bas.h5 Reference Guide

Introduction

This document describes the `bas.h5` file and associated `bax.h5` files. These files are the main output files produced by the primary analysis pipeline on the PacBio® RS II.

The `bas.h5` and associated `bax.h5` files are transferred from the instrument to the customer's offline storage location, and then used as inputs to the SMRT® Analysis software suite to generate alignment, consensus, and variant information.

Prior to the PacBio® RS II, a **single** `bas.h5` file was produced, containing **all** sequence data. Due to the increased throughput and read lengths achieved by the PacBio® RS II upgrade, this information is now contained in **one** `bas.h5` file and **three** `bax.h5` files.

The `bas.h5` file now contains **only** the information necessary to dereference by hole number the ZMW-level data. This information includes:

- A data set [`/MultiPart/Parts`] that provides the names of the `bax.h5` file parts.
- A data set [`/MultiPart/HoleLookup`] that maps hole number to file part index.

The contents of the `bax.h5` files themselves follow the same layout as the original (single-part) `bas.h5` file, modulo additions to the content made in 2.0.

- To view the contents of any `.h5` files (including `bas.h5` and `bax.h5`), use **HDFView**, a free utility available at http://www.hdfgroup.org/hdf-java-html/hdfview/.
- For information about the HDF5 format, see http://www.hdfgroup.org/HDF5/.

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

Hierarchy Layout

At the root level, the content of the `bas.h5` file (and now stored in the `bax.h5` files) consists of two groups: `PulseData` and `ScanData`. (`ScanData` contains instrument-level information needed only for debugging purposes, and is **not** discussed here.)

/PulseData/BaseCalls

BaseCalls are base metrics produced by the PulseToBase pipeline stage for all pulses classified as base-incorporation events by the basecaller. This group contains the following objects:

Name	Data Type	Description
DateCreated	String	Creation date and time of the data group, in ISO 8601 format.
ChangeListID	String	Revision ID of the software that created this data, in format <Major>.<Minor>.<Micro>.<Hotfix>.<PacBioChangeNumber>.
SchemaRevision	String	Version or revision number of the group schema.
QVDecoding	String	Description of the quality-value encoding scheme.
Content	String	Content description of the group, as a data set [name, type].
CountStored	Int32	The number of ZMW reads contained in each data set.
BaseCall	Byte	Called base.
DeletionQV	Byte	Probability of a deletion error prior to the current base. Phred QV.
DeletionTag	Byte	Likely identity of the deleted base, if it exists.
InsertionQV	Byte	Probability that the current base is an insertion. Phred QV.
MergeQV	Byte	Probability of a merged-pulse error at the current base. Phred QV.
PreBaseFrames	UInt16	Duration between the start of a base and the end of the previous base, in Frames.
PulseIndex	Int32	Index into called pulses.
QualityValue	Byte	Probability of a basecalling error at the current base. Phred QV.
SubstitutionQV	Byte	Probability of a substitution error at the current base. Phred QV.
SubstitutionTag	Byte	Most likely alternative base.
WidthInFrames	UInt16	Duration of the base-incorporation event, in Frames.

/PulseData/BaseCalls/ZMW

This group includes ZMW information, and contains the following objects:

Name	Data Type	Description
Content	String	Content description of the group, as a data set [name, type].
HoleChipLook	Int16	Look of ZMW in cell layout. (This is the same as the ‘set’ number for 75k acquisition).
HoleNumber	UInt32	Number assigned to each ZMW on the cell.

Name	Data Type	Description
HoleStatus	Byte	Type of ZMW that produced the data: • 0: SEQUENCING • 1: ANTIHOLE • 2: FIDUCIAL • 3: SUSPECT • 4: ANTIMIRROR • 5: FDZMW • 6: FBZMW • 7: ANTIBEAMLET • 8: OUTSIDEFOV
HoleXY	Int16	Grid coordinates assigned to each ZMW on the cell.
NumEvents	Int32	Event counts per ZMW for all fields in the Event group.

/PulseData/BaseCalls/ZMWMetrics

This group includes ZMW metrics, and contains the following objects:

Name	Data Type	Description
TdmIntervalSec	Single	Time interval used for time-dependent metrics, in seconds.
PauseIpdThreshSec	Single	Threshold on interpulse duration for pause classification, in seconds.
BaseFraction	Single	Base fraction by color channel.
BaseIpd	Single	Robust estimate of the mean inter-pulse distance (IPD) of base-incorporation events, in seconds.
BaseRate	Single	Mean rate of base-incorporation events, in bases per second.
BaseRateVsT	Single	Base rate in HQ (sequencing) region by time interval, in bases per second.
BaseWidth	Single	Mean pulse width of base-incorporation events, in seconds.
CmBasQv	Single	Read-mean of the base QualityValue by color channel.
CmDelQv	Single	Read-mean of the base DeletionQV by color channel.
CmInsQv	Single	Read-mean of the base InsertionQV by color channel.
CmSubQv	Single	Read-mean of the base SubstitutionQV by color channel.
DarkBaseRate	Single	Predicted local base rate when the cell is not illuminated, in bases per second.
HQRegionDyeSpectra	Single	Observed dye spectra in the HQ region as [cam, dye].
HQRegionEndTime	Single	End time of the HQ (sequencing) region, in seconds.
HQRegionEstPkmid	Single	Average estimated intra-pulse amplitude in the HQ region, prior to pulse-calling, in counts.
HQRegionEstPkstd	Single	Average estimated intra-pulse sigma in the HQ region, prior to pulse-calling, in counts.

Name	Data Type	Description
HQRegionIntraPulseStd	Single	Standard deviation of intra-pulse signal in the HQ region, in counts.
HQRegionPkzvar	Single	Ratio of (observed intra-pulse variance) / (model-predicted intra-pulse variance) in the HQ region.
HQRegionSNR	Single	Signal-to-Noise Ratio in the HQ region.
HQRegionStartTime	Single	Start time of the HQ (sequencing) region, in seconds.
LocalBaseRate	Single	Robust estimate (excluding pauses) of the mean base-incorporation rate, in bases per second.
NumBaseVsT	Int16	Number of HQ (sequencing) basecalls by time interval.
NumPauseVsT	Int16	Number of pause events in HQ (sequencing) region by time interval.
Pausiness	Single	Fraction of pause events over the HQ (sequencing) region.
Productivity	Byte	ZMW productivity classifications: • 0: Empty • 1: Productive • 2: Other • 255: Not Defined
ReadScore	Single	Polymerase read accuracy prediction.
ReadType	Byte	ZMW read type classification: • 0: Empty • 1: FullHqRead0 • 2: FullHqRead1 • 3: PartialHqRead0 • 4: PartialHqRead1 • 5: PartialHqRead2 • 6: Multiload • 7: Indeterminate • 255: Not Defined
RmBasQv	Single	Read-mean (over HQ region if present, otherwise global) of the base QualityValue.
RmDelQv	Single	Read-mean of the base DeletionQV.
RmInsQv	Single	Read-mean of the base InsertionQV.
RmSubQv	Single	Read-mean of the base SubstitutionQV.
SpectralDiagRR	Single	Ratio (observed : calibrated) of the spectral matrix (diagonal : dominant-off-diagonal) ratios.

/PulseData/ConsensusBaseCalls

This group includes metrics for circular consensus base calls, and contains the following objects:

Name	Data Type	Description
DateCreated	String	Creation date and time of the data group, in ISO 8601 format.
ChangeListID	String	Revision ID of the software that created this data, in format <Major>.<Minor>.<Micro>.<Hotfix>.<PacBioChangeNumber>.
SchemaRevision	String	Version or revision number of the group schema.
QVDecoding	String	Description of the quality-value encoding scheme.
Content	String	Content description of the group, as a data set [name, type].
CountStored	Int32	The number of ZMW reads contained in each data set.
BaseCall	Byte	Called base.
DeletionQV	Byte	Probability of a deletion error prior to the current base. Phred QV.
DeletionTag	Byte	Likely identity of the deleted base, if it exists.
InsertionQV	Byte	Probability that the current base is an insertion. Phred QV.
QualityValue	Byte	Probability of a basecalling error at the current base, Phred QV.
SubstitutionQV	Byte	Probability of a substitution error at the current base. Phred QV.
SubstitutionTag	Byte	Most likely alternative base.

/PulseData/ConsensusBaseCalls/Passes

This group includes information from Single Molecule Consensus processing of the raw read, and contains the following objects:

Name	Data Type	Description
AdapterHitAfter	Byte	Flag indicating if an adapter hit was detected at the end of this pass. `1` if the pass **began** with an adapter hit, `0` if it didn't.
AdapterHitBefore	Byte	Flag indicating if an adapter hit was detected at the beginning of this pass. `1` if the pass **ended** with an adapter hit, `0` if it didn't.
NumPasses	Int32	The number of passes detected in a ZMW.
PassDirection	Byte	Direction of pass across the SMRTbell™ template. `0` for forward pass, `1` for a reverse pass.
PassNumBases	UInt32	The number of bases in a circular consensus pass.
PassStartBase	UInt32	Index of the first base in a circular consensus pass.

/PulseData/ConsensusBaseCalls/ZMW

This group includes ZMW information and contains the following objects:

Name	Data Type	Description
Content	String	Content description of the group, as a data set [name, type].
HoleChipLook	Int16	Look of ZMW in cell layout. (This is the same as the ‘set’ number for 75k acquisition).
HoleNumber	UInt32	Number assigned to each ZMW on the cell.
HoleStatus	Byte	Type of ZMW that produced the data: • 0: SEQUENCING • 1: ANTIHOLE • 2: FIDUCIAL • 3: SUSPECT • 4: ANTIMIRROR • 5: FDZMW • 6: FBZMW • 7: ANTIBEAMLET • 8: OUTSIDEFOV
HoleXY	Int16	Grid coordinates assigned to each ZMW on the cell.
NumEvents	Int32	Event counts per ZMW for all fields in the Event group.

/PulseData/ConsensusBaseCalls/ZMWMetrics

This group includes ZMW metrics and contains the following objects:

Name	Data Type	Description
InsertReadLength	UInt32	Read length of insert.
PredictedAccuracy	Single	Predicted average accuracy of consensus sequence.

/PulseData/Regions

This group includes information about the Regions table, with columns identified by ColumnNames attribute. This group contains the following objects:

Name	Data Type	Description
RegionTypes	String	Region type lookup table. Example types: `GlobalAccuracy`, `HQRegion`, `Insert`, or `Adapter`. The description and source of `HQRegion` would be "High Quality bases region. Score is 1000 * predicted accuracy, where predicted accuracy is 0 to 1.0" and "Region classifier", respectively.

Name	Data Type	Description
RegionDescriptions	String	Region type description. An array of strings describing the semantics of the region type. The region type index column is an index into this array.
RegionSources	String	Origin or source of the region annotation. An array of strings describing the source of this region annotation type, for example which pipeline stage or algorithm emitted this annotation. The region type index column is an index into this array.
ColumnNames	String	Identification of Regions table columns. • Column 0: The hole number of the trace. • Column 1: The region type index. • Column 2: The region start position in bases, **inclusive** of this base. • Column 3: The region end position in bases, **exclusive** of this base. • Column 4: The region score number, interpreted according to the region type.

P/N 001-496-921-02