

**Introduction** The SMRT Analysis software suite performs assembly and variant detection analysis of sequencing data generated by the PacBio® instrument.

**Installation** For installation instructions, see the document **SMRT Analysis Software Installation**, available on the Pacific Biosciences web site.

**New Features  
in v2.3.0**

**SMRT Analysis**

- **Quiver** is trained for the new P6 Polymerase with C4 Sequencing Chemistry.
- **Long Amplicon Analysis:**
  - HLA class II support
  - Support for amplicons with greater range of lengths (3-9 kb) and mixed amplicon sizes
  - Barcoding - new options:
    - Generate separate FASTA/FASTQ files per barcode.
    - Filter reads based on minimal barcode score.

**BLASR**

- A new clipping mode (`subread`) for SAM output - sequences are soft-clipped within coordinates of subreads instead of unrolled reads.
- A new option (`-printSAMQV`) to print additional quality values to SAM output files, including `InsertionQV`, `DeletionQV`, `SubstitutionQV`, `MergeQV`, `DeletionTag`, and `MergeTag`.

**Iso-Seq™ Software**

- The `P_IsoSeq` module is divided into `P_IsoSeqClassify` and `P_IsoSeqCluster` modules. `Classify` and `Cluster` tasks are handled separately.
- Iso-Seq protocol parameters for `Classify` and `Cluster` algorithms are included in separate panels in SMRT Portal.
- Upgraded GMAP to 2014-08-04.

**Assembly**

- Improved overlap detection in the preassembly process.

**SMRT Pipe**

- SMRT Pipe is refactored from `pbpy` to a separate python package, `pbsmrtpipe`, which enhances the robustness of SMRT Analysis.

**SMRT Portal**

- Added new controls to the `RS_Long_Amplicon_Analysis` Protocol Settings dialog to:
  - Turn on/off clustering

- Turn on/off phasing
- Trim bases off the ends of consensus sequences.
- Added a control to the `RS_IsoSeq` **Protocol Settings** dialog to specify whether or not full-length reads require polyA tails.
- Added a control to specify advanced `pballign` options when using resequencing protocols, including `RS_Resequencing`, `RS_Modification_Detection` and `RS_Modification_and_Motif_Analysis`.
- Troubleshooting tools: Links to download SMRT Analysis and job-specific support files, in zipped format, for use by Pacific Biosciences Technical support.
- The **About** box displays the full installation version.

### Installation/Upgrade

- Implemented a new way to invoke an isolated and controlled SMRT Analysis environment for running SMRT Portal and SMRT Pipe commands. This alleviates some of the problems related to version dependencies for various software packages and permission restrictions for non-privileged users. We **no longer** pass through most environment variables from the user environment, **except**:

`USER`, `LOGNAME`, `PWD`, `TERM`, `TERMCAP`, `HOME`, `WORKSPACE`, `MPLCONFIGDIR`, and all `SMRT_*` variables.

- Added a `smrtwrap` script as the main entry point for using SMRT Analysis scripts.
- Added a `smrtshell` script that mimics `setup.sh` and creates a subshell for execution of the SMRT Pipe analysis.
- Force the "C" (aka "POSIX") locale for all SMRT Analysis tools in `setup.sh`.
- Unset (almost) all user environment variables in `setup.sh`.
- For additional information, see the document **SMRT Pipe Reference Guide**.

### Changes to Protocols in v2.3.0

- New `BAM_Resequencing_Beta` protocol: This is an **experimental** version of the `RS_Resequencing` protocol which uses `BAM` rather than `cmp.h5` as the output file format. The protocol is **faster** than the `RS_Resequencing` protocol for large jobs, but is **not yet** guaranteed to produce identical results as the `RS_Resequencing` protocol.
- The `RS_CeleraAssembler` protocol is no longer included.

### Fixed Issues in v2.3.0

#### SMRT Analysis

- The environmental variable `MPLCONFIGDIR` pointed to `~/matplotlib`. (16052)
- The bundled version of `mysql` did not override the user-level configuration file `~/my.cnf`. (25104)

- The build version number in `config.xml`, `patchnum.txt`, `patchhistory.txt` and `prerun.patchnum.txt` was incorrect. (25350)
- Fixed SMRT Pipe's error detection. (24536)
- `pbtranscript` cluster spawns too many threads at the same time. (24969)

### SMRT Portal

- Users logged in as **Technicians** can now delete their own jobs. (25670)
- Clicking the **Log** button now displays the `master.log` file, which is useful for troubleshooting. (25419)
- Clicking the **H5** button in the **Data Panel** of the **View Data** page now downloads a gzipped directory containing the `metadata.xml`, `bax.h5` and `bax.h5` files. (25417)
- Queued jobs are no longer marked as "FAILED" on clusters with high usage loads. (25465)
- Group names are now correctly exported when clicking **Export Table Data**. (22567)
- When copying an existing job, the **Copy** button is now active **only** if the selected job was created by the running version of SMRT Portal. (22876)
- The **Download** and **Download All** links were removed from the Reports page. (23740)
- Clarified the error message displayed when you create a new job using an invalid group name. (24980)
- The group **all** is now selected by default for new Administrative users. (25144)
- **Cannot** save a job if a job directory with the same name already exists. (25212)

### SMRT Portal Reports

- The **Modifications - Motifs** report includes a meaningful title in the table, and a percentage in the "% Motifs Detected" field. (25530)
- Removed several set-related fields from the **Site Acceptance Test** report. (25541)
- Added a new **Amplicon - Input Metrics** report. Amplicon reports now do **not** display noise and chimeric reads. (25525, 25614)
- Incorrect amplicon lengths were displayed. (24893)
- "Polished Contigs" were not displaying on the **Report Overview** page for `RS_HGAP_Assembly.2` and `RS_HGAP_Assembly.3` jobs. (23675)

### Web Services API

- **All** web services API calls require authentication for reading or downloading data. (25497)

- The `Save User` API function correctly evaluates passwords. (23465)

### SMRT Pipe

- SMRT Pipe creates `TMP` directories on cluster nodes as needed. (24881)
- Tasks core dumps are written to the task log directory. (24947)
- SMRT Cell paths containing white space were reported as not found by `pballign.py`. (25075)
- Removed the `--recover` option from `smrtpipe.py`. (25234)
- SMRT Pipe splits by contig at the merge step rather than waiting for the Quiver step. (25356)
- Changed the default `EXIT_ON_FAILURE` value to `True` so SMRT Pipe exits more quickly after a task failure. (25344)

### SMRT Pipe - Barcoding

- The default barcoding mode is `symmetric`. (25309)

### SMRT Pipe - Long Amplicon Analysis

- Rare alleles were not being consistently detected. (24412)
- Added a maximum subread length filter to help with filtering out concatemer sequences. (24698)
- Added barcode score filtering. (25345)
- The white list option accepted white lists **only** in the form of Subread Ids, but not as ZMWs. (25678)

### SMRT Pipe - Mapping (BLASR)

- Setting the `-concordant` option caused a memory leak. (25618)
- BLASR reported incorrect MapQVs. (24363, 25290)
- BLASR's SAM output conforms to the SAM specification:
  - The `NM` tag now represents the edit distance. (23264)
  - Added `YS`, `YE`, and `ZM` tags to SAM output; changed the SAM header; used specification QV names and tags. (25447)

### SMRT Pipe - Reads of Insert

- Quality scores were not being reversed in the CCS SAM output file. (24006)
- Quality values for CCS Reads were too high. (25113)

### SMRT Pipe - Iso-Seq cDNA Analysis

- Iterative clustering was using too much memory in multiprocessing mode. (25580)
- Added human-readable annotations to unpolished and polished isoform IDs. (25513)

- Added a command-line-only `--detect_chimera_nfl` option to detect chimeric reads among non-full-length reads in `pbtranscript classify`. (25210)
- Added `ice_fa2fq.py` (available from the command-line) to convert an ICE FASTA file containing CCS reads to a FASTQ file. This allows use of input/output FASTQ format with `pbtranscript.py classify`. (25125)
- The polishing step failed, but all jobs actually finished. (25077)
- `pls2fasta` failed with input file paths containing white space. (25204)

### SMRT Pipe - Assembly

- `RS_HGAP_Assembly.2` and `RS_HGAP_Assembly.3` jobs failed when "Use only unambiguously mapped reads" was unchecked. (25070)
- Copying HGAP protocols from an older job added an extra character to the protocol name. (25024)

### Installation/Upgrade

- Improved the help information for `smrtupdater` and its subprograms. (24794)
- The SMRT Portal `TMP` directory might be missing after head node reboot. (25597)
- Log files from patch activity were not generated in `common/log/install`. (25221)
- Improved how `SEYMOUR_HOME` is set and how `setup.sh` is sourced. (22610)
- The "Upgrade and configure" script now checks to see if the execution nodes are the submit nodes. (21962)
- Ensure that any user setting of `JAVA_HOME` is overridden so that the correct version of Java is used. (24657)
- The `smrtupdater`'s `--skip-userquery` option was ignored during upgrades. (24751)
- Fixed the handling of multi-line and comma separated `qconf` lists for SGE settings. (24818)

### Known Issues in v2.3.0

#### SMRT Portal

- Clicking the **Save** button more than once changes some job parameters. (25812)
- Administrator users should be able to select and delete multiple jobs on the **View Data** page. (24912)
- The `RS_Subreads` protocol with barcoding does **not** filter barcoded FASTQ files by quality. (25179)
- The `RS_ReadsOfInsert` protocol with barcoding should include an option to trim barcodes. (24510)
- The `RS_ReadsOfInsert_Mapping` protocol should include barcode support. (25699)

- Every `RS_` protocol should include a spike-in control module. (24163)

### SMRT Portal Reports

- In the **Diagnostic - Loading** Report, overloaded cells should be flagged more clearly and accurately. (23856)
- The **Diagnostic - Adapters** Report underestimates the Adapter Dimers by a large margin. (20357)

### SMRT Pipe

- The pre-filter reads should be High Quality only. (22980)

### SMRT Pipe - Long Amplicon Analysis

- Too many reads are required to generate a reasonable consensus for more than 3 PCR products. (25717)
- In a mixed population of 3 to 5 kb products, the software occasionally truncates a few hundred bases from the 5 kb products. (25688)
- For some specific cases (a large indel, a single base difference, or misalignment), some alleles are missing. (25439, 25078, 25347)

### SMRT Pipe - Base Modification

- The Motif Finder software does **not** work well with high-GC genome base modifications. (24315)

## Fixed Issues in v2.3.0.p1

### Resequencing Protocols

- Released the `BAM_Resequencing_Beta` protocol. This protocol is significantly faster than `RS_Resequencing` and will speed up Quiver.
- A large redundant output file (`aligned_reads.sam`, similar to `aligned_reads.bam`) is no longer produced. This affects the `BAM_Resequencing_Beta`, `RS_Resequencing` and `RS_Resequencing_Barcode` protocols. (22881)

### Installation

- Fixed an issue where Celera Assembler failed because `qsub` was not found when called from Celera Assembler. (25903)
- Fixed an issue that caused installation failure due to DNS/hostname problems. (25891)

### SMRT Pipe - Mapping (BLASR)

- Made enhancement to the read mapping algorithm addressing a case where reads from sub-optimal data were mapping to extended genome coordinates. This fix affects the Resequencing and Base Modification analysis protocols. (25860)

### SMRT Pipe - Long Amplicon Analysis

- Fixed an issue where similar settings with sufficient coverage produced inconsistent results for dinucleotide regions, depending on minor differences in selection of input reads. (25683)
- Enhanced runtime and memory use by modifying the way memory is used by the suffix array. (25932)

### Iso-Seq cDNA Analysis

- `Isoseq_cluster` analysis parameters: Renamed compute parallelization parameter from "Chunks" to "Parallel Tasks" for clarity. (25839)
- Fixed an issue that caused a `pbtranscript.py` exception during `RS_IsoSeq` jobs. (25888)

### Fixed Issues in v2.3.0.p2

#### Iso-Seq cDNA Analysis

- Reduced memory consumption. In addition, the default quality values used throughout Iso-Seq analysis are now the Phred-like FastQ values instead of PacBio quality values used in previous versions. (26047)

**Note:** To avoid out-of-memory conditions, we now limit the number of SMRT Cells that can be included in single analysis job to **12**.

- The `P_IsoSeqCluster.py` script now works correctly in single-node, non-distributed environments. (25943)
- The global `NPROC` value is now correctly applied, and used throughout secondary analysis. (26055)
- The `IcePostQuiver.py` script no longer expects SGE Job Management System output when running under the LSF Job Management System. (25577)

#### BAM Resequencing

- Further optimizations of analysis speed. (25970)
- The correct documentation is now included with the build. (25931)

#### Base Modification Detection

- Fixed a rare failure in low complexity regions. (26065)

#### SMRT Pipe

- Quiver no longer truncates reference names in the `variants.gff` file. (26010)
- `PacBio.Consensus` now loads the correct sequencing chemistry. (25976)
- Temporary directories are now correctly created. (25996)
- The speed of `ConsensusTools` is improved. (25725)
- `cmph5tools.py` select now copies the entire movie table. (25913)



- We **no longer** use local user versions of python packages. Instead, we always use the python version distributed with SMRT Analysis. (26067)

### SMRT Pipe - Long Amplicon Analysis

- The `P_AmpliconAssembly` module now also reports results when only **one** amplicon is found. (24990)

### SMRT Pipe - Reads of Inserts

- Added a command-line option to bypass palindrome filtering; SMRT Pipe now reports **why** reads failed CCS filtering instead of a single count of all failed reads. (26009)

### SMRT Portal/SMRT View

- Updated the security certificate used to sign the code for both SMRT Portal and SMRT View. (26052)
- Unchecking the SMRT Portal **Predict Consensus Isoforms using The ICE algorithm** checkbox now runs the clustering algorithm, as expected. (25963)

## Enhancements in v2.3.0.p3

### Circular Consensus Sequence Analysis

- Enabled the use of higher accuracy data as an input to the CCS analysis: The **Minimum Predicted Accuracy** option now accepts Q30 data.
  - To use this option in SMRT Portal: In the `RS_ReadsOfInsert` Protocol Settings dialog, set the **Minimum Predicted Accuracy** filtering parameter value to `99.9`.
  - To use this option using the command line:

```
% ConsensusTools.sh CircularConsensus .... --  
minPredictedAccuracy=99.9
```
- When run through the command line, CCS analysis now outputs an aggregated report at the end of each run indicating the total yield of CCS reads and the percentage of ZMWs that were filtered out by various criteria.

## Fixed Issues in v2.3.0.p3

### SMRT Pipe - Reads Of Insert (CCS) Analysis

- Enable use of asymmetric adapters - CCS no longer recalls adapters unnecessarily. (25611)
- Made changes to the `RS_ReadsOfInserts` Protocol Settings dialog options:
  - The `RS_ReadsOfInserts` Protocol Settings dialog's `Minimum Full Passes` option now allows you to specify more than 10 full passes. (26109)

### SMRT Pipe - Barcoding

- Set the default minimum barcode score to 22 to reflect the recommended value for the SMRTbell™ Barcoded Adapters and Barcoded Universal Primers. (26343)



- Changed the default trim setting for `pbbarcode emitFastqs` from 20 to 16 to reflect the recommended value for the SMRTbell Barcoded Adapters and Barcoded Universal Primers. (26153)
- The Default file containing the barcode sequences now contains the sequences for the SMRTbell Barcoded Adapters and Barcoded Universal Primers. (26373)

#### Notes:

- The default value is selected to support 16 base pairs barcodes such as those recommended for certain applications of the PacBio instrument. If you use different-length barcodes, change the minimum barcode score value accordingly.
- To view the mapping between a specific well and a barcoded sample on a 96-well plate, click [https://s3.amazonaws.com/files.pacb.com/Barcode\\_General/docs/Barcode\\_Plate\\_Mapping\\_UB.pdf](https://s3.amazonaws.com/files.pacb.com/Barcode_General/docs/Barcode_Plate_Mapping_UB.pdf).

#### Iso-Seq cDNA Analysis

- Upgraded GMAP to version 2014-12-21 to fix an issue that caused Iso-Seq analysis to fail. (26168)

#### SMRT Pipe

- P6 part numbers for binding and sequencing kits are now correctly recognized. (26353)

#### SMRT Pipe - Long Amplicon Analysis

- Analysis now ignores ends when performing checks for duplicate clusters. (26326)

#### SMRT Pipe - Minor Variants

- Lowercase references are now supported. (26238)

#### Installation/Upgrade

- Added new Technical Support scripts to provide troubleshooting ability in case of analysis failure. The scripts enable the collection of data about analysis set-up, data collection, and user environment analysis. The scripts are located in `SMRT_ROOT/current/support`. (25544)

#### Aggregated CCS Report

CCS analysis now generates a table at the end of each run listing the total yield of CCS reads, as well as the number/percentage of ZMWs that were filtered out by various criteria.

Result Report for the 163482 Zmws processed		
Zmw Result	#-Zmws	%-Zmws
Successful - Quiver consensus found	8554	5.23%
Successful - But only 1 region, no true consensus	0	0.00%
Failed - Exception thrown	0	0.00%
Failed - ZMW was not productive	127058	77.72%
Failed - Outside of SNR ranges	355	0.22%
Failed - No insert regions found	3	0.00%
Failed - Not enough full passes	22243	13.61%
Failed - Insert length too small	0	0.00%
Failed - Post POA requirements not met	1952	1.19%
Failed - CCS Read below predicted accuracy	3073	1.88%
Failed - CCS Read was palindrome	36	0.02%
Failed - CCS Read below SNR threshold	0	0.00%
Failed - CCS Read too short or long	208	0.13%

**Note: Not all ZMWs produce CCS reads.** A ZMW's data will **not** be reported as a CCS read if any of the following filtering criteria apply:

- An initial template of sufficient quality could not be generated from the subreads. (This is when data is noisy and no consensus appeared.)
- The read was below predicted accuracy thresholds or user-specified criteria.
- The read appeared to come from 2 enzymes polymerizing from 2 templates in the same ZMW.
- No insert regions were found.
- The template in between inserts was too short (<5 bp).
- The read appeared palindromic - that is, designed to filter out reads with missed adapter calls.
- A rare event caused a program exception while the read was being processed. (This will also generate an error message.)

#### Fixed Issues in v2.3.0.p4

- Fixed an installation problem that resulted in accessibility and activation issues for some reference sequences in SMRT Analysis. (26462)

#### Fixed Issues in v2.3.0.p5

- Fixed an issue where SMRT Analysis did **not** recognize the latest sequencing kit - DNA Sequencing Reagent 4.0 v2 (P/N 100-612-400). This issue caused CCS, Quiver and `kineticsTools` to produce erroneous results. (29386)

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2010-2015, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacb.com/legal-and-trademarks/product-license-and-use-restrictions/>.

Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science, Inc. NGS-go and NGSengine are trademarks of GenDx. All other trademarks are the sole property of their respective owners.

P/N 100-382-300-07