PACBIO®

SMRT Link User Guide

Sequel® System

For Research Use Only. Not for use in diagnostic procedures.

P/N 101-039-100 Version 06 (October 2018)

Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document.

PACIFIC BIOSCIENCES DISCLAIMS ALL WARRANTIES WITH RESPECT TO THIS DOCUMENT, EXPRESS, STATUTORY, IMPLIED OR OTHERWISE, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY, SATISFACTORY QUALITY, NONINFRINGEMENT OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL PACIFIC BIOSCIENCES BE LIABLE, WHETHER IN CONTRACT, TORT, WARRANTY, PURSUANT TO ANY STATUTE, OR ON ANY OTHER BASIS FOR SPECIAL, CONSEQUENTIAL, INCIDENTAL, EXEMPLARY OR INDIRECT DAMAGES IN CONNECTION WITH (OR ARISING FROM) THIS DOCUMENT, WHETHER OR NOT FORESEEABLE AND WHETHER OR NOT PACIFIC BIOSCIENCES IS ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

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 https://www.pacb.com/legal-and-trademarks/terms-and-conditions-of-sale/.

Trademarks:
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. FEMTO Pulse and Fragment Analyzer are trademarks of Advanced Analytical Technologies. All other trademarks are the sole property of their respective owners.

Pacific Biosciences of California, Inc.
1305 O’Brien Drive
Menlo Park, CA 94025
www.pacb.com

SMRT® Link User Guide (v6.0.0)

Introduction

This document describes how to use Pacific Biosciences' SMRT Link software. SMRT Link is the web-based end-to-end workflow manager for Sequel Systems. It also supports analysis and management of data from PacBio RS II systems. SMRT Link includes the following modules:

- **Sample Setup:** Calculate binding and annealing reactions for preparing DNA samples. (See "Sample Setup" on page 7 for details)
- **Run Design**: Design sequencing runs and create and/or import sample sheets. (See "Run Design" on page 10 for details.)
- **Run QC**: Monitor run progress, status and quality metrics. (See "Run QC" on page 17 for details.)
- **Data Management**: Create Projects and Data Sets; generate QC reports for Data Sets; view, import, or delete sequence, reference, and barcode files. (See "Data Management" on page 22 for details.)
- **SMRT Analysis**: Perform secondary analysis on the basecalled data (such as sequence alignment, variant detection, *de novo* assembly, structural variant calling, and RNA analysis) after a run has completed. (See "SMRT® Analysis" on page 32 for details.)

SMRT Link also includes SMRT View, a genome browser that displays sequencing data generated by the PacBio RS II and Sequel Systems. (See "Visualizing Data Using SMRT® View" on page 117 for details.)

This document also describes:

- The data files generated by the Sequel System for each cell that are transferred to network storage. (See "Sequel® System Output Files" on page 120 for details.)
- The data files generated by secondary analysis. (See "Secondary Analysis Output Files" on page 123 for details.)
- Configuration and user management. (See "Configuration and User Management" on page 126 for details.)
- SMRT Link client hardware/software requirements. (See "Hardware/ Software Requirements" on page 129 for details.)

Installation of SMRT Link **Server** software is discussed in the document **SMRT Link Software Installation (v6.0.0)**.

New features, fixed issues and known issues are listed in the document **SMRT Link Release Notes (v6.0.0)**.

PacBio® RS II System Users

SMRT Link's **Data Management** and **SMRT Analysis** modules are compatible with PacBio RS II data; the rest of the SMRT Link modules work **only** with Sequel Systems.

- PacBio RS II users should use **Binding Calculator**, **RS Remote**, and **RS Dashboard** software instead of the SMRT Link **Sample Setup**, **Run Design** and **Run QC** modules.
- To display **only** the modules for the PacBio RS II: Choose **Configure** from the SMRT Link menu, click **Instrument**, then check the **PacBio RS II Only** box. SMRT Link displays only the **Data Management** and **SMRT Analysis** modules.

Contact Information

For additional technical support, contact Pacific Biosciences at support@pacb.com or 1-877-920-PACB (7222).

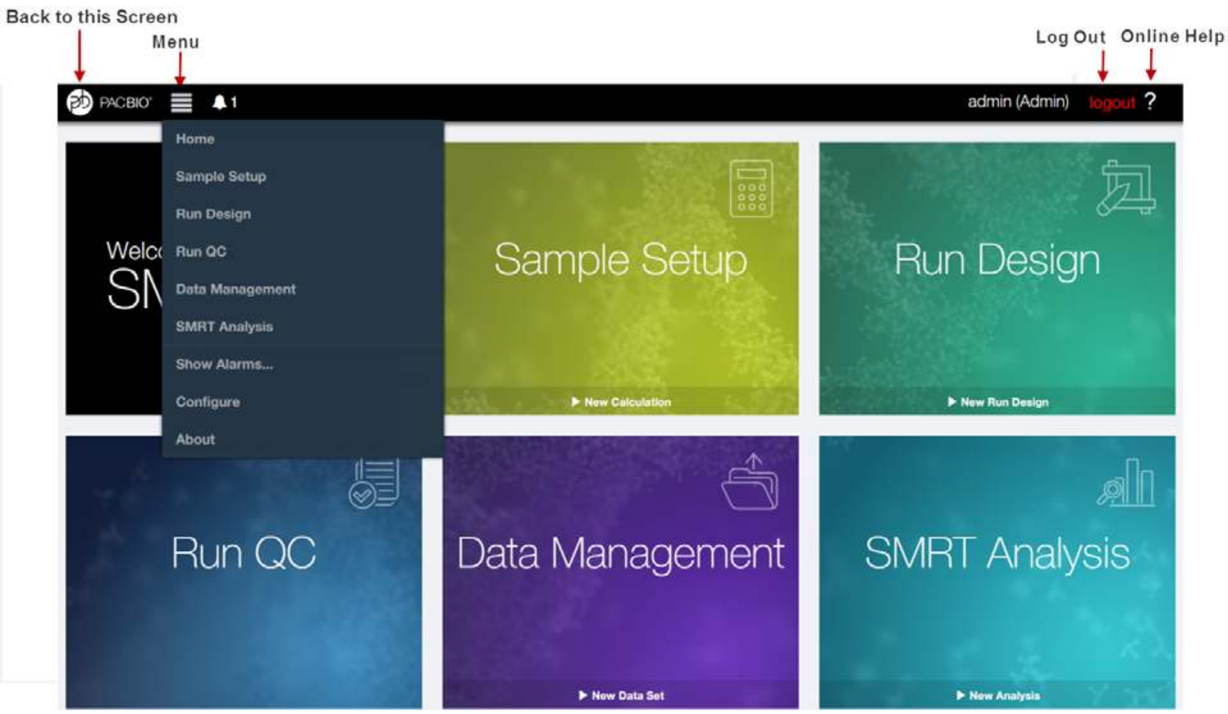
Using SMRT® Link

You access SMRT Link using the Chrome web browser.

- SMRT Link is **not** available on the instrument – it must be accessed from a remote workstation.
- Depending on how SMRT Link was installed at your site, logging in with a user name and password may be required.
- SMRT Link needs a Secure Sockets Layer (SSL) Certificate to ensure a secure connection between the SMRT Link server and your browser using the HTTPS protocol.

If an SSL Certificate is **not** installed with SMRT Link, the application will use the PacBio self-signed SSL Certificate and will use the HTTP protocol. In this case, **each** user will need to accept the browser security warnings described in "Using the PacBio® Self-Signed SSL Certificate" on page 119.

After accessing SMRT Link, the **Home** page displays. (Shown is the interface for the Sequel System.)



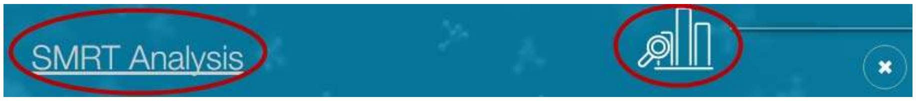
Back to this Screen

Menu

Log Out

Online Help

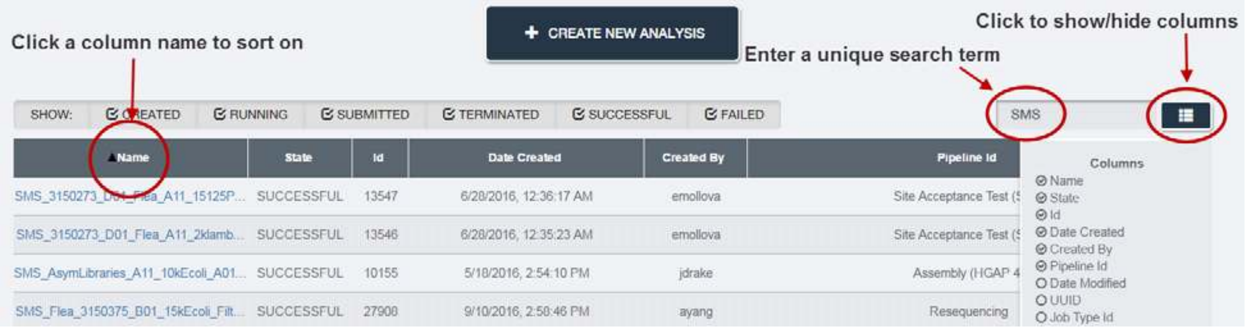
- Click the **menu** to navigate to any of the modules, configure for the PacBio RS II, view version information, or perform administrative functions (Admins **only**).
- Click a module **name** to access that module. **Sample Setup**, **Run Design**, **Data Management** and **SMRT Analysis** include links to create new Calculations, Run Designs, Data Sets, and Analyses.
- Click the **PacBio logo** at the top left to navigate back to the SMRT Link Home page from within the application.
- Click **?** to view the SMRT Link Online help.
- Click **logout** to log out of SMRT Link.



- Within a module: Click the **module name** or the **module design** to navigate back to the module's home screen.

Working with Tables

- To **sort** table columns: Click a **column title**.
- To **search** within a table: Enter a unique search string into the **Search** field.
- To **show or hide** table columns: Click the control, then check or uncheck column name(s) in the dialog box that displays.



▲Name	State	Id	Date Created	Created By	Pipeline Id
SMS_3150273_D01_Flea_A11_15125P...	SUCCESSFUL	13547	6/28/2016, 12:36:17 AM	emollova	Site Acceptance Test (S
SMS_3150273_D01_Flea_A11_2klamb...	SUCCESSFUL	13546	6/28/2016, 12:35:23 AM	emollova	Site Acceptance Test (S
SMS_AsymLibraries_A11_10kEcoli_A01...	SUCCESSFUL	10155	5/18/2016, 2:54:10 PM	jdrake	Assembly (HGAP 4
SMS_Flea_3150375_B01_15kEcoli_Filt...	SUCCESSFUL	27908	9/10/2016, 2:58:46 PM	ayang	Resequencing

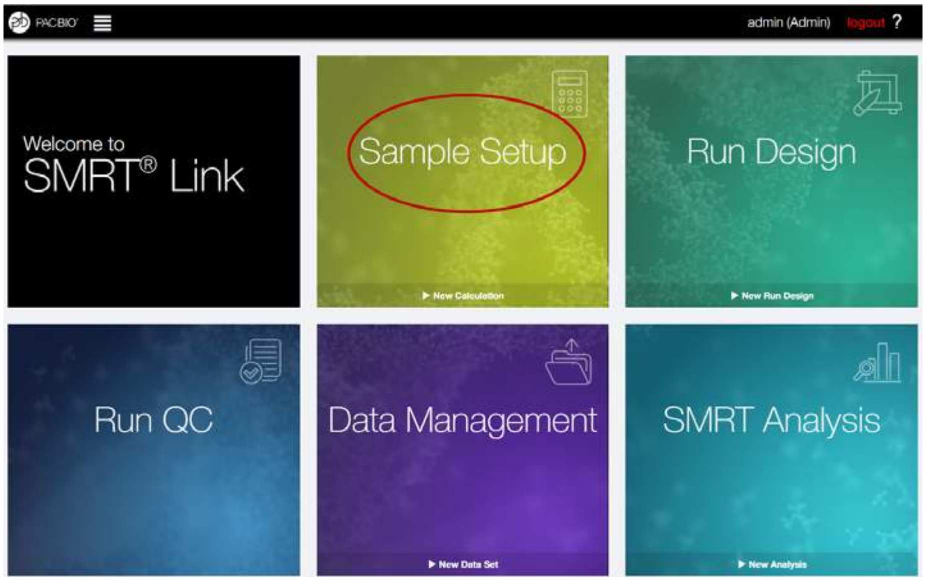
SMRT Link Menu Commands

- **Home**: Displays the SMRT Link Home page.
- **Sample Setup**: Displays the Sample Setup module. (Sequel System **only**.)
- **Run Design**: Displays the Run Design module. (Sequel System **only**.)
- **Run QC**: Displays the Run QC module. (Sequel System **only**.)
- **Data Management**: Displays the Data Management module.
- **SMRT Analysis**: Displays the SMRT Analysis module.
- **Show Alarms...**
  - Displays SMRT Link system-level alarms. To clear alarms, select and click **Clear Alarm** or **Clear All Alarms**.
- **Configure**
  - To display **only** the modules for the PacBio RS II: Click **Instrument**, then check the **PacBio RS II Only** box. SMRT Link displays only the Data Management and SMRT Analysis modules.
  - **Admin users only**: Add/delete SMRT Link users and specify their roles. See “Adding and Deleting SMRT Link Users” on page 127 for details.
- **About**
  - Displays software version information and available space on the server SMRT Link is connected to.
  - Click **Send** to send configuration information to Pacific Biosciences Technical Support for help in troubleshooting failed analyses.
  - **Admin users only**: Update the SMRT Link Chemistry Bundle, which includes kit and DNA Control Complex names used in the Sample Setup and Run Design modules.

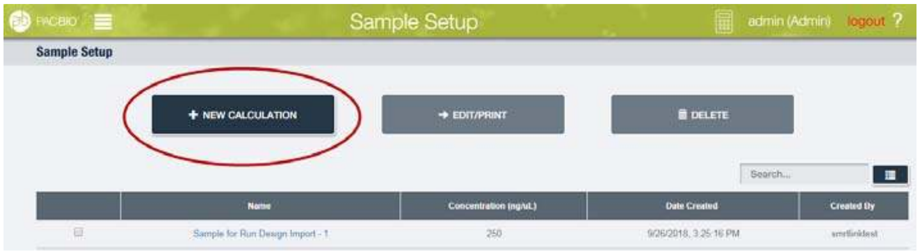
Sample Setup

Before setting up a run, use SMRT Link's **Sample Setup** module to generate a customized protocol for primer annealing and polymerase binding to SMRTbell® templates, with subsequent sample clean-up. These protocols are for use on the Sequel System **only**. You can then print the instructions for use in the lab.

- If you are using SMRT Link with a PacBio RS II, use **Binding Calculator** software.

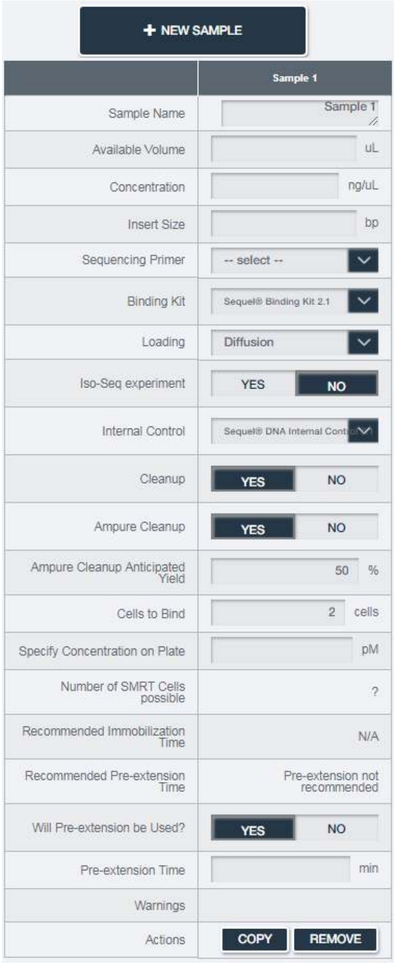


1. Access SMRT Link using the Chrome web browser.
2. Select **Sample Setup**.
3. Click + **New Calculation**.



	Name	Concentration (ng/uL)	Date Created	Created By
□	Sample for Run Design Import - 1	250	9/26/2018, 3:25:16 PM	smrtlinktest

4. Enter the sample **name**.
5. Enter the available sample **volume**, in ul.
6. Enter the sample **concentration**, in ng/ul.
7. Specify an **Insert Size**. Enter the average size of the SMRTbell library.



+ NEW SAMPLE

	Sample 1
Sample Name	Sample 1
Available Volume	uL
Concentration	ng/uL
Insert Size	bp
Sequencing Primer	-- select --
Binding Kit	Sequel® Binding Kit 2.1
Loading	Diffusion
Iso-Seq experiment	YES [✓ NO]
Internal Control	Sequel® DNA Internal Cont
Cleanup	[✓ YES] NO
Ampure Cleanup	[✓ YES] NO
Ampure Cleanup Anticipated Yield	50 %
Cells to Bind	2 cells
Specify Concentration on Plate	pM
Number of SMRT Cells possible	?
Recommended Immobilization Time	N/A
Recommended Pre-extension Time	Pre-extension not recommended
Will Pre-extension be Used?	[✓ YES] NO
Pre-extension Time	min
Warnings	
Actions	COPY REMOVE

8. Select the **Sequencing Primer** to use for this run from the list. (You can also enter the part number for the Sequencing Primer; the Sequencing Primer name displays.)
9. Select the **Binding Kit** to use for this run from the list. (You can also enter the part number for the Binding Kit; the Binding Kit name displays.)
10. Specify what type of **loading** to use. (MagBead or Diffusion loading.)
11. Specify if this is an Iso-Seq experiment.
12. (**Optional**) Select the **Internal Control** to use for this run from the list. (You can also enter the Internal Control part number; the Internal Control name displays.) Pacific Biosciences **highly** recommends using the Internal Control to help distinguish between sample quality and instrument issues in the event of suboptimal sequencing performance. (**Note**: PacBio **requires** the use of the Sequel Internal Control for consumables to be eligible for reimbursement consideration.)
13. Specify whether to **Clean up** the sample (using AMPure PB beads, a MicroSpin column, or Loading Clean-up Beads, depending on insert size) to remove excess primer/polymerase. This results in higher quality data. Note that the Loading Clean-up Beads are **not** equivalent to the MagBeads used in MagBead loading. If AMPure PB beads or

MicroSpin column cleanup is used, you can enter an estimated yield for this step, in percent.

14. Clicking **Yes** for **AMPure Cleanup** specifies that the AMPure PB Bead cleanup protocol should be displayed instead of either the Loading Clean-up Bead or MicroSpin Column protocols. You can then enter the AMPure Cleanup Estimated Yield, in percent.
15. Enter the number of SMRT Cells to bind.
16. Specify the on-plate concentration, in pM.
17. Specify if **Pre-extension** is to be used, and the Pre-extension time, in minutes. This initiates the sequencing reaction prior to data acquisition. After the specified time, the sequencing reagents are removed from the SMRT Cell and replenished with fresh reagents, and data acquisition starts. This feature is useful for short insert sizes (such as amplicons or sheared genomic DNA libraries ≤10 kb) and will provide a significant increase in read length.
    Note: The Pre-extension feature is **not** compatible with Sequencing Kit v1.2 and v1.2.1. If these are used, the run will abort.
18. If either MicroSpin Column or AMPure Cleanup are selected, note that additional data entry is needed to generate the final steps of the protocol. The concentration and volume of the sample will need to be measured and their values input both immediately prior to and following the purification step.
19. Do one of the following:
    - Click **Copy** to start a new sample using the information entered. Then, edit specific fields for each sample.
    - Click **Remove** to delete the current calculation.
    - Click the **+ New Sample** button at the top of the screen to start a new, empty sample.
20. To **print** the calculation(s) and instructions, use the browser's Print command (**Ctrl-P**).

Editing or Printing Calculations

1. On the **Sample Setup** screen, select one or more calculation names.
2. Click **Edit/Print**. (**Note**: If the samples use different versions of chemistry or different magnetic bead protocols, a warning message displays.)
3. Edit the sample(s) as necessary.
4. To print the calculation(s), use the browser's **Print** command (Ctrl-P).

Deleting Calculations

1. On the **Sample Setup** screen, select one or more calculation names to delete.
2. Click **Delete**.

Run Design

Use SMRT Link's **Run Design** module to create, edit, or import Run Designs. A **Run Design** specifies:

- The samples, reagents, and SMRT Cells to include in the sequencing run.
- The run parameters such as movie time and loading to use for the sample.

The Run Design then becomes available from the **Sequel Instrument Control Software (ICS)**, which is the instrument touchscreen software used to select a Run Design, load the instrument, and then start the run.

Run Designs created in SMRT Link are accessible from **all** Sequel Systems linked to the same SMRT Link server

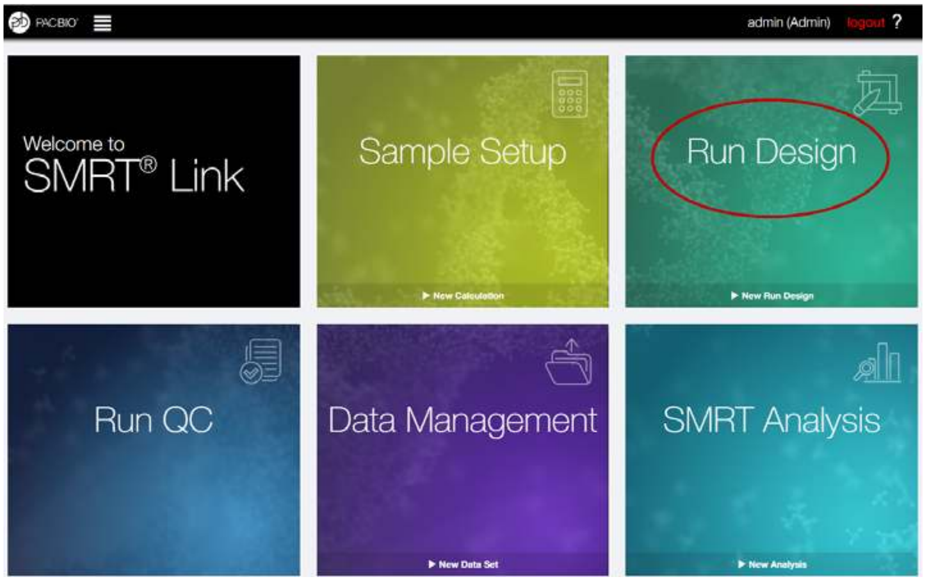
- If you are using SMRT Link with a PacBio RS II, use **RS Remote** software to create Run Designs.

SMRT Link includes two different ways to create a Run Design:

1. Use SMRT Link's **Run Design** module to create a new Run Design.
2. Create a CSV file, then import it using SMRT Link's **Run Design** module.

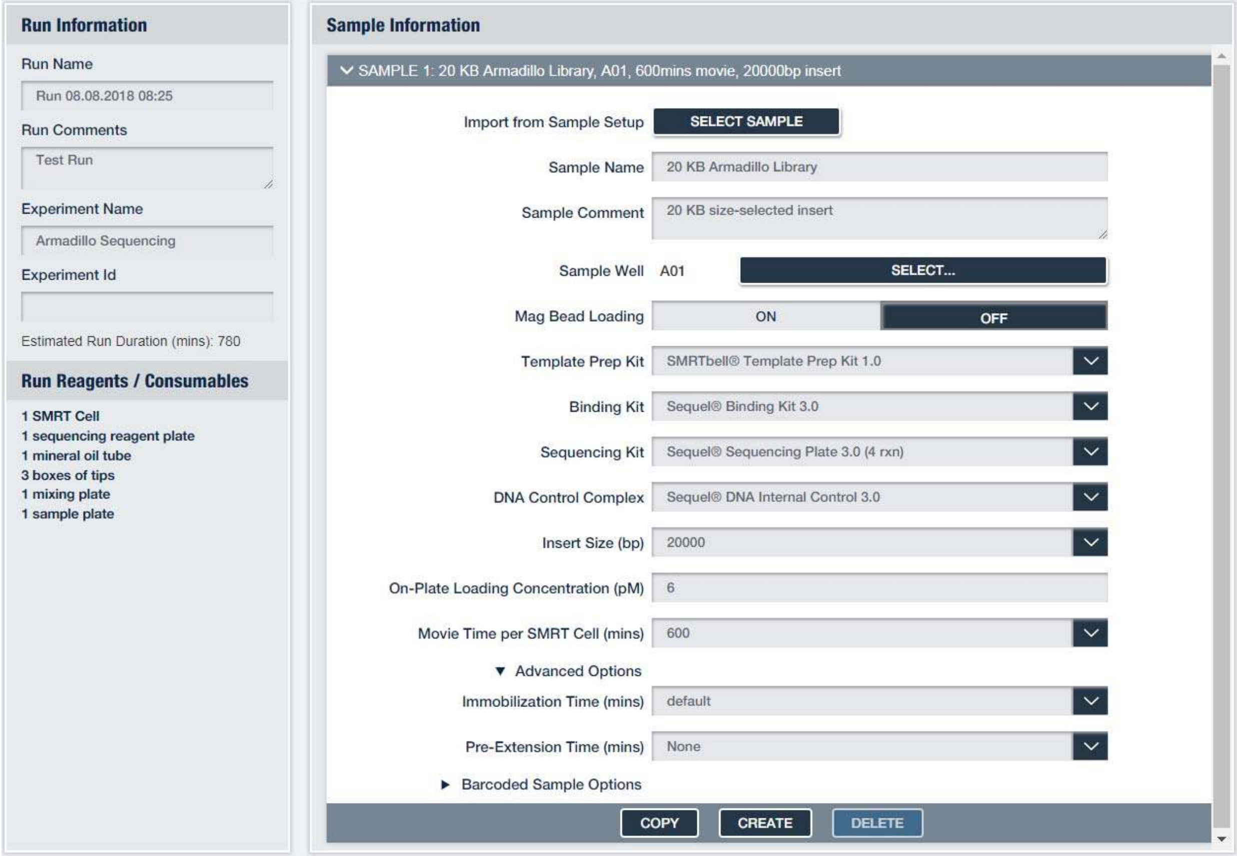
Note: To create a run design, **either** use the Run Design screen, **or** import a CSV file. Do **not** mix the two methods.

Creating a New Run Design



Note: For steps 12-15, you can also enter or scan kit or DNA Control Complex barcode numbers. If the barcode is invalid, "Invalid barcode" displays.

1. Access SMRT Link using the Chrome web browser.
2. Select **Run Design**.
3. Runs Designs can be sorted and searched for:
   - To sort Run Designs, click a **column title**.
   - To search for a Run Design, enter a unique search string into the **Search** field.
4. To initiate a new Run Design, click **New Run Design**.



Run Information

Run Name: Run 08.08.2018 08:25

Run Comments: Test Run

Experiment Name: Armadillo Sequencing

Experiment Id:

Estimated Run Duration (mins): 780

Run Reagents / Consumables

1 SMRT Cell
1 sequencing reagent plate
1 mineral oil tube
3 boxes of tips
1 mixing plate
1 sample plate

Sample Information

SAMPLE 1: 20 KB Armadillo Library, A01, 600mins movie, 20000bp insert

Import from Sample Setup: SELECT SAMPLE

Sample Name: 20 KB Armadillo Library

Sample Comment: 20 KB size-selected insert

Sample Well: A01 SELECT...

Mag Bead Loading: ON [✓ OFF]

Template Prep Kit: SMRTbell® Template Prep Kit 1.0

Binding Kit: Sequel® Binding Kit 3.0

Sequencing Kit: Sequel® Sequencing Plate 3.0 (4 rxn)

DNA Control Complex: Sequel® DNA Internal Control 3.0

Insert Size (bp): 20000

On-Plate Loading Concentration (pM): 6

Movie Time per SMRT Cell (mins): 600

▼ Advanced Options

Immobilization Time (mins): default

Pre-Extension Time (mins): None

▶ Barcoded Sample Options

5. (**Optional**) Click **Select Sample** to import information from a previously created Sample Setup entry. The following fields will be auto-populated as appropriate:
   - Sample Name
   - Mag Bead Loading
   - Binding Kit
   - DNA Control Complex
   - On-Plate Loading Concentration
6. Enter a **Run Name**. (The software creates a new Run Name based on the current date and time; you can edit the name as needed.)

7. (**Optional**) Enter **Run Comments**, **Experiment Name**, and **Experiment ID** as needed. (**Note**: Experiment ID **must** be alphanumeric.)
8. **(Optional)** Enter a **Sample Name**.
9. (**Optional**) Enter **Sample Comments**.
10. Specify the well position used for this sample: Click **Select...** and choose a plate position.
11. Specify whether to use **MagBead** Loading. This immobilizes SMRTbell templates into the ZMWs on the SMRT Cell using MagBeads. When **OFF** is selected, a diffusion run will be performed.
12. Select a **Template Prep Kit** from the list, or type in a kit part number.
13. Select a **Binding Kit** from the list, or type in a kit part number.
14. Select a **Sequencing Kit** from the list, or type in a kit part number.
    - **Note**: If the Sequencing or Binding kit is **incompatible**, an error message displays indicating the obsolete chemistry, and the run is **prevented** from proceeding.
15. (**Optional**) Select a **DNA Control Complex** from the list. Pacific Biosciences **highly** recommends using the Internal Control to help distinguish between sample quality and instrument issues in the event of suboptimal sequencing performance. **(Note**: PacBio **requires** the use of the Sequel Internal Control for consumables to be eligible for reimbursement consideration.)
16. Specify an **Insert Size**, ranging from 500 to 40,000 base pairs. (The Insert Size is the length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters. This matches the average insert size for the sample; the size range boundaries are described in the library preparation protocol and in the **Quick Reference Card - Diffusion Loading and Pre-Extension Time Recommendations for the Sequel System** document.)
17. Specify the **On-plate loading concentration**, in pM.
18. Specify the **Movie time** (**collection time)** per SMRT Cell: 30, 120, 240, 360, 480, 600, 900, or 1200 minutes. **Note**: Movie times greater than 600 minutes require the use of the **SMRT Cell 1M LR** part.
19. (**Optional**) Click **Advanced Options**, then specify the length of time (60, 120 or 240 minutes) for **immobilization** of SMRTbell templates.
    - For **MagBead Loading**, this is the length of time the SMRT Cell is at the MagBead station and the magnet moves the MagBead-bound SMRTbell template across the SMRT Cell to immobilize SMRTbell templates into the ZMWs.
    - For **diffusion**, this is the length of time the SMRT Cell is at the Cell Prep Station to allow diffusion of SMRTbell templates into the ZMWs.
    - PacBio **highly recommends** using the default immobilization time of 120 minutes for **both** loading types.
20. (**Optional**) Click **Advanced Options**, then specify the length of pre-extension time. This initiates the sequencing reaction prior to data acquisition. After the specified time, the sequencing reagents are removed from the SMRT Cell and replenished with fresh reagents, and data acquisition starts. This feature is useful for short inserts (such as ≤10 kb) and provides a significant increase in read length.

Note: This is **not** compatible with Sequencing Kit v1.2 and v1.2.1. If these are used, the run will abort.

21. (**Optional**) If you are using **barcoded samples**, see “Step 1: Specify the Barcode Setup and Sample Names in a Run Design” on page 110 for instructions. For details on secondary analysis of barcoded samples, see “Demultiplex Barcodes Application” on page 61.
22. Sample options:

- Click **Copy**. This starts a new sample, using the values entered in the first sample.
- Click **Create**. This starts a new, empty sample.
- Click **Delete**. This deletes the current sample.

23. After filling in all the samples, click **Save -** this saves the entire Run Design. The new Run Design displays on the main Run Design page.
24. Click **View Summary** to view a table summarizing the entire Run Design. The Run Design file is now imported and available for selection in Sequel ICS on the instrument.

Creating a Run Design by Importing a CSV File

On a remote workstation, open the sample CSV file included with the installation.

To obtain the sample CSV file: Go to the home page URL for your site’s SMRT Link installation and replace the last part of the URL (`#/welcome`) with `docs/xsd-datamodels/run-record/RunDesignTemplate.csv`. Press the Enter key, and the CSV file then downloads to your local machine.

1. Update the CSV file as necessary for the Run Design. (See the definitions of the Run Design attributes in the table below.)
2. Save the edited CSV file.
3. Import the file into Sequel ICS using SMRT Link. To do so, first access SMRT Link using the Chrome web browser.
4. Select **Run Design**.
5. Click **Import Run Design**.
6. Select the saved CSV file designed for the run and click **Open**. The file is now imported and available for selection in Sequel ICS on the instrument.

Run Design Attribute	Required	Description
Experiment Name	No	Enter any ASCII string. Defaults to Run Name. **Example**: Standard_Edna.1

<table>
<tr><th>Run Design Attribute</th><th>Required</th><th>Description</th></tr>
<tr><td><b>Experiment Id</b></td><td>No</td><td>Enter a valid experiment ID. <b>Example</b>: 325/3250057
• Experiment IDs <b>cannot</b> contain the following characters: <, >, :, ", \, |, ?, *, or).
• Experiment IDs <b>cannot</b> start or end with a / and <b>cannot</b> have two adjacent / characters, such as //.
• Experiment IDs <b>cannot</b> contain spaces.
• Specifically, Experiment IDs <b>cannot</b> satisfy the regular expressions:
/[<>:"\\|?*]/g
/(?:^\/)|\/\/|(?:\/$)/
/ /g</td></tr>
<tr><td><b>Experiment Description</b></td><td>No</td><td>Enter any ASCII string. Defaults to Run Description.
<b>Example</b>: 20170530_A6_VVnC_SampleSheet</td></tr>
<tr><td><b>Run Name</b></td><td><b>Yes</b></td><td>Run name <b>must</b> be entered for the first cell and will be applied to the remaining cells in the run. Can be any ASCII string.
<b>Example</b>: 20170530_A6_VVnC_SampleSheet</td></tr>
<tr><td><b>Run Description</b></td><td>No</td><td>Run description <b>must</b> be entered for the first cell and will be applied to the remaining cells in the run. Can be any ASCII string.
<b>Example</b>: ecoliK12_March2018</td></tr>
<tr><td><b>Well No.</b></td><td><b>Yes</b></td><td>Well No. must start with a letter A through H, and end in a number 01 through 12, i.e. A01 through H12.
It must satisfy the regular expression ``/^[A-H](?:0[1-9]|1[0-2])$/``
<b>Example</b>: A01</td></tr>
<tr><td><b>Sample Name</b></td><td><b>Yes</b></td><td>Enter any ASCII string.
<b>Example</b>: A6_3230046_A01_SB_ChemKitv2_8rxnKit</td></tr>
<tr><td><b>Collection Time</b></td><td><b>Yes</b></td><td>Enter a floating point number ≥ 1 and < 1200. Time is in minutes.
<b>Example</b>: 120</td></tr>
<tr><td><b>Sample Description</b></td><td>No</td><td>Enter any ASCII string.
<b>Example</b>: A6_3230046_A01_SB_BindKit_ChemKit</td></tr>
<tr><td><b>Insert Size</b></td><td><b>Yes</b></td><td>Enter an integer > 9. Units are in base pairs. <b>Example</b>: 2000</td></tr>
<tr><td><b>On Plate Loading Concentration</b></td><td>No</td><td>Enter an integer. Units are in pM. <b>Example</b>: 5</td></tr>
<tr><td><b>Size Selection</b></td><td>No</td><td>Enter a Boolean value. Boolean details below. Default is False.</td></tr>
<tr><td><b>DNA Template Prep Kit Box Barcode</b></td><td><b>Yes</b></td><td>Enter or scan a valid template prep barcode. (See barcode details below.)
<b>Working example:</b> DM1117100259100111716</td></tr>
<tr><td><b>Binding Kit Box Barcode</b></td><td><b>Yes</b></td><td>Enter or scan the binding kit barcode. (See barcode details below.)
<b>Working example:</b> DM1117100862200111716</td></tr>
<tr><td><b>Sequencing Kit Box Barcode</b></td><td><b>Yes</b></td><td>Enter or scan the sequencing kit barcode. (See barcode details below.)
<b>Working example:</b> DM0001100861800123120</td></tr>
<tr><td><b>DNA Control Complex Box Barcode</b></td><td>No</td><td>Enter or scan the control DNA barcode. (See barcode details below.)
<b>Working example:</b> DM1234101084300123120</td></tr>
<tr><td><b>Automation Name</b></td><td>No</td><td>Enter diffusion, magbead (<b>not</b> case-sensitive) or a custom script.
A path can also be used, such as /path/to/my/script/my_script.py. The path will <b>not</b> be processed further, so if the full URI is required, it must be provided in the CSV, such as chemistry://path/to/my/script/my_script.py.</td></tr>
</table>

Run Design Attribute	Required	Description
Automation Parameters	No	To enable Pre-Extension time, enter the number of minutes. Example 90 minutes: `ExtensionTime=double:90|ExtendFirst=boolean:True` (**Note**: Leave blank when **not** using Pre-Extension time.)
Primary Analysis	No	Value = Default
Primary Analysis Parameters	No	• The parameters are separated by `|` characters. • Each item in the list follows the format: `[parameter name]=[parameter type]:[parameter value].` • The parameter names and types are **not** case-sensitive. Valid primary analysis parameters: • `Readout, MetricsVerbosity` • `CopyFileTrace, CopyFileBaz` • `CopyFileDarkFrame, CopyStatsH5` Valid parameter types: • `String` • `Int32, UInt32` • `Double, Single` • `Boolean, DateTime` In this example, specify that Trace and H5 files are **not** copied from the instrument: `CopyFileTrace=boolean:false|CopyStatsH5=boolean:false`
Sample is Barcoded	No	Enter a boolean value. Boolean details below. Set to TRUE for a barcoded run. **Example**: TRUE
Barcode Set	No	Must be a UUID for a Barcode Set present in the database. To find the UUID: Click **Data Management > View or Import Sequence Data > Barcode Set**. Click the Barcode Set of interest, then click the small triangle to view additional details. **Example**: dad4949d-f637-0979-b5d1-9777eff62008
Same Barcodes on Both Ends of Sequence	No	Enter a boolean value. Boolean details below. Specify TRUE if symmetric, FALSE if asymmetric. **Example**: TRUE
Barcoded Sample Name	No	Enter barcoded sample names that follow the guidelines below. **Example**: `lbc1--lbc1;sample1|lbc2--lbc2;sample2|lbc3--lbc3;sample3` • Put the entire mapping of barcode name-to-sample name into **one** spreadsheet cell. • Use double hyphens (`--`) to separate the 2 barcodes of each pair, and semicolon to separate the barcode pair and sample name from the next ones. • Barcoded sample names are included in a list separated by `|` characters. Each item in the list follows the format `[barcode name];[biosample name]` • The barcode names **must** be contained within the specified Barcode Set. • A given barcode name **cannot** appear more than once in the list. • The bio sample names can be any ASCII string but **cannot** contain the field separators `|` and `;`. The bio sample names **cannot** be longer than 40 characters. • A maximum of 384 barcodes is permitted per sample.

CSV File General Requirements

- Each line in the CSV file represents 1 sample.
- The CSV file may **only** contain ASCII characters. Specifically, it must satisfy the regular expression /`^[\x00-\x7F]*$/g`

Boolean Values

- Valid boolean values for **true** are: `true, t, yes,` or `y`
- Valid boolean values for **false** are: `false, f, no,` or `n`
- Boolean values are **not** case-sensitive.

Kit Barcode Requirements

Kit barcodes are composed of three parts which are used to make a single string:

1. Lot Number (Example: `DM1234`)
2. Part Number (Example: `100-619-300`)
3. Expiration Date (Example: `2020-12-31`)

For the above example, the full kit barcode would be: `DM1234100619300123120`.

Each kit **must** have a valid Part Number and **cannot** be obsolete. The list of kits can be found through a services endpoint such as:

```
[server name]:[services port number]/smrt-link/bundles/chemistry-pb/active/files/
definitions%2FPacBioAutomationConstraints.xml
```

This services endpoint will list, for each kit, the part numbers (`PartNumber`) and whether it is obsolete (`IsObsolete`).

Dates must also be valid, meaning they must exist on the Gregorian calendar.

Editing or Deleting Run Designs

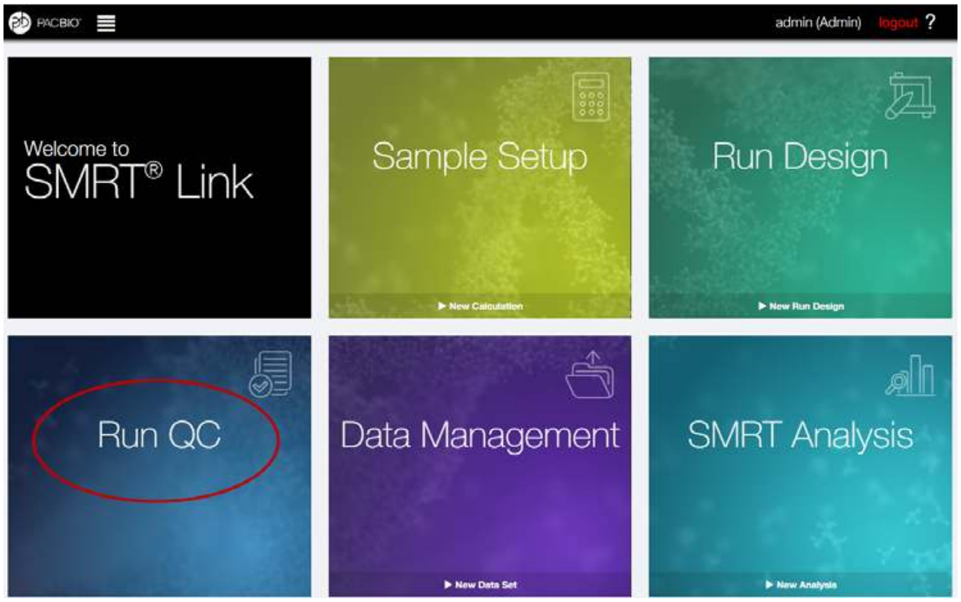
1. On the Home page, select **Run Design**.
2. Click the name of the Run Design to edit or delete.
3. (**Optional**) Click **View Summary** to view a table summarizing the entire Run Design.
4. (**Optional**) Click **Delete** to delete the current Run Design.
5. (**Optional**) Edit any of the fields.
6. Click **Save**.

Run QC

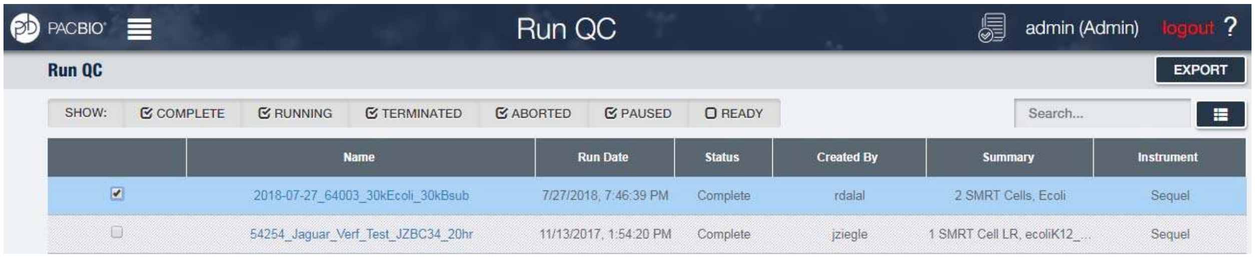
Use SMRT Link's **Run QC** module to monitor performance trends and perform run QC remotely.

Metrics can be reviewed in the Run QC module. **All** Sequel Systems connected to SMRT Link can be reviewed using Run QC.

- If you are using SMRT Link with a PacBio RS II, use **RS Dashboard** software.



1. Access SMRT Link using the Chrome web browser.
2. Select **Run QC**.



Run QC

EXPORT

SHOW: ☑ COMPLETE ☑ RUNNING ☑ TERMINATED ☑ ABORTED ☑ PAUSED □ READY

Search...

	Name	Run Date	Status	Created By	Summary	Instrument
☑	2018-07-27_64003_30kEcoli_30kBsub	7/27/2018, 7:46:39 PM	Complete	rdalal	2 SMRT Cells, Ecoli	Sequel
□	54254_Jaguar_Verf_Test_JZBC34_20hr	11/13/2017, 1:54:20 PM	Complete	jziegle	1 SMRT Cell LR, ecoliK12_...	Sequel

3. Runs can be sorted, searched for, and filtered:
   - To sort runs, click a **column title**.
   - To search for a run, enter a unique search string into the **Search** field.
   - To specify the status of the runs to display, click one or more of the following buttons: **Complete**, **Running**, **Terminated**, **Aborted**, **Paused**, and/or **Ready**.
4. To **export** Run QC data in CSV format: Select one or more runs in the table, then click **Export**.

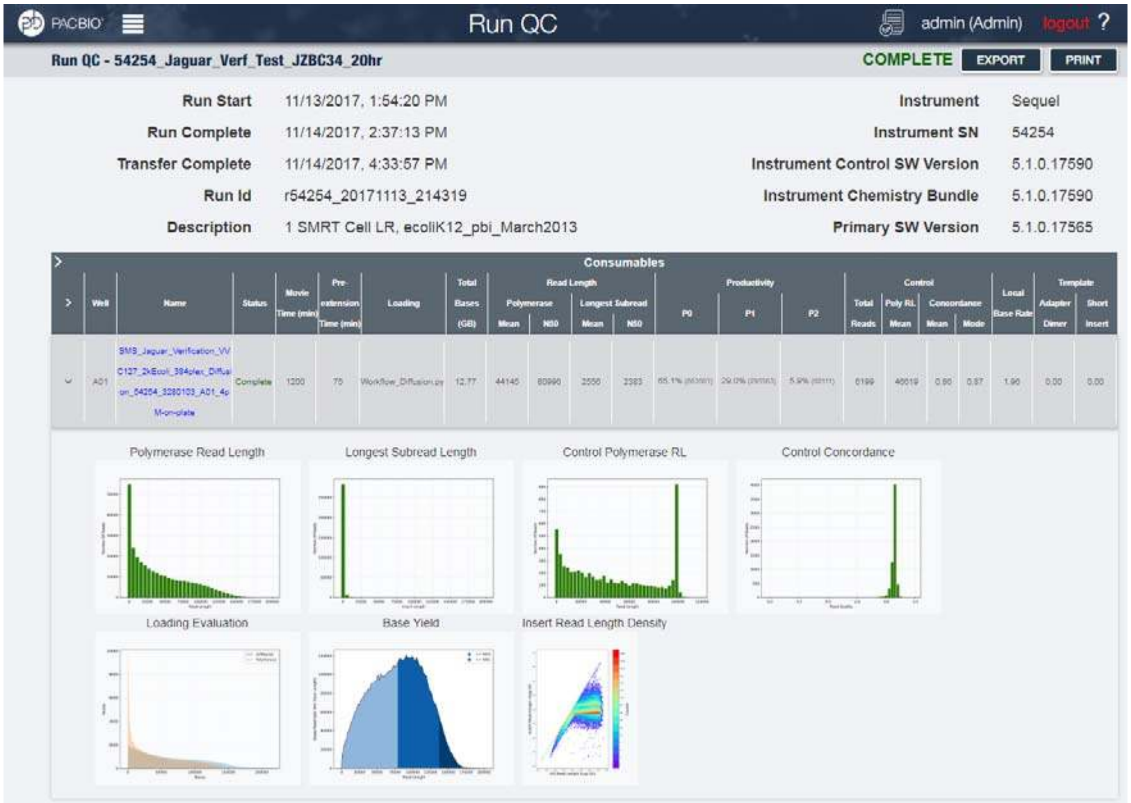
Default Table Fields

- **Name**: A list of all runs for the instruments connected to SMRT Link. Click a run name to view more detailed information on the Individual Run Page.
- **Run Date**: The date and time when the run was started.
- **Status**: The current status of the run. Can be one of the following: Running, Complete, Failed, Terminated, or Unknown.
- **Created By**: The name of the user who created the run.
- **Summary**: A description of the run.
- **Instrument**: The name of the instrument.

Additional Table Fields

- **Instrument SN**: The instrument's serial number.
- **Instrument SW**: The version of Sequel Instrument Control Software (ICS) installed on the instrument.
- **Primary Analysis SW**: The version of Primary Analysis software installed on the instrument.
- **Run ID**: An internally-generated ID number identifying the run.
- **UUID**: Another internally-generated ID number identifying the run.
- **Completion Date**: The date and time the run was completed.
- **Transferred Date**: The date and time the run results were transferred to the network.
- **Total Cells**: The total number of SMRT Cells used in the run.
- **Completed Cells**: The number of SMRT Cells that generated data for the run.
- **Failed Cells**: The number of SMRT Cells that failed to generate data during the run.

5. Click the **Run name** of interest. Following are the fields and metrics displayed.



Run QC - 54254_Jaguar_Verf_Test_JZBC34_20hr

Run Start	11/13/2017, 1:54:20 PM	Instrument	Sequel
Run Complete	11/14/2017, 2:37:13 PM	Instrument SN	54254
Transfer Complete	11/14/2017, 4:33:57 PM	Instrument Control SW Version	5.1.0.17590
Run Id	r54254_20171113_214319	Instrument Chemistry Bundle	5.1.0.17590
Description	1 SMRT Cell LR, ecoliK12_pbi_March2013	Primary SW Version	5.1.0.17565

> Consumables

>	Well	Name	Status	Movie Time (min)	Pre-extension Time (min)	Loading	Total Bases (GB)	Read Length				Productivity			Control				Local Base Rate	Template	
								Polymerase		Longest Subread		P0	P1	P2	Total Reads	Poly RL	Concordance			Adapter Dimer	Short Insert
								Mean	N50	Mean	N50					Mean	Mean	Mode			
⌄	A01	SMS_Jaguar_Verification_VVC127_2kEcoli_384plex_Diffusion_54254_3280103_A01_4pM-on-plate	Complete	1200	75	Workflow_Diffusion.py	12.77	44145	80990	2556	2383	65.1% ([illegible])	29.0% ([illegible])	5.9% ([illegible])	6199	40619	0.90	0.87	1.96	0.00	0.00

- **Run Start**: The date and time when the run was started.
- **Run Complete**: The date and time the run was completed.
- **Transfer Complete**: The date and time that the run data was successfully transferred to the network.
- **Run ID:** An internally-generated ID number identifying the run.
- **Description**: The description, as defined when creating the run.
- **Instrument**: The name of the instrument.
- **Instrument SN**: The serial number of the instrument.
- **Instrument Control SW Version**: The versions of Sequel Instrument Control Software (ICS) installed on the instrument.
- **Instrument Chemistry Bundle:** The version of the Chemistry Bundle installed on the instrument when the run was initiated.
- **Primary SW Version**: The versions of Primary Analysis software installed on the instrument.

Consumables Table

6. Click the **>** arrow at the top of the **Consumables** table to see the consumable type, lot number, and expiration date.

- **Well**: The ID of an individual well used for this sample.

- **Name**: The sample name, as defined when creating the run. Clicking the name will take you to the corresponding entry in the **Data Management** module.
- **Status**: The current collection status for the SMRT Cell. This can be one of the following: **Complete**, **Collecting**, **Aborted**, **Failed**, **In Progress**, or **Pending**.
- **Movie Time (min)**: The length of the movie associated with this SMRT Cell.
- **Pre-extension Time (min)**: The pre-extension time used in the collection, if any.
- **Loading**: Whether MagBeads loading or Diffusion loading was used for the run.
- **Total Bases (GB)**: Calculated by multiplying the number of **productive** (P1) ZMWs by the mean polymerase read length; displayed in Gigabases.
- **Read Length**: Polymerase reads are trimmed to the high quality region and include bases from adapters, as well as potentially multiple passes around a SMRTbell template.
  - **Polymerase Mean:** The mean high-quality read length of all polymerase reads. The value includes bases from adapters as well as multiple passes around a circular template.
  - **Polymerase N50**: 50% of all read bases came from polymerase reads longer than this value.
  - **Longest Subread Mean**: The mean subread length, considering only the longest subread from each ZMW.
  - **Longest Subread N50:** 50% of all read bases came from subreads longer than this value when considering only the longest subread from each ZMW.
- **Productivity**
  - **P0**: Empty ZMW; no signal detected.
  - **P1**: ZMW with a high quality read detected.
  - **P2**: Other, signal detected but no high quality read.
- **Control**
  - **Total Reads**: The number of control reads obtained.
  - **Poly RL Mean**: The mean polymerase read length of the control reads.
  - **Concordance Mean**: The average concordance (agreement) between the control raw reads and the control reference sequence.
  - **Concordance Mode**: The median concordance (agreement) between the control raw reads and the control reference sequence.
- **Local Base Rate**: The average base incorporation rate, excluding polymerase pausing events.
- **Template**
  - **Adapter Dimer:** The % of pre-filter ZMWs which have observed inserts of 0-10 bp. These are likely adapter dimers.
  - **Short Insert**: The % of pre-filter ZMWs which have observed inserts of 11-100 bp. These are likely short fragment contamination.

7. Click the **>** arrow to expand rows to view plots for each SMRT Cell where data was successfully transferred. Clicking on an individual plot displays an expanded view. These plots include:
   - **Polymerase Read Length**: Plots the number of reads against the polymerase read length.
   - **Longest Subread Length**: Plots the number of reads against the insert length.
   - **Control Polymerase RL**: Displays the Polymerase read length distribution of the control, if used.
   - **Control Concordance**: Maps control reads against the known control reference and reports the concordance.
   - **Loading Evaluation**: Displays the length distribution of unfiltered and filtered (polymerase) reads.
   - **Base Yield Density**: Displays the number of bases sequenced in the collection, according to the length of the read in which they were observed. Values displayed are per unit of read length (i.e. the base yield density) and are averaged over 2000 bp windows to gently smooth the data. Regions of the graph corresponding to bases found in reads longer than the N50 and N95 values are shaded in medium and dark blue, respectively.
   - **Insert Read Length Density:** Displays a density plot of reads, hexagonally binned according to their HQ Read Length and median subread length. For very large insert libraries, most reads consist of a single subread and will fall along the diagonal. For shorter inserts, subreads will be shorter than the HQ read length, and will appear as horizontal features. This plot is useful for quickly visualizing aspects of library quality, including insert size distributions, reads terminating at adapters, and missing adapters.

Data Management

Use the **Data Management** module to:

- Create and manage Data Sets,
- View Data Set information,
- Create and manage Projects,
- View, import, or delete sequence, reference, and barcode data.

What is a Data Set?

Data Sets are logical collections of sequencing data (basecalled or analyzed) that are analyzed together, and for which reports are created.

Data Sets:

- Help to **organize** and **manage** basecalled and analyzed data. This is especially valuable when dealing with large amounts of data collected from different sequencing runs from one or more instruments.
- Are the way that sequence data is represented and manipulated in SMRT Link. Sequence data from the instrument is organized in Data Sets. Data from **each** cell or collection is a Data Set.
- Can be used to collect data and summarize performance characteristics, such as data throughput, while an experiment is in progress.
- Can be used to generate reports about data, and to exchange reports with collaborators and customers.
- Can be used to start an analysis. (See "Starting an Analysis from a Data Set" on page 24 for details.)

A Data Set can contain sequencing data from **one** or **multiple** SMRT Cells or collections from different runs, or a portion of a collection with multiplexed samples.

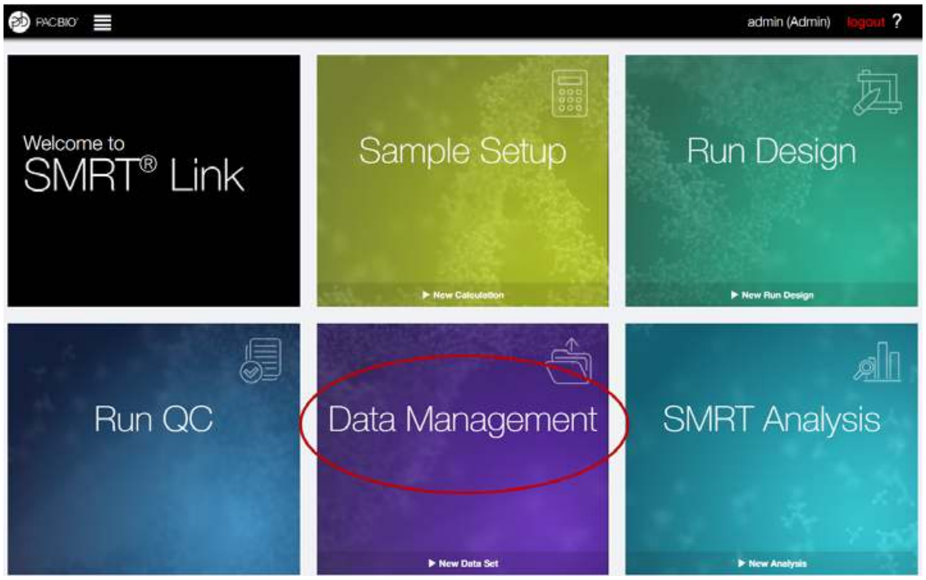
In SMRT Link, movies, cells/collections, context names and well samples are all in one-to-one relationships and can be used more or less interchangeably. That is, a Data Set from a single cell or collection will also be from a single collection derived from DNA from a single well sample. Data produced by SMRT Cells, however, can be used by **multiple** Data Sets, so that data may have a many-to-one relationship with collections.

Some Data Sets can contain **basecalled** data, while others can contain **analyzed** data:

- **Basecalled data** Data Sets contain sequence data from one or multiple cells or collections.
- **Analyzed data** Data Sets contain data from previous analyse(s).

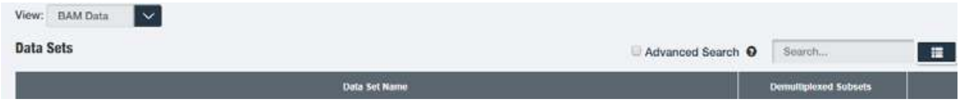
Elements within a Data Set are of the same data type, typically subreads or consensus reads, in aligned or unaligned format.

Creating a Data Set

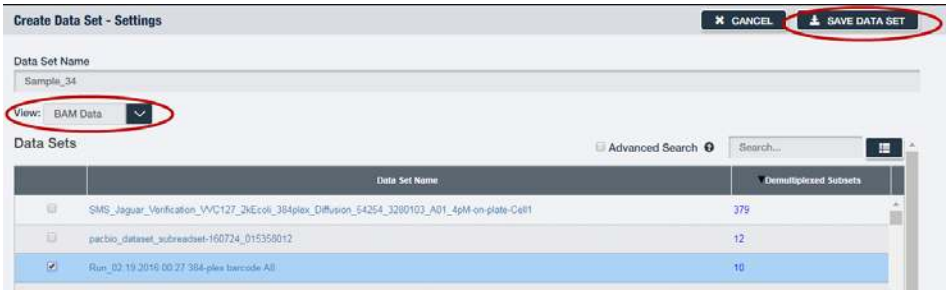


1. Access SMRT Link using the Chrome web browser.
2. Select **Data Management**.
3. Data Sets can be sorted and searched for:

- To sort Data Sets, click a **column title**.
- To search for a Data Set **locally** or **remotely**, use the Search function. See "Appendix C - Data Set/Analysis Search" on page 134 for details.



4. Click **+ Create Data Set**.
5. Enter a name for the new Data Set.



Create Data Set - Settings

Data Set Name: Sample_34

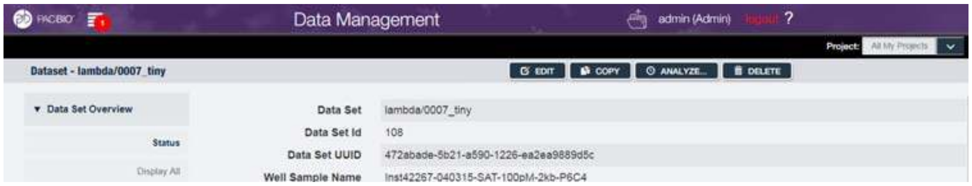
View: BAM Data

Data Sets

	Data Set Name	Demultiplexed Subsets
□	SMS_Jaguar_Verification_VVC127_2kEcoli_384plex_Diffusion_54254_3280103_A01_4pM-on-plate-Cell1	379
□	pacbio_dataset_subreadset-160724_015358012	12
☑	Run_02 19 2016 00 27 384-plex barcode All	10

6. Select the type of data to include in the new Data Set: **BAM-format** data generated on the Sequel System, or **PacBio RS II** data.
7. In the **Data Sets** box, select one or more sets of sequence data.

8. (**Optional**) Use the Search function to search for specific Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
9. (**Optional**) Click the **Filter reads by subread length** box beneath the Data Set list. Enter the minimum and/or maximum subread length to retain in the new Data Set.
10. Click **Save Data Set**. The new Data Set becomes available for starting analyses, viewing, or generating reports.
11. After the Data Set is created, click its name in the main Data Management screen to see reports, metrics, and charts describing the data included in the Data Set.



Data Set lambda/0007_tiny
Data Set Id 108
Data Set UUID 472abade-5b21-a590-1226-ea2ea9889d5c
Well Sample Name Inst42267-040315-SAT-100pM-2kb-P6C4

Starting an Analysis from a Data Set

From the Data Set page, an analysis can be started using the Data Set.

1. Click **Analyze**, then follow the instructions starting at Step 8 of “Creating and Starting an Analysis” on page 32.

Note: To analyze PacBio RS II data, first convert it using the **Convert RS to BAM** application. After converting the data to the BAM file format, analyzing the data using **all** the other applications is enabled.

Copying a Data Set

1. On the Home Page, select **Data Management**.
2. Specify what type of Data Set to copy: BAM-format data (generated by the Sequel System) or data generated on a PacBio RS II System.
3. (**Optional**) Use the Search function to search for Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
4. Click the name of the Data Set to copy. The Data Set Reports page displays.
5. Click **Copy**. The main Data Management page displays; the new Data Set has (**copy**) appended to the name.

Deleting a Data Set

Note: SMRT Link's Delete Data Set functionality **only** deletes the Data Set from the SMRT Link interface, **not** from your server.

It is good practice to export Data Sets you no longer need to a backup server, then delete them from SMRT Link. This frees up space in the SMRT Link interface.

1. On the Home Page, select **Data Management**.
2. Specify what type of Data Set to delete: BAM-format data (generated by the Sequel System) or data generated on a PacBio RS II System.
3. (**Optional**) Use the Search function to search for Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
4. Click the name of the Data Set to delete.
5. Click **Delete**. Note that this **only** deletes the Data Set from the SMRT Link interface; **not** from your server. To delete the Data Set from your server, **manually** delete it from the disk.
6. Click **Yes**. The Data Set is no longer available from SMRT Link.

Exporting Data Sets

Use this procedure to compress one or more Data Sets in ZIP format and export them to a directory of your choice.

1. On the Home Page, select **Data Management**.
2. Click **Export Data Sets**.
3. Specify what type of data to export: BAM-format data (generated by the Sequel System) or data generated on a PacBio RS II System.
4. Select one or more sets of sequence data to export. (Multiple Data Sets are combined as one ZIP file for export.)
5. (**Optional**) Use the Search function to search for Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
6. Click **Export Selected Data**.
7. Navigate to the export destination directory.
8. (**Optional**) Click **Delete data set files after export** to delete the Data Set(s) you selected from the SMRT Link installation. (Exporting, then deleting, Data Sets is useful for archiving Data Sets you no longer need.)
9. Click **Export**.

Viewing Data Set Information

1. On the Home Page, select **Data Management**.
2. Click **View > BAM Data** or **View > RS II Data**.
3. (**Optional**) Use the Search function to search for Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
4. Click the name of the Data Set to see information about the sequence data included in the Data Set, as well as QC reports.

Editing Data Set Information

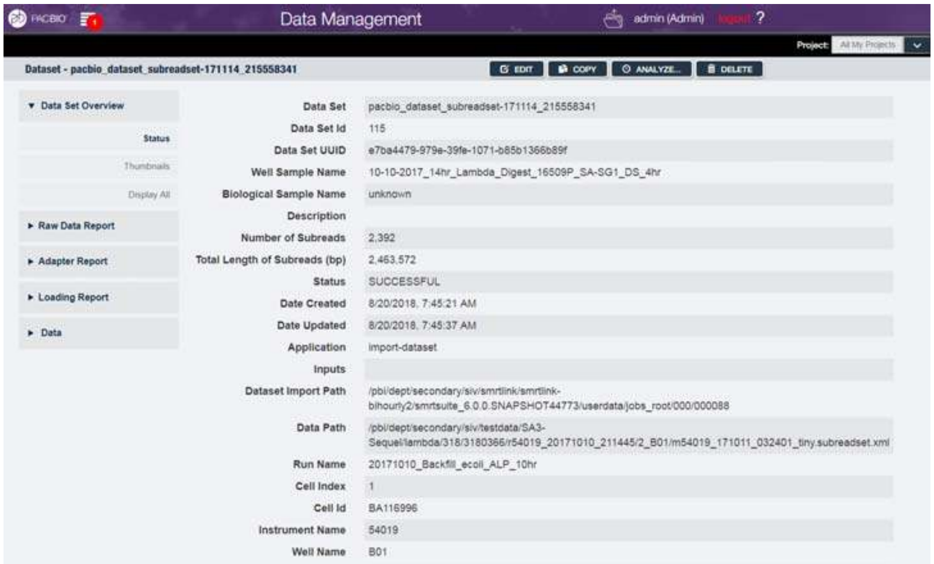
1. On the Home Page, select **Data Management**.
2. (**Optional**) Use the Search function to search for Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
3. Click the name of the Data Set to edit.
4. Click **Edit**.

5. Edit the Well Sample Name and/or the Biological Sample Name.
6. Click **Save**.

Data Set QC Reports

The Data Set QC Reports are generated when you create a new Data Set or update the data contained in existing Data Sets. These reports are designed to provide all relevant information about the data included in the Data Set as it comes from the instrument prior to data analysis, and are useful for data QC purposes.

The following reports are generated by default, for data generated by a **Sequel System**:



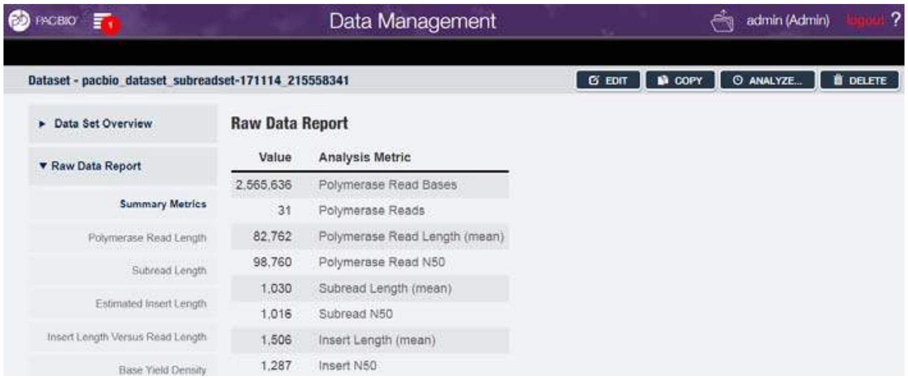
Dataset - pacbio_dataset_subreadset-171114_215558341

Data Set	pacbio_dataset_subreadset-171114_215558341
Data Set Id	115
Data Set UUID	e7ba4479-979e-39fe-1071-b85b1366b89f
Well Sample Name	10-10-2017_14hr_Lambda_Digest_16509P_SA-SG1_DS_4hr
Biological Sample Name	unknown
Description	
Number of Subreads	2,392
Total Length of Subreads (bp)	2,463,572
Status	SUCCESSFUL
Date Created	8/20/2018, 7:45:21 AM
Date Updated	8/20/2018, 7:45:37 AM
Application	import-dataset
Inputs	
Dataset Import Path	/pbi/dept/secondary/siv/smrtlink/smrtlink-bihourly2/smrtsuite_6.0.0.SNAPSHOT44773/userdata/jobs_root/000/000088
Data Path	/pbi/dept/secondary/siv/testdata/SA3-Sequel/lambda/318/3180366/r54019_20171010_211445/2_B01/m54019_171011_032401_tiny.subreadset.xml
Run Name	20171010_Backfill_ecoli_ALP_10hr
Cell Index	1
Cell Id	BA116996
Instrument Name	54019
Well Name	B01

Data Set Overview > Status

Displays the following information about the Data Set:

- The Data Set Name, ID, description, and when it was created.
- The number of subreads and their total length in base pairs.
- The names of the run and instrument that generated the data.
- The names of the well and sample used to generate the data.
- The name of the analysis application used to generate the data.



Dataset - pacbio_dataset_subreadset-171114_215558341

Raw Data Report

Value	Analysis Metric
2,565,636	Polymerase Read Bases
31	Polymerase Reads
82,762	Polymerase Read Length (mean)
98,760	Polymerase Read N50
1,030	Subread Length (mean)
1,016	Subread N50
1,506	Insert Length (mean)
1,287	Insert N50

Raw Data Report > Summary Metrics

- **Polymerase Read Bases**: The total number of polymerase read bases in the Data Set.
- **Polymerase Reads**: The total number of polymerase reads in the Data Set.
- **Polymerase Read Length (mean)**: The mean read length of all polymerase reads in the Data Set.
- **Polymerase Read N50**: The read length at which 50% of all the bases in the Data Set are in polymerase reads longer than, or equal to, this value.
- **Insert Length (mean)**: The mean length of all the inserts in the Data Set.

The following report is generated when an `.sts.xml` file is **not** available, when the other reports fail to generate, or the data was generated by a **PacBio RS II** instrument:

Simple Dataset Report > Summary Metrics

- **Total Length**: The total length (in base pairs) of all the sequences in the Data Set.
- **Num Records**: The total number of sequence records in the Data Set.

What is a Project?

- Projects are collections of Data Sets, and can be used to restrict access to Data Sets to a subset of SMRT Link users.
- By default, **all** Data Sets and data belong to the **General** project and are accessible to **all** users of SMRT Link.
- **Any** SMRT Link user can create a Project and be the owner. Projects must have an owner, and can have **multiple** owners.
- Unless a Project is shared with other SMRT Link users, it is **only** accessible by the owner.
- Only owner(s) can delete a Project; deleting a Project deletes **all** Data Sets and analyses that are part of the Project.

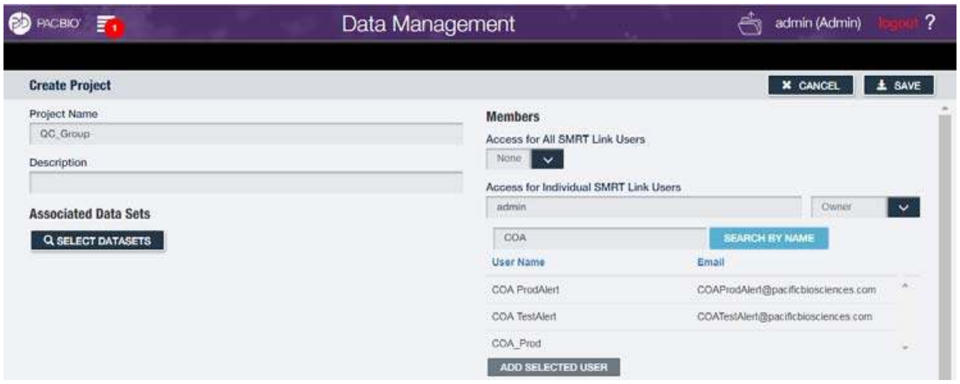
Projects include:

- One or more Data Sets and associated Quality Control information.
- One or more analysis results and the associated Data Sets, including information for all analysis parameters and reference sequence (if used).

Data Sets and Projects

- Once created, a Data Set **always** belongs to at least **one** project; either the **General** project or another project the user has access to.
- Data Sets can be associated with **multiple** projects.
- The data represented by a Data Set can be copied into **multiple** projects using the Data Management Report page **Copy** button. Any changes made to a particular copy of a Data Set affect **only** that copy, **not** any other copies in other Projects. If a Data Set is to be used with multiple Projects, Pacific Biosciences recommends that you make a **separate copy** for each Project.

Creating a Project



Create Project

Project Name: QC_Group

Description

Associated Data Sets

Members

Access for All SMRT Link Users: None

Access for Individual SMRT Link Users: admin — Owner

COA

User Name	Email
COA ProdAlert	COAProdAlert@pacificbiosciences.com
COA TestAlert	COATestAlert@pacificbiosciences.com
COA_Prod	

1. Access SMRT Link using the Chrome web browser.
2. Select **Data Management**.
3. Click **+ Create Project**.
4. Enter a name for the new project.
5. (**Optional**) Enter a description for the project.
6. Click **Select Data Sets** and select one or more sets of sequence data to associate with the project.
7. (**Optional**) Share the Project with other SMRT Link users. (**Note**: Unless a Project is shared, it is **only** visible to the owner.) There are two ways to specify who can access the new Project:
   - For **all** SMRT Link Users: **None -** No one can access the project other than the user who created it; **View -** Everyone can view the Project; **View/Edit**: Everyone can see and edit the Project.
   - **To give individual users access to the Project**: Enter a user name and click **Search By Name**. Choose **Owner**, **View**, or **View/Edit**, then click **Add Selected User**.

- **Notes:** A) Projects can have **multiple** owners. B) If you enable **all** SMRT Link users to have **View/Edit** access, you cannot change an individual member's access to **View**.

8. Click **Save**. The new project becomes available for SMRT Link users who now have access.

Editing a Project

1. On the Home Page, select **Data Management**.

2. Click **View > Projects**.
3. Projects can be sorted and searched for:
   - To sort Projects, click a **column title**.
   - To search for a Project, enter a unique search string into the **Search** field.
4. Click the name of the project to edit.
   - (**Optional**) Edit the project name or description.
   - (**Optional**) Delete a Data Set associated with the project by clicking **X**.
   - (**Optional**) Add one or more sets of sequence data to the project by clicking **Select Data Sets** and selecting one or more Data Sets to add.
   - (**Optional**) Delete members: Click **X** next to a Project member's name to delete that user from the Project.
   - (**Optional**) Add members: See Step 7 in **Creating a Project**.
   - Click **Save**. The modified Project is saved.

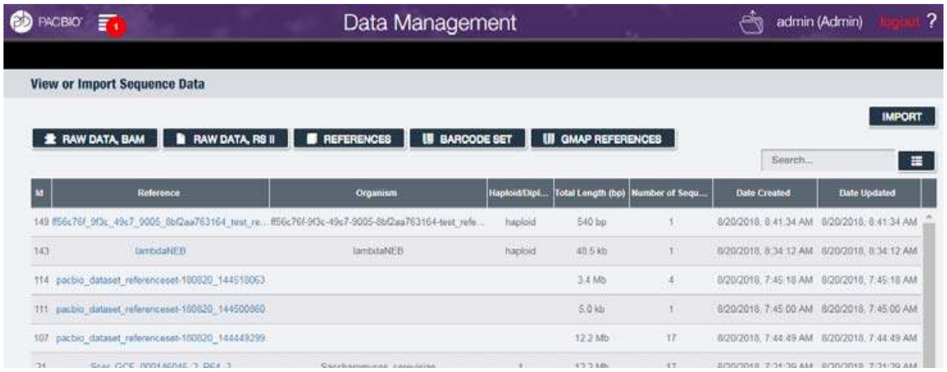
Deleting a Project

1. On the Home Page, select **Data Management**.
2. Click **View > Projects**.
3. Click the name of the project to delete.
4. Click **Delete**. (This deletes **all** Data Sets and analyses that are part of the Project from SMRT Link, but **not** from the server.)

Viewing/Deleting Sequence, Reference and Barcode Data

On the **View or Import Sequence Data** page, information on available sequence data, reference sequence files and barcode files can be accessed. Note that this data can only be **viewed** or **deleted**, not modified.

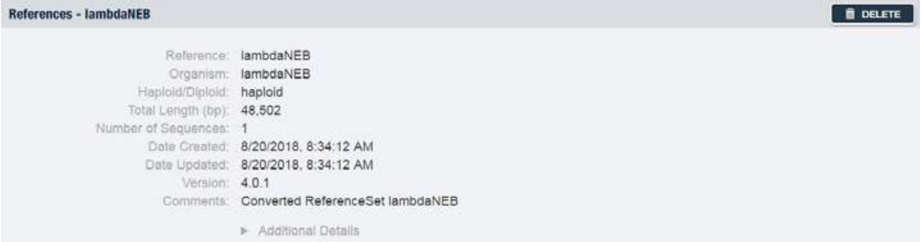
1. On the Home Page, select **Data Management**.
2. Click **View or Import Sequence Data**.
3. Click the appropriate button to see information on:
   - **Raw Data, BAM:** Sequencing data in BAM format; the Sequel System generates data in this format.
   - **Raw Data, RS II**: Sequencing data generated by a PacBio RS II instrument.
   - **References**: Reference sequence FASTA files used when creating certain analyses.
   - **Barcode Set**: Barcodes from barcoded samples.
   - **GMAP Reference**: Reference sequence files used for creating Iso-Seq 1 with Mapping and Iso-Seq 3 with Mapping analyses.
4. Click on the link for the data selected to see details about the sequence data, reference sequence, or barcode file.



View or Import Sequence Data

Id	Reference	Organism	Haploid/Dipl...	Total Length (bp)	Number of Sequ...	Date Created	Date Updated
149	ff56c76f_9f3c_49c7_9005_8bf2aa763164_test_re...	ff56c76f-9f3c-49c7-9005-8bf2aa763164-test_refe...	haploid	540 bp	1	8/20/2018, 8:41:34 AM	8/20/2018, 8:41:34 AM
143	lambdaNEB	lambdaNEB	haploid	48.5 kb	1	8/20/2018, 8:34:12 AM	8/20/2018, 8:34:12 AM
114	pacbio_dataset_referenceset-180820_144518063			3.4 Mb	4	8/20/2018, 7:45:18 AM	8/20/2018, 7:45:18 AM
111	pacbio_dataset_referenceset-180820_144500860			5.0 kb	1	8/20/2018, 7:45:00 AM	8/20/2018, 7:45:00 AM
107	pacbio_dataset_referenceset-180820_144449299			12.2 Mb	17	8/20/2018, 7:44:49 AM	8/20/2018, 7:44:49 AM
21	Scer_GCF_000146045_2_R64_2	Saccharomyces_cerevisiae	1	12.2 Mb	17	8/20/2018, 7:21:29 AM	8/20/2018, 7:21:29 AM

5. Click the name of the sequence, reference or barcode of interest. Details for that sequence, reference sequence file or barcode file display.



References - lambdaNEB

Reference: lambdaNEB
Organism: lambdaNEB
Haploid/Diploid: haploid
Total Length (bp): 48,502
Number of Sequences: 1
Date Created: 8/20/2018, 8:34:12 AM
Date Updated: 8/20/2018, 8:34:12 AM
Version: 4.0.1
Comments: Converted ReferenceSet lambdaNEB

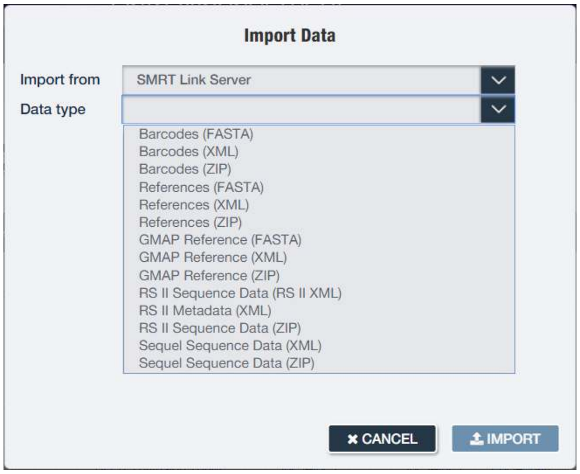
6. (**Optional**) To delete the sequence data, reference sequence, or barcode file, click **Delete**.

Importing Data

Note: If your Sequel instrument is linked to the SMRT Link software during the instrument installation, your Sequel System data will be **automatically** imported into SMRT Link.

Several types of sequence data, as well as barcode files, can be imported for use in SMRT Link.

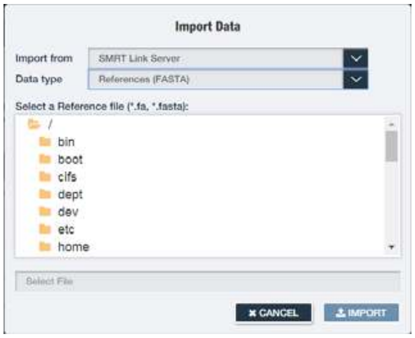
1. On the Home Page, select **Data Management**.
2. Click **View or Import Sequence Data**.
3. Click **Import**, then select whether to import data from the **SMRT Link Server**, or from a **Local File System**. (**Note**: **Only** references and barcodes are available if you select **Local File System**.)



Import from: SMRT Link Server

Data type:

4. Select the data type to import:
   - **Barcodes**: FASTA (`.fa` or `.fasta`), XML (`.barcodeset.xml`), or ZIP files containing barcodes.
   - **References**: FASTA (`.fa` or `.fasta`), XML (`.referenceSet.xml`), or ZIP files containing a reference sequence for use in starting analyses.
   - **GMAP References:** FASTA (`.fa` or `.fasta`), XML (`.xml`), or ZIP files containing a GMAP reference sequence for use in starting Iso-Seq 1 with Mapping or Iso-Seq 3 with Mapping analyses. **Note**: The XML file (`gmapreferenceset.xml`) points to a GMAP reference FASTA file and indices.
   - **RS II Sequence Data**: RS II XML (`.hdfsubreadset.xml`), RS II Metadata XML (`.metadata.xml`), or RS II Sequence Data (ZIP). These files contain information about PacBio RS II sequence data from a **single cell**.
   - **Sequel Sequence Data (XML)**: XML file (`.subreadset.xml`) containing information about Sequel sequence data, such as paths to the BAM files.
   - **Sequel Sequence Data (ZIP):** ZIP file containing information about Sequel sequence data, such as paths to the BAM files.
5. Navigate to the appropriate file and click **Import**. The sequence data, reference, or barcodes are imported and becomes available in SMRT Link.



Import from: SMRT Link Server

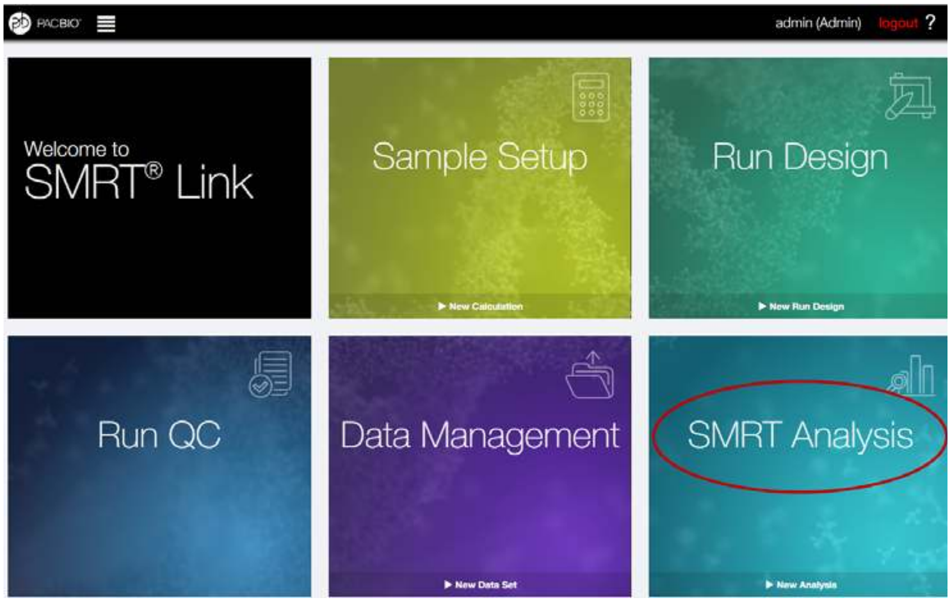
Data type: References (FASTA)

Select a Reference file (*.fa, *.fasta):

SMRT® Analysis

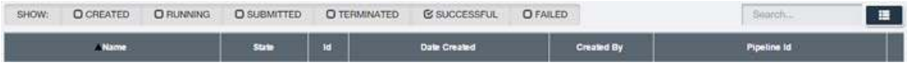
After a run has completed, use SMRT Link's **SMRT Analysis** module to perform **secondary analysis** of the data.

Creating and Starting an Analysis



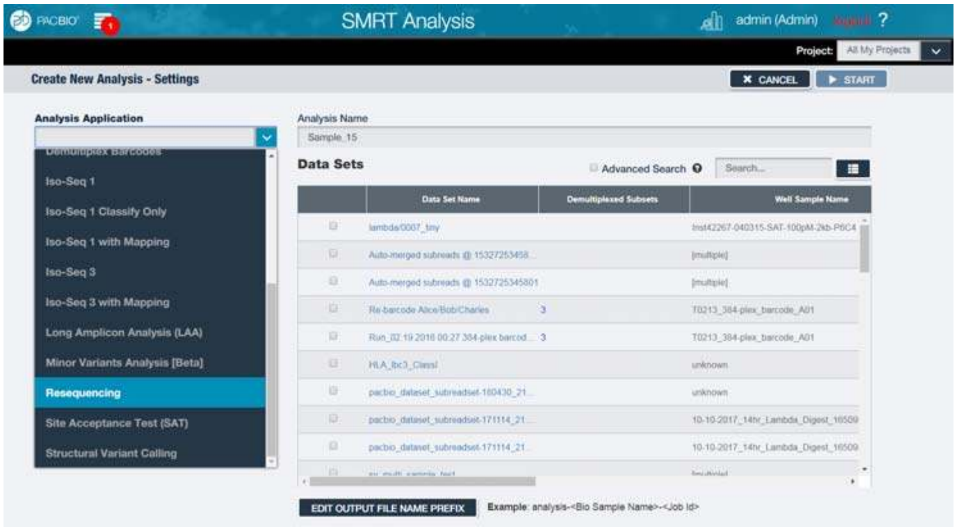
1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
3. Analyses can be sorted, searched for, and filtered:

- To sort analyses, click a **column title**.
- To search for an analysis, enter a unique search string into the **Search** field.
- To filter the list of analyses to display, click one or more of the following buttons: **Created**, **Running**, **Submitted**, **Terminated**, **Successful**, and/or **Failed**.



4. Click **+ Create New Analysis**.

5. Select a secondary analysis application from the droplist.



Create New Analysis - Settings

Analysis Name: Sample_15

Data Sets

☐ Advanced Search

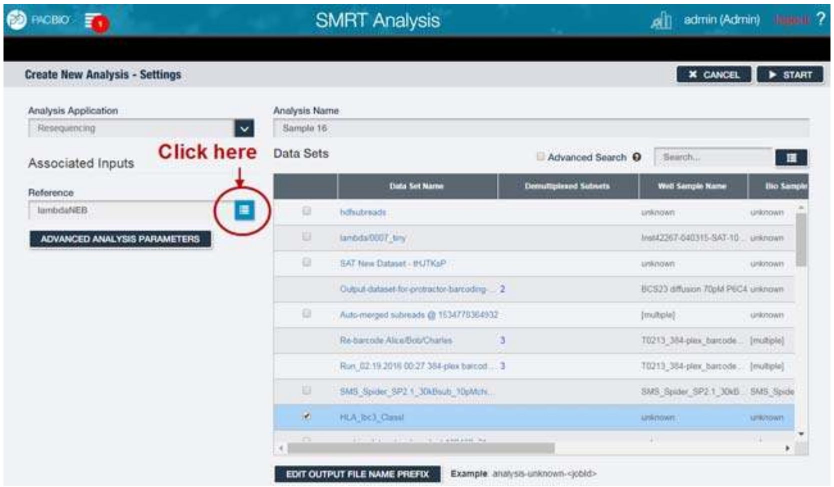
	Data Set Name	Demultiplexed Subsets	Well Sample Name
☐	lambda/0007_tiny		Inst42267-040315-SAT-100pM-2kb-P6C4
☐	Auto-merged subreads @ 15327253458...		[multiple]
☐	Auto-merged subreads @ 1532725345801		[multiple]
☐	Re-barcode Alice/Bob/Charles	3	T0213_384-plex_barcode_A01
☐	Run_02.19.2016 00:27 384-plex barcod...	3	T0213_384-plex_barcode_A01
☐	HLA_bc3_ClassI		unknown
☐	pacbio_dataset_subreadset-180430_21...		unknown
☐	pacbio_dataset_subreadset-171114_21...		10-10-2017_14hr_Lambda_Digest_16509
☐	pacbio_dataset_subreadset-171114_21...		10-10-2017_14hr_Lambda_Digest_16509

Example: analysis-<Bio Sample Name>-<Job Id>

- Each of the secondary analysis applications has **required parameters** that are displayed. Please review the default values shown.
- Secondary analysis applications also have **advanced analysis parameters**. These are set to default values, and need only be changed when analyzing data generated in non-standard experimental conditions.

The **Resequencing** application will be used as an example. This application maps sequencing reads against a reference sequence. It identifies the consensus sequence and performs variant detection.

6. Click the **Reference** field and select a reference sequence from the dialog. (The reference sequences available in SMRT Link and displayed in the dialog were imported into SMRT Analysis. See "Importing Data" on page 30 for details.)



Create New Analysis - Settings

Analysis Application: Resequencing

Analysis Name: Sample 16

Associated Inputs

Reference: lambdaNEB

Data Sets

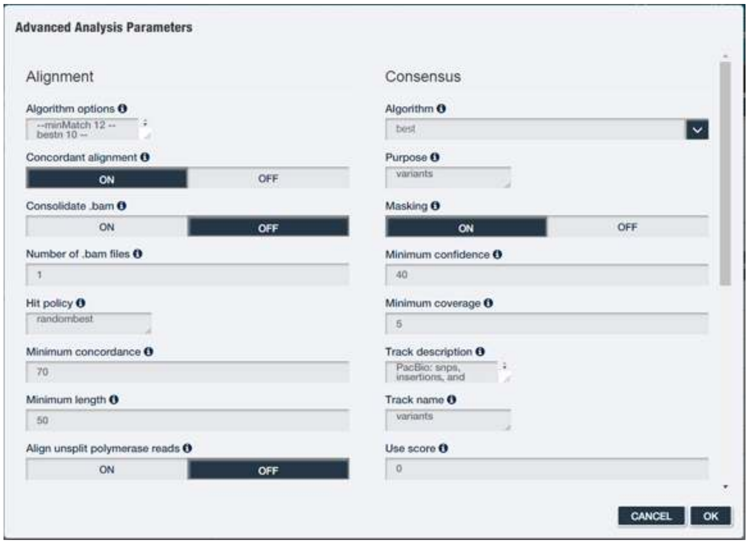
☐ Advanced Search

	Data Set Name	Demultiplexed Subsets	Well Sample Name	Bio Sample
☐	hdfsubreads		unknown	unknown
☐	lambda/0007_tiny		Inst42267-040315-SAT-10...	unknown
☐	SAT New Dataset - tHJTKsP		unknown	unknown
	Output-dataset-for-protractor-barcoding-...	2	BCS23 diffusion 70pM P6C4	unknown
☐	Auto-merged subreads @ 1534778364932		[multiple]	unknown
	Re-barcode Alice/Bob/Charles	3	T0213_384-plex_barcode...	[multiple]
	Run_02.19.2016 00:27 384-plex barcod...	3	T0213_384-plex_barcode...	[multiple]
☐	SMS_Spider_SP2.1_30kBsub_10pMchi...		SMS_Spider_SP2.1_30kB...	SMS_Spide
☑	HLA_bc3_ClassI		unknown	unknown

Example analysis-unknown-<jobId>

7. **(Optional)** Click **Advanced Analysis Parameters** and specify the values of the parameters you would like to change. Click **OK** when finished.

   - To see information about parameters for **all** secondary analysis applications provided by Pacific Biosciences, see “PacBio® Secondary Analysis Applications” on page 39.



Advanced Analysis Parameters

Alignment

Algorithm options
--minMatch 12 --bestn 10 --

Concordant alignment
[✓ ON] OFF

Consolidate .bam
ON [✓ OFF]

Number of .bam files
1

Hit policy
randombest

Minimum concordance
70

Minimum length
50

Align unsplit polymerase reads
ON [✓ OFF]

Consensus

Algorithm
best

Purpose
variants

Masking
[✓ ON] OFF

Minimum confidence
40

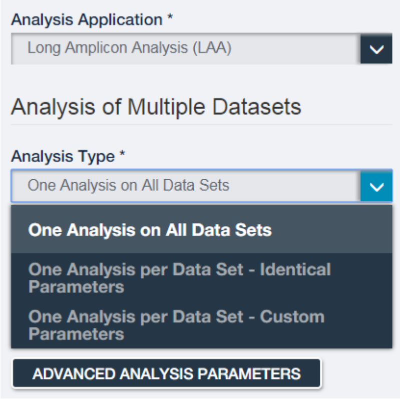
Minimum coverage
5

Track description
PacBio: snps, insertions, and

Track name
variants

Use score
0

8. Enter a **name** for the analysis.
9. In the **Data Sets** box, select one or more sets of data to be analyzed together. (Use the Search function to search for Data Sets **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.)
10. (**Optional**) If you selected **multiple** Data Sets as input for the analysis, additional **Analysis Type** options become available. You can select from the following options:



Analysis Application *
Long Amplicon Analysis (LAA)

Analysis of Multiple Datasets

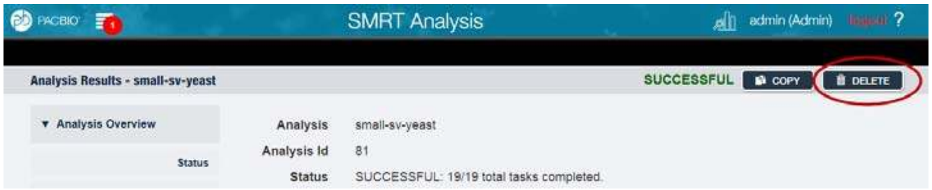
Analysis Type *
One Analysis on All Data Sets

One Analysis on All Data Sets
One Analysis per Data Set - Identical Parameters
One Analysis per Data Set - Custom Parameters
ADVANCED ANALYSIS PARAMETERS

- **One Analysis on All Data Sets:** Runs **one** analysis using all the selected Data Sets.

- **One Analysis per Data Set - Identical Parameters:** Runs one separate analysis for **each** of the selected Data Sets, using the **same** parameters. Optionally click **Advanced Analysis Parameters** and modify parameters.
- **One Analysis per Data Set - Custom Parameters:** Runs one separate analysis for **each** of the selected Data Sets, using **different** parameters for each Data Set. Click **Advanced Analysis Parameters** and modify parameters. Then click **Start and Create Next**. You can then specify parameters for each of the included Data Sets.

11. (**Optional**) Specify prefixe(s) used in the names of files generated by the analysis and displayed on the **Files Download** page. Example: **Run Name** can be included in the name of every file generated by the analysis. Click **Edit Output File Name Prefix**, check the type(s) of information to append to the file names, then click **Save**.
12. **(Optional)** Specify the **Project** that this analysis will be associated with. **All Projects**: This analysis will be visible to **all** SMRT Link users. **All My Projects**: This analysis will be visible **only** to users who have access to Projects that you are a member of. **Note**: To **restrict** access to an analysis, make sure to select a project limited to the appropriate users **before** starting the analysis.
13. Click **Start**.
14. Click **SMRT Analysis** to navigate to the main SMRT Analysis screen. There, the status of the analysis displays. When the analysis has **completed**, click on its name - reports are available for the completed analysis.
15. (**Optional**) To **delete** the completed analysis: Click **Delete**, then click **Yes** in the confirmation dialog. The analysis is deleted.



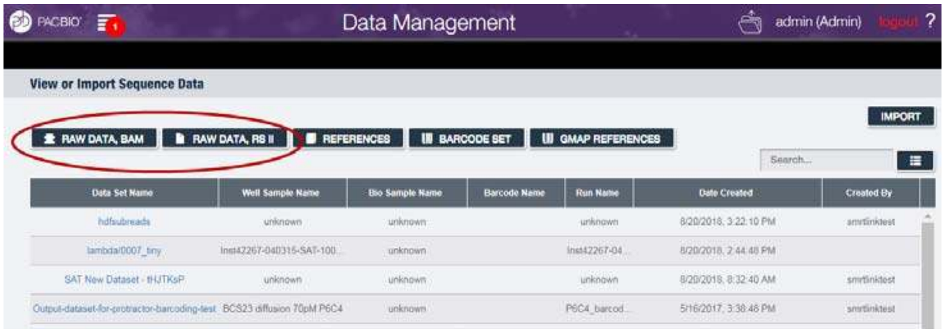
Analysis Results - small-sv-yeast

Analysis small-sv-yeast
Analysis Id 81
Status SUCCESSFUL: 19/19 total tasks completed.

Starting an Analysis After Viewing Sequence Data

An analysis can be started by **first** viewing information about specific sequence data:

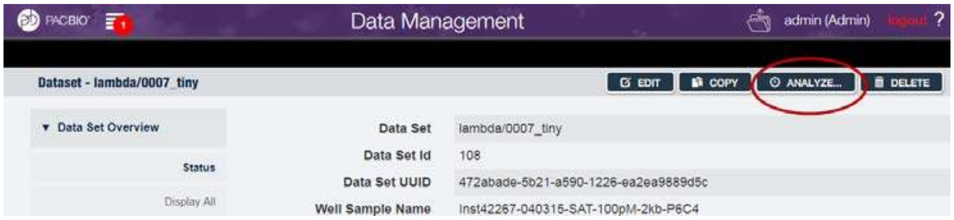
1. On the Home Page, select **Data Management**.
2. Click **View or Import Sequence Data**.
3. Click **Raw Data, BAM** or **Raw Data, RS II** to specify what type of sequence data to view. (To **narrow** the list of sequence data, enter the first few characters of the desired Data Set name in the Search field.)



View or Import Sequence Data

Data Set Name	Well Sample Name	Bio Sample Name	Barcode Name	Run Name	Date Created	Created By
hdfsubreads	unknown	unknown		unknown	8/20/2018, 3:22:10 PM	smrtlinktest
lambda/0007_tiny	Inst42267-040315-SAT-100...	unknown		Inst42267-04...	8/20/2018, 2:44:48 PM	
SAT New Dataset - tHJTKsP	unknown	unknown		unknown	8/20/2018, 8:32:40 AM	smrtlinktest
Output-dataset-for-protractor-barcoding-test	BCS23 diffusion 70pM P6C4	unknown		P6C4_barcod...	5/16/2017, 3:38:46 PM	smrtlinktest

4. In the **Data Set Name** column, click the name of the sequence data of interest. Details for the selected sequence data display.



Dataset - lambda/0007_tiny

Data Set lambda/0007_tiny

Data Set Id 108

Data Set UUID 472abade-5b21-a590-1226-ea2ea9889d5c

Well Sample Name Inst42267-040315-SAT-100pM-2kb-P6C4

5. To **start** an analysis using this sequence data, click **Analyze**, then follow the instructions starting at Step 5 of "Creating and Starting an Analysis" on page 32.

Canceling a Running Analysis

1. On the Home Page, select **SMRT Analysis**.
2. Click the **Running** button to see only running analyses.
3. Select a currently-running analysis to cancel.
4. Click **Cancel**.
5. Click **Yes** in the confirmation dialog.

Viewing Analysis Results

1. On the Home Page, select **SMRT Analysis**. You see a list of **all** analyses.
2. (**Optional**) Click the **Successful** button to see only successfully-completed analyses.
3. (**Optional**) Use the Search function to search for specific analyses **locally** or **remotely**. See "Appendix C - Data Set/Analysis Search" on page 134 for details.
4. Click the analysis link of interest.
5. Click **Analysis Overview > Status** to see analysis information status, including which application was used for the analysis, and the inputs used.
6. Click Analysis **Overview > Thumbnails** or **Display All** to view thumbnails of the reports generated for the analysis. Click a thumbnail to see a larger image.
7. Depending on the application used for the analysis, different analysis-specific reports are available.

- For mapping applications **only**: Click **Mapping Report > Summary Metrics** to see an overall summary of the mapping data.
- For information on the reports and data files produced by analysis applications, see “PacBio® Secondary Analysis Applications” on page 39.

8. To download data files created by SMRT Link: You can use these data files as input for further downstream processing, pass on to collaborators, or upload to public genome sites. Click **Data > File Downloads**, then click the appropriate file. The file is downloaded according to your browser settings.
9. To view analysis log details: Click **Data > SMRT Link Log**.
10. To visualize the secondary analysis results, click the SMRT View button. SMRT View is a genome browser that displays sequencing data generated by the Sequel System. (See “Visualizing Data Using SMRT® View” on page 117 for details.)

Copying and Running an Existing Analysis

If you run very similar analyses, you can **copy** an existing analysis, rename it, optionally modify one or more parameters, then run it.

1. On the Home Page, select **SMRT Analysis**. You see a list of **all** analyses.
2. (**Optional**) Click the **Successful** button to see only successfully-completed analyses.
3. (**Optional**) Use the Search function to search for specific analyses **locally** or **remotely**. See “Appendix C - Data Set/Analysis Search” on page 134 for details.
4. Click the analysis link of interest.
5. Click **Copy** - this creates a copy of the analysis, named `Copy of <analysis name>`, using the same parameters.
6. Edit the name of the analysis.
7. (**Optional**) Edit any other parameter. See “PacBio® Secondary Analysis Applications” on page 39 for further details.
8. Click **Start**.

Exporting an Analysis

You can export the entire contents of an analysis directory, including the input sequence files, as a ZIP file. Afterwards, deleting the analysis saves room on the SMRT Link server; you can also later reimport the exported analysis into SMRT Link if necessary.

1. On the Home Page, select **SMRT Analysis**.
2. Click **Export Analysis**.
3. Select one or more analyses to export. This exports the entire contents of the analysis directory. To **also** export the input sequence data files associated with the analyses, select **Include Input Sequence Data**.
4. Click **Export Selected Data**.
5. Select the output directory for the analysis data and click **Export**.

Importing an Analysis

Note: You can **only** import an analysis that was created in SMRT Link, then exported.

1. On the Home Page, select **SMRT Analysis**.
2. Click **Import Analysis**.
3. Select a ZIP file containing the analysis to import.
4. Click **Import**. The analysis is imported and is available on the main SMRT Analysis page.

PacBio® Secondary Analysis Applications

Following are the secondary analysis applications provided with SMRT Analysis v6.0.0. Each application is described later, including all parameters and the reports and data files output by the application.

Assembly (HGAP 4)

- Generate *de novo* assemblies of genomes.
- See "Assembly (HGAP 4) Application" on page 41 for details.

Base Modification Detection

- Identify putative sites of base modification as well as common bacterial base modifications (6mA, 4mC).
- See "Base Modification Detection Application" on page 47 for details.

Base Modification and Motif Analysis

- Identify putative sites of base modification as well as common bacterial base modifications (6mA, 4mC), and then analyze the methyltransferase recognition motifs.
- See "Base Modification and Motif Analysis Application" on page 50 for details.

CCS Mapping

- Generate consensus sequences from single molecules, and map these consensus sequences to a user-provided reference sequence.
- See "CCS Mapping Application" on page 54 for details.

Circular Consensus Sequences (CCS)

- Identify consensus sequences for single molecules.
- See "Circular Consensus Sequences (CCS) Application" on page 58 for details.

Convert BAM to FASTX

- Convert sequence data in BAM file format to the FASTX file format.
- For **barcoded** runs, you must **first** run the **Demultiplex Barcodes** application to create BAM files **before** using this application.
- See "Convert BAM to FASTX Application" on page 60 for details.

Convert RS to BAM

- Convert sequence data generated on a PacBio RS II system in HDF5 file format to the BAM file format, compatible with SMRT Analysis v4.0.0 and later.
- See "Convert RS to BAM Application" on page 60 for details.

Demultiplex Barcodes

- Separate reads by barcode.
- See "Demultiplex Barcodes Application" on page 61 for details.

Iso-Seq® 1 Analysis, Iso-Seq® 3 Analysis

- Characterize full-length transcripts.
- See “Iso-Seq® 1 Analysis Application” on page 65 and “Iso-Seq® 3 Analysis Application” on page 69 for details.

Iso-Seq® 1 Analysis Classify Only

- This analysis includes only the **Classify** step of the Iso-Seq 1 Analysis application. Sequencing reads are classified into full length or non-full length reads.
- See “Iso-Seq® 1 Analysis Classify Only Application” on page 73 for details.

Iso-Seq® 1 Analysis with Mapping, Iso-Seq® 3 Analysis with Mapping

- Characterize full-length transcripts, then map the transcripts back to the reference genome. A GMAP reference genome is required for alignment.
- See “Iso-Seq® 1 Analysis with Mapping Application” on page 76 and “Iso-Seq® 3 Analysis with Mapping Application” on page 81 for details.

Long Amplicon Analysis (LAA)

- Identify phased consensus sequences from a heterogeneous pool of amplicons.
- See “Long Amplicon Analysis (LAA) Application” on page 87 for details.

Minor Variants Analysis [Beta]

- Identify and phase minor single nucleotide substitution variants in complex populations.
- See “Minor Variants Analysis Application [Beta]” on page 90 for details.

Resequencing

- Map sequencing reads against a reference sequence and identify variants.
- See “Resequencing Application” on page 97 for details.

Site Acceptance Test (SAT)

- Generate a report displaying instrument acceptance test metrics. (The application is designed **only** for analysis of Site Acceptance data.)
- See “Site Acceptance Test (SAT) Application” on page 102 for details.

Structural Variant Calling

- Identify structural variants (Default: ≥20 bp) in a sample or set of samples relative to a reference.
- See “Structural Variant Calling Application” on page 107 for details.

Assembly (HGAP 4) Application

Use this application (**H**ierarchical **G**enome **A**ssembly **P**rocess) to generate high quality *de novo* assemblies of genomes, using PacBio data.

- HGAP 4 includes pre-assembly, *de novo* assembly and assembly polishing steps.
- HGAP 4 uses Falcon for *de novo* assembly and Arrow for polishing.

Genome Length: (Required; Default = 5,000,000):

- The approximate number of base pairs expected in the genome. Other parameters are set automatically based on this value.

Parameters

Advanced Analysis Parameters	Default Value	Description
Minimum Subread Length	0	The minimum length of subreads to use in the assembly.
Filters to Add to the Data Set	rq ≥ 0.7	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Assembly	Default Value	Description
Aggressive Option	OFF	If **ON**, allows more overlaps to be detected and reported, which creates longer preads that go into assembly. This can be useful when a Data Set assembles poorly using the defaults, possibly due to lower quality input subreads. The default is **OFF** as this is not as well tested as the default options and may cause side-effects on larger, more complex genomes.
FALCON cfg Overrides	NONE	Allows PacBio Support engineers to override the configuration file generated from other options. This is a semicolon-separated list of KEY= VALUE pairs. Newline characters are accepted but ignored.
Seed Length Cutoff	-1	Only reads as long as this value will be used as seeds in the draft assembly. `-1` means this will be calculated automatically so that the total number of seed bases equals (`Genome Length` times `Seed Coverage`.)
Seed Coverage	30	A target value for the total number of "raw" postprimary reads, divided by the total number of seed reads. Valid values are `20` to `100`.

Advanced Analysis Parameters - Alignment	Default Value	Description
Minimum Concordance	70	The minimum required alignment concordance, in percent.
Concordant Alignment	OFF	Specify whether to map subreads of a ZMW to the same genomic location.
Align Unsplit Polymerase Reads	OFF	Do not split reads into subreads even if subread regions are available.
Minimum Length	50	The minimum required alignment length, in base pairs.

Advanced Analysis Parameters - Alignment	Default Value	Description
Hit Policy	randombest	Specify how to treat multiple hits: • **random**: Selects a random hit. • **all**: Selects all hits. • **allbest**: Selects all the best score hits. • **randombest**: Selects a random hit from all best score hits. • **leftmost**: Selects a hit which has the best score and the smallest mapping coordinate in any reference.
Algorithm Options		List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0`

Advanced Analysis Parameters - Consensus	Default Value	Description
Use Score	0	Specify the score to use in the display.
Minimum Confidence	40	The minimum confidence for a variant call to be output to the file `variants.gff`.
Track Description	NONE	Description to display in the header.
Track Name	variants	Name to display in the header.
Purpose	variants	Specify the run mode - `variants` or `coverage`.
Masking	ON	During the polish step, omit regions of reads that have low concordance with the template.
Algorithm	best	• **Quiver** is a variant-calling algorithm that operates on RS II data **only.** • **Arrow** is a more sophisticated algorithm that provides additional information about each read, allowing more accurate consensus calls. Arrow does **not** use the alignment provided by the mapper except for determining how to group reads together at the gross level. Arrow implicitly performs its own realignment, so it is highly sensitive to all variant types, including indels. • **Plurality** is a very simple variant-calling algorithm which does **not** perform any local realignment. It is heavily biased by the alignment produced by the mapper, and it is **insensitive** at detecting indels. • **Best** is the best algorithm based on the data provided.
Minimum Coverage	5	The minimum site coverage that must be achieved for variant calls and consensus to be calculated for a site.

Advanced Analysis Parameters - Reports	Default Value	Description
Number of Regions	1000	Specify the number of genome regions in the summary statistics. (This is used for guidance, and is **not** strict.)
Region Size	0	If specified, use a fixed region size.
Maximum Region Size	100,000	The upper limit for region size. This is ignored if **Region Size** is set explicitly.
Force the Number of Regions	OFF	If **ON**, try to use this number of regions per reference. Otherwise, the Coverage Summary Report will optimize the number of regions in the case of many references. This is **not** compatible with fixed region sizes.

Advanced Analysis Parameters - Reports	Default Value	Description
Maximum Number of Contigs to Plot	25	The maximum number of contigs to plot in the Coverage Report.

Reports and Data Files

The Assembly (HGAP 4) application generates the following reports:

Polished Assembly > Summary Metrics

Displays statistics on the contigs from the *de novo* assembly that were corrected by Arrow.

- **Polished Contigs**: The number of polished contigs.
- **Maximum Contig Length**: The length of the longest contig.
- **N50 Contig Length**: 50% of the contigs are longer than this value.
- **Sum of Contig Lengths**: Total length of all the contigs.
- **E-size (sum of squares/sum)**: The expected contig size for a random base in the polished contigs.

Polished Assembly > Contig Coverage vs Confidence

- Maps the mean confidence (Quality Value) against the mean coverage depth.

Coverage Report > Summary Metrics

Displays depth of coverage across references, as well as depth of coverage distribution.

- **Mean Coverage:** The mean depth of coverage across the reference sequence.
- **Missing Bases**: The percentage of the reference sequence that has zero coverage.

Coverage > Coverage across Reference

- Maps coverage of the reference against the reference start position.

Coverage > Depth of coverage Distribution

- Histogram distribution of the reference regions by the coverage.

Realignment to Draft Assembly > Summary Metrics

Displays statistics on reads that realigned to the draft assembly.

- **Percent Realigned Bases**: The number of subread bases that realigned to the draft assembly, divided by the total number of bases in the BAM file.
- **Mean Concordance (realigned)**: The mean concordance of subreads that realigned to the draft assembly.
- **Number of Subreads (realigned):** The number of subreads that realigned to the draft assembly.
- **Number of Subread Bases (realigned)**: The number of subread bases that realigned to the draft assembly.
- **Subread Length Mean (realigned)**: The mean length of the mapped portion of subreads that realigned to the draft assembly.

- **Subread Length N50 (realigned)**: The subread length at which 50% of the bases realigned to the draft assembly are in subreads longer than, or equal to, this value.
- **Subread Length 95% (realigned)**: The 95th percentile of length of subreads that realigned to the draft assembly.
- **Subread Length Max (realigned)**: The maximum length of subreads that realigned to the draft assembly.
- **Number of Polymerase Reads (realigned)**: The number of polymerase reads that realigned to the draft assembly. This includes adapters.
- **Polymerase Read Length Mean (realigned)**: The mean read length of polymerase reads that realigned to the draft assembly, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (realigned)**: The read length at which 50% of the bases realigned to the draft assembly are in polymerase reads longer than, or equal to, this value.
- **Polymerase Read Length 95% (realigned)**: The 95th percentile of read length of polymerase reads that realigned to the draft assembly.
- **Polymerase Read Length Max (realigned)**: The maximum length of polymerase reads that realigned to the draft assembly.

Realignment to Draft Assembly > Realignment Statistics Summary

Displays, per movie, statistics on reads that realigned to the draft assembly.

- **Movie**: Movie name for which the following metrics apply.
- **Number of Polymerase Reads (realigned)**: The number of polymerase reads that realigned to the draft assembly. This includes adapters.
- **Polymerase Read Length Mean (realigned):** The mean read length of polymerase reads that realigned to the draft assembly, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (realigned)**: The read length at which 50% of the bases realigned to the draft assembly are in polymerase reads longer than, or equal to, this value.
- **Number of Subreads (realigned)**: The number of subreads that realigned to the draft assembly.
- **Number of Subread Bases (realigned)**: The number of subread bases that realigned to the draft assembly.
- **Subread Length Mean (realigned)**: The mean length of the mapped portion of subreads that realigned to the draft assembly.
- **Mean Concordance (realigned)**: The mean concordance of subreads that realigned to the draft assembly.

Realignment to Draft Assembly > Realigned Polymerase Read Length

- Histogram distribution of the number of reads by read length.

Realignment to Draft Assembly > Realigned Subread Concordance

- Histogram distribution of the number of subreads against the percent concordance with the subreads that realigned to the draft assembly.

Realignment to Draft Assembly > Realigned Subread Length

- Histogram distribution of the number of subread by the subread length.

Realignment to Draft Assembly > Realigned Concordance vs Read Length

- Maps the percent concordance with the reference sequence against the subread length, in base pairs.

Preassembly > Summary Metrics

Displays statistics on the pre-assembly process.

- **Genome Length (user input)**: The number of base pairs expected in the genome.
- **Number of Filtered Subreads**: The total number of filtered subreads used as initial input for the pre-assembly.
- **Filtered Subread Length Mean**: The mean length of the filtered subreads used as initial input for pre-assembly.
- **Filtered Subread Length (N50)**: 50% of the filtered subreads used as initial input are longer than this value.
- **Filtered Subread Length 95%**: The 95th percentile of the length of the filtered subreads used as initial input.
- **Filtered Subread E-Size**: The expected contig size for a random base in the filtered subreads.
- **Number of Filtered Subread Bases**: The total number of bases included in the filtered subreads used as initial input for pre-assembly.
- **Filtered Subread Coverage**: The number of filtered subread bases divided by the number of base pairs expected in the genome.
- **Length Cutoff (user input or auto-calc)**: The minimum length for a raw read to be used as a seed read for pre-assembly. Raw reads shorter than this value are filtered out.
- **Number of Seed Reads**: The number of reads longer than the length cutoff used in the pre-assembly.
- **Seed Read Length Mean**: The mean length of all the seed reads used in the pre-assembly.
- **Seed Read Length (N50)**: 50% of the seed reads used in the pre-assembly are longer than this value.
- **Seed Read Length 95%**: The 95th percentile of the length of the seed reads used in the pre-assembly.
- **Seed Read E-Size**: The expected contig size for a random base in the seed reads.
- **Number of Seed Bases (total)**: The total number of bases included in the seed reads used in the pre-assembly.
- **Seed Coverage (bases/genome_size)**: The number of seed bases divided by the number of base pairs expected in the genome.
- **Number of Pre-Assembled Reads**: The number of reads output by the pre-assembler. Pre-assembled reads are very long, highly accurate reads that can be used as input to a *de novo* assembler.
- **Pre-Assembled Read Length Mean**: The mean length of the pre-assembled reads.
- **Pre-Assembled Read Length (N50)**: 50% of the pre-assembled reads are longer than this value.
- **Pre-Assembled Read Length 95%**: The 95th percentile of the length of the reads output by the pre-assembler.
- **Pre-Assembled E-size (sum of squares/sum)**: The expected contig size for a random base in the pre-assembled reads.
- **Number of Pre-Assembled Bases (total)**: The total number of bases output by the pre-assembler.

- **Pre-Assembled Coverage (bases/genome_size)**: The number of bases output by the pre-assembler divided by the number of base pairs expected in the genome.
- **Pre-Assembled Yield (bases/seed_bases):** The percentage of seed read bases that were successfully aligned to generate pre-assembled reads.
- **Average Number of Reads that Each Seed is Broken Into**: The average number of preliminary reads that each seed is broken into. (Preliminary reads are derived from seeds using error correction; some portions of seeds might be too "noisy" to use.)
- **Average Number of Bases Lost from Each Seed:** The average number of bases from each seed that were completely discarded.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Alignments**: Data Set containing the alignment results.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference.
- **Polished Assembly**: The final polished assembly, in Data Set, FASTA and FASTQ formats.
- **Draft Assembly**: The unpolished draft assembly.

Base Modification Detection Application

Use this application to identify putative sites of base modification as well as common bacterial base modifications (6mA, 4mC). Detection can use an in-silico control consisting of expected kinetic signals.

Reference (Required):

- Specify a reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.

Parameters

Advanced Analysis Parameters - Alignment	Default Value	Description
Number of .bam files	1	Number of .bam files to create in consolidate mode.
Minimum Concordance	70	The minimum required alignment concordance, in percent.
Consolidate .bam	OFF	Specify whether to merge chunked/gathered .bam files.
Concordant Alignment	ON	Specify whether to map subreads of a ZMW to the same genomic location.
Align Unsplit Polymerase Reads	OFF	Do not split reads into subreads even if subread regions are available.
Minimum Length	50	The minimum required alignment length, in base pairs.
Hit Policy	randombest	Specify how to treat multiple hits: • **random**: Selects a random hit. • **all**: Selects all hits. • **allbest**: Selects all the best score hits. • **randombest**: Selects a random hit from all best score hits. • **leftmost**: Selects a hit which has the best score and the smallest mapping coordinate in any reference.
Algorithm Options		List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0`

Advanced Analysis Parameters - Base Mods	Default Value	Description
Compute Methyl Fraction	OFF	When identifying specific modifications (6mA and/or 4mC), enabling this option will estimate the methylated fraction, along with 95% confidence interval bounds.
P-Value	0.001	The probability value cutoff.
Maximum Sequence Length	2,112,827,392	The maximum number of bases to process per contig.
Identify Basemods	NONE	A comma-separated list of the modifications to identify. Currently, this includes 6mA and 4mC.

Advanced Analysis Parameters - Reports	Default Value	Description
Number of Regions	1000	Specify the number of genome regions in the summary statistics. (This is used for guidance, and is **not** strict.)
Region Size	0	If specified, use a fixed region size.
Maximum Region Size	100,000	The upper limit for region size. This is ignored if **Region Size** is set explicitly.

Advanced Analysis Parameters - Reports	Default Value	Description
Force the Number of Regions	OFF	If **ON**, try to use this number of regions per reference. Otherwise, the Coverage Summary Report will optimize the number of regions in the case of many references. This is **not** compatible with fixed region sizes.

Reports and Data Files

The Base Modification Detection application generates the following reports:

Base Modifications > Per-Base Kinetic Detections

- Maps the modification QV against per-strand coverage.

Base Modifications > Kinetic Detections Histogram

- Histogram distribution of the number of bases by modification QV.

Mapping Report > Summary Metrics

Mapping is local alignment of a read or subread to a reference sequence.

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Subread Length N50 (mapped)**: The subread length at which 50% of the mapped bases are in subreads longer than, or equal to, this value.
- **Subread Length 95% (mapped)**: The 95th percentile of length of subreads that mapped to the reference sequence.
- **Subread Length Max (mapped)**: The maximum length of subreads that mapped to the reference sequence.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Polymerase Read Length 95% (mapped)**: The 95th percentile of read length of polymerase reads that mapped to the reference sequence.
- **Polymerase Read Length Max (mapped)**: The maximum length of polymerase reads that mapped to the reference sequence.

Mapping Report > Mapping Statistics Summary

Displays mapping statistics per movie.

- **Movie**: Movie name for which the following metrics apply.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.

- **Polymerase Read Length Mean (mapped)**: The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.

Mapping Report > Mapped Polymerase Read Length

- Histogram distribution of the number of reads by read length.

Mapping Report > Mapped Subread Concordance

- Histogram distribution of the number of subreads by the percent concordance with the reference sequence.

Mapping Report > Mapped Subread Length

- Histogram distribution of the number of subreads by the subread length.

Mapping Report > Mapped Concordance vs Read Length

- Maps the percent concordance with the reference sequence against the subread length, in base pairs.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Alignments**: Data Set of alignment results.
- **IPD Ratios**: BigWig file containing encoded base IPD ratios.
- **Modifications**: Duplicate of the modification summary file.
- **Full Kinetics Summary**: HDF5 file containing per-base information.

Base Modification and Motif Analysis Application

Use this application to identify putative sites of base modifications, as well as common bacterial base modifications (6mA, 4mC), and then analyze the methyltransferase recognition motifs.

- Filters reads by length and quality, maps them against a provided reference sequence, and then calls variants.
- Detection can use an in-silico control consisting of expected kinetic signals.

Reference (Required):

- Specify a reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.

Parameters

Advanced Analysis Parameters - Alignment	Default Value	Description
Number of .bam files	1	Number of .bam files to create in consolidate mode.
Minimum Concordance	70	The minimum required alignment concordance, in percent.
Concordant Alignment	ON	Specify whether to map subreads of a ZMW to the same genomic location.
Align Unsplit Polymerase Reads	OFF	Do not split reads into subreads even if subread regions are available.
Minimum Length	50	The minimum required alignment length, in base pairs.
Consolidate .bam	OFF	Specify whether to merge chunked/gathered .bam files.
Hit Policy	randombest	Specify how to treat multiple hits: • **random**: Selects a random hit. • **all**: Selects all hits. • **allbest**: Selects all the best score hits. • **randombest**: Selects a random hit from all best score hits. • **leftmost**: Selects a hit which has the best score and the smallest mapping coordinate in any reference.
Algorithm Options		List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0`

Advanced Analysis Parameters - Base Mods	Default Value	Description
Compute Methyl Fractions	OFF	When identifying specific modifications (6mA and/or 4mC), enabling this option will estimate the methylated fraction, along with 95% confidence interval bounds.
P-Value	0.001	The probability value cutoff.
Maximum Sequence Length	2,112,827,392	The maximum number of bases to process per contig.
Identify Basemods	NONE	A comma-separated list of the modifications to identify. Currently, this includes 6mA and/or 4mC.

Advanced Analysis Parameters - Reports	Default Value	Description
Number of Regions	1000	Specify the number of genome regions in the summary statistics. (This is used for guidance, and is **not** strict.)
Region Size	0	If specified, use a fixed region size.
Maximum Number of Motifs in QV Plot	10	Specify the number of motifs whose QV are plotted in the report.
Maximum Region Size	100,000	The upper limit for region size. This is ignored if **Region Size** is set explicitly.
Force the Number of Regions	OFF	If **ON**, try to use this number of regions per reference. Otherwise, the Coverage Summary Report will optimize the number of regions in the case of many references. This is **not** compatible with fixed region sizes.

Advanced Analysis Parameters - Motifs	Default Value	Description
Minimum Methylated Fraction	0.3	The minimum methylated fraction to identify a motif.
Minimum Qmod Score	30	The minimum QMod score used to identify a motif.

Reports and Data Files

The Base Modification and Motif Analysis application generates the following reports:

Modified Base Motifs > Modified Base Motifs

Displays statistics for the methyltransferase recognition motifs detected.

- **Motif**: The nucleotide sequence of the methyltransferase recognition motif, using the standard IUPAC nucleotide alphabet.
- **Modified Position**: The position within the motif that is modified. The first base is 1. Example: The modified adenine in GATC is at position 2.
- **Modification Type**: The type of chemical modification most commonly identified at that motif. These are: 6mA, 4mC, 5mC, or `modified_base` (modification not recognized by the software.)
- **% of Motifs Detected**: The percentage of times that this motif was detected as modified across the entire genome.
- **# of Motifs Detected**: The number of times that this motif was detected as modified across the entire genome.
- **# of Motifs In Genome**: The number of times this motif occurs in the genome.
- **Mean QV**: The mean modification QV for all instances where this motif was detected as modified.
- **Mean Coverage**: The mean coverage for all instances where this motif was detected as modified.
- **Partner Motif**: For motifs that are not self-palindromic, this is the complementary sequence.
- **Mean IPD Ratio**: The mean inter-pulse duration. An IPD ratio greater than 1 means that the sequencing polymerase slowed down at this base position, relative to the control. An IPD ratio less than 1 indicates speeding up.
- **Group Tag**: The motif group of which the motif is a member. Motifs are grouped if they are mutually or self reverse-complementary. If the motif isn’t complementary to itself or another motif, the motif is given its own group.

- **Objective Score**: For a given motif, the objective score is defined as (`fraction methylated)*(sum of log-p values of matches`).

Modified Base Motifs > Modification QVs

- Maps motif sites against Modification QV.

Mapping Report > Summary Metrics

Mapping is local alignment of a read or subread to a reference sequence.

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Subread Length N50 (mapped)**: The subread length at which 50% of the mapped bases are in subreads longer than, or equal to, this value.
- **Subread Length 95% (mapped)**: The 95th percentile of length of subreads that mapped to the reference sequence.
- **Subread Length Max (mapped)**: The maximum length of subreads that mapped to the reference sequence.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Polymerase Read Length 95% (mapped)**: The 95th percentile of read length of polymerase reads that mapped to the reference sequence.
- **Polymerase Read Length Max (mapped)**: The maximum length of polymerase reads that mapped to the reference sequence.

Mapping Report > Mapping Statistics Summary

Displays mapping statistics per movie.

- **Movie**: Movie name for which the following metrics apply.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped)**: The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.

Mapping Report > Mapped Polymerase Read Length

- Histogram distribution of the number of reads by read length.

Mapping Report > Mapped Subread Concordance

- Histogram distribution of the number of subreads by the percent concordance with the reference sequence.

Mapping Report > Mapped Subread Length

- Histogram distribution of the number of subreads by the subread length.

Mapping Report > Mapped Concordance vs Read Length

- Maps the percent concordance with the reference sequence against the subread length, in base pairs.

Base Modifications > Per-Base Kinetic Detections

- Maps the modification QV against per-strand coverage.

Base Modifications > Kinetic Detections Histogram

- Histogram distribution of the number of bases by modification QV.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Alignments**: Data Set of alignment results.
- **IPD Ratios**: BigWig file containing encoded base IPD ratios.
- **Motifs and Modifications**: Summary of analysis results for each `kinModCall` with motif information.
- **Motifs Summary**: Summary of analysis results for each motif.
- **Full Kinetics Summary**: HDF5 file containing per-base information.

CCS Mapping Application

Use this application to generate consensus sequences from single molecules, and map these consensus sequences to a user-provided reference sequence.

The CCS Mapping application:

- Generates consensus sequences from single molecules.
- Maps consensus sequences to a provided reference sequence, and then identifies consensus and variants against this reference.
- Haploid variants and small indels, but **not** diploid variants, are called as a result to alignment to the reference sequence.

CCS Mapping takes multiple subreads of the same SMRTbell template and combines them to produce one high quality consensus sequence. The Circular Consensus Sequences are then mapped to a reference sequence.

Reference (Required):

- Specify a reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.

Minimum Number of Passes:

- The minimum number of **full** passes for a ZMW to be emitted. Full passes **must** have an adapter hit before and after the insert sequence and so do not include any partial passes at the start and end of the sequencing reaction.

Minimum Predicted Accuracy:

- The **minimum** predicted accuracy of a read, ranging from 0 to 1. (`0.99` indicates that only reads expected to be 99% accurate are emitted.)

Parameters

Advanced Analysis Parameters - Alignment	Default Value	Description
Consolidate .bam	OFF	Specify whether to merge chunked/gathered .bam files.
Number of .bam files	1	Number of .bam files to create in consolidate mode.
Minimum Concordance	70	The minimum required alignment concordance, in percent.
Minimum Length	50	The minimum required alignment length, in base pairs.
Hit Policy	randombest	Specify how to treat multiple hits: • **random**: Selects a random hit. • **all**: Selects all hits. • **allbest**: Selects all the best score hits. • **randombest**: Selects a random hit from all best score hits. • **leftmost**: Selects a hit which has the best score and the smallest mapping coordinate in any reference.

Advanced Analysis Parameters - Alignment	Default Value	Description
Algorithm Options		List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0`

Advanced Analysis Parameters - CCS	Default Value	Description
By Strand CCS	OFF	For each ZMW, generate two CCS sequences - one for each strand.
Maximum Dropped Fraction	0.34	The maximum fraction of subreads that can be dropped before giving up.
Maximum Subread Length	21,000	The maximum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	10	The minimum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimal Read Score	0.65	The minimum read score of input subreads.
Minimum SNR	3.75	The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR <3.75 is typically considered lower quality.
Minimum Z Score	-3.4	The minimum Z-Score for a subread to be included in the consensus-generating process.
Polish CCS	ON	Specify whether to polish CCS sequences using Arrow.
Report File Output		Specify the name of the report file to output. (Default = `ccs_report_txt`)
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` “rich” quality tracks.
Filters to add to the Dataset	NONE	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Reports	Default Value	Description
Number of Regions	1000	Specify the number of genome regions in the summary statistics. (This is used for guidance, and is **not** strict.)
Region Size	0	If specified, use a fixed region size.
Maximum Region Size	100,000	The upper limit for region size. This is ignored if **Region Size** is set explicitly.
Force the Number of Regions	OFF	If **ON**, try to use this number of regions per reference. Otherwise, the Coverage Summary Report will optimize the number of regions in the case of many references. This is **not** compatible with fixed region sizes.
Maximum Number of Contigs to Plot	25	The maximum number of contigs to plot in the coverage report.

Reports and Data Files

The CCS Mapping application generates the following reports:

Mapping Report > Summary Metrics

Mapping is local alignment of a read to a reference sequence.

- **Mapped CCS Read Mean Concordance**: The mean concordance of the CCS reads that mapped to the reference sequence.
- **Number of CCS Reads (mapped)**: The number of CCS reads that mapped to the reference sequence.
- **Number of CCS Bases (mapped)**: The number of bases in the CSS reads that mapped to the reference sequence.
- **CCS Read Length Mean (mapped)**: The mean length of CCS reads that mapped to the reference sequence.
- **CCS Read Length N50 (mapped)**: The read length at which 50% of the bases are in reads longer than, or equal to, this value.
- **CCS Read Length 95% (mapped)**: The 95^{th} percentile of length of CCS reads that mapped to the reference sequence.
- **CCS Read Length Max (mapped)**: The maximum length of CCS reads that mapped to the reference sequence.

Mapping Report > CCS Mapping Statistics Summary

Displays CCS mapping statistics per movie.

- **Movie**: Movie name for which the following metrics apply.
- **Number of CCS Reads (mapped)**: The number of CCS reads that mapped to the reference sequence.
- **CCS Read Length Mean (mapped)**: The mean length of CCS reads that mapped to the reference sequence.
- **CCS Read Length N50 (mapped)**: The read length at which 50% of the bases are in reads longer than, or equal to, this value.
- **Number of CCS Bases (mapped)**: The number of bases in the CSS reads that mapped to the reference sequence.
- **Mapped CCS Read Mean Concordance**: The mean concordance of the CCS reads that mapped to the reference sequence.

Mapping Report > Mapped CCS Read Length

- Histogram distribution of the mapped CCS reads by the read length.

Mapping Report > Mapped CCS Read Concordance

- Histogram distribution of the mapped CCS reads by their concordance with the reference sequence.

Mapping Report > Mapped Concordance vs Read Length

- Maps the percent concordance with the reference sequence against CCS read length.

Mapping Report > Mapped QV Calibration

- Maps the percent concordance with the reference sequence against predicted accuracy.

Coverage > Summary Metrics

- **Mean Coverage**: The mean depth of coverage across the reference sequence.
- **Missing Bases (%)**: The percentage of the reference sequence without coverage.

Coverage > Coverage across a Reference

- Maps coverage of the user-selected reference against the reference start position.

Coverage > Depth of Coverage Distribution

- Maps the reference regions against the percent coverage.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.
- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.)
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram distribution of the CCS reads by the read length.

CCS Report > CCS Read Score

- Maps CCS reads against their quality (Read Score).

CCS Report > Number of Passes

- Maps CCS reads against the number of complete subreads per CCS read.

CCS Report > Number of Passes vs Read Score

- Maps the number of complete subreads per CCS read against the read scores (as Phred QV).

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Alignments**: Data Set containing alignment results.
- **Consensus Sequences**: Consensus sequences generated from CCS.
- **CCS Statistics**: Summary of CCS performance and yield.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference.
- **FASTQ File(s), FASTA File(s)**: Consensus sequences generated from CCS, in FASTA or FASTQ format.

Circular Consensus Sequences (CCS) Application

Use this application to identify consensus sequences for single molecules.

Minimum Number of Passes:

- The minimum number of **full** passes for a ZMW to be emitted. Full passes **must** have an adapter hit before and after the insert sequence and so do not include any partial passes at the start and end of the sequencing reaction.

Minimum Predicted Accuracy:

- The **minimum** predicted accuracy of a read, ranging from 0 to 1. (`0.99` indicates that only reads expected to be 99% accurate are emitted.)

Parameters

Advanced Analysis Parameters - CCS	Default Value	Description
By Strand CCS	OFF	For each ZMW, generate two CCS sequences - one for each strand.
Maximum Dropped Fraction	0.34	The maximum fraction of subreads that can be dropped before giving up.
Maximum Subread Length	21,000	The maximum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	10	The minimum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimal Read Score	0.65	The minimum read score of input subreads.
Minimum SNR	3.75	The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR <3.75 is typically considered lower quality.
Minimum Z Score	-3.4	The minimum Z-Score for a subread to be included in the consensus generating process.
Polish CCS	ON	Specify whether to polish CCS sequences using Arrow.
Report File Output		Specify the name of the report file to output. (Default = `ccs_report_txt`)
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` “rich” quality tracks.
Filters to add to the Dataset	NONE	A comma-separated list of additional Data Set filters to use.

Reports and Data Files

The Circular Consensus Sequences (CCS) application generates the following reports:

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.
- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.

- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.)
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram distribution of the CCS reads by the read length.

CCS Report > CCS Read Score

- Maps CCS reads against their quality (Read Score).

CCS Report > Number of Passes

- Maps CCS reads against the number of complete subreads per CCS read.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Sequences**: Consensus sequences generated from CCS.
- **CCS Statistics**: Summary of CCS performance and yield.
- **FASTQ File(s), FASTA File(s)**: Consensus sequences generated from CCS, in FASTA or FASTQ format.

Convert BAM to FASTX Application

Use this application to convert sequence data in BAM file format to the FASTX file format.

- For **barcoded** runs, you must **first** run the **Demultiplex Barcodes** application to create BAM files **before** using this application.
- This application does **not** generate any reports.

Parameters

Advanced Analysis Parameters	Default Value	Description
Filters to Add to the Data Set	NONE	A comma-separated list of additional Data Set filters to use.
Minimum Subread Length	0	The minimum length of subreads to write out to FASTA/FASTQ files.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **FASTA file**(s): Sequence data converted to FASTA format.
- **FASTQ file**(s): Sequence data converted to FASTQ format.

Convert RS to BAM Application

Use this application to convert sequence data generated on a PacBio RS II system in HDF5 file format to the BAM file format, compatible with SMRT Analysis v4.0.0 and later.

- This application does **not** generate any reports.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Workflow Task Reports**: Contains attributes of workflow tasks of the analysis job.

Demultiplex Barcodes Application

Use this application to separate sequence reads by barcode. (See "Working with Barcoded Data" on page 110 for more details.)

Note: To demultiplex Iso-Seq samples in the SMRT Link (GUI), **always** choose the Iso-Seq 3 or Iso-Seq 3 with Mapping applications, **not** the Demultiplex Barcodes application.

- Barcoded SMRTbell templates are SMRTbell templates with adapters flanked by barcode sequences, located on both ends of an insert.
- For **symmetric** and **tailed** library designs, the **same** barcode is attached to both sides of the insert sequence of interest. The only difference is the orientation of the trailing barcode. For **asymmetric** designs, **different** barcodes are attached to the sides of the insert sequence of interest.
- Barcode names and sequences, independent of orientation, **must** be unique.
- Most-likely barcode sequences per SMRTbell are identified using a FASTA-format file.

Given an input set of barcodes and a BAM Data Set, the Demultiplex Barcodes application produces:

- A set of BAM files whose reads are annotated with the barcodes;
- A `subreadset` file that contains the file paths of that collection of barcode-tagged BAM files and their related files.

Barcode Set (Required):

- Specify a barcode sequence file to separate the reads.

Name of Output Data Set (Required)

- Specify the name for the new demultiplexed Data Set that will display in SMRT Link.

Same Barcodes on Both Ends of Sequence

- Specify **On** to retain all the reads with the **same** barcodes on both ends of the insert sequence, such as symmetric and tailed designs. (See "Working with Barcoded Data" on page 110 for information on barcode designs.)
- Specify **Off** to specify asymmetric designs where the barcodes are **different** on each end of the insert.

Minimum Barcode Score

- A **barcode score** measures the alignment between a barcode attached to a read and an ideal barcode sequence, and is an indicator how well the chosen barcode pair matches. It ranges between `0` (no match) and `100` (a perfect match). Specify that reads with barcode scores **below** this minimum value are **not** included in downstream analysis.

Infer Barcodes Used

- The barcoding algorithm can detect the set of barcodes used. It infers the barcodes used by looking at the first 35,000 ZMWs, then selecting barcodes with ≥10 counts **and** mean scores ≥45. Specify **ON** to use this mode.

Reports and Data Files

The Demultiplex Barcodes application generates the following reports:

Barcodes > Summary Metrics

- **Unique Barcodes:** The number of unique barcodes in the sequence data.
- **Barcoded Reads:** The number of barcoded reads in the sequence data.
- **Mean Reads:** The mean number of reads per barcode.
- **Max. Reads:** The maximum number of reads per barcode.
- **Min. Reads:** The minimum number of reads per barcode.
- **Mean Read Length:** The mean read length of reads per barcode.
- **Mean Longest Subread Length:** The mean length of the longest subread in each barcoded sample.
- **Unbarcoded Reads:** The number of reads without barcodes in the sequence data.

Barcodes > Barcode Data

- **Bio Sample Name:** The name of the biological sample associated with the barcode.
- **Barcode Index:** The index number associated with the barcode.
- **Barcode Name**: A string containing the pair of barcode indices for which the following metrics apply.
- **Polymerase Reads:** The number of polymerase reads associated with the barcode.
- **Subreads**: The number of subreads associated with the barcode.
- **Bases**: The number of bases associated with the barcode.
- **Mean Read Length:** The mean read length of reads associated with the barcode.
- **Longest Subread Length:** The longest subread length associated with the barcode.
- **Mean Barcode Quality:** The mean barcode quality associated with the barcode.
- **Rank Order (Num. Reads)**: The rank order of this barcode in terms of the number of reads.

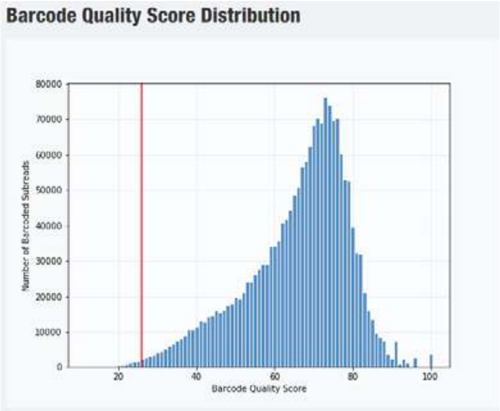
Barcodes > Barcoded Read Statistics

- **Number of Reads per Barcode**: Line graph displays the number of sorted reads per barcode.
  - **Good performance**: The Number of Reads per Barcode line (blue) should be mostly linear. Note that this depends on the choice of Y-axis scale. The mean Number of Reads per Barcode line (red) should be near the middle of the graph and should not be skewed by samples with too many or too few barcodes.
  - **Questionable performance**: A sharp discontinuity in the blue line, followed by no yield, with the red line way off center. This indicates that the user should allow the software to infer the barcodes.
- **Barcode Frequency Distribution**: Histogram distribution of read counts per barcode.

  - **Good performance**: A uniform distribution, which is most often a fairly tight symmetric normal distribution, with few barcodes in the tails.
  - **Questionable performance**: A large peak at zero indicates that the user should rerun the **Demultiplex Barcodes** application with the **Infer Barcodes** option set to **On**.
- **Mean Read Length Distribution**: Histogram distribution of the mean polymerase read length for all samples.
  - **Good performance**: The distribution should be normal with a relatively tight range.
  - **Questionable performance**: A spread out distribution, with a mode towards the low end.

Barcodes > Barcode Quality Scores

- **Barcode Quality Score Distribution:** Histogram distribution of barcode Quality scores. The scores range from 0-100, with 100 being a perfect match. Any significant modes or accumulation of scores <40 suggests issues with some of the barcode analyses. The red line is set at 26 – the minimum default barcode score.
  - **Good performance**: Distributions with a mode >65 and the low-end tail tapering off below 40.



  - **Questionable performance**: A bimodal distribution with a large second peak usually indicates that some barcodes that were sequenced were **not** included in the barcode scoring set.

Barcodes > Barcoded Read Binned Histograms

- **Read Length Distribution By Barcode**: Histogram distribution of the Polymerase read length by barcode. Each column of rectangles is similar to a read length histogram rotated vertically, seen from the top. Each sample should have similar Polymerase read length distribution. Non-smooth changes in the pattern looking from left to right might indicate suboptimal performance.
- **Barcode Quality Distribution By Barcode**: Histogram distribution of the per-barcode version of the **Read Length Distribution by Barcode** histogram. The histogram should contain a single cluster of hot spots in each column. All barcodes should also have similar profiles; significant differences in the pattern moving from left to right might indicate suboptimal performance.
  - **Good performance**: All columns show a single cluster of hot spots.
  - **Questionable performance**: A bimodal distribution would indicate missing barcodes in the scoring set.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Barcode Files**: Barcoded subread Data Sets; one file per barcode.
- **Barcode Report Details**: Data displayed in the reports, in CSV format.

Note: You can get the demultiplexed BAM files using the Data Management module's "Export Data Sets" feature. In the demultiplexed BAM output file a tag `bc` is added for each read, indicating the assigned barcode. The `bc` tag is the zero-based index of the barcodes in the FASTA file. For example, when using the barcodes `RSII_96_barcodes`, a subread with barcode `lbc1` identified on both sides will have the tag `bc:B:S,0,0` in the BAM output file.

A second `bq` tag corresponds to the barcode quality (0-100).

Iso-Seq® 1 Analysis Application

Use this application to characterize full-length transcripts. The analysis is performed *de novo*, without a reference genome.

The Iso-Seq 1 application enables analysis and functional characterization of transcript isoforms for sequencing data generated on PacBio instruments.

This application generates full-length transcripts, eliminating the need for computational reconstruction.

The Iso-Seq 1 application provides accurate information about alternatively spliced exons and transcriptional start and end sites.

The application includes three main steps:

1. **CCS**: Build Circular Consensus Sequences (CCSs) from the sequencing subreads.
2. **Classify**: Classify CCS reads in two groups – full length and non-full length. Identify and remove polyA/T tails, remove primers, and identify read-strandedness. Also remove artificial concatemers but do **not** remove PCR chimeras.
3. **Cluster**: Perform *de novo* clustering and consensus calling. Output polished, full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on predicted accuracies.

Parameters

Advanced Analysis Parameters - CCS	Default Value	Description
By Strand CCS	OFF	For each ZMW, generate two CCS sequences - one for each strand.
Maximum Dropped Fraction	0.8	The maximum fraction of subreads that can be dropped before giving up.
Maximum Subread Length	15,000	The maximum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	50	The minimum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Number of Passes	0	The minimum number of **full** passes for a ZMW to be emitted. Full passes **must** have an adapter identified before and after the insert sequence and so do not include any partial passes at the start and end of the sequencing reaction. For the Iso-Seq 1 application, a 0-pass read can still be full-length.
Minimum Predicted Accuracy	0.8	The minimum predicted accuracy of a read, ranging from 0 to 1. (`0.99` indicates that only reads expected to be 99% accurate are emitted.)
Minimal Read Score	0.65	The minimum read score of input subreads.
Minimum SNR	3.75	The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR <3.75 is typically considered lower quality.

Advanced Analysis Parameters - CCS	Default Value	Description
Minimum Z Score	-9999	The minimum Z-Score for a subread to be included in the consensus-generating process.
Polish CCS	OFF	Specify whether to polish CCS sequences using Arrow.
Report File Output		Specify the name of the report file to output. (Default = `ccs_report_txt`)
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` “rich” quality tracks.
Filters to add to the Dataset	NONE	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Iso-Seq	Default Value	Description
Bin by Primer	OFF	Specify binning reads by Primers. This overwrites the **Bin by Read Length Manually** and **Bin by Read Length in KB** options.
Bin by Read Length Manually	NONE	Specify binning reads by manually inputting read length bins. (Example: `[0, 2, 3, 5]` means binning reads into 0-2 kb, 2-3 kb, 3-5 kb, and above 5 kb bins.) This overwrites the **Bin By Read Length in KB** option.
Bin by Read Length in KB	1	Specify binning reads by the read length, in kb.
Minimum Accuracy of Polished Isoforms	0.99	Specify the minimum predicted consensus accuracy to classify an isoform as high-quality (HQ). All isoforms below this cutoff are considered low-quality (LQ).
Minimum Sequence Length	50	The minimum sequence length to output.
Customer Primer Sequences	NONE	Specify any custom primers used with the cDNA sample preparation. (The PacBio cDNA protocol uses the Clontech SMARTer primers.) Copy and paste custom primer sequences, in FASTA format, into the field. See “Appendix A - Barcoded Primers” on page 130 for details.
Trim QVs 3'	30	Specify the number of bases whose Quality Value to ignore in the 3' end.
Trim QVs 5'	100	Specify the number of bases whose Quality Value to ignore in the 5' end.
Require PolyA	ON	**ON** means that polyA tails are required for a sequence to be considered full length. **OFF** means sequences do not need polyA tails to be considered full length.
Sample Name	NONE	The name of the input sample. A random string is used when Sample Name is NONE.

Reports and Data Files

The Iso-Seq 1 application generates the following reports:

Transcript Clustering > Summary Metrics

- **Number of unpolished consensus isoforms**: The number of consensus isoform reads, both high and low-quality.
- **Number of polished high-quality isoforms**: The number of consensus isoforms that have an estimated accuracy above the specified threshold.
- **Number of polished low-quality isoforms**: The number of consensus isoforms that have an estimated accuracy below the specified threshold.

- **Mean unpolished consensus isoforms read length**: The mean read length of the consensus isoform reads, both high and low-quality.

Transcript Clustering > Read Length of Consensus Isoforms Reads

- Maps the read length of consensus isoform reads against the number of reads.

Transcript Clustering > Average Quality Value of HQ and LQ Isoforms

- Maps the High Quality/Low Quality Isoform average QV against the number of Isoforms with greater than the average QV.

Transcript Classification > Summary Metrics

- **Number of consensus reads**: The number of consensus isoform reads.
- **Number of five prime reads**: The number of CCS reads with 5' primer detected.
- **Number of three prime reads**: The number of CCS reads with 3' primer detected.
- **Number of poly-A reads**: The number of CCS reads with polyA tail and 3' primer detected.
- **Number of filtered short reads**: The number of reads whose read length is less than the specified Minimum Sequence Length.
- **Number of non-full-length reads**: The number of non-full-length CCS reads missing the polyA tail and/or a terminal signal. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length reads**: The number of full-length CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length non-chimeric reads**: The number of full-length non-artificial-concatemer CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length non-chimeric bases**: The total number of bases in full-length non-artificial-concatemer CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Mean full-length non-chimeric read length**: The mean length of full-length, non-artificial-concatemer CCS reads.

Transcript Classification > Iso-Seq Transcript Classification

- Displays the same information as the **Transcript Classification > Summary Metrics** report.

Transcript Classification > Read Length of Full-Length Non-Chimeric Reads

- Histogram distribution of the number of full-length non-chimeric reads by the read length.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.
- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.) For the Iso-Seq 1 application, the default option for "Polish CCS" is **OFF**, which results in a read score of 0.

- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram distribution of the CCS reads by the read length.

CCS Report > Number of Passes

- Maps CCS reads against the number of complete subreads per CCS read.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Isoforms**: Consensus Isoforms produced by clustering FLNC reads using the ICE algorithm.
- **Consensus Sequences**: Consensus sequences generated from CCS, in Data Set, FASTA, or FASTQ format.
- **Draft Isoforms**: Intermediate Data Set used to get full length reads.
- **Full-Length Non-Chimeric Reads**: Full-length non-chimeric subreads generated from `pbtranscript` classify.
- **High-Quality Isoforms**: Data Set or FASTQ file of isoforms with high consensus accuracy.
- **Low-Quality Isoforms**: Data Set or FASTQ file of isoforms with low consensus accuracy.
- **Non-Full-Length Reads**: Non-full-Length reads generated from `pbtranscript` classify.
- **Primer Info**: Per-CCS read annotation and classification results.

Iso-Seq® 3 Analysis Application

Use this application to characterize full-length transcripts. The analysis is performed *de novo*, without a reference genome.

The Iso-Seq 3 application enables analysis and functional characterization of transcript isoforms for sequencing data generated on PacBio instruments.

This application generates full-length transcript isoforms, eliminating the need for computational reconstruction.

The Iso-Seq 3 application provides accurate information about alternatively spliced exons and transcriptional start and end sites.

The application includes three main steps:

1. **CCS**: Build Circular Consensus Sequences (CCSs) from each sequencing ZMW.
2. **Classify**: Identify and remove primers (which includes cDNA primers and optionally barcodes). Identify strandedness based on the 5' and 3' primers.
3. **Cluster**: Trim off polyA tails. Also remove artificial concatemers but do **not** remove PCR chimeras. Perform *de novo* clustering and consensus calling. Output full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on estimated accuracies.

For barcoded samples: The Iso-Seq 3 application pools all demultiplexed reads from the Classify step and outputs only one set of consensus isoforms after the Cluster step. This is suitable for samples that are from the same species but different tissues, or samples of the same genes but different individuals.

If you have samples from **different** species and need to run the Cluster step separately for each primer, you can either run the Iso-Seq 1 application, or run the Iso-Seq 3 Cluster step on the command line.

If you only want to obtain full-length non-concatemer (FLNC) reads and do **not** wish to complete the Cluster step, you can run the Iso-Seq 1 Classify Only application, or, run the Iso-Seq 3 Cluster step from the command line and terminate after the FLNC BAM file is generated.

Iso-Seq 3 determines two FLNC reads to be the same isoform, and will place them in the same cluster, if the two reads:

- Differ less than 100 bp on the 5' end.
- Differ less than 30 bp on the 3' end.
- Have no internal gaps that exceed 10 bp.

Iso-Seq 3 will only output clusters that have at least two FLNC reads.

Primer Set (Required):

- Specify a primer sequence file in FASTA format to identify cDNA primers for removal. The primer sequence includes the 5' and 3' cDNA primers and (if applicable) barcodes.
- Primer IDs must be specified using the suffix `_5p` to indicate 5' cDNA primers and the suffix `_3p` to indicate 3' cDNA primers. The 3' cDNA primer should **not** include the Ts and is written in reverse complement (see examples below).
- If barcodes were used, they should be included.
- Each primer sequence must be **unique**.

Example 1: The Clontech primer set.

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGG
>3p
GTACTCTGCGTTGATACCACTGCTT
```

Example 2: 4 tissues were multiplexed using barcodes on the 3' end only.

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGGG
>tissue1_3p
atgacgcatcgtctgaGTACTCTGCGTTGATACCACTGCTT
>tissue2_3p
gcagagtcatgtatagGTACTCTGCGTTGATACCACTGCTT
>tissue3_3p
gagtgctactctagtaGTACTCTGCGTTGATACCACTGCTT
>tissue4_3p
catgtactgatacacaGTACTCTGCGTTGATACCACTGCTT
```

Parameters

Advanced Analysis Parameters - CCS	Default Value	Description
Maximum Subread Length	15,000	The maximum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	50	The minimum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Polish CCS	OFF	Specify whether to polish CCS sequences. The default is OFF as this reduces run time and does not impact output quality because Polishing is done at the end of the Cluster step.
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` "rich" quality tracks. The default is OFF as **Polish CCS** is OFF by default.
Require and Trim Poly(A) Tail	ON	**ON** means that polyA tails are required for a sequence to be considered full length, and are to be removed. **OFF** means sequences do not need polyA tails to be considered full length. (Example: PCR amplicons with no polyA tail.)
QV Cutoff for HQ Transcripts	0.99	Specify the minimum estimated accuracy for a transcript to be considered "High-Quality". Transcripts below the cutoff will be "Low-Quality".

Reports and Data Files

The Iso-Seq 3 application generates the following reports:

Transcript Clustering > Summary Metrics

- **Number of polished high-quality isoforms**: The number of consensus isoforms that have an estimated accuracy above the specified threshold. (This is set by the **QV Cutoff for HQ Transcripts** option in the **Advanced Parameters** dialog.)
- **Number of polished low-quality isoforms**: The number of consensus isoforms that have an estimated accuracy below the specified threshold. (This is set by the **QV Cutoff for HQ Transcripts** option in the **Advanced Parameters** dialog.)

Transcript Clustering > Read Length of Consensus Isoforms Reads

- Histogram of the consensus isoform lengths and the distribution of isoforms exceeding a read length cutoff.

Transcript Clustering > Average Quality Value of HQ and LQ Isoforms

- Histogram of the consensus isoform QVs and the distribution of isoforms exceeding a QV cutoff.

CCS Read Classification > Summary Metrics

- **Reads:** The total number of CCS reads.
- **Reads with 5’ and 3’ Primers:** The number of CCS reads with 5’ and 3’ primer detected.
- **Non-Concatemer Reads with 5’ and 3’ Primers:** The number of non-concatemer CCS reads with 5’ and 3’ primer detected.
- **Non-Concatemer Reads with 5’ and 3’ Primers and Poly-A Tail**: The number of non-concatemer CCS reads with 5’ and 3’ primer and polyA tail detected. This is usually the number for full-length, non-concatemer (FLNC) reads, unless polyA tails are not present in the sample.
- **Unique Primers**: The number of unique primers in the sequence.
- **Mean Reads per Primer**: The mean number of CCS reads per primer.
- **Max. Reads per Primer**: The maximum number of CCS reads per primer.
- **Min. Reads per Primer**: The minimum number of CCS reads per primer.
- **Reads without Primers**: The number of CCS reads without a primer.

CCS Read Classification > Primer Data

- **Bio Sample Name**: The name of the biological sample associated with the primer.
- **Primer Index**: The index number associated with the primer.
- **Primer Name**: A string containing the pair of primer indices associated with this biological sample.
- **CCS Reads**: The number of CCS reads associated with this primer.
- **Mean Primer Quality**: The mean primer quality associated with the primer.
- **Rank Order (Num. Reads)**: The rank order of this primer, sorted by number of reads.

CCS Read Classification > Primer Read Statistics

- Number of reads per primer, sorted by ranking.

CCS Read Classification > Primer Quality Scores

- Histogram of primer scores.

CCS Read Classification > Primer Reads Binned Histograms > Read Length Distribution By Primer

- Heat map of read lengths, sorted by ranking.

CCS Read Classification > Primer Reads Binned Histograms > Primer Quality Distribution By Primer

- Heat map of number of reads by primer scores, sorted by ranking.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.
- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.) For the Iso-Seq 3 application, the default option for Polish CCS is **OFF**, which results in a read score of 0.
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram of the CCS read lengths.

CCS Report > Number of Passes

- Histogram of the number of complete subreads in CCS reads.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Low-Quality Transcripts**: Isoforms with low consensus accuracy, in FASTQ format. We recommend that you work only with High-Quality Transcripts unless there are specific reasons to analyze Low-Quality Transcripts.
- **High-Quality Transcripts**: Isoforms with high consensus accuracy, in FASTQ format. This is the recommended output file to work with.
- **Polished Report**: Read count information for each isoform.
- **Cluster Report**: Report of each full-length read into isoform clusters.
- **Full-Length Non-Concatemer CCS**: Full-length reads that have primers and polyA tails removed, in BAM format.
- **CCS FASTQ**: Circular Consensus Sequences in FASTQ format.

Iso-Seq® 1 Analysis Classify Only Application

This analysis includes only the Classify step of the Iso-Seq 1 algorithm. The application classifies sequencing reads into full-length or non-full length reads.

The Iso-Seq 1 Classify Only application runs **much faster** than the full Iso-Seq 1 application, and can be used to QC the data in advance of running the full Iso-Seq 1 application. Example: The number of full-length non-chimeric reads is a good indicator of data quality.

This application generates full-length cDNA sequences, eliminating the need for computational reconstruction.

Parameters

Advanced Analysis Parameters - CCS	Default Value	Description
By Strand CCS	OFF	For each ZMW, generate two CCS sequences - one for each strand.
Maximum Dropped Fraction	0.8	The maximum fraction of subreads that can be dropped before giving up.
Maximum Subread Length	15,000	The maximum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	50	The minimum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Number of Passes	0	The minimum number of **full** passes for a ZMW to be emitted. Full passes **must** have an adapter identified before and after the insert sequence and so do not include any partial passes at the start and end of the sequencing reaction. For the Iso-Seq 1 application, a 0-pass read can still be full-length.
Minimum Predicted Accuracy	0.8	The **minimum** predicted accuracy of a read, ranging from 0 to 1. (`0.99` indicates that only reads expected to be 99% accurate are emitted.)
Minimal Read Score	0.65	The minimum read score of input subreads.
Minimum SNR	3.75	The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR <3.75 is typically considered lower quality.
Minimum Z Score	-9999	The minimum Z-Score for a subread to be included in the consensus-generating process.
Polish CCS	OFF	Specify whether to polish CCS sequences using Arrow.
Report File Output		Specify the name of the report file to output. (Default = `ccs_report_txt`)
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` "rich" quality tracks.
Filters to add to the Dataset	NONE	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Iso-Seq	Default Value	Description
Customer Primer Sequences	NONE	Specify any custom primers used with the cDNA sample preparation. (The PacBio cDNA protocol uses the Clontech SMARTer primers.) Copy and paste custom primer sequences, in FASTA format, into the field. See "Appendix A - Barcoded Primers" on page 130 for details.
Require PolyA	ON	**ON** means that polyA tails are required for a sequence to be considered full length. **OFF** means sequences do not need polyA tails to be considered full length.
Minimum Sequence Length	50	The minimum sequence length to output.

Reports and Data Files

The Iso-Seq 1 Algorithm Classify Only application generates the following reports:

Transcript Classification > Summary Metrics

- **Number of consensus reads**: The number of consensus isoform reads.
- **Number of five prime reads:** The number of CCS reads with 5' primer detected.
- **Number of three prime reads:** The number of CCS reads with 3' primer detected.
- **Number of poly-A reads:** The number of CCS reads with polyA tail and 3' primer detected.
- **Number of filtered short reads:** The number of reads whose read length is less than the specified Minimum Sequence Length.
- **Number of non-full-length reads:** The number of non-full-length CCS reads missing the polyA tail and/or a terminal signal. (Full-length reads are reads which have both primer and polyA detected.)
- **Number of full-length reads:** The number of full-length CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length non-chimeric reads:** The number of full-length non-artificial-concatemer CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length non-chimeric bases:** The total number of bases in full-length non-artificial-concatemer CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Mean full-length non-chimeric read length**: The mean length of full-length, non-artificial-concatemer CCS reads.

Transcript Classification > Iso-Seq Transcript Classification

- Displays the same information as the **Transcript Classification > Summary Metrics** report.

Transcript Classification > Read Length of Full-Length Non-Chimeric Reads

- Histogram distribution of the number of full-length non-chimeric reads by the read length.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.

- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.) For the Iso-Seq 1 application, the default option for **Polish CCS** is **OFF**, which will result in a read score of 0.
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram distribution of the CCS reads by the read length.

CCS Report > Number of Passes

- Maps CCS reads against the number of passes.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Sequences**: Consensus sequences generated from CCS.
- **Consensus Sequences**: Consensus sequences generated from CCS, in FASTA or FATSQ format.
- **Draft Isoforms**: Intermediate Data Set used to get full length reads.
- **Full-Length Non-Chimeric Reads**: Full-length non-chimeric subreads generated from `pbtranscript` classify.
- **Non-Full-Length Reads**: Non-full-Length reads generated from `pbtranscript` classify.
- **Primer Info**: Per-CCS read annotation and classification results.

Iso-Seq® 1 Analysis with Mapping Application

Use this application to characterize full-length transcripts, then map the transcripts back to the reference genome. A GMAP reference genome is required for alignment.

The Iso-Seq 1 application enables analysis and functional characterization of transcript isoforms for sequencing data generated on PacBio instruments.

This application generates full-length transcripts, eliminating the need for computational reconstruction.

The Iso-Seq 1 application provides accurate information about alternatively spliced exons and transcriptional start and end sites.

The application includes three main steps:

1. **CCS**: Build Circular Consensus Sequences (CCSs) from the sequencing subreads.
2. **Classify**: Classify CCS reads in two groups – full length and non-full length. Identify and remove polyA/T tails, remove primers, and identify read-strandedness. Also remove artificial concatemers but do **not** remove PCR chimeras.
3. **Cluster**: Perform *de novo* clustering and consensus calling. Output polished, full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on predicted accuracies.

GMAP Reference (Required):

- Specify a GMAP reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.

Parameters

Advanced Analysis Parameters - CCS	Default Value	Description
By Strand CCS	OFF	For each ZMW, generate two CCS sequences - one for each strand.
Maximum Dropped Fraction	0.8	The maximum fraction of subreads that can be dropped before giving up.
Maximum Subread Length	15,000	The maximum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	50	The minimum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Number of Passes	0	The minimum number of **full** passes for a ZMW to be emitted. Full passes **must** have an adapter identified before and after the insert sequence and so do not include any partial passes at the start and end of the sequencing reaction. For the Iso-Seq 1 application, a 0-pass read can still be full-length.

Advanced Analysis Parameters - CCS	Default Value	Description
Minimum Predicted Accuracy	0.8	The minimum predicted accuracy of a read, ranging from 0 to 1. (`0.99` indicates that only reads expected to be 99% accurate are emitted.)
Minimal Read Score	0.65	The minimum read score of input subreads.
Minimum SNR	3.75	The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR <3.75 is typically considered lower quality.
Minimum Z Score	-9999	The minimum Z-Score for a subread to be included in the consensus-generating process.
Polish CCS	OFF	Specify whether to polish CCS sequences using Arrow.
Report File Output		Specify the name of the report file to output. (Default = `ccs_report_txt`)
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` "rich" quality tracks.
Filters to add to the Dataset	NONE	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Iso-Seq	Default Value	Description
Allow Extra 5 Exon	OFF	If **ON**, collapse shorter 5' transcripts; if **OFF**, don't collapse them.
Bin by Primer	OFF	Specify binning reads by Primers. This overwrites the **Bin by Read Length Manually** and **Bin by Read Length in KB** options.
Bin by Read Length Manually	NONE	Specify binning reads by manually inputting read length bins. (Example: `[0, 2, 3, 5]` means binning reads into 0-2 kb, 2-3 kb, 3-5 kb, and above 5 kb bins.) This overwrites the **Bin By Read Length in KB** option.
Bin by Read Length in KB	1	Specify binning reads by the read length, in kb.
GMAP nproc	24	The number of processing threads used to run the GMAP aligner. Adjust this value based on your processor power and reference genome size.
Minimum Quiver/Arrow Accuracy	0.99	Specify the minimum predicted consensus accuracy to classify an isoform as high-quality (HQ). All isoforms below this cutoff are considered low-quality (LQ).
Maximum Fuzzy Junction	5	The maximum edit distance between mergeable fuzzy junctions.
Minimum FL Count	2	The minimum FL count to **not** filter a collapsed isoform.
Minimum GMAP aln Coverage	0.99	The minimum query coverage to analyze a GMAP alignment.
Minimum GMAP aln Identity	0.95	The minimum identity to analyze a GMAP alignment.
Minimum Sequence Length	50	The minimum sequence length to output.
Customer Primer Sequences	NONE	Specify any custom primers used with the cDNA sample preparation. (The PacBio cDNA protocol uses the Clontech SMARTer primers.) Copy and paste custom primer sequences, in FASTA format, into the field. See "Appendix A - Barcoded Primers" on page 130 for details.
Trim QVs 3'	30	Specify the number of bases whose Quality Value to ignore in the 3' end.

Advanced Analysis Parameters - Iso-Seq	Default Value	Description
Trim QVs 5'	100	Specify the number of bases whose Quality Value to ignore in the 5' end.
Require PolyA	ON	**ON** means that polyA tails are required for a sequence to be considered full length. **OFF** means sequences do not need polyA tails to be considered full length.
Sample Name	NONE	The name of the input sample. A random string is used when Sample Name is NONE.

Reports and Data Files

The Iso-Seq 1 with Mapping application generates the following reports:

Transcript Clustering > Summary Metrics

- **Number of unpolished consensus isoforms**: The number of consensus isoform reads, both high and low-quality.
- **Number of polished high-quality isoforms**: The number of consensus isoforms that have an estimated accuracy above the specified threshold.
- **Number of polished low-quality isoforms**: The number of consensus isoforms that have an estimated accuracy below the specified threshold.
- **Mean unpolished consensus isoforms read length**: The mean read length of the consensus isoform reads, both high and low-quality.

Transcript Clustering > Read Length of Consensus Isoforms Reads

- Maps the read length of consensus isoform reads against the number of reads.

Transcript Clustering > Average Quality Value of HQ and LQ Isoforms

- Maps the High Quality/Low Quality Isoform average QV against the number of Isoforms with greater than the average QV.

Transcript Classification > Summary Metrics

- **Number of consensus reads**: The number of consensus isoform reads.
- **Number of five prime reads**: The number of CCS reads with 5’ primer detected.
- **Number of three prime reads**: The number of CCS reads with 3’ primer detected.
- **Number of poly-A reads**: The number of CCS reads with polyA tail and 3’ primer detected.
- **Number of filtered short reads**: The number of reads whose read length is less than the specified Minimum Sequence Length.
- **Number of non-full-length reads**: The number of non-full-length CCS reads missing the polyA tail and/or a terminal signal. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length reads**: The number of full-length CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length non-chimeric reads**: The number of full-length non-artificial-concatemer CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)
- **Number of full-length non-chimeric bases**: The total number of bases in full-length non-artificial-concatemer CCS reads. (Full-length reads are reads which have both primers and polyA tail detected.)

- **Mean full-length non-chimeric read length**: The mean length of full-length, non-artificial-concatemer CCS reads.

Transcript Classification > Iso-Seq Transcript Classification

- Displays the same information as the **Transcript Classification > Summary Metrics** report.

Transcript Classification > Read Length of Full-Length Non-Chimeric Reads

- Maps the number of full-length non-chimeric reads against the read length.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.
- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.) For the Iso-Seq 1 application, the default option for “Polish CCS” is **OFF**, which results in a read score of 0.
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram distribution of the CCS reads by the read length.

CCS Report > Number of Passes

- Maps CCS reads against the number of complete subreads per CCS read.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Isoforms**: Consensus Isoforms produced by clustering FLNC reads using the ICE algorithm.
- **Consensus Sequences**: Consensus sequences generated from CCS, in FASTA and FASTQ format.
- **CSV Report (TXT)**: Summary of CCS performance and yield.
- **Draft Isoforms**: Intermediate Data Set used to get full-length reads.
- **Full-Length Non-Chimeric Reads**: Full-length non-chimeric subreads generated from `pbtranscript` classify.
- **High-Quality Isoforms**: Data Set or FASTQ file of isoforms with high consensus accuracy.

- **Low-Quality Isoforms**: Data Set or FASTQ file of isoforms with low consensus accuracy.
- **Non-Full-Length Reads**: Non-full-Length reads generated from `pbtranscript` classify.
- **Primer Info**: Per-CCS read annotation and classification results.
- **Collapsed Isoform Groups**: Displays how redundant HQ isoforms were collapsed into transcripts, and transcripts grouped into gene families.
- **Isoform Abundance**: Counts for each transcript.
- **Collapsed Filtered Isoforms**: Collapsed filtered isoforms, in GFF or FASTQ format.
- **FLnFL Reads Status**: Status of FLNC and NFL reads associated with collapsed isoforms.
- **Gmap SAM Mapping HQ isoforms to Genome:** High-quality isoforms mapped to the genome, in SAM format.
- **Clustering Results**: Report of each full-length read into isoform clusters.

Iso-Seq® 3 Analysis with Mapping Application

Use this application to characterize full-length transcripts, then map the transcripts back to the reference genome. A GMAP reference genome is required for alignment.

The Iso-Seq 3 application enables analysis and functional characterization of transcript isoforms for sequencing data generated on PacBio instruments.

This application generates full-length transcript isoforms, eliminating the need for computational reconstruction.

The Iso-Seq 3 application provides accurate information about alternatively spliced exons and transcriptional start and end sites.

The application includes four main steps:

1. **CCS**: Build Circular Consensus Sequences (CCSs) from each sequencing ZMW.
2. **Classify**: Identify and remove primers (which includes cDNA primers and optionally barcodes). Identify strandedness based on the 5' and 3' primers.
3. **Cluster**: Trim off polyA tails. Also remove artificial concatemers but do **not** remove PCR chimeras. Perform *de novo* clustering and consensus calling. Output full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on estimated accuracies.
4. **Mapping and Collapsing**: Map the high-quality (HQ) transcripts to the user-specified genome using GMAP. Filter out alignments based on user-defined coverage and identity cutoffs. Collapse redundant transcripts that have the same exonic structures into the final set of Collapsed Filtered Isoforms. Output in GFF and FASTQ format.

For barcoded samples: The Iso-Seq 3 application currently pools all demultiplexed reads from the Classify step and outputs only one set of consensus isoforms after the Cluster step. This is suitable for samples that are from the same species but different tissues, or samples of the same genes but different individuals.

If you have samples from **different** species and need to run the Cluster step separately for each primer, you can either run the Iso-Seq 1 application, or run the Iso-Seq 3 Cluster step from the command line.

If you only want to obtain full-length non-concatemer (FLNC) reads and do **not** wish to complete the Cluster step, you can run the Iso-Seq 1 Classify Only application, or, run the Iso-Seq 3 Cluster step from the command line and terminate after the FLNC BAM file is generated.

Iso-Seq 3 determines two FLNC reads to be the same isoform, and will place them in the same cluster, if the two reads:

- Differ less than 100 bp on the 5’ end.
- Differ less than 30 bp on the 3’ end.
- Have no internal gaps that exceed 10 bp.

The Cluster step will only output clusters that have at least two FLNC reads.

Use this analysis application **only** if you have a reference genome. If you do **not** have a reference genome, use the Iso-Seq 3 application instead.

The mapped, (collapsed) unique isoforms are available in the **Data > File Downloads** section in GFF and FASTQ format with the names “Collapsed Filtered Isoforms FASTQ/GFF”. Download this file to obtain the number of final, unique isoforms.

Primer Set (Required):

- Specify a primer sequence file in FASTA format to identify cDNA primers for removal. The primer sequence includes the 5’ and 3’ cDNA primers and (if applicable) barcodes.
- Primer IDs must be specified using the suffix `_5p` to indicate 5’ cDNA primers and the suffix `_3p` to indicate 3’ cDNA primers. The 3’ cDNA primer should **not** include the Ts and is written in reverse complement (see examples below).
- If barcodes were used, they should be included.
- Each primer sequence must be **unique**.

Example 1: The Clontech primer set.

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGG
>3p
GTACTCTGCGTTGATACCACTGCTT
```

Example 2: 4 tissues were multiplexed using barcodes on the 3’ end only.

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGGG
>tissue1_3p
atgacgcatcgtctgaGTACTCTGCGTTGATACCACTGCTT
>tissue2_3p
gcagagtcatgtatagGTACTCTGCGTTGATACCACTGCTT
>tissue3_3p
gagtgctactctagtaGTACTCTGCGTTGATACCACTGCTT
>tissue4_3p
catgtactgatacacaGTACTCTGCGTTGATACCACTGCTT
```

GMAP Reference (Required):

- Specify a GMAP reference genome to align the high-quality (HQ) transcript sequences reads to obtain final set of mapped, unique isoforms.

Parameters

Advanced Analysis Parameters - CCS	Default Value	Description
Maximum Subread Length	15,000	The maximum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates
Minimum Subread Length	50	The minimum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates
Polish CCS	OFF	Specify whether to polish CCS sequences. The default is **OFF** as this reduces run time and does not impact output quality because Polishing is done at the end of the Cluster step.
Emit Individual QVs	OFF	Specify whether to emit `dq`, `iq`, and `sq` "rich" quality tracks. The default is **OFF** as **Polish CCS** is OFF by default.

Advanced Analysis Parameters - Transcript Mapping	Default Value	Description
Allow 5' Shorter Transcripts	OFF	If **ON**, collapse shorter 5' transcripts; if **OFF**, don't collapse them.
Maximum Edit Distance	5	The maximum edit distance between mergeable fuzzy junctions.
Minimum GMAP Alignment Query Coverage	0.99	The minimum query coverage to analyze a GMAP alignment.
Minimum GMAP Alignment Identity	0.95	The minimum identity to analyze a GMAP alignment.
QV cutoff for HQ Transcripts	0.99	Specify the minimum estimated accuracy for a transcript to be considered "High-Quality". Transcripts below the cutoff will be "Low-Quality".
Require and Trim Poly(A) Tail	ON	**ON** means that polyA tails are required for a sequence to be considered full length, and are to be removed. **OFF** means sequences do not need polyA tails to be considered full length. (Example: PCR amplicons with no polyA tail.)

Reports and Data Files

The Iso-Seq 3 with Mapping application generates the following reports:

Transcript Clustering > Summary Metrics

- **Number of polished high-quality isoforms**: The number of consensus isoforms that have an estimated accuracy above the specified threshold. (This is set by the **QV Cutoff for HQ Transcripts** option in the **Advanced Parameters** dialog.)
- **Number of polished low-quality isoforms**: The number of consensus isoforms that have an estimated accuracy below the specified threshold. (This is set by the **QV Cutoff for HQ Transcripts** option in the **Advanced Parameters** dialog.)

Transcript Clustering > Read Length of Consensus Isoforms Reads

- Histogram of the consensus isoform lengths and the distribution of isoforms exceeding a read length cutoff.

Transcript Clustering > Average Quality Value of HQ and LQ Isoforms

- Histogram of the consensus isoform QVs and the distribution of isoforms exceeding a QV cutoff.

CCS Read Classification > Summary Metrics

- **Reads:** The total number of CCS reads.
- **Reads with 5' and 3' Primers:** The number of CCS reads with 5' and 3' primer detected.
- **Non-Concatemer Reads with 5' and 3' Primers:** The number of non-concatemer CCS reads with 5' and 3' primer detected.
- **Non-Concatemer Reads with 5' and 3' Primers and Poly-A Tail**: The number of non-concatemer CCS reads with 5' and 3' primer and polyA tail detected. This is usually the number for full-length, non-concatemer (FLNC) reads, unless polyA tails are not present in the sample.
- **Unique Primers**: The number of unique primers in the sequence.
- **Mean Reads per Primer**: The mean number of CCS reads per primer.
- **Max. Reads per Primer**: The maximum number of CCS reads per primer.
- **Min. Reads per Primer**: The minimum number of CCS reads per primer.
- **Reads without Primers**: The number of CCS reads without a primer.

CCS Read Classification > Primer Data

- **Bio Sample Name**: The name of the biological sample associated with the primer.
- **Primer Index**: The index number associated with the primer.
- **Primer Name**: A string containing the pair of primer indices associated with this biological sample.
- **CCS Reads**: The number of CCS reads associated with this primer.
- **Mean Primer Quality**: The mean primer quality associated with the primer.
- **Rank Order (Num. Reads)**: The rank order of this primer, sorted by number of reads.

CCS Read Classification > Primer Read Statistics

- Number of reads per primer, sorted by ranking.

CCS Read Classification > Primer Quality Scores

- Histogram of primer scores.

CCS Read Classification > Primer Reads Binned Histograms > Read Length Distribution By Primer

- Heat map of read lengths, sorted by ranking.

CCS Read Classification > Primer Reads Binned Histograms > Primer Quality Distribution By Primer

- Heat map of number of reads by primer scores, sorted by ranking.

CCS Read Classification > Primer Reads Binned Histograms > Read Length Distribution By Primer

- Histogram distribution of the primer rank order (by read count) by read length.

CCS Read Classification > Primer Reads Binned Histograms > Primer Quality Distribution By Primer

- Histogram distribution of read primer quality by read count.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.
- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.) For the Iso-Seq 3 application, the default option for **Polish CCS** is **OFF**, which results in a read score of 0.
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram of the CCS read lengths.

CCS Report > Number of Passes

- Histogram of the number of complete subreads in CCS reads.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Read Stat TXT**: Report of full-length read association with collapsed filtered isoforms.
- **Abundance TXT**: Report of read count information for each collapsed filtered isoform.
- **Group TXT:** Report of isoforms mapped into collapsed filtered isoforms.
- **Collapsed Filtered Isoforms GFF:** Mapped, unique isoforms, in GFF format. This is the Mapping step output that is the recommended output file to work with.
- **Collapsed Filtered Isoforms FASTQ:** Mapped, unique isoforms, in FASTQ format. This is the Mapping step output that is recommended output file to work with.
- **Gmap SAM Mapping HQ isoforms to Genome:** High-quality isoforms mapped to the genome, in SAM format.
- **Low-Quality Transcripts:** Isoforms with low consensus accuracy, in FASTQ format. We recommend that you work only with High-Quality Transcripts unless there are specific reasons to analyze Low-Quality Transcripts.
- **High-Quality Transcripts:** Isoforms with high consensus accuracy, in FASTQ format. This is the Cluster step output that is not yet mapped to the genome.

- **Polished Report**: Read count information for each isoform.
- **Cluster Report**: Report of each full-length read into isoform clusters.
- **Full-Length CCS**: Full-length reads that have primers and polyA tails removed, in BAM format.
- **CCS FASTQ**: Circular Consensus Sequences in FASTQ format.

Long Amplicon Analysis (LAA) Application

Use this application to determine phased consensus sequences for pooled amplicon data. The LAA application:

- Allows for accurate allelic phasing and variant calling in large genomic amplicons.
- Supports the phasing and consensus of novel haplotypes in loci of biomedical interest, such as the HLA genes in the MHC region of the human genome.
- Can pool more than 5 distinct diploid amplicons. Reads are clustered into high-level groups, then each group is phased and a consensus generated for each resulting phase using the Arrow algorithm.

The application includes five main steps:

1. **Coarse clustering**: Group reads from different amplicons into different clusters; detect read-to-read similarities and build a graph with the results, then cluster and break the graph into groups of similar reads.
2. **Waterfall**: Align additional reads against a rough consensus sequence generated from each coarse cluster, adding the reads to the cluster that they have the greatest similarity to.
3. **Phasing**: Load the reads for each cluster into the Arrow consensus software. Identify high scoring mutations with Arrow and recursively look for groups of mutations that can separate reads into different haplotypes representing alleles or other PCR products.
4. **Consensus**: Generate a final polished consensus for each haplotype or PCR product using the Arrow model.
5. **Post-Processing Filters**: Detect and separate PCR artifacts from other consensus results. Duplicate sequences are removed, chimeric sequences are identified using the UCHIME algorithm, and other PCR artifacts are identified by overall consensus quality.

Minimum Subread Length (Required):

- The minimum length of subreads to use.

Parameters

Advanced Analysis Parameters - LAA	Default Value	Description
Chimera Filter	ON	Specify whether to activate the chimera filter and separate all consensus chimeric outputs.
Clustering	ON	Specify whether to activate the coarse clustering phase.
Filter Input Reads by Presence of Both Flanking Barcodes	OFF	Specify whether to filter the input reads if both flanking barcodes are present.
Ignore End-Bases	0	When splitting, ignore `N` bases at the end. This prevents excessive splitting caused by degenerate primers.
Maximum Clustering Reads	400	The maximum number of input reads to cluster per barcode.
Maximum Subread Length	0	The maximum length of input reads to use. To **disable**, set to `0`.

Advanced Analysis Parameters - LAA	Default Value	Description
Maximum Phasing Reads	500	The maximum number of input reads to use for phasing and consensus.
Maximum Reads	2000	The maximum number of input reads to cluster per barcode.
Minimum Barcode Score	26	The minimum average barcode score required for subreads.
Minimum Predicted Accuracy	0.95	The minimum predicted consensus accuracy below which a consensus is treated as noise.
Minimum Allele/Haplotype Read Fraction	0.10	The minimum fraction of reads favoring the minor phase required to split a haplotype.
Minimum Allele/Haplotype Reads	20	The minimum number of reads favoring the minor phase required to split a haplotype.
Phasing	ON	Specify that the fine phasing step take place.
Random Number Generator Seed	42	Modulates the reservoir filtering of seeds.
Take Top N Sequences	0	Report only the top `N` consensus sequences for each barcode. To **disable**, use a number less than 1.
Trim Sequence Ends	0	Specify the number of bases to trim from each end of each consensus sequence.

Reports and Data Files

The Long Amplicon Analysis (LAA) application generates the following reports:

Amplicon Inputs > Amplicon Input Molecule Summary

Displays statistics on the type of input molecules seen, summarized by barcode.

- **Barcode Name**: A string containing the pair of barcode names (or indices if not available) for which the following metrics apply.
- **Good**: The number of subreads used in a consensus sequence not categorized as Chimeric or Noise.
- **Good (%)**: The percentage of subreads used in a consensus sequence not categorized as Chimeric or Noise.
- **Chimeric**: The number of subreads used in a consensus sequence flagged as likely coming from PCR cross-over events.
- **Chimeric (%)**: The percentage of subreads used in a consensus sequence flagged as likely coming from PCR cross-over events.
- **Noise**: The number of subreads used in a consensus sequence that has a very low predicted accuracy (<95%) despite sufficient coverage (>20 reads and >10% of all sequences in the current bin) to be called a novel allele.
- **Noise (%)**: The percentage of subreads used in a consensus sequence that has a very low predicted accuracy (<95%) despite sufficient coverage (>20 reads and >10% of all sequences in the current bin) to be called an novel allele.

Amplicon Consensus > Amplicon Consensus Summary

Displays summary statistics of all output consensus sequences and the results of all post-processing filters.

- **Barcode Name**: A string containing the pair of barcode names (or indices if not available) for which the following metrics apply.
- **Sequence Cluster**: An identifying number given to the cluster of sequences from which this consensus sequence was generated, roughly corresponding to one locus or amplicon.
- **Sequence Phase**: An identifying number given to each phased haplotype within a sequence cluster.
- **Length (Bp)**: The length of the consensus amplicon sequence.
- **Estimated Accuracy**: The estimated accuracy of the consensus amplicon sequence.
- **Subreads Coverage**: The number of subreads used to call consensus for this sequence.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Sequences:** Consensus amplicons that passed all sequence quality filters, in FASTQ and zipped-FASTQ format.
- **Chimeric/Noise Consensus Sequences:** Consensus amplicons that failed one or more sequence quality filters, in FASTQ and zipped-FASTQ format.
- **Consensus Sequences Summary:** Combined consensus sequences, summary information and sample map as a single ZIP file for ease of importing into third-party applications for sequence typing.

Minor Variants Analysis Application [Beta]

Use this application to identify and phase minor single nucleotide substitution variants in complex populations. This application is powered by the Juliet SMRT Analysis tool and features:

- Reference-based codon amino acid-calling (indel variants not called) in amplicons ≤4kb, fully spanned by long reads.
- Extensive application reports for the HIV pol coding region, including drug resistance annotation from publicly-available databases.
- Reliable 1% minor variant detection with 6000 high-quality CCS reads with predicted accuracy of ≥0.99 per sample.
- The current version of this application provides additional reports for the HIV pol coding region, but it can be configured for any target organism or gene.

Reference (Required):

- Specify a reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.

Target Config:

- Defines genes of interest within the reference and, optionally, drug resistance mutations for specific variants. Minor Variants Analysis contains one predefined target configuration for HIV HXB2. To specify this target configuration, enter `HIV_HXB2` into the **Target Config** field. To specify a **custom** target configuration for any organism or gene other than HIV HXB2: Enter **either** the path to the target configuration JSON file on the SMRT Link server, **or** the entire content of the JSON file.

Parameters

Advanced Analysis Parameters - Minor Variants	Default Value	Description
Maximum Variant Frequency to Report (%)	100	Specify that only variants whose percentage of the population is **less** than this value be reported. Lowering this value helps to phase low-frequency variants when the highest frequency variant is different from the reference.
Minimum Variant Frequency to Report (%)	0.1	Specify that only variants whose percentage of the population is **greater** than this value be reported. Increasing this value helps to reduce PCR noise.
Phase Variants	ON	Specify whether to phase variants and cluster haplotypes.
Only Report Variants in Target Config	OFF	Specify whether to only report variants that confer drug resistance, as listed in the target configuration file.
Region of Interest	NONE	Specify genomic regions of interest; reads will be clipped to that region. If not specified, specifies **all** reads.
Filters to add to the Data Set	NONE	A comma-separated list of filters to add to the Data Set.

Advanced Analysis Parameters - CCS	Default Value	Description
Maximum Subread Length	21,000	The maximum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Subread Length	10	The minimum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Minimum Predicted Accuracy	0.99	The minimum predicted accuracy of a read, defined as the expected percentage of matches in an alignment of the consensus sequence to the true read. A value of `0.99` indicates that only reads expected to be 99% accurate are emitted.

Reports and Data Files

The Minor Variants Analysis [Beta] application generates the following reports:

Minor Variants > Summary

- **Barcode Name**: The pair of barcode indices for which the following metrics apply. If this was a single-sample analysis, this section of the report will display `NA`.
- **Median Coverage**: The median read coverage across all observed variant positions.
- **Number of Variants**: The number of variants found in the sample.
- **Number of Genes**: The number of genes observed in the sample.
- **Number of Affected Drugs**: The number of drugs to which resistance is conferred by variants in the sample.
- **Number of Haplotypes**: The number of haplotypes with different co-occurring variants found in the sample.
- **Maximum Frequency Haplotypes (%)**: The maximum haplotype frequency reconstructed from the sample.

Minor Variants > Details

- **Barcode Name**: The pair of barcode indices for which the following metrics apply. If this was a single-sample analysis, this section of the report will display `NA`.
- **Position**: The amino acid position of the minor variant, with respect to the current gene.
- **Reference Codon**: The reference codon of the minor variant.
- **Variant Codon**: The mutated codon for the minor variant.
- **Variant Frequency (%)**: The frequency of the minor variant, in percent.
- **Coverage**: The read coverage at the position of the codon.
- **ORF**: The name of the open reading frame/gene.
- **Affected Drugs**: Drugs to which resistance is conferred by the minor variant, according to a database specified in the configuration file.
- **Haplotypes**: The haplotypes associated with this variant.
- **Haplotype Frequencies (%)**: The cumulative haplotype frequencies associated with the variant.

CCS Report > Summary Metrics

- **CCS reads**: The total number of CCS reads.

- **Number of CCS bases**: The total number of consensus bases in the CCS reads.
- **CCS Read Length (mean)**: The mean read length of the CCS reads.
- **CCS Read Score (mean)**: The mean Read Score for the analysis. (The Read Score is a *de novo* prediction of the mapped accuracy of subreads from a single ZMW.)
- **Number of Passes (mean)**: The mean number of complete subreads per CCS read, rounded to the nearest integer.

CCS Report > By Movie

- Lists the same information as the **CCS Report > Summary Metrics** report, but per movie.

CCS Report > CCS Read Length

- Histogram distribution of the CCS reads by the read length.

CCS Report > CCS Read Score

- Maps CCS reads against their quality (Read Score).

CCS Report > Number of Passes

- Maps CCS reads against the number of complete subreads per CCS read.

CCS Report > Number of Passes vs Read Score

- Maps the number of complete subreads per CCS read against the read scores (as Phred QV).

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

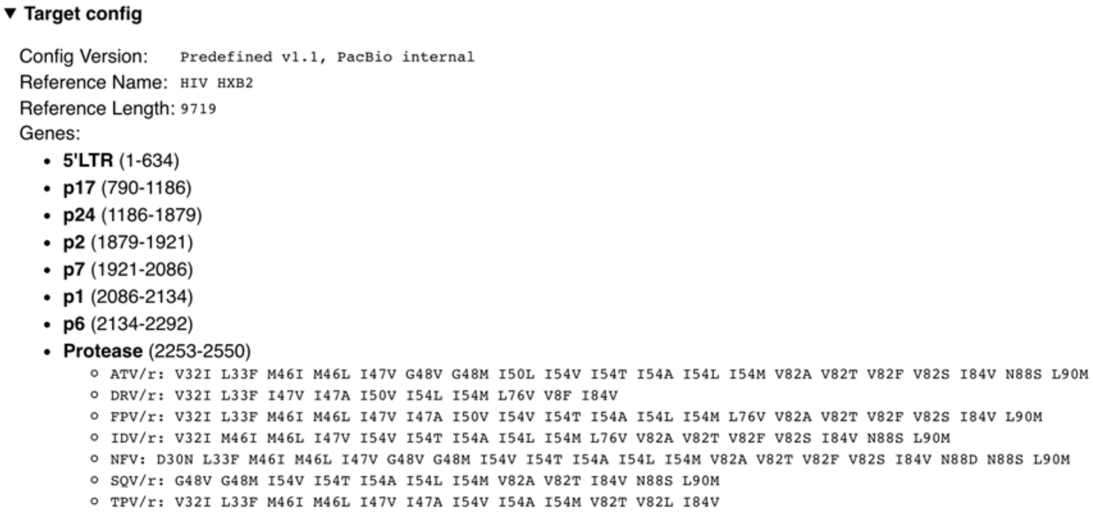
- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Minor Variants Table**: Minor variants report detail information, in CSV format.
- **Minor Variants Report**: Minor variants report information generated; contains the **full** report in JSON format.
- **JSON Results**: Per-sample information on all the samples, in JSON format.
- **Alignments**: Data Set containing alignment results.
- **Consensus Sequences**: Consensus sequences generated from CCS.
- **FASTQ File(s), FASTA File(s)**: Consensus sequences generated from CCS, in FASTA or FASTQ format.
- **Per-Variant Table**: Contains key attributes of each variant called in the sample, as well as phasing of the variants into haplotypes.
- **Minor Variants HTML Reports**: Minor variants report information generated, as a ZIP-compressed HTML file. This includes the **full** report, in human-readable format, and contains four sections:

1. Input Data

Summarizes the data provided, the exact call for `juliet`, and `juliet` version for traceability purposes.

2. Target Config

Summarizes details of the provided target configuration for traceability. This includes the configuration version, reference name and length, and annotated genes. Each gene name (in bold) is followed by the reference start, end positions, and possibly known drug resistance mutations.



▼ **Target config**

Config Version: `Predefined v1.1, PacBio internal`
Reference Name: `HIV HXB2`
Reference Length: `9719`
Genes:

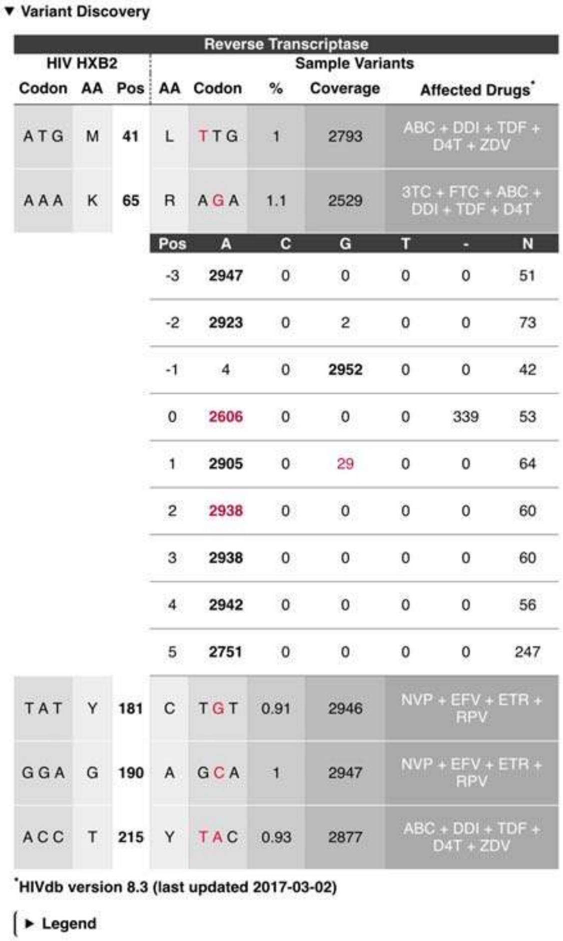
- **5'LTR** (1-634)
- **p17** (790-1186)
- **p24** (1186-1879)
- **p2** (1879-1921)
- **p7** (1921-2086)
- **p1** (2086-2134)
- **p6** (2134-2292)
- **Protease** (2253-2550)
    - `ATV/r: V32I L33F M46I M46L I47V G48V G48M I50L I54V I54T I54A I54L I54M V82A V82T V82F V82S I84V N88S L90M`
    - `DRV/r: V32I L33F I47V I47A I50V I54L I54M L76V V8F I84V`
    - `FPV/r: V32I L33F M46I M46L I47V I47A I50V I54V I54T I54A I54L I54M L76V V82A V82T V82F V82S I84V L90M`
    - `IDV/r: V32I M46I M46L I47V I54V I54T I54A I54L I54M L76V V82A V82T V82F V82S I84V N88S L90M`
    - `NFV: D30N L33F M46I M46L I47V G48V G48M I54V I54T I54A I54L I54M V82A V82T V82F V82S I84V N88D N88S L90M`
    - `SQV/r: G48V G48M I54V I54T I54A I54L I54M V82A V82T I84V N88S L90M`
    - `TPV/r: V32I L33F M46I M46L I47V I47A I54V I54A I54M V82T V82L I84V`

3. Variant Discovery

For each gene/open reading frame, there is one overview table.

Each row represents a variant position. Each variant position consists of the reference codon, reference amino acid, relative amino acid position in the gene, mutated codon, percentage, mutated amino acid, coverage, and possible affected drugs.

Clicking the row displays counts of the multiple-sequence alignment counts of the -3 to +3 context positions.



▼ Variant Discovery

Reverse Transcriptase							
HIV HXB2			Sample Variants				
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*
ATG	M	41	L	TTG	1	2793	ABC + DDI + TDF + D4T + ZDV
AAA	K	65	R	AGA	1.1	2529	3TC + FTC + ABC + DDI + TDF + D4T

Pos	A	C	G	T	-	N
-3	2947	0	0	0	0	51
-2	2923	0	2	0	0	73
-1	4	0	2952	0	0	42
0	2606	0	0	0	339	53
1	2905	0	29	0	0	64
2	2938	0	0	0	0	60
3	2938	0	0	0	0	60
4	2942	0	0	0	0	56
5	2751	0	0	0	0	247

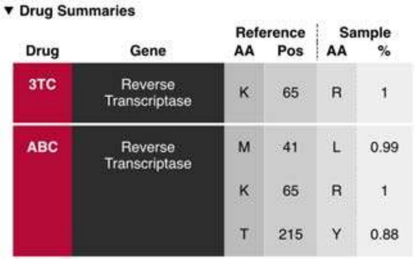
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*
TAT	Y	181	C	TGT	0.91	2946	NVP + EFV + ETR + RPV
GGA	G	190	A	GCA	1	2947	NVP + EFV + ETR + RPV
ACC	T	215	Y	TAC	0.93	2877	ABC + DDI + TDF + D4T + ZDV

*HIVdb version 8.3 (last updated 2017-03-02)

▸ Legend

4. Drug Summaries

Summarizes the variants grouped by annotated drug mutations:

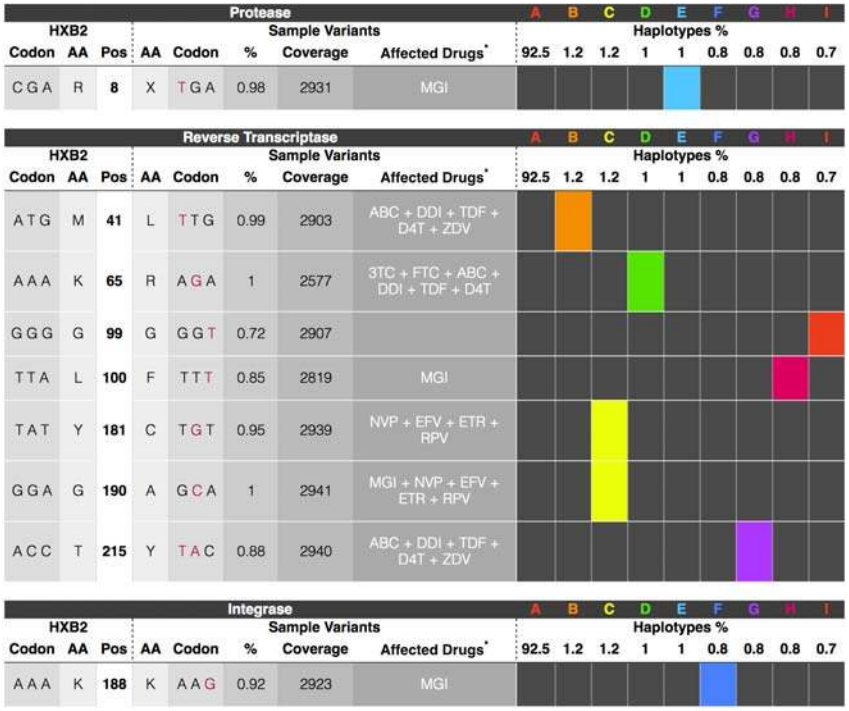


▼ Drug Summaries

		Reference		Sample	
Drug	Gene	AA	Pos	AA	%
3TC	Reverse Transcriptase	K	65	R	1
ABC	Reverse Transcriptase	M	41	L	0.99
		K	65	R	1
		T	215	Y	0.88

Phasing

The default mode is to call amino-acid/codon variants independently. Setting the **Phase Variants** parameter to **On**, variant calls from distinct haplotypes are clustered and visualized in the HTML output.



Protease								A	B	C	D	E	F	G	H	I
HXB2			Sample Variants					Haplotypes %								
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*	92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
CGA	R	**8**	X	TGA	0.98	2931	MGI					🟦				

Reverse Transcriptase								A	B	C	D	E	F	G	H	I
HXB2			Sample Variants					Haplotypes %								
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*	92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
ATG	M	**41**	L	TTG	0.99	2903	ABC + DDI + TDF + D4T + ZDV		🟧							
AAA	K	**65**	R	AGA	1	2577	3TC + FTC + ABC + DDI + TDF + D4T				🟩					
GGG	G	**99**	G	GGT	0.72	2907										🟥
TTA	L	**100**	F	TTT	0.85	2819	MGI								🟥	
TAT	Y	**181**	C	TGT	0.95	2939	NVP + EFV + ETR + RPV			🟨						
GGA	G	**190**	A	GCA	1	2941	MGI + NVP + EFV + ETR + RPV			🟨						
ACC	T	**215**	Y	TAC	0.88	2940	ABC + DDI + TDF + D4T + ZDV							🟪		

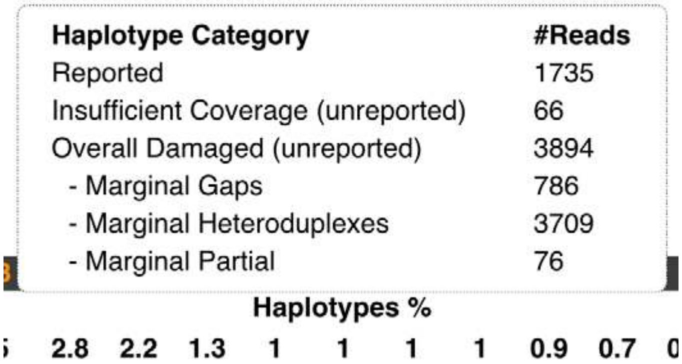
Integrase								A	B	C	D	E	F	G	H	I
HXB2			Sample Variants					Haplotypes %								
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*	92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
AAA	K	**188**	K	AAG	0.92	2923	MGI						🟦			

- The row-wise variant calls are "transposed" onto per-column haplotypes. Each haplotype has an ID: `[A-Z]{1}[a-z]?`.
- For each variant, colored boxes in this row mark haplotypes that contain this variant.
- Colored boxes per haplotype/column indicate variants that co-occur. Wild type (no variant) is represented by plain dark gray. A color palette helps to distinguish between columns.
- The JSON variant positions has an additional `haplotype_hit` boolean array with the length equal to the number of haplotypes. Each entry indicates if that variant is present in the haplotype. A haplotype block under the root of the JSON file contains counts and read names. The order of those haplotypes matches the order of all `haplotype_hit` arrays.

There are two types of tooltips in the haplotype section of the table.

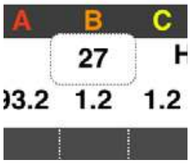
The first tooltip is for the **Haplotypes %** and shows the number of reads that count towards (a) actually reported haplotypes, (b) haplotypes that have less than 10 reads and are not being reported, and (c) haplotypes that are not suitable for phasing. Those first three categories are mutually exclusive and their sum is the total number of reads going into `juliet`. For (c), the three different marginals provide insights into the sample quality; as they are marginals, they are not exclusive and can overlap.

The following image shows a sample with bad PCR conditions:



Haplotype Category	#Reads
Reported	1735
Insufficient Coverage (unreported)	66
Overall Damaged (unreported)	3894
- Marginal Gaps	786
- Marginal Heteroduplexes	3709
- Marginal Partial	76

The second type of tooltip is for each haplotype percentage and shows the number of reads contributing to this haplotype:



Resequencing Application

Use this application to map length and quality-filtered reads against a reference sequence, then to identify consensus and variant sequences.

The Resequencing application:

- Can be used for whole-genome or targeted resequencing analysis.
- Filters reads, maps them to a provided reference sequence, and identifies SNPs.
- Uses BAM as the output file format.

Reference (Required):

- Specify a reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.

Parameters

Advanced Analysis Parameters	Default Value	Description
Minimum Subread Length	0	The minimum length of subreads to use in the assembly.
Filters to Add to the Data Set	NONE	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Consensus	Default Value	Description
Use Score	0	Specify the score to use in the display.
Minimum Confidence	40	The minimum confidence for a variant call to be output to the file `variants.gff`.
Track Description	NONE	Description to display in the header.
Track Name	variants	Name to display in the header.
Purpose	variants	Specify the run mode - `variants` or `coverage`.
Masking	ON	During the polish step, omit regions of reads that have low concordance with the template.
Algorithm	best	• **Quiver** is a variant-calling algorithm that operates on RS II data **only.** • **Arrow** is a more sophisticated algorithm that provides additional information about each read, allowing more accurate consensus calls. Arrow does **not** use the alignment provided by the mapper except for determining how to group reads together at the gross level. Arrow implicitly performs its own realignment, so it is highly sensitive to all variant types, including indels. • **Plurality** is a very simple variant-calling algorithm which does **not** perform any local realignment. It is heavily biased by the alignment produced by the mapper, and it is **insensitive** at detecting indels. • **Best** is the best algorithm based on the data provided.
Minimum Coverage	5	The minimum site coverage that must be achieved for variant calls and consensus to be calculated for a site.

Advanced Analysis Parameters - Alignment	Default Value	Description
Number of .bam files	1	Number of .bam files to create in consolidate mode.
Minimum Concordance	70	The minimum required alignment concordance, in percent.
Consolidate .bam	OFF	Specify whether to merge chunked/gathered .bam files.
Concordant Alignment	ON	Specify whether to map subreads of a ZMW to the same genomic location.
Align Unsplit Polymerase Reads	OFF	Do not split reads into subreads even if subread regions are available.
Minimum Length	50	The minimum required alignment length, in base pairs.
Hit Policy	randombest	Specify how to treat multiple hits: • **random**: Selects a random hit. • **all**: Selects all hits. • **allbest**: Selects all the best score hits. • **randombest**: Selects a random hit from all best score hits. • **leftmost**: Selects a hit which has the best score and the smallest mapping coordinate in any reference.
Algorithm Options		List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0` **Note**: By default, BLASR places gap inconsistently when aligning a sequence and its reverse complement sequence. It is preferable to place gap **consistently** to call a consensus sequence from multiple alignments or call single nucleotide variants (SNPs), as the output alignments will make it easier for variant callers to call variants. To do so, enter `--placeGapConsistently.`

Advanced Analysis Parameters - Reports	Default Value	Description
Maximum Number of Contigs to Plot	25	The maximum number of contigs to plot in the coverage report.
Number of Regions	1000	The number of genome regions in the summary statistics. (This is used for guidance, and is **not** strict.)
Number of Variants	100	The number of top variants to display.
Region Size	0	If specified, use a fixed region size.
Batch Sort Size	10,000	This is an intermediate sort size parameter.
Maximum Region Size	100,000	The upper limit for region size. This is ignored if **Region Size** is set explicitly.
Force the Number of Regions	OFF	If **ON**, try to use this number of regions per reference. Otherwise, the Coverage Summary Report will optimize the number of regions in the case of many references. This is **not** compatible with fixed region sizes.

Reports and Data Files

The Resequencing application generates the following reports:

Consensus Variants > Summary Metrics

- **Reference Consensus Concordance (mean)**: The percent concordance of the consensus sequence compared to the reference.

- **Reference Contig Length (mean)**: The mean length of contigs in the reference sequence.
- **Longest Reference Contig**: The name (FASTA header ID) of the longest reference contig.
- **Percent Reference Bases Called (mean)**: The percentage of the reference sequence for which consensus bases were called.
- **Reference Coverage (mean)**: The mean depth of coverage across the reference sequence.

Consensus Variants > Consensus Calling Results

- **Reference**: The name of the reference sequence.
- **Reference Contig Length**: The length of the reference sequence.
- **Percent Reference Bases Called**: The percentage of reference sequence that has ≥1-fold coverage.
- **Reference Consensus Concordance**: The concordance of the consensus sequence compared to the reference.
- **Reference Coverage**: The depth of coverage across the reference sequence.

Consensus Variants > Observed variants across Reference

- Maps the number of variants across the user-selected reference against the reference start position.

Top Variants > High-Confidence Variance Calls

Displays the position, type and coverage of the top 100 variants, sorted on confidence.

- **Sequence**: The name of the reference sequence.
- **Position**: The position of the variant along the reference sequence.
- **Variant**: The variant position, type, and affected nucleotide.
- **Type**: The variant type: `Insertion`, `Deletion`, or `Substitution`.
- **Coverage**: The coverage at position.
- **Confidence**: The confidence of the variant call.
- **Genotype**: Includes the full number of chromosomes (diploid) or half the number (haploid).

Coverage > Summary Metrics

Displays depth of coverage across references, as well as depth of coverage distribution.

- **Mean Coverage**: The mean depth of coverage across the reference sequence.
- **Missing Bases**: The percentage of the reference sequence without coverage.

Coverage > Coverage across Reference

- Maps coverage of the reference against the reference start position.

Coverage > Depth of Coverage Distribution

- Histogram distribution of the reference regions by the percent coverage.

Mapping Report > Summary Metrics

Mapping is local alignment of a read or subread to a reference sequence.

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Subread Length N50 (mapped)**: The subread length at which 50% of the mapped bases are in subreads longer than, or equal to, this value.
- **Subread Length 95% (mapped)**: The 95th percentile of length of subreads that mapped to the reference sequence.
- **Subread Length Max (mapped)**: The maximum length of subreads that mapped to the reference sequence.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Polymerase Read Length 95% (mapped)**: The 95th percentile of read length of polymerase reads that mapped to the reference sequence.
- **Polymerase Read Length Max (mapped)**: The maximum length of polymerase reads that mapped to the reference sequence.

Mapping Report > Mapping Statistics Summary

Displays mapping statistics per movie.

- **Movie**: Movie name for which the following metrics apply.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped)**: The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.

Mapping Report > Mapped Polymerase Read Length

- Histogram distribution of the number of mapped reads by read length.

Mapping Report > Mapped Subread Concordance

- Histogram distribution of the number of mapped subreads by the percent concordance with the reference sequence.

Mapping Report > Mapped Subread Length

- Histogram distribution of the number of mapped subreads by the subread length.

Mapping Report > Mapped Concordance vs Read Length

- Maps the percent concordance with the reference sequence against the subread length, in base pairs.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Sequences**: Data Set containing consensus sequences.
- **Consensus Contigs**: Consensus contigs in FASTQ format.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference.
- **Coverage and Variant Call Summary**: Coverage and variant call summary for regions (bins) spanning the reference.
- **Variant Calls**: List of variants from the reference, in BED, GFF or VCF format.
- **Alignments**: Data Set containing alignment results.

Site Acceptance Test (SAT) Application

Use this application to generate a report displaying site acceptance test metrics. This application is used to validate all new PacBio systems upon installation, and is designed to be run using specific lambda sequencing data (**lambda/007_tiny**) included with the instrument.

Reference (Required):

- Specify the Lambda NEB reference sequence (included with the installation) to align the SMRT Cells reads to and to produce a consensus sequence.

Parameters

Advanced Analysis Parameters	Default Value	Description
Minimum Subread Length	0	The minimum length of subreads to use in the assembly.
Filters to Add to the Data Set	NONE	A comma-separated list of additional Data Set filters to use.

Advanced Analysis Parameters - Consensus	Default Value	Description
Use Score	0	Specify the score to use in the display.
Minimum Confidence	40	The minimum confidence for a variant call to be output to the file `variants.gff`.
Track Description	NONE	Description to display in the header.
Track Name	variants	Name to display in the header.
Purpose	variants	Specify the run mode - `variants` or `coverage`.
Masking	ON	During the polish step, omit regions of reads that have low concordance with the template.
Algorithm	plurality	• **Quiver** is a variant-calling algorithm that operates on RS II data **only.** • **Arrow** is a more sophisticated algorithm that provides additional information about each read, allowing more accurate consensus calls. Arrow does **not** use the alignment provided by the mapper except for determining how to group reads together at the gross level. Arrow implicitly performs its own realignment, so it is highly sensitive to all variant types, including indels. • **Plurality** is a very simple variant-calling algorithm which does **not** perform any local realignment. It is heavily biased by the alignment produced by the mapper, and it is **insensitive** at detecting indels. • **Best** is the best algorithm based on the data provided.
Minimum Coverage	5	The minimum site coverage that must be achieved for variant calls and consensus to be calculated for a site.

Advanced Analysis Parameters - Alignment	Default Value	Description
Number of .bam files	1	Number of .bam files to create in consolidate mode.
Minimum Concordance	70	The minimum required alignment concordance, in percent.

Advanced Analysis Parameters - Alignment	Default Value	Description
Consolidate .bam	OFF	Specify whether to merge chunked/gathered .bam files.
Concordant Alignment	ON	Specify whether to map subreads of a ZMW to the same genomic location.
Align Unsplit Polymerase Reads	OFF	Do **not** split reads into subreads even if subread regions are available.
Minimum Length	50	The minimum required alignment length, in base pairs.
Hit Policy	randombest	Specify how to treat multiple hits: • **random**: Selects a random hit. • **all**: Selects all hits. • **allbest**: Selects all the best score hits. • **randombest**: Selects a random hit from all best score hits. • **leftmost**: Selects a hit which has the best score and the smallest mapping coordinate in any reference.
Algorithm Options		List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0`

Advanced Analysis Parameters - Reports	Default Value	Description
Number of Regions	1000	Specify the number of genome regions in the summary statistics. (This is used for guidance, and is **not** strict.)
Number of Variants	100	Specify the number of top variants to display.
Region Size	0	If specified, use a fixed region size.
Batch Sort Size	10,000	This is an intermediate sort size parameter.
Maximum Region Size	100,000	The upper limit for region size. This is ignored if **Region Size** is set explicitly.
Force the Number of Regions	OFF	If **ON**, try to use this number of regions per reference. Otherwise, the Coverage Summary Report will optimize the number of regions in the case of many references. This is **not** compatible with fixed region sizes.

Reports and Data Files

The Site Acceptance Test (SAT) application generates the following reports:

Site Acceptance Test Report > Summary Metrics

- **Instrument ID**: The ID number of the Sequel or PacBio RS II instrument on which the Site Acceptance Test is running.
- **Genome Coverage**: The percent of the genome for which consensus bases were called.
- **Consensus Concordance**: The percent concordance of the consensus sequence compared to the reference.
- **Polymerase Read Length Mean (mapped)**: The mean length of polymerase reads that mapped to the reference sequence, including adapters and other unmapped regions.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that could be mapped to the reference genome.

Mapping Report > Summary Metrics

Mapping is local alignment of a read or subread to a reference sequence.

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Subread Length N50 (mapped)**: The subread length at which 50% of the mapped bases are in subreads longer than, or equal to, this value.
- **Subread Length 95% (mapped)**: The 95th percentile of length of subreads that mapped to the reference sequence.
- **Subread Length Max (mapped)**: The maximum length of subreads that mapped to the reference sequence.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Polymerase Read Length 95% (mapped)**: The 95th percentile of read length of polymerase reads that mapped to the reference sequence.
- **Polymerase Read Length Max (mapped)**: The maximum length of polymerase reads that mapped to the reference sequence.

Mapping Report > Mapping Statistics Summary

Displays mapping statistics per movie.

- **Movie**: Movie name for which the following metrics apply.
- **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped)**: The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.
- **Polymerase Read N50 (mapped)**: The read length at which 50% of the mapped bases are in polymerase reads longer than, or equal to, this value.
- **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Subread Length Mean (mapped)**: The mean length of the mapped portion of subreads that mapped to the reference sequence.
- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.

Mapping Report > Mapped Polymerase Read Length

- Histogram distribution of the number of mapped reads by read length.

Mapping Report > Mapped Subread Concordance

- Histogram distribution of the number of mapped subreads by the percent concordance with the reference sequence.

Mapping Report > Mapped Subread Length

- Histogram distribution of the number of mapped subreads by the subread length.

Mapping Report > Mapped Concordance vs Read Length

- Maps the percent concordance with the reference sequence against the subread length, in base pairs.

Consensus Variants > Summary Metrics

- **Reference Consensus Concordance (mean)**: The percent concordance of the consensus sequence compared to the reference.
- **Reference Contig Length (mean)**: The mean length of contigs in the reference sequence.
- **Longest Reference Contig**: The name (FASTA header ID) of the longest reference contig.
- **Percent Reference Bases Called (mean)**: The percentage of the reference sequence for which consensus bases were called.
- **Reference Coverage (mean)**: The mean depth of coverage across the reference sequence.

Consensus Variants > Consensus Calling Results

- **Reference**: The name of the reference sequence.
- **Reference Contig Length**: The length of the reference sequence.
- **Percent Reference Bases Called**: The percentage of reference sequence that has ≥1-fold coverage.
- **Reference Consensus Concordance**: The concordance of the consensus sequence compared to the reference.
- **Reference Coverage**: The depth of coverage across the reference sequence.

Consensus Variants > Observed variants across Reference

- Maps the number of variants across the user-selected reference against the reference start position.

Top Variants > High-Confidence Variance Calls

Displays the position, type and coverage of the top 100 variants, sorted on confidence.

- **Sequence**: The name of the reference sequence.
- **Position**: The position of the variant along the reference sequence.
- **Variant**: The variant position, type, and affected nucleotide.
- **Type**: The variant type: `Insertion`, `Deletion`, or `Substitution`.
- **Coverage**: The coverage at position.
- **Confidence**: The confidence of the variant call.
- **Genotype**: Includes the full number of chromosomes (diploid) or half the number (haploid).

Coverage > Summary Metrics

Displays depth of coverage across references, as well as depth of coverage distribution.

- **Mean Coverage**: The mean depth of coverage across the reference sequence.
- **Missing Bases**: The percentage of the reference sequence without coverage.

Coverage > Coverage across Reference

- Maps coverage of the reference against the reference start position.

Coverage > Depth of Coverage Distribution

- Histogram distribution of the reference regions by the percent coverage.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Consensus Contigs**: Consensus contigs in FASTQ format.
- **Consensus Sequences**: Data Set containing consensus sequences.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference.
- **Coverage and Variant Call Summary**: Coverage and variant call summary for regions (bins) spanning the reference.
- **Variant Calls**: List of variants from the reference, in BED, GFF or VCF format.
- **Alignments**: Data Set containing alignment results.

Structural Variant Calling Application

Use this application to identify structural variants (Default: ≥20 bp) in a sample or set of samples relative to a reference. Variant types identified are insertions, deletions, inversions, and translocations.

Reference (Required):

- Specify a reference genome against which to align the reads and call variants.

Minimum Reads That Support Variant (Total over all samples) (Required, Default = 2):

- Ignore variant calls supported by fewer than this number of reads summed across all samples.
- Separate subreads from a ZMW count only once.

Minimum Length of Structural Variant (bp) (Required, Default = 20):

- Specify the minimum length of structural variants to be output, in base pairs.

Parameters

Advanced Analysis Parameters	Default Value	Description
Minimum Reads That Support Variant (Any one sample)	2	Ignore calls supported by fewer than this number of reads in every sample.
Minimum % of Reads That Support Variant (Any one sample)	20	Ignore calls supported by fewer than this percentage of reads in every sample.

To Launch a Multi-Sample Analysis

1. Click + **Create New Analysis**.
2. Select **Structural Variant Calling** from the Analysis Application list.
3. Select all the Data Sets for all the input samples.
4. In the **Analysis of Multiple Data Sets** > **Analysis Type** list, select **One Analysis on All Data Sets**.

Note: The Data Set field **Bio Sample Name** identifies which Data Sets belong to which biological samples.

- The Bio Sample name is **strongly recommended**. To add or edit this information, see “Editing Data Set Information” on page 25.
- If **multiple** Data Sets with the same Bio Sample Name are selected and submitted, the Structural Variant Calling application **merges** those Data Sets as belonging to the same sample.
- If any input Data Sets do **not** have a Bio Sample Name specified, they are merged (if there are multiple such Data Sets) and their Bio Sample Name is set to `UnnamedSample` in the analysis results.

Reports and Data Files

The Structural Variant Calling application generates the following reports:

Report > Count by Sample (SV Type)

This table describes the type of called variants broken down by individual sample. For each sample, only variants for which the sample has a heterozygous (“0/1”) or homozygous alternative (“1/1”) genotype are considered.

- **Insertions (total bp):** The count and total length (in base pairs) of all called insertions in the sample.
- **Deletions (total bp):** The count and total length (in base pairs) of all called deletions in the sample.
- **Inversions (total bp):** The count and total length (in base pairs) of all called inversions in the sample.
- **Translocations:** The count of all called translocations in the sample.
- **Total Variants (total bp):** The count and total length (in base pairs) of all variants in the sample.

Report > Count by Sample (Genotype)

This table describes the genotype of called variants broken down by individual sample. For each sample, only variants for which the sample has a heterozygous (“0/1”) or homozygous alternative (“1/1”) genotype are considered.

- **Homozygous Variants**: The count of homozygous variants called in the sample.
- **Heterozygous Variants**: The count of heterozygous variants called in the sample.
- **Total Variants:** The count of all called variants in the sample.

Report > Count by Annotation

This table describes the called variants broken down by a set of repeat annotations. Each variant is counted once (regardless of sample genotypes) and assigned to exactly **one** annotation category. Only insertion and deletion variants are considered in this report.

– **Tandem repeat**: Variant sequence is a short pattern repeated directly next to itself.
– **ALU**: Variant sequence matches the ALU SINE repeat consensus.
– **L1**: Variant sequence matches the L1 LINE repeat consensus.
– **SVA**: Variant sequence matches the SVA LINE repeat consensus.
– **Unannotated**: Variant sequence does **not** match any of the above patterns.
– **Total**: The sum of variants from all annotations.

Report > Structural Variants > Length Histogram

- Histogram of the distribution of variant lengths, in base pairs, broken down by individual. For each individual, separate distributions are provided for variants between 10-99 base pairs, 100-999 base pairs, and ≥ 1 kilobase pairs. Each variant is counted once, regardless of sample genotypes.

Data > File Downloads

The following files are available on the Analysis Results page. Additional files are available on the SMRT Link server, in the analysis output directory.

- **Analysis Log**: Log information for the analysis workflow.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Aligned Reads (per sample)**: Aligned reads, in BAM format, separated by individual.
- **BAI Index of Aligned Reads (per sample)**: BAM index files associated with the Aligned Reads BAM files.
- **Structural Variants**: All the structural variants, in VCF format.
- **Merged Sorted Alignment BAM**: Aligned reads, in BAM format, for all individuals.
- **BAI Index**: BAM index file for the Merged Sorted Alignment BAM file.

Working with Barcoded Data

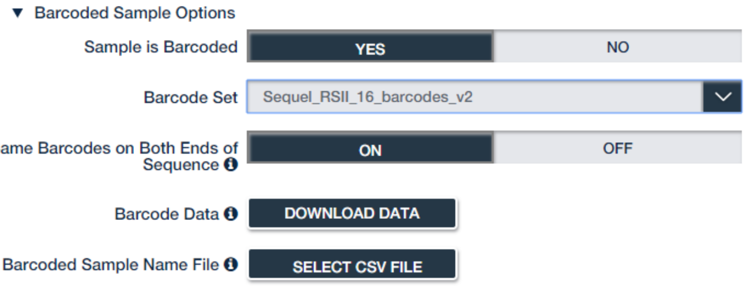
This section describes how to use SMRT Link to work with barcoded data.

The canned data provided with SMRT Link v6.0.0 includes 8 barcode sets:

- `IsoSeqPrimers`
- `RSII_384_barcodes`
- `RSII_96_barcodes`
- `Sequel_RSII_16_Barcodes_Overhang_Adapter`
- `Sequel_RSII_16_barcodes_v1`
- `Sequel_RSII_16_barcodes_v2`
- `Sequel_RSII_384_barcodes_v1`
- `Sequel_RSII_96_barcodes_v1`

Step 1: Specify the Barcode Setup and Sample Names in a Run Design

1. In SMRT Link, create a new run design as described in "Creating a New Run Design" on page 10. **Before** you finish the new Run Design, perform the following steps.



▼ Barcoded Sample Options

Sample is Barcoded: YES / NO (YES selected)

Barcode Set: Sequel_RSII_16_barcodes_v2

Same Barcodes on Both Ends of Sequence: ON / OFF (ON selected)

Barcode Data: DOWNLOAD DATA

Barcoded Sample Name File: SELECT CSV FILE

2. Click **Barcoded Sample Options** and then click **Yes** for **Sample is Barcoded**. Additional fields related to barcoding will be displayed.
3. Specify a **Barcode Set** using the dropdown list.
4. Specify if the **same** barcodes are used on both ends of the sequences.
   - Selecting **On** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
   - Selecting **Off** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such experiments retains any barcode pair combination identified in the Data Set.
5. SMRT Link **automatically** creates a CSV-format **Autofilled Barcode Name File**. The barcode name is populated based on your choice of barcode set, and if the barcodes are the same at both ends of the sequence. The file includes a blank column for the biological sample names.

- (**Optional**) If you want to specify the biological sample names corresponding to each barcode, click **Download Data**, enter the biological sample names associated with the barcodes (Maximum: 40 characters) in the second column, and save the file. If you did **not** use all barcodes in the Autofilled Barcode Name file in the sequencing run, either leave the biological sample name column blank for those barcodes, or delete those rows.
- If you **don't** specify the biological sample name, it will automatically be set to the same value as the barcode name in SMRT Link.
- **Note**: Open the CSV file in a text editor and check that the columns are separated by **commas**, not semicolons.

6. (**Optional**) Select the **Barcoded Sample Name File** you edited in **Step 5**. If you do **not** upload a Barcoded Sample Name File, the biological sample names for those barcodes will automatically be set to the barcode names.
7. Click **Create**.

Note: You can also create a new **Barcode Sample Name File** (**not** recommended):

1. Create a CSV file containing 2 columns.
2. The contents of the first row must be in the form of "Barcode Name,Bio Sample Name". (Valid characters: Alphanumeric; space; dot; underscore; hyphen.)
3. Each row **must** contain a pair of barcode names that exist in the selected barcode set, separated by 2 hyphens. The Bio Sample name is entered after a comma.
   Example: `bc1001--bc1001,biological sample name 1`

Step 2: Perform the Sequencing Run

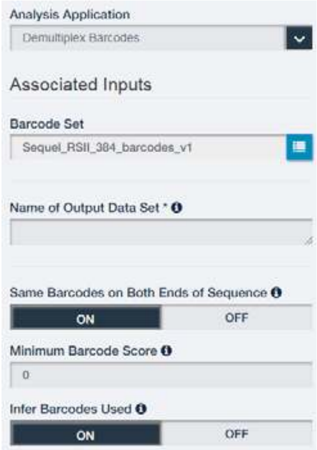
Load the samples and perform the sequencing run, using the Run Design you created in Step 1. The demultiplexing analysis is performed automatically on the SMRT Link Server once the data is transferred from the Sequel System. This creates an analysis of type `Demultiplex Barcodes (Auto)` in the SMRT Analysis module. You can click to select this analysis and review the reports and data created. If everything looks fine, you can continue to **Step 4** and use the demultiplexed Data Set(s) created by the run as input to further analysis.

Note: By default, `Demultiplex Barcodes (Auto)` runs with the **Infer Barcodes Used** option switched on, and creates **one** Data Set per autodetected barcode within the selected barcode set. It also applies a Data Set filter of a minimum barcode score greater than 26 for optimal results in secondary analyses. If used, the analysis parameter **Filters to add to the DataSet** overrides other barcode filtering even if the barcode score set with it is lower than 26.

Step 3: (Optional) Run the Demultiplex Barcodes Application

(**Optional**) If you did **not** specify the barcode setup in the Run Design, or if you need to change any of the parameters used in the `Demultiplex Barcodes` analysis automatically launched from Run Design, run the **Demultiplex Barcodes** application. This application separates reads by barcode and creates a new demultiplexed Data Set that you can then use as input to other secondary analysis applications.

1. Click **+ Create New Analysis**.
2. Enter a **name** for the analysis.
3. Select **Demultiplex Barcodes** from the Applications list.



Analysis Application: Demultiplex Barcodes

Associated Inputs

Barcode Set: Sequel_RSII_384_barcodes_v1

Name of Output Data Set *

Same Barcodes on Both Ends of Sequence: ON OFF

Minimum Barcode Score: 0

Infer Barcodes Used: ON OFF

4. Specify a barcode sequence file.
5. Specify the name for the new demultiplexed Data Set that will display in SMRT Link.
6. Specify if the **same** barcodes are used on both ends of the sequences.
   - Selecting **On** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
   - Selecting **Off** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such data retains any barcode pair combination identified in the Data Set.
7. Specify the **Minimum Barcode Score**: Reads with barcode scores below the value are **not** included in downstream analysis. We recommend that you set this value to `26` for **all** applications.
8. Specify if you want to infer which barcodes were used:
   - **On** infers which subset of barcodes from the selected barcode set were used, and outputs **one** data set for **each** of those inferred barcodes.
   - **Off** outputs **one** data set with all barcodes in the selected barcode set.
9. Click **Start**. After the analysis is finished, a new demultiplexed Data Set is available.

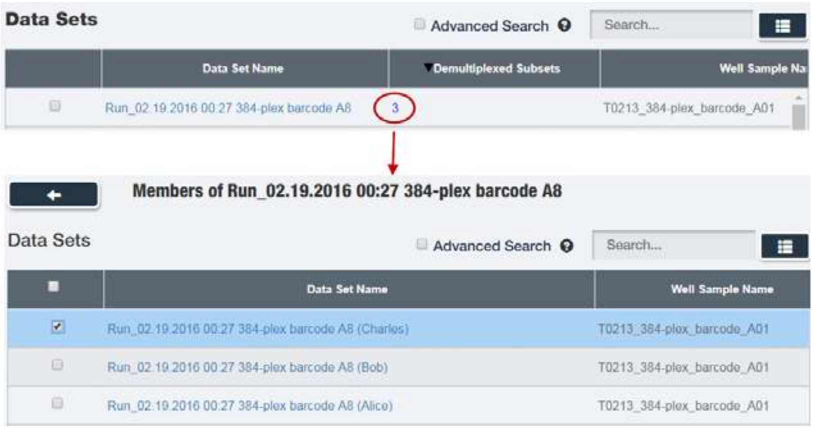
Note: For information about the reports generated by the Demultiplex Barcodes application, see “Reports and Data Files” on page 62.

Step 4: Run Applications Using the Demultiplexed Data as Input

All secondary analysis applications except **Demultiplex Barcodes** and **Structural Variant Calling** can take demultiplexed Data Sets as input.

Note: For **Iso-Seq** analysis using barcoded samples, use the appropriate Iso-Seq application instead of the Demultiplex Barcodes application.

1. Select the secondary analysis application to use.
2. Click the number in the **Demultiplexed Subsets** column, then select the demultiplexed Data Set to use as input:



Data Sets

□ Advanced Search

Search...

	Data Set Name	▼Demultiplexed Subsets	Well Sample Na
□	Run_02.19.2016 00:27 384-plex barcode A8	3	T0213_384-plex_barcode_A01

Members of Run_02.19.2016 00:27 384-plex barcode A8

Data Sets

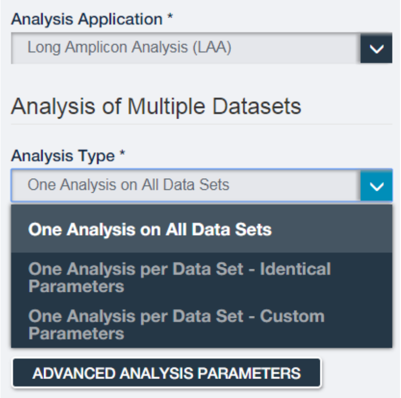
□ Advanced Search

Search...

□	Data Set Name	Well Sample Name
☑	Run_02.19.2016 00:27 384-plex barcode A8 (Charles)	T0213_384-plex_barcode_A01
□	Run_02.19.2016 00:27 384-plex barcode A8 (Bob)	T0213_384-plex_barcode_A01
□	Run_02.19.2016 00:27 384-plex barcode A8 (Alice)	T0213_384-plex_barcode_A01

- You can select the **entire** Data Set as input, or one or more specific outputs from selected barcodes, to a maximum of 16 sub-Data Sets.

3. Additional **Analysis Type** options become available. You can select from the following options:



Analysis Application *

Long Amplicon Analysis (LAA)

Analysis of Multiple Datasets

Analysis Type *

One Analysis on All Data Sets

One Analysis on All Data Sets
One Analysis per Data Set - Identical Parameters
One Analysis per Data Set - Custom Parameters
ADVANCED ANALYSIS PARAMETERS

- **One Analysis on All Data Sets:** Runs **one** analysis using all the selected barcode Data Sets for a maximum of 30 Data Sets.

- **One Analysis per Data Set - Identical Parameters:** Runs a separate analysis for **each** of the selected barcode Data Sets, using the **same** parameters, for a maximum of 384 Data Sets. Optionally click **Advanced Analysis Parameters** and modify parameters.
- **One Analysis per Data Set - Custom Parameters:** Runs a separate analysis for **each** of the selected barcode Data Sets, using **different** parameters for each Data Set, for a maximum of 16 Data Sets. Click **Advanced Analysis Parameters** and modify parameters. Then click **Start and Create Next**. You can then specify parameters for each of the included barcode Data Sets.

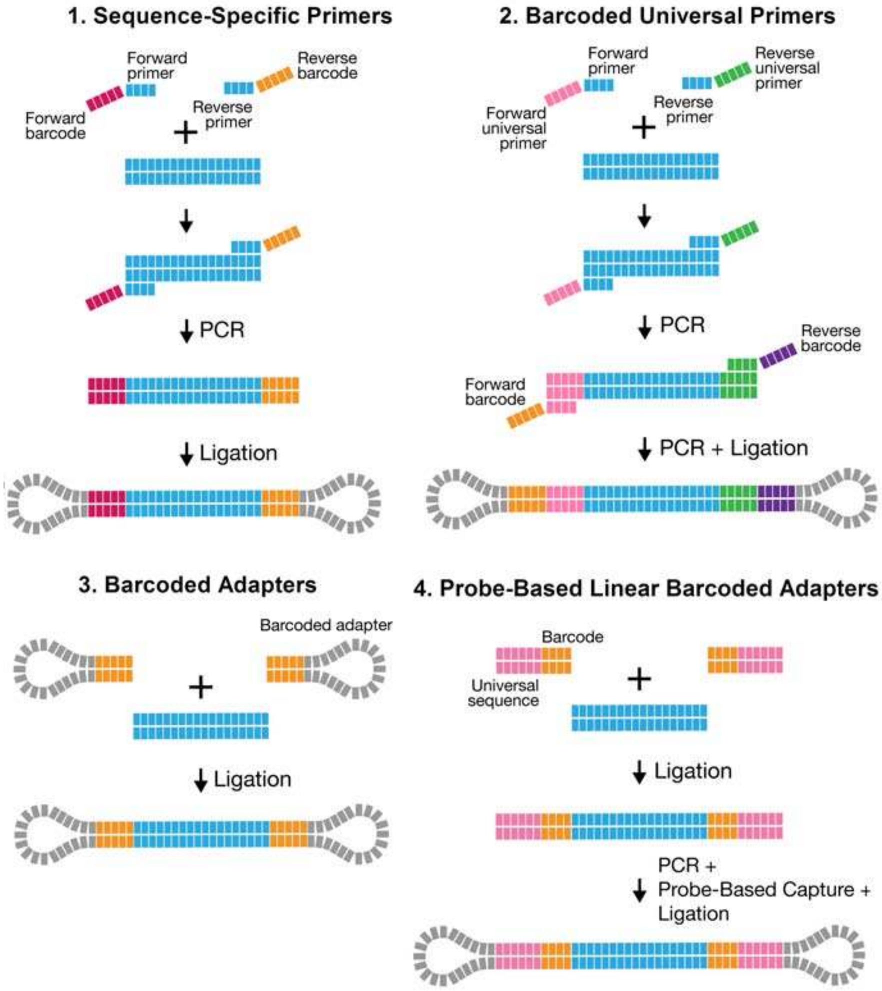
4. Click **Start** to submit the analysis.

Demultiplex Barcodes Application Details

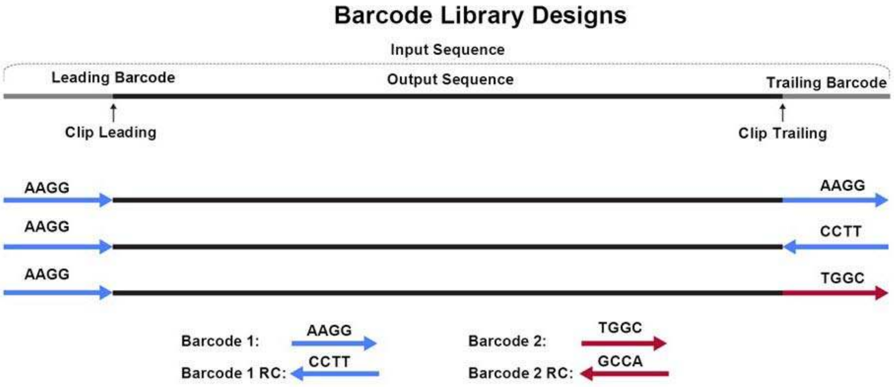
The **Demultiplex Barcodes** application identifies barcode sequences in PacBio single-molecule sequencing data. It **replaced** `pbbarcode` and `bam2bam` for demultiplexing, starting with SMRT Analysis v5.1.0.

Demultiplex Barcodes can demultiplex samples that have a unique per-sample barcode pair and were pooled and sequenced on the same SMRT Cell. There are four different methods for barcoding samples with PacBio technology:

1. Sequence-specific primers
2. Barcoded universal primers
3. Barcoded adapters
4. Probe-based linear barcoded adapters



In addition, there are three different barcode library designs.



Terminology

library	Demultiplex Barcodes
tailed	same
symmetric	same
asymmetric	different

The **Demultiplex Barcodes** application in SMRT Link supports demultiplexing of subreads. The following terminology is based on the per (sub-) read view.

Demultiplexing of CCS reads is possible on the command line.

Symmetric Mode

For **symmetric** and **tailed** library designs, the **same** barcode is attached to both sides of the insert sequence of interest. The only difference is the orientation of the trailing barcode. For barcode identification, one read with a single barcode region is sufficient. This is most commonly the case when using barcoded SMRTbell adapters and for target enrichment (non-hairpin) adapters. This is also the default scoring mode in SMRT Link v6.0.0 and later.

Asymmetric Mode

Barcode sequences are **different** on the forward and reverse ends of the insert. Asymmetric mode is most commonly used when appending barcodes during a single round of PCR with barcoded primers. Pacific Biosciences recommends using this mode **only** for cases when both ends of the insert are expected to be sequenced for most molecules in the SMRT Cell.

When running the **Demultiplex Barcodes** applications in SMRT Link, set the **Same Barcodes on Both Ends of the Sequence** option to **Off**.

Mixed Mode

Libraries that use symmetric and asymmetric labeling are **not** supported.

Workflow

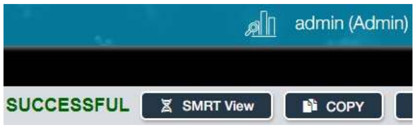
By default, **Demultiplex Barcodes** processes input reads grouped by ZMW, **except** if the `--per-read` option is used. All barcode regions along the read are processed individually. The final per-ZMW result is a summary over all barcode regions. Each ZMW is assigned to a pair of selected barcodes from the provided set of candidate barcodes. Subreads from the same ZMW will have the same barcode and barcode quality. For a particular target barcode region, every barcode sequence gets aligned as given and as reverse-complement, and higher scoring orientation is chosen. This results in a list of scores over all candidate barcodes.

Visualizing Data Using SMRT® View

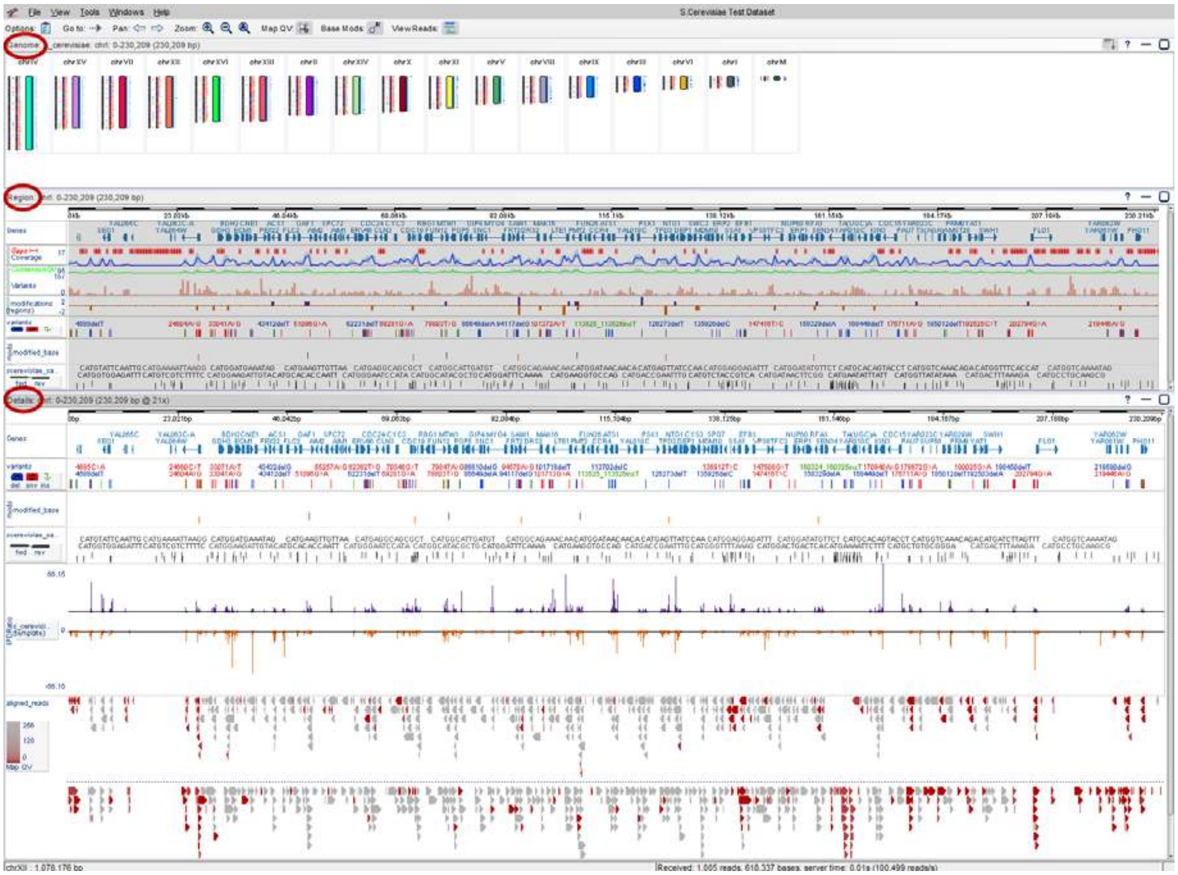
Once an analysis has successfully completed, visualize the results using **SMRT View**; a genome browser that displays sequencing data generated by the Sequel System.

Using SMRT View

1. In SMRT Link, select **SMRT Analysis**. A list of **all** analyses displays.
2. (**Optional**) Click the **Successful** button to see only successfully-completed analyses.
3. Click the name of a successfully-completed analysis to visualize. (**Note**: The analysis **must** have produced alignments as output for the SMRT View button to display.)
4. Click the **SMRT View** button located at the upper-right of the page.



5. The SMRT View application downloads to your computer and displays the data in three panels: **Genome**, **Region**, and **Details**.


File View Tools Windows Help
S.Cerevisiae Test Dataset
Options: Go to: Pan: Zoom: Map QV Base Mods View Reads
Genome
Region
Details
Genes
Coverage
Variants
modified_base
aligned_reads
Map QV
chrXII : 1,078,176 bp
Received: 1,005 reads, 618,337 bases, server time: 0.01s (100,499 reads/s)

- The **Genome** panel displays whole chromosomes or DNA segments, along with significant points of interest. The panel displays **only** if the secondary analysis data includes multiple genomes, chromosomes, or segments.
- The **Region** panel acts as a summary of the data. It displays metrics such as coverage and variants, and allows fast navigation across data to identify regions of interest.
- The **Details** panel allows drilling down to base-level resolution and visualizing SNPs, indels and kinetics used for base modification detection.

6. (**Optional**) **Click** a genome, chromosome or DNA segment to select it. Or, click and drag to select a section of interest. The Region panel displays the selection in greater detail.
7. In the **Region** panel, click and drag to select a smaller section. That section displays in the Details panel.
8. In the **Details** panel, click and drag to view the smallest area, down to the individual bases.

Note: To run SMRT View, 64-bit Java (Version 8 or later) **must** be installed on your local Windows or Macintosh computer.

Installing 64-bit Java 8 on a Windows Operating System

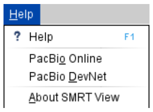
1. Use **Control Panel > Programs and Features** to check for and uninstall **all** existing versions of Java software.
2. Go to **http://www.java.com/en/download/manual.jsp**.
3. Click **Windows Offline (64-bit)**. This downloads a `x64.exe` file. (**Note**: Other Java versions are 32-bit, and will **not** work with SMRT View.)
4. Double-click the `.exe` file to start the Java installer, and follow the installer directions.
5. After the installation is finished, restart the browser.

Installing 64-bit Java 8 on Mac OS

Note: This requires Mac OS 10.7.3 or later.

1. Use the Finder to search for **all** existing versions of Java software, then drag them to the Trash to uninstall.
2. Go to **http://www.java.com/en/download/manual.jsp**.
3. Click **Mac OS X**. This downloads a `x64.dmg` file.
4. Double-click the `.dmg` file to mount the installer volume.
5. Double-click the Java icon to start the Java installer, and follow the installer directions.
6. After the installation is finished, restart the browser.

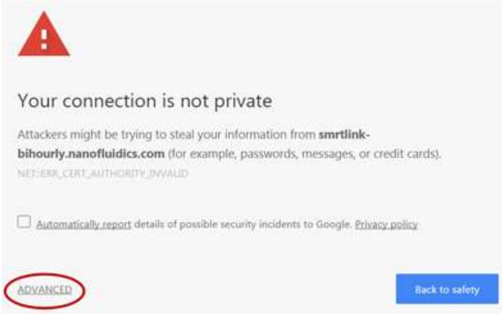
For more information on visualizing data, see the **SMRT® View Online Help**. To access the help, choose **Help > Help**.



Using the PacBio® Self-Signed SSL Certificate

SMRT Link v6.0.0 ships with a PacBio Self-Signed SSL Certificate. If this is used at your site, security messages display when you try to login to SMRT Link for the **first time** using the Chrome browser. These messages may also display **other times** when accessing SMRT Link.

1. The first time you start SMRT Link after installation, you see the following. Click the **Advanced** link.



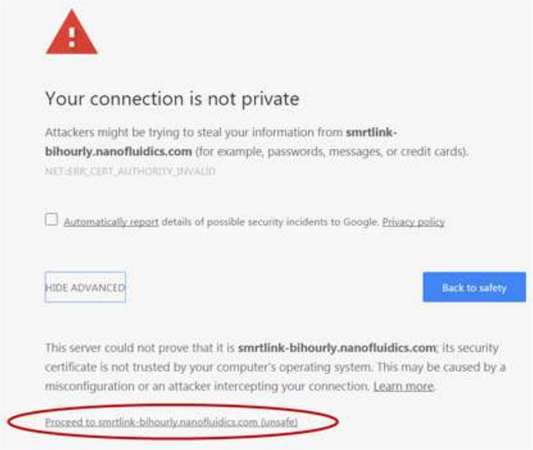
Your connection is not private

Attackers might be trying to steal your information from **smrtlink-bihourly.nanofluidics.com** (for example, passwords, messages, or credit cards).

NET::ERR_CERT_AUTHORITY_INVALID

□ Automatically report details of possible security incidents to Google. Privacy policy

2. Click the **Proceed...** link. (You may need to scroll down.)



Your connection is not private

Attackers might be trying to steal your information from **smrtlink-bihourly.nanofluidics.com** (for example, passwords, messages, or credit cards).

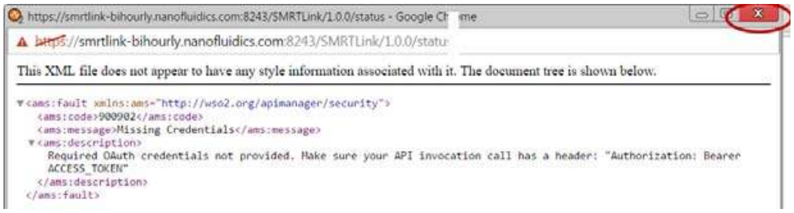
NET::ERR_CERT_AUTHORITY_INVALID

□ Automatically report details of possible security incidents to Google. Privacy policy

This server could not prove that it is **smrtlink-bihourly.nanofluidics.com**; its security certificate is not trusted by your computer's operating system. This may be caused by a misconfiguration or an attacker intercepting your connection. Learn more.

Proceed to smrtlink-bihourly.nanofluidics.com (unsafe)

3. Close the window by clicking the **Close** box in the corner.



https://smrtlink-bihourly.nanofluidics.com:8243/SMRTLink/1.0.0/status - Google Ch me

https://smrtlink-bihourly.nanofluidics.com:8243/SMRTLink/1.0.0/statu

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<ams:fault xmlns:ams="http://wso2.org/apimanager/security">
   <ams:code>900902</ams:code>
   <ams:message>Missing Credentials</ams:message>
  ▼<ams:description>
     Required OAuth credentials not provided. Make sure your API invocation call has a header: "Authorization: Bearer
     ACCESS_TOKEN"
   </ams:description>
 </ams:fault>
```

The **Login** dialog displays, where you enter the User Name and Password. The next time you access SMRT Link, the Login dialog displays **directly**.

Sequel® System Output Files

This section describes the data generated by the PacBio Sequel System for each SMRT Cell transferred to network storage.

File Structure

Following is a sample of the file and directory structure output by the Sequel System:

```
<your_specified_output_directory>/r54008_20160116_003347/1_A01
|-- m54008_160116_003634.scraps.bam
|-- m54008_160116_003634.scraps.bam.pbi
|-- m54008_160116_003634.subreads.bam
|-- m54008_160116_003634.subreads.bam.pbi
|-- m54008_160116_003634.subreadset.xml
|-- m54008_160116_003634.sts.xml
|-- m54008_160116_003634.transferdone
|-- m54008_160116_003634.adapters.fasta
```

In this example, `/r54008_20160116_003347` is a directory containing the output files associated with **one** run.

- – `r54008` is the instrument ID number.
- – `20160116_003347` is the run **date**, in `YYYYMMDD` format, and **time**, in UTC format.
- The run directory includes a subdirectory for **each** collection/cell associated with a sample well - in this case `1_A01`. The collection/cell subdirectory contains output files of interest, described in this document.

Subreads.BAM File

The Sequel System outputs **one** `subreads.bam` file per collection/cell, which contains unaligned base calls from high-quality regions. This file is transferred from the instrument to network storage, then is used as **input** for secondary analysis by Pacific Biosciences' SMRT Analysis software.

Data in a `subreads.bam` file is analysis-ready; all of the data present should be quality-filtered for downstream analyses. Subreads that contain information such as double-adapter inserts or single-molecule artifacts are **not** used in secondary analysis, and are excluded from this file and placed in `scraps.bam`.

- The BAM format is a binary, compressed, record-oriented container format for raw or aligned sequence reads. The associated SAM format is a text representation of the same data. The BAM specifications are maintained by the SAM/BAM Format Specification Working Group.

- BAM files produced by the Sequel System are **fully compatible** with the BAM specification.

  For more information on the BAM file format specifications, see. **http://pacbiofileformats.readthedocs.io/en/5.1/BAM.html**

BAM.PBI File

Pacific Biosciences' previous alignment file format (`cmp.h5`) contained a data table (the **alignment index**) that recorded auxiliary identifying information and precomputed summary statistics per aligned read. This table:

- Enabled fast random access to aligned reads satisfying fairly complex searches, for example, reads from a specific list of ZMWs which had unambiguous mapping (MapQV==254), or a read with a given read name.
- Allowed summary reports (read length, mapped identity/accuracy, and so on) to be constructed by quick operations over the alignment index instead of loading all of the sequence reads for each analysis.

To provide backwards-compatibility with the APIs enabled for accessing the `cmp.h5` file, a new BAM companion file was created - the **PacBio BAM index**, which supports the two use cases above.

For more information on the Pacific Biosciences BAM.PBI file format specifications, see **http://pacbiofileformats.readthedocs.io/en/3.0/ PacBioBamIndex.html**.

Other Output Files

- `.subreadset.xml`: This file is needed to import data into SMRT Link.
- `.scraps.bam` and `.scraps.bam.pbi`: These files contain sequence data outside of the High Quality region, rejected subreads, excised adapter and possible barcode sequences, as well as spike-in control sequences. (The basecaller marks regions of single molecule sequence activity as high-quality.)

  Note: This applies to files generated by Sequel Instrument Control Software (ICS) v3.1.0 or later.
- `.sts.xml`: Contains summary statistics about the collection/cell and its post-processing.
- `.control` files: Contains sequence data for spike-in-control reads.
- `.transferdone`: Contains a list of files successfully transferred.

Frequently Asked Questions
What are the minimum files needed to analyze data on SMRT Link? • `.bam` file • `bam.pbi` file • `subreadset.xml` file
What is the average size of the file bundle for a 6-hour movie? Approximately 5 Gb.
What is the difference between a regular .bam file and an aligned.bam file? The `subreads.bam` file contains all the subreads sequences, while the `aligned.bam` file additionally contains the genomic coordinates of the reads mapped to a reference sequence. The `subreads.bam` file is created by the PacBio Sequel System, while the `aligned.bam` file is created by SMRT Link after running Resequencing or Mapping analysis applications.

Secondary Analysis Output Files

This is data produced by secondary analysis, which is performed on the primary analysis data generated by the instrument.

- All files for a specific analysis reside in **one** directory named according to the analysis job ID number.
- Every analysis result has the following file structure. **Example**:

```
$SMRT_ROOT/userdata/jobs_root/000/000000/
├── html/
│   └── <...>
├── logs/
│   ├── master.log
│   └── pbsmrtpipe.log
├── tasks/
│   ├── pbproject.tasks.toolname-0
│   │   ├── stdout
│   │   ├── stderr
│   │   ├── run.sh
│   │   ├── cluster.sh
│   │   ├── cluster.stdout
│   │   ├── cluster.stderr
│   │   └── <...>
│   └── pbproject.tasks.toolname-1
│       └── <...>
├── workflow/
│   └── <...>
├── index.html
├── job.stderr
├── job.stdout
├── pbscala-job.sh
├── pbscala-job.stderr
├── pbscala-job.stdout
└── preset.json
```

- `html/`: Contains web-browsable analysis metadata and summary statistics.
- `logs/`: Contains log files for the analysis job.
  - `master.log` : Global logs entries of each significant step in the analysis and snippets from a task's `stderr` output if the analysis failed.
  - `pbsmrtpipe.log`: Global logs similar to `master.log,` but less verbose.
- `tasks/`: Contains subdirectories for each workflow task, along with executable scripts, tools contracts, and `stderr/stdout` for the analysis job.
  - `pbproject.tasks.toolname-0/`: Sample name with a numeric end specifying the chunk instance. (This is replaced with `<task_id>` below)
  - `<task_id>/stdout`: General task `stdout` log collection.
  - `<task_id>/stderr`: General task `stderr` log collection.
  - `<task_id>/run.sh`: The SMRT Tools command for the given analysis task.

  - `<task_id>/cluster.sh`: The JMS submission script wrapping `run.sh`. (This is generated in distributed mode only.)
  - `<task_id>/cluster.stdout`: The `stdout` collection for the `cluster.sh` script. (This requires distributed mode.)
  - `<task_id>/cluster.stderr:` The `stderr` collection for the `cluster.sh` script. (This requires distributed mode.)
- `workflow/`: Contains JSON files for analysis settings and workflow diagrams.
- `index.html`: The index page for the HTML files in the `html` directory.
- `job.stderr`: The `pbsmrtpipe` command `stderr` collection.
- `job.stdout`: The `pbsmrtpipe` command `stdout` collection.
- `pbscala-job.sh`: The `pbsmrtpipe` command used to launch the analysis.
- `pbscala-job.stderr`: Log collection of `stderr` output from `pbscala`.
- `pbscala-job.stdout`: Log collection of `stdout` output from `pbscala`.
- `preset.xml`: Contains global, job-distribution, and task-specific parameters.

Within the various `task` directories are several types of output files. You can use these data files as input for further downstream processing, pass on to collaborators, or upload to public genome sites. Depending on the analysis application being used, the `task` directories contain files in the following formats:

- **BAM**: Binary version of the Sequence Alignment Map (SAM) format. (See http://genome.ucsc.edu/goldenPath/help/bam.html for details.)
- **BAI**: The index file for a file generated in the BAM format. (This is a non-standard file type.)
- **FASTA**: Sequence files that contains either nucleic acid sequence (such as DNA) or protein sequence information. FASTA files store multiple sequences in a single file. (See http://en.wikipedia.org/wiki/FASTA_format for details.)
- **GFF**: General Feature Format, used for describing genes and other features associated with DNA, RNA and Protein sequences. (See http://genome.ucsc.edu/FAQ/FAQformat#format3 for details.)
- **VCF**: Variant Call Format, for use with the molecular visualization and analysis program VMD. (See http://en.wikipedia.org/wiki/Variant_Call_Format for details.)
- **BED**: Format that defines the data lines displayed in an annotation track. (See http://genome.ucsc.edu/FAQ/FAQformat#format1 for details.)
- **CSV**: Comma-Separated Values file. Can be viewed using Microsoft Excel or a text editor.

To Download Data Files Created by SMRT Link:

1. On the Home Page, select **SMRT Analysis**. You see a list of **all** analyses.
2. Click the analysis link of interest.
3. Click **Data > File Downloads**, then click the appropriate file. The file is downloaded according to your browser settings.
   - (**Optional**) Click the small icon to the right of the file name to copy the file’s path to the Clipboard.

Configuration and User Management

LDAP

SMRT Link supports the use of LDAP for user login and authentication. **Without** LDAP integration with SMRT Link, only **one** user (with the login `admin/admin`) is enabled. SMRT Link **must** be integrated and configured to work with LDAP at your site **before** you can add SMRT Link users, or modify their roles.

- For details on integrating LDAP and SMRT Link, see the document **SMRT Link Software Installation (v6.0.0)**.

SSL

SMRT Link allows the use of Secure Sockets Layer (SSL) to enable access via HTTP over SSL (HTTPS), so that SMRT Link logins and data are encrypted during transport to and from SMRT Link. SMRT Link includes an Identity Server, which can be configured to integrate with your LDAP/AD servers and enable user authentication using your organizations' user name and password. To ensure a secure connection between the SMRT Link server and your browser, the SSL Certificate can be installed **after** completing SMRT Link installation.

It is important to note that PacBio will **not** provide a Signed SSL Certificate, however – once your site has obtained one – PacBio tools can be used to install it and configure SMRT Link to use it. You will need a certificate issued by a Certificate Authority (CA, sometimes referred to as a 'certification authority'). PacBio has tested SMRT Link with certificates from the following certificate vendors: VeriSign, Thawte and digicert.

Note: Pacific Biosciences recommends that you consult your IT administrator about obtaining an SSL Certificate.

Alternatively, you can use your site's Self-Signed Certificate.

SMRT Link ships with a PacBio self-signed SSL Certificate. If used, **each** user will need to accept the browser warnings related to access in an insecure environment. Otherwise, your IT administrator can configure desktops to **always** trust the provided self-signed Certificate. Note that SMRT Link is installed within your organization's secure network, behind your organization's firewall.

- For details on updating SMRT Link to use an SSL Certificate, see the document **SMRT Link Software Installation (v6.0.0)**.

The following procedures are available **only** for SMRT Link users whose role is **Admin**.

Adding and Deleting SMRT Link Users

1. Choose **Configure > User Management**.
2. There are 2 ways to find users:
   - To display **all** SMRT Link users: Click **Display all Enabled Users**.
   - To find a specific user: Enter a user name, or partial name, and click **Search By Name**.
3. Click the desired user. If the user status is **Enabled**, the user has access to SMRT Link; **Disabled** means the user **cannot** access SMRT Link.
   - To **add** a SMRT Link user: Click the **Enabled** button, then assign a role. (See below for details.)
   - To **delete** a SMRT Link user: Click the **Disabled** button.
4. Click **Save**.

Assigning User Roles

SMRT Link supports three user roles: **Admin**, **Lab Tech**, and **Bioinformatician**. (A fourth role, **Instrument**, displays in the User Management page. The Sequel Instrument Control Software uses this role to communicate with SMRT Link. Do **not** assign any SMRT Link users to this role.) Roles define which SMRT Link modules a user can access. The following table lists the privileges associated with the three user roles:

Tasks/Privileges	Admin	Lab Tech	Bioinformatician
Add/Delete SMRT Link Users	Y	N	N
Assign roles to SMRT Link users	Y	N	N
Update SMRT Link software	Y	N	N
Access Sample Setup Module	Y	Y	N
Access Run Design Module	Y	Y	N
Access Run QC Module	Y	Y	Y
Access Data Management Module	Y	Y	Y
Access SMRT Analysis Module	Y	Y	Y

1. Choose **Configure > User Management**.
2. There are 2 ways to find users:
   - To display **all** SMRT Link users: Click **Display all Enabled Users**.
   - To find a specific user: Enter a user name, or partial name, and click **Search By Name**.
3. Click the desired user.

4. Click the **Role** field and select one of the three roles. (A **blank** role means that this user **cannot** access SMRT Link.)
   - **Note**: There can be **multiple** users with the Admin role; but there **must** always be at least **one** Admin user.
5. Click **Save**.

Hardware/Software Requirements

Client Hardware Requirements

SMRT Link requires a minimum screen resolution of 1600 by 900 pixels.

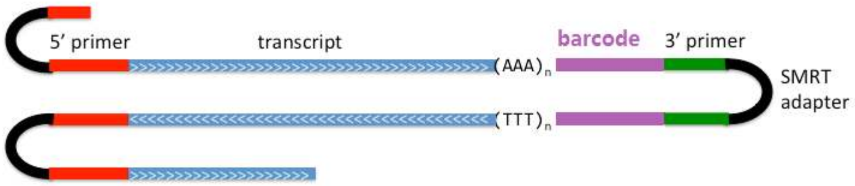
Client Software Requirements

- SMRT Link **requires** the Google® Chrome web browser, version 64 or later.
- PacBio **recommends** 64-bit Java (Version 8 or later) installed on your local Windows or Mac OS host to run SMRT View.

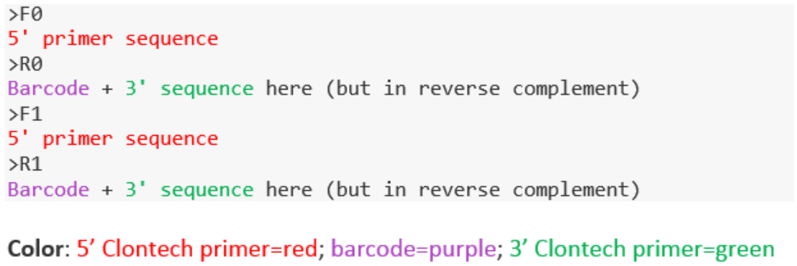
Note: SMRT Link **Server** hardware and software requirement are listed in the document **SMRT Link Software Installation (v6.0.0)**.

Appendix A - Barcoded Primers

SMRTbell® templates of transcripts with barcoded 3' primers look like this:



To use barcoded primers, first create a text primer file using the following format:



```
>F0
5' primer sequence
>R0
Barcode + 3' sequence here (but in reverse complement)
>F1
5' primer sequence
>R1
Barcode + 3' sequence here (but in reverse complement)
```

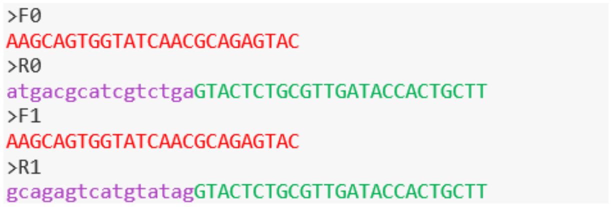
Color: 5' Clontech primer=red; barcode=purple; 3' Clontech primer=green

You can add additional barcoded primers – just name them `F0/R0`, `F1/R1`, `F2/R2`, and so on. **Note**: The `F0/R0`, `F1/R1`, `F2/R2`...naming system is **required**. If you do not follow this format, starting from `F0/R0`, the analysis will **fail**.

For example, if you used the following barcoded oligo-dTs in 3' for your experiments:

Oligo	**Sequence**
dT_BC1	AAGCAGTGGTATCAACGCAGAGTACtcagacgatgcgtcatTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
dT_BC2	AAGCAGTGGTATCAACGCAGAGTACctatacatgactctgcTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

Then the primer file should look like this:



```
>F0
AAGCAGTGGTATCAACGCAGAGTAC
>R0
atgacgcatcgtctgaGTACTCTGCGTTGATACCACTGCTT
>F1
AAGCAGTGGTATCAACGCAGAGTAC
>R1
gcagagtcatgtatagGTACTCTGCGTTGATACCACTGCTT
```

To use the primers when creating an **Iso-Seq**, **Iso-Seq Classify Only**, or **Iso-Seq with Mapping** analysis: Copy and paste the custom primer text into the **Advanced Analysis Parameters** dialog's **Customer Primer Sequences** field.

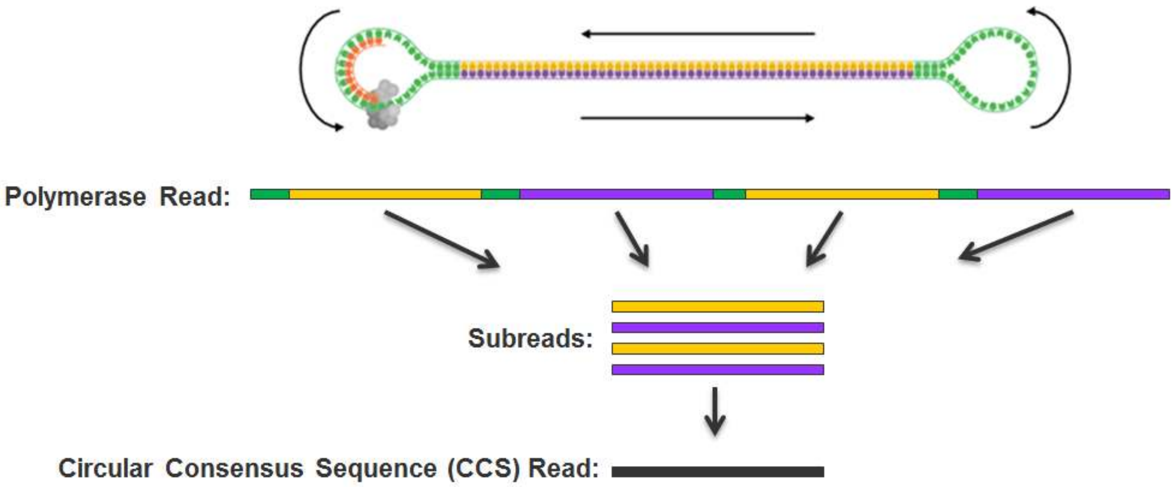
Appendix B - Pacific Biosciences Terminology

General Terminology

- **SMRT® Cell**: Consumable substrates comprising arrays of zero-mode waveguide nanostructures. SMRT Cells are used in conjunction with the DNA Sequencing Kit for on-instrument DNA sequencing.
- **SMRTbell template**: A double-stranded DNA template capped by hairpin adapters (i.e., SMRTbell adapters) at both ends. A SMRTbell template is topologically circular and structurally linear, and is the library format created by the DNA Template Prep Kit.
- **collection**: The set of data collected during real-time observation of the SMRT Cell; including spectral information and temporal information used to determine a read.
- **Zero-mode waveguide (ZMW)**: A nanophotonic device for confining light to a small observation volume. This can be, for example, a small hole in a conductive layer whose diameter is too small to permit the propagation of light in the wavelength range used for detection. Physically part of a SMRT Cell.
- **Run Design**: Specifies
  - The samples, reagents, and SMRT Cells to include in the sequencing run.
  - The run parameters such as movie time and loading to use for the sample.

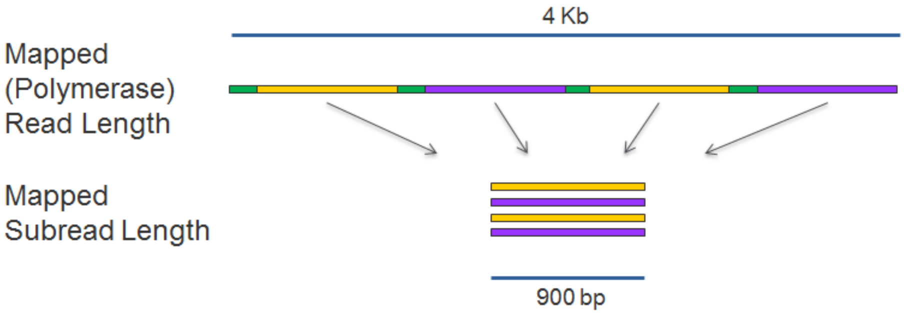
Read Terminology

- **polymerase read**: A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell template. They can include sequences from adapters and from one or multiple passes around a circular template, which includes the insert of interest. Polymerase reads are most useful for quality control of the instrument run. Polymerase read metrics primarily reflect movie length and other run parameters rather than insert size distribution. Polymerase reads are trimmed to include only the high-quality region. **Note**: Sample quality is a major factor in polymerase read metrics.
- **subreads**: Each polymerase read is partitioned to form one or more subreads, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template and no adapter sequences. The subreads contain the full set of quality values and kinetic measurements. Subreads are useful for applications such as *de novo* assembly, resequencing, base modification analysis, and so on.
- **longest subread length**: The mean of the maximum subread length per ZMW.
- **Circular consensus (CCS) read**: The consensus sequence resulting from alignment between subreads taken from a single ZMW. Note that generation of the CCS read does not include or require alignment against a reference sequence. Generation of CCS reads using the CCS algorithm requires at least two full-pass subreads from the insert.



Read Length Terminology

- **mapped polymerase read length**: Approximates the sequence produced by a polymerase in a ZMW. The total number of bases along a read from the first adapter of aligned subread to the last adapter or aligned subread.
- **mapped subread length**: The length of the subread alignment to a target reference sequence. This does **not** include the adapter sequence.



Secondary Analysis Terminology

- **secondary analysis**: Follows primary analysis and uses basecalled data. It is application-specific, and may include:
  - Filtering/selection of data that meets a desired criteria (such as quality, read length, and so on).
  - Comparison of reads to a reference or between each other for mapping and variant calling, consensus sequence determination, alignment and assembly (*de novo* or reference-based), variant identification, and so on.
  - Quality evaluations for a sequencing run, consensus sequence, assembly, and so on.
  - PacBio's SMRT Analysis contains a variety of secondary analysis applications including RNA and Epigenomics analysis tools.

- **secondary analysis application** (Formerly “Secondary analysis protocol”): A secondary analysis workflow that may include multiple analysis steps. Examples include *de novo* assembly, resequencing, RNA and epigenomics analysis.
- **consensus**: Generation of a consensus sequence from multiple-sequence alignment.
- **filtering**: Removes reads that do not meet the Read Length criteria set by the user.
- **mapping**: Local alignment of a read or subread to a reference sequence.

Accuracy Terminology

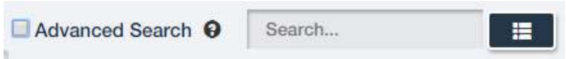
- **circular consensus accuracy**: Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.
- **consensus accuracy**: Accuracy based on aligning multiple sequencing reads or subreads together.
- **polymerase read quality**: A trained prediction of a read’s mapped accuracy based on its pulse and base file characteristics (peak signal-to-noise ratio, inter-pulse distance, and so on).

Appendix C - Data Set/Analysis Search

Use this function to search for Data Sets or analyses **locally** and **remotely**. You can search for Data Sets/analyses two ways:

A **local** search performs a keyword search over all fields in the set of records displayed in the table. This searches the most recent 10,000 Data Sets or 6,000 analyses.

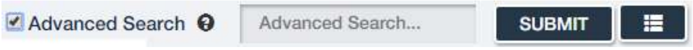
- To perform a **local** search: Enter a search term into the Search box. The Data Set/analysis table is automatically filtered.



An **advanced** search uses the SMRT Link API to perform a structured search over **all** Data Set or analysis records on the SMRT Link server.

To perform an **advanced** search:

1. Click the **Advanced Search** box.
2. Enter a search query (described below) into the Advanced Search box.
3. Click **Submit**.



Advanced search queries consist of one or more search entries separated by the AND operator. (OR is not available.)

Search clauses format: `<field> <operator> <value>`

To search over **multiple** choices for <`value`> use the expression `<field>=in:<comma-separated values>`

- **Note**: `in` searches are only supported for **string** (not date or numeric) fields.

Example Data Set search queries:

- `name=TH_RC0_UnivABC_2pM_Diff-Cell1 (small)`
- `createdAt>2018-04-07`
- `numChildren>=2 AND isActive=true`
- `name=in:mito_mixtures_9plex,Duke_gDNA_SCNAcap_3`

Example analysis search queries:

- `name=Demultiplexing of Sample 1`
- `createdAt>2018-04-07`

- `createdAt>2018-04-07 AND state = SUCCESSFUL`
- `createdBy=in:mcantor,smrtlinktest`

The tables below list allowed Data Set/analysis fields and their corresponding allowed operators and values.

- Date values **must** be specified using the ISO8601 date format. (https://en.wikipedia.org/wiki/ISO_8601)
- Boolean values **must** be specified as either `true` or `false` (case-insensitive.)
- Numeric values **must** be integers.
- Field names are **case-sensitive**.

Data Set Fields

Field	Operator	Value
`uuid`	`=`	String
`name`	`=`	String
`path`	`=`	String
`parentUuid`	`=`	String
`version`	`=`	String
`id`	`=, >, <, <=, >=`	Number
`jobId`	`=, >, <, <=, >=`	Number
`projectId`	`=, >, <, <=, >=`	Number
`numChildren`	`=, >, <, <=, >=`	Number
`numRecords`	`=, >, <, <=, >=`	Number
`totalLength`	`=, >, <, <=, >=`	Number
`createdAt`	`=, >, <, <=, >=`	Date (ISO8601 format)
`importedAt`	`=, >, <, <=, >=`	Date (ISO8601 format)
`updatedAt`	`=, >, <, <=, >=`	Date (ISO8601 format)
`isActive`	`=`	Boolean

Analysis Fields

Field	Operator	Value
`uuid`	`=`	String
`name`	`=`	String
`comment`	`=`	String
`path`	`=`	String

Field	Operator	Value
createdBy	=	String
createdByEmail	=	String
smrtLinkVersion	=	String
errorMessage	=	String
state	=	String
subJobTypeId	=	String
id	=, >, <, <=, >=	Number
projectId	=, >, <, <=, >=	Number
parentMultiJobId	=, >, <, <=, >=	Number
createdAt	=, >, <, <=, >=	Date (ISO8601 format)
importedAt	=, >, <, <=, >=	Date (ISO8601 format)
updatedAt	=, >, <, <=, >=	Date (ISO8601 format)
jobUpdatedAt	=, >, <, <=, >=	Date (ISO8601 format)