conflicts with sections used by third-party software.

Example for TruSeq Amplicon:

[Header]
Date,2017-04-05
Workflow,Custom Amplicon
Application,TruSeq Amplicon
Assay,TruSeq Amplicon
Description,
Chemistry,Amplicon

[Manifests]
A,TruSeqAmpliconManifest-1.txt
B,TruSeqAmpliconManifest-2.txt

[Reads]
151
151

[Settings]
VariantFilterQualityCutoff,30
outputgenomevcf,FALSE

[Data]
Sample_ID,Sample_Name,I7_Index_ID,index,I5_Index_ID,index2,Manifest,GenomeFolder
A10001,Sample_A,A701,ATCACGAC,A501,TGAACCTT,A,Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFasta
A10002,Sample_B,A702,ACAGTGGT,A501,TGAACCTT,A,Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFasta
A10003,Sample_C,A703,CAGATCCA,A501,TGAACCTT,B,Bos_taurus\Ensembl\UMD3.1\Sequence\WholeGenomeFasta
A10004,Sample_D,A704,ACAAACGG,A501,TGAACCTT,B,Bos_taurus\Ensembl\UMD3.1\Sequence\WholeGenomeFasta

Example for bcl2fastq/FASTQ generation:

```
[Header]
Date,2017-04-05
Workflow,GenerateFASTQ
Application,FASTQ Only
Assay,TruSeq HT
Description,
Chemistry,Amplicon

[Reads]
151
151

[Settings]
Adapter,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGT

[Data]
Sample_ID,Sample_Name,I7_Index_ID,index,I5_Index_ID,index2
A10001,Sample_A,D701,ATTACTCG,D501,TATAGCCT
A10002,Sample_B,D702,TCCGGAGA,D501,TATAGCCT
A10003,Sample_C,D703,CGCTCATT,D501,TATAGCCT
A10004,Sample_D,D704,GAGATTCC,D501,TATAGCCT
```

Learn More

For more information about Illumina Experiment Manager support, visit: support.illumina.com/sequencing/sequencing_software/experiment_manager.html

For more information about bcl2fastq software support, visit: support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html

For troubleshooting, contact Illumina technical support at: techsupport@illumina.com

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

illumina®