Illumina SRF Barcode Submissions (Deprecated)

Created: October 27, 2010; Updated: April 12, 2011.

<table>
<thead>
<tr>
<th>Status</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Date</td>
<td>2010-10-27</td>
</tr>
<tr>
<td>Inactive Date</td>
<td>2011-04-12</td>
</tr>
<tr>
<td>Scope</td>
<td>INSDC SRA Illumina SRF barcodes</td>
</tr>
</tbody>
</table>

1 Overview

This application note describes the method for specifying metadata that will allow for loading of Illumina barcoded data delivered in SRF format. Future changes will deprecate this application note as well as simplify the submission process for Illumina barcoded data.

1.1 Related Documents

- SRA Barcoding Guide

2 Problem Statement

We have updated the current Illumina SRF loader to accept Illumina barcode submissions. We expect these submissions will be de-multiplexed such that each SRA Run is associated with a specific member name. One feature of this data is that the Illumina barcode sequence occurs between the two 'application' reads within each spot. Although this spot construction allows for the correct member name assignment, it can lead to an issue with dividing the spot into its individual reads. When an EXPECTED_BASECALL_TABLE is present and a member name assigned, our SRF loader will use the BASECALL value or values associated with that member name to partition the spot sequence from left to right using the criteria specified in the 'match_edge', 'max_mismatch', and 'min_match' attributes. A match found within the first 'application' read results in an incorrectly partitioned spot.

We have a solution on the way in the form of a schema update that will allow:

1. Specification of a BASE_COORD value at the same time as an EXPECTED_BASECALL_TABLE, and
2. Incorporation of a SPOT_DESCRIPTOR construct within the Run XML.

NLM Citation: SRA Application Notes [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Illumina SRF Barcode Submissions (Deprecated). 2010 Oct 27 [Updated 2011 Apr 12].
Support for this schema update within the SRF loader will not be immediate. Until this support is in place, the near-term solution is to replace the EXPECTED_BASECALL_TABLE with a BASE_COORD value and to transfer information encapsulated in the EXPECTED_BASECALL_TABLE to the EXPERIMENT_ATTRIBUTES area of the Experiment XML. Additionally, you must remove the ‘read_group_tag’ within the READ_LABEL for each MEMBER within the POOL construct. We expect integration of the new schema into the SRF loader will occur by early 2011.

3 Treatment

In order to apply our near-term solution the following MEMBER contained within a sample POOL:

```
<MEMBER accession="SRS066103" refcenter="BI" refname="35956.0" member_name="tagged_109_ACAGGTAT">
    <READ_LABEL read_group_tag="tagged_109">barcode</READ_LABEL>
</MEMBER>
```

Must have the ‘read_group_tag’ attribute removed:

```
<MEMBER accession="SRS066103" refcenter="BI" refname="35956.0" member_name="tagged_109_ACAGGTAT">
    <READ_LABEL>barcode</READ_LABEL>
</MEMBER>
```

Additionally, this EXPECTED_BASECALL_TABLE for the second read within a SPOT_DESCRIPTOR:

```
<EXPECTED_BASECALL_TABLE>
    <BASECALL match_edge="full" max_mismatch="1" min_match="7" read_group_tag="tagged_109">ACAGGTAT</BASECALL>
    <BASECALL match_edge="full" max_mismatch="1" min_match="7" read_group_tag="tagged_110">ACAGTTGA</BASECALL>
    <BASECALL match_edge="full" max_mismatch="1" min_match="7" read_group_tag="tagged_117">ACCAACTG</BASECALL>
</EXPECTED_BASECALL_TABLE>
```

Transforms into this BASECOORD specification:

```
<Base_COORD>69</Base_COORD>
```

An example of the EXPERIMENT_ATTRIBUTE to add in order to capture information contained in the EXPECTED_BASECALL_TABLE is:

```
<EXPERIMENT_ATTRIBUTE>
    <TAG>EXPECTED_BASECALL_TABLE</TAG>
    <VALUE>
        match_edge="full" max_mismatch="1" min_match="7"
        read_group_tag="tagged_109" member_name="tagged_109_ACAGGTAT" ACAGGTAT
        read_group_tag="tagged_110" member_name="tagged_110_ACAGTTGA" ACAGTTGA
    </VALUE>
</EXPERIMENT_ATTRIBUTE>
```
Note that we added 'member_name' to permit reinstating the original EXPECTED_BASECALL_TABLE and to maintain the association between this barcode sequence and the member name. Once you have made these modifications to the Experiment XML, the Run should load successfully. A caveat to this near-term solution is that a single BASE_COORD value may not be appropriate for all Runs associated with an Experiment. In this situation, you will need to create additional Experiments, each of which represents a different BASE_COORD value.

Please add a RUN_ATTRIBUTE, ExpectedBasecallTableInAttribute, with a value of yes, to the Run xml in order to record the presence of an EXPECTED_BASECALL_TABLE attribute.

When all Runs associated with an Experiment are loaded, you can reinstate the EXPECTED_BASECALL_TABLE in the Experiment XML with the qualification that a reload of one of these Runs could cause the underlying spots to partition incorrectly.

4 Example

A screen capture of the desired result for SRR057667 is below.

Note that the barcode sequence, 'acagttga', occurs in the first application read and would have resulted in an incorrectly partitioned spot if the EXPECTED_BASECALL_TABLE were present.
>gnl|SRA|SRR057667.1211.1 61GK7AAXX100301:5:1:15088:3374 Application Read (Forward)

GCCATTCATTAAGCCCAATAGCTAAAACCAGTTGCAACAGTGGAAATCGGAATAACATGATAAACGCA