Statistics Output Guide

Introduction

This document describes the file `stats.xml`, which is produced by the primary analysis pipeline. This file packages summary statistics from a **single** movie acquisition.

The latest version of this documentation is available from the PacBio Developer's Network, at **http://www.pacbiodevnet.com/Tech_Lib**.

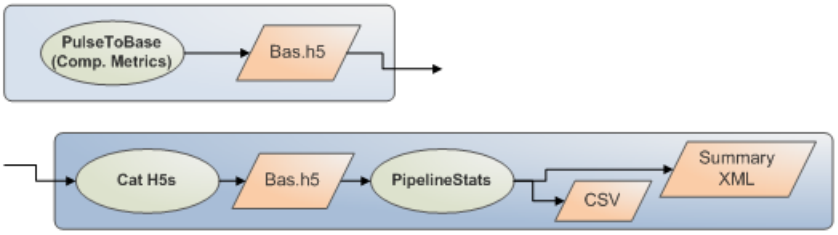
The PipelineStats primary analysis component computes and packages summary statistics from a single movie acquisition on the instrument. It gathers the ZMW metrics defined in the table below, computes all summary statistics, and writes them out to the XML summary statistics file `stats.xml`.

This information is easily consumed by other parts of the system, including instrument management or LIMS systems, without the need for more complex or low-level data APIs.

Metrics are computed by the appropriate pipeline stage:

- **PulseToBase** is the pipeline stage that implements the core basecalling algorithm:
  - Takes the input pulse stream for each ZMW.
  - Classifies pulses as bases (or not).
  - Assigns quality values to bases.
  - Produces the `<>.bas.h5` file.
- **TraceToPulse** is the pipeline stage that implements trace signal processing, pulse detection and pulse classification:
  - Takes an input DWS trace and trace metadata for each ZMW.
  - Estimates and subtracts baseline.
  - Detects and classifies pulses.
  - Produces the `<>.pls.h5` file.

A single PipelineStats component reads metrics through the API for the full movie, computes and writes summary statistics to the XML file, and writes a CSV (comma-separated value) file as an auxiliary result.



ZMW Metrics in stats.xml

Note: P2B = PulseToBase, T2P = TraceToPulse

ZMW Metric Name	Definition	Source
HQRegion	After basecalling, an algorithm determines the contiguous region of the trace that contains high quality sequence data. This region is named **HQRegion**. There will always be 0 or 1 HQRegions found in each trace. The ZMW metrics listed below that are generated by PulseToBase (P2B) are defined only on the bases falling **within** the HQRegion. The extent of the HQRegion is defined in bases. For example, the **HQRegionStart** and **HQRegionEnd** fields in the `sts.csv` file give the indices of the first and last basecall in the HQRegion.	P2B
Base Fraction	The fraction of total basecalls made in each base channel.	P2B
Base Rate	The mean "global" base rate, in bases per seconds. The time window for rate calculation is **[time(end of last base) – time(start of first base)]**.	P2B
Base Width	The mean pulse width of pulses called as bases, in seconds.	P2B
Baseline Level	A vector-valued hole metric that is an estimate of the mean value of the background (characterized by trace regions with no pulse events present), over the duration of the movie, in each of the DWS trace channels. The units used are photo-electrons. (**DWS** is dye-weighted sum, referring to the canonical trace representation.)	T2P
Baseline Sigma	A vector-valued hole metric that is an estimate of the standard deviation of the background (after some locally "smooth" mean-background subtraction), over the duration of the movie, in each of the DWS trace channels. The units used are photo-electrons.	T2P
Number of Pulses	The total number of pulses identified and input to PulseToBase.	T2P
Productivity	A prediction that the ZMW was in one of three possible states for the duration of the movie: **0**: Empty - No enzyme complex. **1**: Productive - Sequencing, single enzyme complex. **2**: Other - Multiple occupation or other condition resulting in low-quality data.	P2B
Pulse Rate	The mean "global" pulse rate, in pulses per seconds. The time window for rate calculation is **[time(end of last base) – time(start of first base)]**.	T2P
Pulse Width	The mean pulse width, in seconds.	T2P
Read Length	The number of called bases in a ZMW read	P2B
Read Score	A scalar-valued metric on a ZMW that predicts accuracy. The metric takes on values in the range 0-1, where 0.95 predicts an overall accuracy of 95% as measured by an alignment of the read to its true template sequence.	P2B
SNR	**Signal-to-Noise Ratio**: **(median called-base pkmid) / (channel baseline sigma)**, computed for each channel. For holes and/or channels where no bases are called, or for which all called bases have no pkmid defined, the SNR is defined as 0.	T2P
HQRegionSnr	The Signal-to-Noise ratio computed as above, but limited to only the inside the HQRegion. For holes **without** an HQRegion, the SNR is defined as 0.	P2B
CmBasQv	**Channel-mean base quality value** A vector-valued metric defined as the mean QualityValue over bases called in each of (A,C,G,T) channels.	P2B
CmDelQv	A vector-valued metric defined as the mean base Deletion QV over bases called in each of (A,C,G,T) channels.	P2B

ZMW Metric Name	Definition	Source
CmInsQv	A vector-valued metric defined as the mean base Insertion QV over bases called in each of (A,C,G,T) channels.	P2B
CmSubQv	A vector-valued metric defined as the mean base Substitution QV over bases called in each of (A,C,G,T) channels.	P2B
RmBasQv	**Read-mean base quality value** A scalar-valued metric defined as the mean (total) Quality Value over all called bases in the read.	P2B
RmDelQv	A scalar-valued metric defined as the mean base Deletion QV over all called bases in the read.	P2B
RmInsQv	A scalar-valued metric defined as the mean base Insertion QV over all called bases in the read.	P2B
RmSubQv	A scalar-valued metric defined as the mean base Substitution QV over all called bases in the read.	P2B

Movie Metrics in stats.xml

The following definitions specify movie metrics. These comprise the reduced or summary values (statistics) computed and reported by PipelineStats. In most cases, summary metrics (such as the mean, median, and so on) are packaged in a representation of the distribution of the full sample or a filtered sample; for example, only productive holes.

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 or classification; 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.

Movie Metric Name	Definition
Total Base Fraction [A,C,G,T]	The total fraction of A, C, G, and T basecalls for all productive reads. The calculation is made from the input data by computing the number of each base type called by a ZMW (such as BaseFraction_A * ReadLength) and then summing over the ZMWs that pass the “Productivity = 1” filter.
Baseline Level Distribution [A,C,G,T]	The distribution of Baseline Level for all holes, in all channels (independently).
Baseline Sigma Distribution [A,C,G,T]	The distribution of Baseline Sigma for all holes, in all channels (independently).
Movie Read Quality Distribution	The distribution of Read Quality for all holes in the movie.
Movie Score	A score in the range 0-1 that can be used to form a quantitative assessment of overall movie quality.
Productivity Distribution	The distribution of the productivity classification.
Pulse Rate Distribution	The distribution of Pulse Rate for productive holes.
Pulse Width Distribution	The distribution of Pulse Width for productive holes.
Read Length Distribution	The distribution of Read Length for productive holes.
Read Quality Distribution	The distribution of Read Quality for productive holes.
RmBasQv Distribution	The distribution of the RmBasQv (read-mean base quality value) metric over productive holes.
SNR Distribution [A,C,G,T]	The distribution of the SNR (signal-to-noise-ratio) over productive holes for channels A,C,G,T (independently).
High Quality Base PkMid Distribution [A,C,G,T]	The distribution of PkMid from bases in HQRegions (over any holes containing HQRegions – possibly non-productive holes) for channels A,C,G,T (independently).
Baseline Level Sequencing ZMWs [A,C,G,T]	The distribution of the Baseline levels over holes designated as sequencing ZMWs for channels A,C,G,T (independently).
Baseline Level Antiholes [A,C,G,T]	The distribution of the Baseline levels over antiholes for channels A,C,G,T (independently). An **antihole** is a grid position containing a micro mirror without a ZMW. This is useful for measuring certain components of background.
Baseline Level Antimirrors [A,C,G,T]	The distribution of the Baseline levels over antimirrors for channels A,C,G,T (independently). An **antimirror** is a grid position with no micro mirror and no ZMW. This is useful for measuring certain components of background, and is also used for gridding.
Baseline Level Fiducial ZMWs [A,C,G,T]	The distribution of the Baseline levels over fiducial ZMWs for channels A,C,G,T (independently). A **fiducial** is a non-standard ZMW; possibly larger than normal.

Abbreviations Used in stats.xml

The following abbreviations and conventions are used in code variables and text file headings to convert the full names defined in the previous table 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 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 quantities that are vector-valued, where the vector is a 4-vector corresponding to base channels, have component quantities identified by appending ‘_’ followed by the corresponding base label.

Examples:

Abbreviation	Definition
ReadLenMed	The median read length of productive holes.
ReadLenDist	The distribution of the read length of productive holes.
ProductivityDist	The distribution of the productivity classification.
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.

P/N 001-137-969-01