Statistics Output Guide

Introduction

This document describes the file `sts.xml`, which is produced by the Sequel® System's primary analysis pipeline. This XML file contains the chip-level sample distribution and related summary statistics for specified per-ZMW metrics, with the ZMW sample set determined by filtering according to specific (metric-dependent) criteria.

- `sts.xml` packages statistics from a single movie acquisition, and is an independent summary of metrics.
- Movie metrics are summarized in terms of whole-chip sample distributions, including various sample distribution statistics.
- Metrics are generated during sequencing and aggregated into the `sts.xml` file.
- SMRT® Link's Run QC and the Data Management Data Set reports both use `sts.xml` for histogram calculations. `sts.xml` is also used as input by LIMS systems.

In most cases, summary metrics (such as the mean, median, and so on) are packaged in a representation of the distribution of a sample; for example, only productive ZMWs.

The representation of distributions of continuous metrics includes:

- A sample histogram.
- The total number of counts in the sample.
- A computation of the mean, the median and the standard deviation of the sample.
- Relevant information about how the histogram is formed: bin intervals, outliers, and so on.
- A "presentation-ready" description of the metric; for example, to label a plot.

The representation of distributions of discrete metrics (hole classifications) includes:

- A name for each outcome.
- The total number of counts in the sample.
- The counts for each outcome over the sample.

Sample sts.xml File

```
<?xml version="1.0" encoding="UTF-8"?>
<PipeStats Version="4.0.1" xmlns:ns="http://pacificbiosciences.com/PacBioBaseDataModel.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://pacificbiosciences.com/PacBioPipelineStats.xsd">
    <MovieName>/data/pa/m54088_171021_061044.baz</MovieName>
    <MovieLength>600</MovieLength>
    <NumSequencingZmws>1019235</NumSequencingZmws>
    <NumFailedSnrFilterZmws>172</NumFailedSnrFilterZmws>
    <NumFailedDmeZmws>0</NumFailedDmeZmws>
  - <TotalBaseFractionPerChannel Channel="A">
        <TotalBaseFractionValue>0.285157</TotalBaseFractionValue>
    </TotalBaseFractionPerChannel>
  - <TotalBaseFractionPerChannel Channel="C">
        <TotalBaseFractionValue>0.209694</TotalBaseFractionValue>
    </TotalBaseFractionPerChannel>
  - <TotalBaseFractionPerChannel Channel="G">
        <TotalBaseFractionValue>0.260973</TotalBaseFractionValue>
    </TotalBaseFractionPerChannel>
  - <TotalBaseFractionPerChannel Channel="T">
        <TotalBaseFractionValue>0.244175</TotalBaseFractionValue>
    </TotalBaseFractionPerChannel>
  - <BaselineLevelDist Channel="A">
        <ns:SampleSize>53990</ns:SampleSize>
        <ns:SampleMean>73.0882</ns:SampleMean>
        <ns:SampleMed>72.4</ns:SampleMed>
        <ns:SampleStd>10.3345</ns:SampleStd>
        <ns:Sample95thPct>87.4</ns:Sample95thPct>
        <ns:NumBins>30</ns:NumBins>
      - <ns:BinCounts>
            <ns:BinCount>2</ns:BinCount>
            <ns:BinCount>5</ns:BinCount>
            <ns:BinCount>8</ns:BinCount>
            <ns:BinCount>12</ns:BinCount>
            <ns:BinCount>27</ns:BinCount>
            <ns:BinCount>36</ns:BinCount>
            <ns:BinCount>56</ns:BinCount>
            <ns:BinCount>130</ns:BinCount>
            <ns:BinCount>331</ns:BinCount>
            <ns:BinCount>906</ns:BinCount>
            <ns:BinCount>2215</ns:BinCount>
            <ns:BinCount>4281</ns:BinCount>
            <ns:BinCount>6861</ns:BinCount>
            <ns:BinCount>8456</ns:BinCount>
```

Abbreviations Used in sts.xml

The following abbreviations and conventions are used in code variables and text file headings to convert the full names defined in the document to a short form.

Abbreviation	Definition
`Dist`	Distribution
`Len`	Length
`Max`	Maximum
`Med`	Median
`Min`	Minimum
`Num`	Number of
`Prod`	Productive
`Qual`	Quality
`Std`	Standard Deviation

- All quantities that are acronyms, such as signal/noise ratio (SNR), are modified to standard “sentence-case” form for labels. Example: `Snr`.
- For sample statistics, the identity of the statistic is appended (**not** prepended) to the metric name. Example: `ReadLenMed`, **not** `MedReadLen`.
- All vector-valued quantities associated with base channels are identified by appending ‘_’, followed by the corresponding base label.

Examples:

Abbreviation	Definition
`ReadLenDist`	The distribution of the read length of productive holes.
`ProductivityDist`	The distribution of the Productivity metric.
`SnrMean_A`	The mean SNR of the A channel for all holes.
`BaselineSigmaStd_T`	The standard deviation of the Baseline Sigma metric, for all holes in the T channel.

Data Type:

The `ContinuousDist` or `CDist` data type is used to summarize the chip-wide distributions of real- or integer-valued ZMW metrics. The latest specification of this data type is provided by the data model XSD.

Metrics in the sts.xml File

Singleton Metrics - Single Value for the Entire Chip

Metric Name	Type	Description
`MovieName`	String	The name of the movie, using the following format: `<year><month><date>_<time>_<instrument>_ <SMRTCellbarcode>_<setnumber>_<partnumber>`
`MovieLength`	Integer	The length of the movie (in minutes), based on the number of frames and the acquisition frame rate.
`NumSequencingZmws`	Integer	The number of ZMWs on the cell that are type “Sequencing ZMW.”
`AdapterDimerFraction`	Float	The percentage of ZMWs identified as adapter dimers. The adapter dimer classification is made when the median insert length < 10 bp. An adapter dimer is a SMRTbell™ whose inserts flanked by adapters have median length below some predefined threshold; currently hard-coded as `10`.
`ShortInsertFraction`	Float	The percentage of ZMWs identified as short inserts. The short-insert classification is made when `(10 <= median (insert-length) < 100 bp)`. Inserts are regions flanked by adapter sequence. `10` and `100` bp are hard-coded constants.
`IsReadsFraction`	Float	The percentage of ZMWs with at least 100 bases per hour across all 4 base channels (A,C,G or T) for the entire duration of movie. Example: For a 6-hour movie, a ZMW contributes as `IsRead` if it has at least 2400 bp and at least 600 bp in each channel.

Per-ZMW, Per Analog Metrics - 4 Values Per ZMW

Metric Name	Type	Description
TotalBaseFractionPerChannel	Float	The fraction of called bases by analog in the High-Quality region.
SnrDist	CDist	The signal/noise ratio (SNR) computed on base calls over the entire trace. The calculation is made as the weighted mean of block-SNR values, weighted by the number of pkmid frames in each block. Block-SNR values are computed as `(mean DWS pkmid)/(DWS baseline sigma)`. DWS: Dye weighed sum PKMID: Mean of the per-frame DWS signal estimate (above background) over interior frames. This is undefined if the width is fewer than 3 frames.
HqRegionSnrDist	CDist	The signal/noise ratio (SNR) computed on base calls in the High-Quality region. The SNR calculation is made as described above.
HqBasPkMidDist	CDist	The mean dye weighed sum PKMID computed over base calls in the High-Quality region.

Per-ZMW Metrics - 1 Value Per ZMW

Metric Name	Type	Description
PausinessDist	CDist	The percentage of paused bases. A paused base is defined as a base whose inter-pulse distance (IPD) is greater than 2.5 seconds. (This is a fixed constant.)
ControlReadQualDist	CDist	The mapped accuracy versus the control reference sequence.
ControlReadLenDist	CDist	The mapped length of the read versus the control reference sequence.
ProductivityDist	CDist	The outcome of the Productivity classification.
ReadTypeDist	CDist	The outcome of the ReadType classification.
MovieReadQualDist	CDist	The full (ZMW) read quality score.
PulseRateDist	CDist	The mean pulse rate over the High-Quality region.
PulseWidthDist	CDist	The mean pulse width over the High-Quality region.
BaseRateDist	CDist	The mean base rate over the High-Quality region.
BaseWidthDist	CDist	The mean base width over the High-Quality region.
BaseIpdDist	CDist	The mean base inter-pulse distance (distance between two successive bases; IPD) over the High-Quality region.
LocalBaseRateDist	CDist	An estimate of the local base rate (excluding pauses) over the High-Quality region. Pause exclusion is achieved by using a robust estimate of the mean inter-pulse distance (IPD), modeled as an exponential distribution: `mean(IPD)~median(IPD)/log(2)`.
NumUnfilteredBasecallsDist	CDist	The raw or ZMW read length.
ReadLenDist	CDist	The High-Quality region length.
ReadQualDist	CDist	The read quality score computed over the High-Quality region. As of the Sequel v4.0.0 software, this is a place-holder value.
InsertReadLenDist	CDist	The maximum subread length from the ZMW.

Metric Name	Type	Description
InsertReadQualDist	CDist	As of the Sequel v4.0.0 software, this is a place-holder value.
MedianInsertDist	CDist	The median insert length using subreads flanked by both adapters.
HqBaseFractionDist	CDist	The fraction of bases of ZMW in the High-Quality region.

Per-ZMW, Per Sensor-Channel Metrics – 2 Unique Values Per ZMW (one for each sensor channel)

Baseline level and sigma values are computed as the median over block estimates.

As of the Sequel v4.0.0 software, each distribution is reported by base channel (A and C, duplicates of the red-dye channel; G and T, duplicates of the green-dye channel).

- Reporting baseline metrics in the dye weighed sum (DWS) representation, using analog-based spectra, is a hold-over from the RS convention.

Metric Name	Type	Description
DmeAngleEstDist	CDist	The dye angle estimate over the High-Quality region.
BaselineLevelDist	CDist	The dye weighed sum (DWS) baseline level over the High-Quality region.
BaselineStdDist	CDist	The dye weighed sum (DWS) baseline sigma over the High-Quality region.
BaselineLevelSequencingDist	CDist	The dye weighed sum (DWS) baseline level over full ZMW read for sequencing (non-fiducial) ZMWs.
BaselineLevelNoZmwsNoAperturesDist	CDist	The dye weighed sum (DWS) baseline level for unit cells containing no ZMWs and no apertures.
BaselineLevelNoZmwsWithAperturesDist	CDist	The dye weighed sum (DWS) baseline level for unit cells containing no ZMWs and apertures.
BaselineLevelZmwGreenFilterOnlyDist	CDist	The dye weighed sum (DWS) baseline level for unit cells containing only the green filter.
BaselineLevelZmwRedFilterOnlyDist	CDist	The dye weighed sum (DWS) baseline level for unit cells containing only the red filter.
BaselineLevelZmwFLTIDist	CDist	The dye weighed sum (DWS) baseline level for unit cells marked as FLTI.
BaselineLevelScatteringMetrologyDist	CDist	The dye weighed sum (DWS) baseline level for unit cells marked as scattering metrology.

P/N 001-137-969 Version 02 (March 2018)