SMRT Link User Guide
Sequel® System
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>PacBio® RS II System Users</td>
<td>3</td>
</tr>
<tr>
<td>Contact Information</td>
<td>4</td>
</tr>
<tr>
<td>Using SMRT® Link</td>
<td>4</td>
</tr>
<tr>
<td>Working with Tables</td>
<td>5</td>
</tr>
<tr>
<td>SMRT Link Menu Commands</td>
<td>6</td>
</tr>
<tr>
<td>Sample Setup</td>
<td>7</td>
</tr>
<tr>
<td>Editing or Printing Existing Calculations</td>
<td>9</td>
</tr>
<tr>
<td>Deleting Existing Calculations</td>
<td>9</td>
</tr>
<tr>
<td>Run Design</td>
<td>10</td>
</tr>
<tr>
<td>Creating a New Run Design</td>
<td>10</td>
</tr>
<tr>
<td>Creating a Run Design by Importing a CSV File</td>
<td>13</td>
</tr>
<tr>
<td>Importing a Run Design</td>
<td>13</td>
</tr>
<tr>
<td>Editing or Deleting Run Designs</td>
<td>14</td>
</tr>
<tr>
<td>Run QC</td>
<td>15</td>
</tr>
<tr>
<td>Default Table Fields</td>
<td>16</td>
</tr>
<tr>
<td>Additional Table Fields</td>
<td>16</td>
</tr>
<tr>
<td>Consumables Table</td>
<td>17</td>
</tr>
<tr>
<td>Data Management</td>
<td>20</td>
</tr>
<tr>
<td>What is a Data Set?</td>
<td>20</td>
</tr>
<tr>
<td>What's in a Data Set?</td>
<td>20</td>
</tr>
<tr>
<td>Creating a Data Set</td>
<td>21</td>
</tr>
<tr>
<td>Starting an Analysis from a Data Set</td>
<td>22</td>
</tr>
<tr>
<td>Copying a Data Set</td>
<td>22</td>
</tr>
<tr>
<td>Deleting a Data Set</td>
<td>22</td>
</tr>
<tr>
<td>Exporting Data Sets</td>
<td>23</td>
</tr>
<tr>
<td>Viewing Data Set Information</td>
<td>23</td>
</tr>
<tr>
<td>Editing Data Set Information</td>
<td>23</td>
</tr>
<tr>
<td>Data Set QC Reports</td>
<td>24</td>
</tr>
<tr>
<td>What is a Project?</td>
<td>25</td>
</tr>
<tr>
<td>What's in a Project?</td>
<td>25</td>
</tr>
<tr>
<td>Data Sets and Projects</td>
<td>26</td>
</tr>
<tr>
<td>Creating a Project</td>
<td>26</td>
</tr>
<tr>
<td>Editing a Project</td>
<td>26</td>
</tr>
<tr>
<td>Deleting a Project</td>
<td>27</td>
</tr>
<tr>
<td>Viewing/Deleting Sequence, Reference and Barcode Data</td>
<td>27</td>
</tr>
<tr>
<td>Importing Data</td>
<td>28</td>
</tr>
<tr>
<td>SMRT® Analysis</td>
<td>30</td>
</tr>
<tr>
<td>Creating and Starting an Analysis</td>
<td>30</td>
</tr>
<tr>
<td>Starting an Analysis After Viewing Sequence Data</td>
<td>33</td>
</tr>
<tr>
<td>Canceling a Running Analysis</td>
<td>34</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Viewing Analysis Results</td>
<td>34</td>
</tr>
<tr>
<td>Copying and Running an Existing Analysis</td>
<td>35</td>
</tr>
<tr>
<td>Exporting an Analysis</td>
<td>35</td>
</tr>
<tr>
<td>Importing an Analysis</td>
<td>36</td>
</tr>
<tr>
<td><strong>PacBio® Secondary Analysis Applications</strong></td>
<td>37</td>
</tr>
<tr>
<td>Assembly (HGAP 4) Application</td>
<td>39</td>
</tr>
<tr>
<td>Base Modification Detection Application</td>
<td>45</td>
</tr>
<tr>
<td>Base Modification and Motif Analysis Application</td>
<td>48</td>
</tr>
<tr>
<td>CCS Mapping Application</td>
<td>52</td>
</tr>
<tr>
<td>Circular Consensus Sequences (CCS) Application</td>
<td>56</td>
</tr>
<tr>
<td>Convert BAM to FASTX Application</td>
<td>58</td>
</tr>
<tr>
<td>Convert RS to BAM Application</td>
<td>58</td>
</tr>
<tr>
<td>Demultiplex Barcodes Application</td>
<td>59</td>
</tr>
<tr>
<td>Iso-Seq®/Iso-Seq 2 Algorithm Application</td>
<td>63</td>
</tr>
<tr>
<td>Iso-Seq® Algorithm Classify Only Application</td>
<td>68</td>
</tr>
<tr>
<td>Iso-Seq®/Iso-Seq 2 Algorithm with Mapping Application</td>
<td>71</td>
</tr>
<tr>
<td>Long Amplicon Analysis (LAA 2) Application</td>
<td>76</td>
</tr>
<tr>
<td>Minor Variants Analysis Application [Beta]</td>
<td>79</td>
</tr>
<tr>
<td>Resequencing Application</td>
<td>85</td>
</tr>
<tr>
<td>Site Acceptance Test (SAT) Application</td>
<td>90</td>
</tr>
<tr>
<td>Structural Variant Calling Application</td>
<td>95</td>
</tr>
<tr>
<td><strong>Working with Barcoded Data</strong></td>
<td>98</td>
</tr>
<tr>
<td>Step 1: Specify the Barcode Setup and Sample Names in a Run Design</td>
<td>98</td>
</tr>
<tr>
<td>Step 2: Perform the Sequencing Run</td>
<td>99</td>
</tr>
<tr>
<td>Step 3: (Optional) Run the Demultiplex Barcodes Application</td>
<td>100</td>
</tr>
<tr>
<td>Step 4: Run Applications Using the Demultiplexed Data as Input</td>
<td>101</td>
</tr>
<tr>
<td>Demultiplex Barcodes Application Details</td>
<td>102</td>
</tr>
<tr>
<td><strong>Visualizing Data Using SMRT® View</strong></td>
<td>105</td>
</tr>
<tr>
<td>Installing 64-bit Java 8 on a Windows Operating System</td>
<td>106</td>
</tr>
<tr>
<td>Installing 64-bit Java 8 on macOS</td>
<td>106</td>
</tr>
<tr>
<td><strong>Using the PacBio® Self-Signed SSL Certificate</strong></td>
<td>107</td>
</tr>
<tr>
<td><strong>Sequel® System Output Files</strong></td>
<td>108</td>
</tr>
<tr>
<td><strong>Configuration and User Management</strong></td>
<td>111</td>
</tr>
<tr>
<td>LDAP</td>
<td>111</td>
</tr>
<tr>
<td>SSL</td>
<td>111</td>
</tr>
<tr>
<td>Adding and Deleting SMRT Link Users</td>
<td>112</td>
</tr>
<tr>
<td>Assigning User Roles</td>
<td>112</td>
</tr>
<tr>
<td><strong>Hardware/Software Requirements</strong></td>
<td>114</td>
</tr>
<tr>
<td><strong>Appendix A - Barcoded Primers</strong></td>
<td>115</td>
</tr>
<tr>
<td><strong>Appendix B - Pacific Biosciences Terminology</strong></td>
<td>116</td>
</tr>
</tbody>
</table>
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 15 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 20 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 30 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 105 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 108 for details.)
- Configuration and user management. (See “Configuration and User Management” on page 111 for details.)
- SMRT Link client hardware/software requirements. (See “Hardware/Software Requirements” on page 114 for details.)

Installation of SMRT Link Server software is discussed in the document SMRT Link Software Installation (v5.1.0).

New features, fixed issues and known issues are listed in the document SMRT Link Release Notes (v5.1.0).

**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 107.

After accessing SMRT Link, the Home page displays. (Shown is the interface for the Sequel System.)
• 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 X 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.
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 112 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 the **Binding Calculator**.

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

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 size of the SMRTbell library.
8. Select the Sequencing Primer to use for this run. (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. (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. (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 unusually low yields.

13. Specify whether to Clean up the sample (through a MicroSpin column, or loading clean-up beads, depending on insert size) to remove excess primer/polymerase. This results in higher quality data. Please note that the Loading Clean-up Beads are not equivalent to the MagBeads used in MagBead loading. If 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 protocol is listed instead of either the TC6 or Spin 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 targeted at short inserts (such as amplicons <5 kb) and will provide 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.
18. If either 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 Existing Calculations**
1. On the **Sample Setup** screen, select one or more existing 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 Existing Calculations**
1. On the **Sample Setup** screen, select one or more existing 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 Sequel Instrument Control Software (ICS), 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 SMRT Link.

- If you are using SMRT Link with a PacBio RS II, use RS Remote.

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 11-14, 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**.

5. 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.)
6. **(Optional)** Enter **Run Comments**, **Experiment Name**, and **Experiment ID** as needed. (**Note:** Experiment ID must be alphanumeric.)
7. Next, specify information for the first sample that will be part of the run. Enter a **Sample Name**.
8. **(Optional)** Enter **Sample Comments**.
9. Specify the well position used for this sample: Click **Select...** and choose a plate position.
10. 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.
11. Select a **Template Prep Kit** from the list, or type in a kit part number.
12. Select a **Binding Kit** from the list, or type in a kit part number.
13. 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.

14. **(Optional)** Select a **DNA Control Complex** from the list. PacBio **highly recommends** using a DNA Control Complex.

15. Specify an **Insert Size**, ranging from 500 to 50,000 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 table of loading recommendations.)

16. 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 **SMRT Cell LR** part.

17. **(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.

18. **(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 targeted at short inserts (e.g., amplicons <5 kb) and will provide 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.

19. **(Optional)** If you are using **barcoded samples**, see “Step 1: Specify the Barcode Setup and Sample Names in a Run Design” on page 98 for instructions. For details on secondary analysis of barcoded samples, see “Demultiplex Barcodes Application” on page 59.

20. 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.

21. 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.

22. 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. (Contact your FAS for a copy of the file if it is not available.)

1. Update the 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.

<table>
<thead>
<tr>
<th>Run Design Attribute</th>
<th>Required</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run Name</td>
<td>Yes</td>
<td>Run name must be entered for the first cell and will be applied to the remaining cells in the run. Use only alphanumeric characters.</td>
</tr>
<tr>
<td>Run Description</td>
<td>No</td>
<td>Run description must be entered for the first cell and will be applied to the remaining cells in the run. No character limitations; use only alphanumeric characters.</td>
</tr>
<tr>
<td>Well No.</td>
<td>Yes</td>
<td>Format as follows: A01, B12. Not acceptable: A1,B3 and so on. Use only alphanumeric characters.</td>
</tr>
<tr>
<td>Sample Name</td>
<td>Yes</td>
<td>Enter sample name using only alphanumeric characters.</td>
</tr>
<tr>
<td>Sample Description</td>
<td>No</td>
<td>Enter sample description using only alphanumeric characters.</td>
</tr>
<tr>
<td>Insert Size</td>
<td>Yes</td>
<td>Enter values rounded to the nearest kb. Example: 1000, 2000, 10000, and so on.</td>
</tr>
<tr>
<td>Size Selection</td>
<td>Yes</td>
<td>Enter “Yes” or “No”.</td>
</tr>
<tr>
<td>Automation Name</td>
<td>Yes</td>
<td>Enter &quot;Magbead&quot; or &quot;Diffusion&quot;.</td>
</tr>
<tr>
<td>Collection Time</td>
<td>Yes</td>
<td>Enter the desired length of time for the cell run, in minutes, between 30 and 1200.</td>
</tr>
<tr>
<td>DNA Template Prep Kit Box Barcode</td>
<td>Yes</td>
<td>Enter or scan the template prep barcode. Working example: DM1117100259100111716</td>
</tr>
<tr>
<td>DNA Control Complex Box Barcode</td>
<td>No</td>
<td>Enter or scan the control DNA barcode. Working example: LXXXXXX101084300060817</td>
</tr>
<tr>
<td>Binding Kit Box Barcode</td>
<td>Yes</td>
<td>Enter or scan the binding kit barcode. Working example: DM1117100619300111716</td>
</tr>
<tr>
<td>Sequencing Kit Box Barcode</td>
<td>Yes</td>
<td>Enter or scan the sequencing kit barcode. Working example: DM1117100620000111716</td>
</tr>
<tr>
<td>Automation Parameters</td>
<td>No</td>
<td>To enable Pre-Extension time, enter the number of minutes. Example 90 minutes: &quot;ExtensionTime=double:90</td>
</tr>
</tbody>
</table>

Importing a Run Design

After creating the CSV file, import the file into Sequel ICS using SMRT Link.

1. Access SMRT Link using the Chrome web browser.
2. Select Run Design.
3. Click Import Run Design.
4. 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.
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.

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

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**.

5. To **print** Run QC information about the selected run(s), click **Print**.

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

- **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**

7. 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.
• **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.

• **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 the trimmed read length of all polymerase reads are longer than this value. The value includes bases from adapters as well as multiple passes around a circular template.
  - **Longest Subread Mean**: The mean subread length, considering only the longest subread from each ZMW.
  - **Longest Subread N50**: 50% of the subreads are longer than this value when considering only the longest subread from each ZMW.

• **Productivity**
  - **P0**: Empty ZMW; no high quality read 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 read length of the polymerase control.
  - **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.

8. Click the > arrow to expand rows to view histograms for each SMRT Cell where data was successfully transferred. These include:
  - **Polymerase Read Length**: Maps the number of reads against the read length.
  - **Longest Subread Length**: Maps the number of reads against the insert length.
- **Control Polymerase RL**: Displays the Polymerase read length 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.
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 22 for details.)

**What’s in a Data Set?**

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, enter a unique search string into the **Search** field.

4. **Click + Create Data Set.**
5. 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.
6. In the **Data Sets** box, select one or more sets of sequence data.
7. Click **Save Data Set**. The new Data Set becomes available for starting analyses, viewing, or generating reports.
8. 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.

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 30.

**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)** Search for a specific Data Set: Enter any search string into the Search field.
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)** Search for a specific Data Set: Enter any search string into the Search field.
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)** Search for specific sequence data: Enter any search string into the Search field.
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)** Search for a specific Data Set: Enter any search string into the Search field.
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)** Search for a specific Data Set: Enter any search string into the Search field.
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:

![Image of Data Management interface]

### 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.
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.

What’s in a Project?

- 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

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.
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.
   • **Barcodes**: Barcodes from barcoded samples.
4. Click on the link for the data selected to see details about the sequence data, reference sequence, or barcode file.
5. In the Name column, click the name of the Sequence, Reference or Barcode of interest. Details for that sequence, reference sequence file or barcode file display.

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.

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

- **Reference sequence files** - FASTA sequence files (.fa or .fasta) containing a reference sequence for use in starting analyses.
- **Sequel sequence data** - A .subreadset.xml file containing information about Sequel sequence data, such as paths to the BAM files.
- **RS II sequence data** - A .metadata.xml file containing information about PacBio RS II sequence data from a single cell.
- **Barcodes/Barcodes (FASTA)** - A .barcodeset.xml or FASTA-format (.fa or .fasta) files containing barcodes.
1. On the Home Page, select **Data Management**.
2. Click **View or Import Sequence Data**.
3. Click **Import**, then select either **References**, **Sequel Sequence Data**, **RS II Sequence Data**, **Barcodes**, or **Barcodes (FASTA)**.

4. Navigate to the appropriate file and click **Import**. The sequence data, reference, or barcodes are imported and becomes available in SMRT Link.
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 specify the status of the 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.

- 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 using the Arrow algorithm.

6. Click the **References** 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 28 for details.)

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 37.

8. Enter a name for the analysis.

9. Specify whether to use BAM-format data generated on the Sequel System, or PacBio RS II data.

   Note: If PacBio RS II data was selected for the analysis, the only application available is Convert RS to BAM. After converting the data to the BAM file format, analyzing the data using all the other analysis applications is enabled.

10. In the Data Sets box, select one or more sets of data to be analyzed together.

11. (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:

   - 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.

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.

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.

---

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

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 30.

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) Search for a specific analysis: Enter any search string into the Search field.
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 37.

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 > Analysis 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 105 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) Search for a specific analysis: Enter any search string into the Search field.
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 “Creating and Starting an Analysis” on page 30 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 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 v5.1.0. 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 39 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 45 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 48 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 52 for details.

Circular Consensus Sequences (CCS)
- Identify consensus sequences for single molecules.
- See “Circular Consensus Sequences (CCS) Application” on page 56 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 58 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 v5.1.0.
- See “Convert RS to BAM Application” on page 58 for details.

Demultiplex Barcodes
- Separate reads by barcode.
- See “Demultiplex Barcodes Application” on page 59 for details.
**Iso-Seq®/Iso-Seq 2® Algorithm**
- Characterize transcripts and splice variants.
- See “Iso-Seq® Iso-Seq 2 Algorithm Application” on page 63 for details.

**Iso-Seq® Algorithm Classify Only**
- This analysis includes only the **Classify** step of the Iso-Seq algorithm. Sequencing reads are classified into full length or non-full length reads.
- See “Iso-Seq® Algorithm Classify Only Application” on page 68 for details.

**Iso-Seq®/Iso-Seq 2® Algorithm with Mapping**
- Characterize transcripts and splice variants, then map the transcripts back to the reference genome. A GMAP reference genome is required for alignment.
- See “Iso-Seq® Iso-Seq 2 Algorithm with Mapping Application” on page 71 for details.

**Long Amplicon Analysis (LAA 2)**
- Identify phased consensus sequences from a heterogeneous pool of amplicons.
- See “Long Amplicon Analysis (LAA 2) Application” on page 76 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 79 for details.

**Resequencing**
- Map sequencing reads against a reference sequence and identify variants.
- See “Resequencing Application” on page 85 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 90 for details.

**Structural Variant Calling**
- Identify large structural variant insertions and deletions (default: ≥50 bp) in a sample or set of samples relative to a reference.
- See “Structural Variant Calling Application” on page 95 for details.
Assembly (HGAP 4) Application

Use this application (Hierarchical Genome Assembly Process) 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.

Parameters

<table>
<thead>
<tr>
<th>Required Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Length</td>
<td>5,000,000</td>
<td>The approximate number of base pairs expected in the genome. Other parameters are set automatically based on this value.</td>
</tr>
</tbody>
</table>

Advanced Analysis Parameters

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Subread Length</td>
<td>0</td>
<td>The minimum length of subreads to use in the assembly.</td>
</tr>
<tr>
<td>Filters to Add to the Data Set</td>
<td>rq ≥ 0.7</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

Advanced Analysis Parameters - Assembly

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Assembly</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive Option</td>
<td>OFF</td>
<td>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.</td>
</tr>
<tr>
<td>FALCON cfg Overrides</td>
<td>NONE</td>
<td>Allows 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.</td>
</tr>
<tr>
<td>Seed Length Cutoff</td>
<td>-1</td>
<td>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).</td>
</tr>
<tr>
<td>Seed Coverage</td>
<td>30</td>
<td>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.</td>
</tr>
<tr>
<td>Save Output for Unzip</td>
<td>OFF</td>
<td>If ON, saves a set of large HGAP 4 output files for use as input for later downstream command-line analysis using Falcon Unzip.</td>
</tr>
</tbody>
</table>

Advanced Analysis Parameters - Alignment

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

<table>
<thead>
<tr>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>randombest</td>
<td>Specify how to treat multiple hits:</td>
</tr>
<tr>
<td></td>
<td>• random: Selects a random hit.</td>
</tr>
<tr>
<td></td>
<td>• all: Selects all hits.</td>
</tr>
<tr>
<td></td>
<td>• allbest: Selects all the best score hits.</td>
</tr>
<tr>
<td></td>
<td>• randombest: Selects a random hit from all best score hits.</td>
</tr>
<tr>
<td></td>
<td>• leftmost: Selects a hit which has the best score and the smallest mapping coordinate in any reference.</td>
</tr>
</tbody>
</table>

### Algorithm Options

List of space-separated arguments passed to BLASR. Default: `-minMatch 12 -bestn 10 -minPctIdentity 70.0`

### Advanced Analysis Parameters - Alignment

<table>
<thead>
<tr>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Score</td>
<td>0 Specify the score to use in the display.</td>
</tr>
<tr>
<td>Minimum Confidence</td>
<td>40 The minimum confidence for a variant call to be output to the file variants.gff.</td>
</tr>
<tr>
<td>Track Description</td>
<td>NONE Description to display in the header.</td>
</tr>
<tr>
<td>Track Name</td>
<td>variants Name to display in the header.</td>
</tr>
<tr>
<td>Purpose</td>
<td>variants Specify the run mode - Variants or Coverage.</td>
</tr>
<tr>
<td>Masking</td>
<td>ON During the polish step, omit regions of reads that have low concordance with the template.</td>
</tr>
<tr>
<td>Algorithm</td>
<td>best • Quiver is a variant-calling algorithm that operates on RS II data only.</td>
</tr>
<tr>
<td></td>
<td>• 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.</td>
</tr>
<tr>
<td></td>
<td>• 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.</td>
</tr>
<tr>
<td></td>
<td>• Best is the best algorithm based on the data provided.</td>
</tr>
<tr>
<td>Minimum Coverage</td>
<td>5 The minimum site coverage that must be achieved for variant calls and consensus to be calculated for a site.</td>
</tr>
</tbody>
</table>

### Advanced Analysis Parameters - Consensus

### Advanced Analysis Parameters - Reports

<table>
<thead>
<tr>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>1000 Specify the number of genome regions in the summary statistics. (This is used for guidance, and is not strict.)</td>
</tr>
<tr>
<td>Region Size</td>
<td>0 If specified, use a fixed region size.</td>
</tr>
<tr>
<td>Maximum Region Size</td>
<td>1,000,000 The upper limit for region size. This is ignored if Region Size is set explicitly.</td>
</tr>
<tr>
<td>Force the Number of Regions</td>
<td>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.</td>
</tr>
</tbody>
</table>
The Assembly (HGAP 4) application generates the following reports:

**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 95\textsuperscript{th} percentile of the length of the filtered subreads used as initial input.
- **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 95\textsuperscript{th} percentile of the length of the seed reads used in the pre-assembly.
- **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 95\textsuperscript{th} percentile of the length of the reads output by the pre-assembler.
- **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.

### 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**

• Maps the number of reads against the read length.

**Realignment to Draft Assembly > Realigned Subread Concordance**

• Maps the number of subreads against the percent concordance with the subreads that realigned to the draft assembly.

**Realignment to Draft Assembly > Realigned Subread Length**

• Maps the number of subreads against 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.

**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 lambda_NEB3011**

• Maps coverage of the lambda_NEB3011 reference against the reference start position.

**Coverage > Depth of coverage Distribution**

• Maps the reference regions over the coverage.

**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.
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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
- **Alignments**: Data Set containing the alignment results.
- **Consensus GFF**: A list of variants from the reference.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference.
- **INI file**: Smaller job configuration file.
- **Run Configuration**: Job configuration file.
- **Polished Assembly**: The final polished assembly.
- **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**

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Alignment</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of .bam files</td>
<td>1</td>
<td>Number of .bam files to create in consolidate mode.</td>
</tr>
<tr>
<td>Minimum Concordance</td>
<td>70</td>
<td>The minimum required alignment concordance, in percent.</td>
</tr>
<tr>
<td>Consolidate .bam</td>
<td>OFF</td>
<td>Specify whether to merge chunked/gathered .bam files.</td>
</tr>
<tr>
<td>Concordant Alignment</td>
<td>ON</td>
<td>Specify whether to map subreads of a ZMW to the same genomic location.</td>
</tr>
<tr>
<td>Align Unsplit Polymerase Reads</td>
<td>OFF</td>
<td>Do not split reads into subreads even if subread regions are available.</td>
</tr>
<tr>
<td>Minimum Length</td>
<td>50</td>
<td>The minimum required alignment length, in base pairs.</td>
</tr>
<tr>
<td>Hit Policy</td>
<td>randombest</td>
<td>Specify how to treat multiple hits:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• random: Selects a random hit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• all: Selects all hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• allbest: Selects all the best score hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• randombest: Selects a random hit from all best score hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• leftmost: Selects a hit which has the best score and the smallest mapping coordinate in any reference.</td>
</tr>
</tbody>
</table>

**Algorithm Options**

List of space-separated arguments passed to BLASR. Default: -minMatch 12 -bestn 10 -minPctIdentity 70.0

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

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Reports</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>1000</td>
<td>Specify the number of genome regions in the summary statistics. (This is used for guidance, and is not strict.)</td>
</tr>
<tr>
<td>Region Size</td>
<td>0</td>
<td>If specified, use a fixed region size.</td>
</tr>
<tr>
<td>Maximum Region Size</td>
<td>1,000,000</td>
<td>The upper limit for region size. This is ignored if Region Size is set explicitly.</td>
</tr>
</tbody>
</table>
Reports and Data Files
The Base Modification Detection application generates the following reports:

**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**
• Maps the number of reads against the read length.

**Mapping Report > Mapped Subread Concordance**
• Maps the number of subreads against the percent concordance with the reference sequence.

**Mapping Report > Mapped Subread Length**
• Maps the number of subreads against 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**
• Maps the number of bases against 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 Details Log**: Log information for the analysis workflow.
• **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
• **Alignments**: 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.
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

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

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

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

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**
• Maps the number of reads against the read length.

**Mapping Report > Mapped Subread Concordance**
• Maps the number of subreads against the percent concordance with the reference sequence.

**Mapping Report > Mapped Subread Length**
• Maps the number of subreads against 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**
• Maps the number of bases against modification QV.

**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 $(\text{fraction methylated}) \times (\text{sum of log-p values of matches})$.

### Modified Base Motifs > Modification QVs

• Maps motif sites against 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 Details Log**: Log information for the analysis workflow.

• **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)

• **Alignments**: 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.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Alignment</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consolidate .bam</td>
<td>OFF</td>
<td>Specify whether to merge chunked/gathered .bam files.</td>
</tr>
<tr>
<td>Number of .bam files</td>
<td>1</td>
<td>Number of .bam files to create in consolidate mode.</td>
</tr>
<tr>
<td>Minimum Concordance</td>
<td>70</td>
<td>The minimum required alignment concordance, in percent.</td>
</tr>
<tr>
<td>Minimum Length</td>
<td>50</td>
<td>The minimum required alignment length, in base pairs.</td>
</tr>
<tr>
<td>Hit Policy</td>
<td>randombest</td>
<td>Specify how to treat multiple hits:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• random: Selects a random hit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• all: Selects all hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• allbest: Selects all the best score hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• randombest: Selects a random hit from all best score hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• leftmost: Selects a hit which has the best score and the smallest mapping coordinate in any reference.</td>
</tr>
<tr>
<td>Algorithm Options</td>
<td></td>
<td>List of space-separated arguments passed to BLASR. Default: -minMatch 12 -bestn 10 -minPctIdentity 70.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - CCS</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Strand CCS</td>
<td>OFF</td>
<td>For each ZMW, generate two CCS sequences - one for each strand.</td>
</tr>
<tr>
<td>Maximum Dropped Fraction</td>
<td>0.8</td>
<td>The maximum fraction of subreads that can be dropped before giving up.</td>
</tr>
</tbody>
</table>
The CCS Mapping 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.

### Advanced Analysis Parameters - CCS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Subread Length</td>
<td>15,000</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Subread Length</td>
<td>50</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Number of Passes</td>
<td>3</td>
<td>The minimum number of <strong>full</strong> passes for a ZMW to be emitted. Full passes <strong>must</strong> have an adapter hit before and after the insert sequence and so does not include any partial passes at the start and end of the sequencing reaction.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.9</td>
<td>The <strong>minimum</strong> predicted accuracy of a read, ranging from 0 to 1. (0.99 indicates that only reads expected to be 99% accurate are emitted.)</td>
</tr>
<tr>
<td>Minimal Read Score</td>
<td>0.65</td>
<td>The minimum read score of input subreads.</td>
</tr>
<tr>
<td>Minimum SNR</td>
<td>3.75</td>
<td>The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR &lt;3.75 is typically considered lower quality.</td>
</tr>
<tr>
<td>Minimum Z Score</td>
<td>-9999</td>
<td>The minimum Z-Score for a subread to be included in the consensus generating process.</td>
</tr>
<tr>
<td>Polish CCS</td>
<td>OFF</td>
<td>Specify whether to polish CCS sequences using Arrow.</td>
</tr>
<tr>
<td>Report File Output</td>
<td></td>
<td>Specify the name of the report file to output. (Default = ccs_report_txt)</td>
</tr>
<tr>
<td>Emit Individual QVs</td>
<td>OFF</td>
<td>Specify whether to emit dq, iq, and sq &quot;rich&quot; quality tracks.</td>
</tr>
<tr>
<td>Filters to add to the Dataset</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

### Advanced Analysis Parameters - Reports

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>1000</td>
<td>Specify the number of genome regions in the summary statistics. (This is used for guidance, and is <strong>not</strong> strict.)</td>
</tr>
<tr>
<td>Region Size</td>
<td>0</td>
<td>If specified, use a fixed region size.</td>
</tr>
<tr>
<td>Maximum Region Size</td>
<td>1,000,000</td>
<td>The upper limit for region size. This is ignored if Region Size is set explicitly.</td>
</tr>
<tr>
<td>Force the Number of Regions</td>
<td>OFF</td>
<td>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 <strong>not</strong> compatible with fixed region sizes.</td>
</tr>
<tr>
<td>Maximum Number of Contigs to Plot</td>
<td>25</td>
<td>The maximum number of contigs to plot in the coverage report.</td>
</tr>
</tbody>
</table>
• 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
• Maps CCS reads against 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).

Mapping Statistics 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 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.
• CCS Read Length 95% (mapped): The 95th 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**

• Maps CCS reads against read length.

**Mapping Report > Mapped CCS Read Concordance**

• Maps CCS reads against 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 lambda_NEB3011**

• Maps coverage of the lambda_NEB3011 reference against the reference start position.

**Coverage > Depth of Coverage Distribution**

• Maps the reference regions against 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 Details Log:** Log information for the analysis workflow.

• **SMRT Link Job Log:** Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)

• **Alignments:** Data Set containing alignment results.

• **Consensus Sequences:** Consensus sequences generated from CCS.

• **CCS Performance Results:** Summary of CCS performance and yield.

• **Coverage Summary:** Coverage summary for regions (bins) spanning the reference.

• **Consensus Sequences:** Consensus sequences generated from CCS, in FASTA or FASTQ format.
Use this application to identify consensus sequences for single molecules.

**Parameters**

<table>
<thead>
<tr>
<th>Required Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Number of Passes</td>
<td>3</td>
<td>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 does not include any partial passes at the start and end of the sequencing reaction.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.9</td>
<td>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.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - CCS</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Strand CCS</td>
<td>OFF</td>
<td>For each ZMW, generate two CCS sequences - one for each strand.</td>
</tr>
<tr>
<td>Maximum Dropped Fraction</td>
<td>0.8</td>
<td>The maximum fraction of subreads that can be dropped before giving up.</td>
</tr>
<tr>
<td>Maximum Subread Length</td>
<td>15,000</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Subread Length</td>
<td>50</td>
<td>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.</td>
</tr>
<tr>
<td>Minimal Read Score</td>
<td>0.65</td>
<td>The minimum read score of input subreads.</td>
</tr>
<tr>
<td>Minimum SNR</td>
<td>3.75</td>
<td>The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR &lt;3.75 is typically considered lower quality.</td>
</tr>
<tr>
<td>Minimum Z Score</td>
<td>-9999</td>
<td>The minimum Z-Score for a subread to be included in the consensus generating process.</td>
</tr>
<tr>
<td>Polish CCS</td>
<td>OFF</td>
<td>Specify whether to polish CCS sequences using Arrow.</td>
</tr>
<tr>
<td>Report File Output</td>
<td></td>
<td>Specify the name of the report file to output. (Default = ccs_report_txt)</td>
</tr>
<tr>
<td>Emit Individual QVs</td>
<td>OFF</td>
<td>Specify whether to emit dq, iq, and sq “rich” quality tracks.</td>
</tr>
<tr>
<td>Filters to add to the Dataset</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

**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**
• Maps CCS reads against 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 Details Log:** Log information for the analysis workflow.
• **SMRT Link Job Log:** Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
• **Consensus Sequences:** Consensus sequences generated from CCS.
• **CCS Performance Results:** Summary of CCS performance and yield.
• **Consensus Sequences:** 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**

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

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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis 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 v5.1.0.

- 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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
Use this application to separate sequence reads by barcode. (See “Working with Barcoded Data” on page 98 for more details.)

- 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 Sequences**

- 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 98 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.
- **Unbarcoded Reads**: The number of reads without barcodes 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.

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.

  ![Barcode Quality Score Distribution](image)

  – **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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
- **Barcode Files**: Barcoded subread Data Sets; one XML 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).
Use this application to characterize transcripts and splice variants. (Iso-Seq analysis is performed *de novo*, without a reference genome.)

The Iso-Seq application enables analysis and functional characterization of transcripts and alternative splice variants for sequencing data generated on PacBio instruments.

This application generates full-length cDNA sequences, eliminating the need for transcriptome reconstruction using isoform-inference algorithms.

The Iso-Seq 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.

**Iso-Seq/Iso-Seq 2 application**:

- The Iso-Seq 2 application is currently a **beta** version, and uses a slightly different algorithm than the Iso-Seq application. Compared to the Iso-Seq application, the Iso-Seq 2 application:
  - Is faster and has better specificity.
  - Generates fewer false positive isoforms predicted based on evaluation of synthetic RNA.
  - Filters out more chimeras.
- Use the Iso-Seq application if you have barcoded cDNA samples in one SMRT Cell and want to make consensus isoforms of each barcoded sample **separately** instead of jointly. The Iso-Seq 2 application currently supports only calling isoforms from barcoded cDNA samples jointly, and not separately.
- Certain Iso-Seq application parameters are now obsolete and removed in the Iso-Seq 2 application. (For more details, see the **Parameters** section.)

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - CCS</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Strand CCS</td>
<td>OFF</td>
<td>For each ZMW, generate two CCS sequences - one for each strand.</td>
</tr>
</tbody>
</table>
### Advanced Analysis Parameters - CCS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Dropped Fraction</td>
<td>0.8</td>
<td>The maximum fraction of subreads that can be dropped before giving up.</td>
</tr>
<tr>
<td>Maximum Subread Length</td>
<td>15,000</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Subread Length</td>
<td>50</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Number of Passes</td>
<td>0</td>
<td>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 does not include any partial passes at the start and end of the sequencing reaction. For the Iso-Seq/Iso-Seq 2 application, a 0-pass read can still be full-length.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.8</td>
<td>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.)</td>
</tr>
<tr>
<td>Minimal Read Score</td>
<td>0.65</td>
<td>The minimum read score of input subreads.</td>
</tr>
<tr>
<td>Minimum SNR</td>
<td>3.75</td>
<td>The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR &lt;3.75 is typically considered lower quality.</td>
</tr>
<tr>
<td>Minimum Z Score</td>
<td>-9999</td>
<td>The minimum Z-Score for a subread to be included in the consensus generating process.</td>
</tr>
<tr>
<td>Polish CCS</td>
<td>OFF</td>
<td>Specify whether to polish CCS sequences using Arrow.</td>
</tr>
<tr>
<td>Report File Output</td>
<td></td>
<td>Specify the name of the report file to output. (Default = ccs_report_txt)</td>
</tr>
<tr>
<td>Emit Individual QVs</td>
<td>OFF</td>
<td>Specify whether to emit dq, iq, and sq “rich” quality tracks.</td>
</tr>
<tr>
<td>Filters to add to the Dataset</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

### Advanced Analysis Parameters - Iso-Seq

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin by Primer (Iso-Seq application only)</td>
<td>OFF</td>
<td>Specify binning reads by Primers. This overwrites the Bin by Read Length Manually and Bin by Read Length in KB options. This option is removed in the Iso-Seq 2 application, where binning is done automatically.</td>
</tr>
<tr>
<td>Bin by Read Length Manually (Iso-Seq application only)</td>
<td>NONE</td>
<td>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. This option is removed in the Iso-Seq 2 application, where binning is done automatically.</td>
</tr>
<tr>
<td>Bin by Read Length in KB (Iso-Seq application only)</td>
<td>1</td>
<td>Specify binning reads by the read length, in kb. This option is removed in the Iso-Seq 2 application, where binning is done automatically.</td>
</tr>
<tr>
<td>Minimum Quiver/Arrow Accuracy</td>
<td>0.99</td>
<td>Specify the minimum predicted consensus accuracy to classify an isoform as high-quality (HQ). All isoforms below this cutoff are considered low-quality (LQ).</td>
</tr>
<tr>
<td>Minimum Sequence Length</td>
<td>50</td>
<td>The minimum sequence length to output.</td>
</tr>
</tbody>
</table>
The Iso-Seq/Iso-Seq 2 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.) For the Iso-Seq/Iso-Seq 2 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
- Maps CCS reads against 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).

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.

**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.

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

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

The Iso-Seq Classify Only application runs much faster than the full Iso-Seq application, and can be used to QC the data in advance of running the full Iso-Seq 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 transcriptome reconstruction using isoform-inference algorithms.

Parameters

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - CCS</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Strand CCS</td>
<td>OFF</td>
<td>For each ZMW, generate two CCS sequences - one for each strand.</td>
</tr>
<tr>
<td>Maximum Dropped Fraction</td>
<td>0.8</td>
<td>The maximum fraction of subreads that can be dropped before giving up.</td>
</tr>
<tr>
<td>Maximum Subread Length</td>
<td>15,000</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Subread Length</td>
<td>50</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Number of Passes</td>
<td>0</td>
<td>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 does not include any partial passes at the start and end of the sequencing reaction. For the Iso-Seq/Iso-Seq 2 application, a 0-pass read can still be full-length.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.8</td>
<td>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.)</td>
</tr>
<tr>
<td>Minimal Read Score</td>
<td>0.65</td>
<td>The minimum read score of input subreads.</td>
</tr>
<tr>
<td>Minimum SNR</td>
<td>3.75</td>
<td>The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR &lt;3.75 is typically considered lower quality.</td>
</tr>
<tr>
<td>Minimum Z Score</td>
<td>-9999</td>
<td>The minimum Z-Score for a subread to be included in the consensus generating process.</td>
</tr>
<tr>
<td>Polish CCS</td>
<td>OFF</td>
<td>Specify whether to polish CCS sequences using Arrow.</td>
</tr>
<tr>
<td>Report File Output</td>
<td></td>
<td>Specify the name of the report file to output. (Default = ccs_report_txt)</td>
</tr>
<tr>
<td>Emit Individual QVs</td>
<td>OFF</td>
<td>Specify whether to emit dq, iq, and sq &quot;rich&quot; quality tracks.</td>
</tr>
<tr>
<td>Filters to add to the Dataset</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>
Reports and Data Files

The Iso-Seq Algorithm Classify Only 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.) For the Iso-Seq/Iso-Seq 2 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**
- Maps CCS reads against 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).

**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.

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Iso-Seq</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Primer Sequences</td>
<td>NONE</td>
<td>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 &quot;Appendix A - Barcoded Primers&quot; on page 115 for details.</td>
</tr>
<tr>
<td>Require PolyA</td>
<td>ON</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Sequence Length</td>
<td>50</td>
<td>The minimum sequence length to output.</td>
</tr>
</tbody>
</table>
- **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-concatemter 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-concatemter 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-concatemter 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.

**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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
- **Consensus Sequences (XML)**: Consensus sequences generated from CCS.
- **CSV Report (TXT)**: Summary of CCS performance and yield.
- **Draft Isoforms (XML)**: Intermediate Data Set used to get full length reads.
- **Full-Length Non-Chimeric Reads (XML)**: Full-length non-chimeric subreads generated from pbtranscript classify.
- **Consensus Sequences**: Consensus sequences generated from CCS, in FASTA or FATSQ format.
- **Non-Full-Length Reads (XML)**: Non-full-Length reads generated from pbtranscript classify.
- **Primer Info**: Per-CCS read annotation and classification results.
Use this application to characterize transcripts and splice variants, then map the transcripts back to the reference genome. A GMAP reference genome is required for alignment.

The Iso-Seq application enables analysis and functional characterization of transcripts and alternative splice variants for sequencing data generated on PacBio instruments.

This application generates full-length cDNA sequences, eliminating the need for transcriptome reconstruction using isoform-inference algorithms.

The Iso-Seq 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.

**Iso-Seq/Iso-Seq 2 application:**

- The Iso-Seq 2 application is currently a **beta** version, and uses a slightly different algorithm than the Iso-Seq application. Compared to the Iso-Seq application, the Iso-Seq 2 application:
  - Is faster and has better specificity.
  - Generates fewer false positive isoforms predicted based on evaluation of synthetic RNA.
  - Filters out more chimeras.
- Use the Iso-Seq application if you have barcoded cDNA samples in one SMRT Cell and want to make consensus isoforms of each barcoded sample **separately** instead of jointly. The Iso-Seq 2 application currently supports only calling isoforms from barcoded cDNA samples jointly, and not separately.
- Certain Iso-Seq application parameters are now obsolete and removed in the Iso-Seq 2 application. (For more details, see the **Parameters** section.)

**Gmap Reference (Required):**

- Specify a Gmap reference sequence to align the SMRT Cells reads to and to produce a consensus sequence.
### Advanced Analysis Parameters - CCS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Strand CCS</td>
<td>OFF</td>
<td>For each ZMW, generate two CCS sequences - one for each strand.</td>
</tr>
<tr>
<td>Maximum Dropped Fraction</td>
<td>0.8</td>
<td>The maximum fraction of subreads that can be dropped before giving up.</td>
</tr>
<tr>
<td>Maximum Subread Length</td>
<td>15,000</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Subread Length</td>
<td>50</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Number of Passes</td>
<td>0</td>
<td>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 does not include any partial passes at the start and end of the sequencing reaction. For the Iso-Seq/Iso-Seq 2 application, a 0-pass read can still be full-length.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.8</td>
<td>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.)</td>
</tr>
<tr>
<td>Minimal Read Score</td>
<td>0.65</td>
<td>The minimum read score of input subreads.</td>
</tr>
<tr>
<td>Minimum SNR</td>
<td>3.75</td>
<td>The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR &lt;3.75 is typically considered lower quality.</td>
</tr>
<tr>
<td>Minimum Z Score</td>
<td>-9999</td>
<td>The minimum Z-Score for a subread to be included in the consensus generating process.</td>
</tr>
<tr>
<td>Polish CCS</td>
<td>OFF</td>
<td>Specify whether to polish CCS sequences using Arrow.</td>
</tr>
<tr>
<td>Report File Output</td>
<td></td>
<td>Specify the name of the report file to output. (Default = ccs_report_txt)</td>
</tr>
<tr>
<td>Emit Individual QVs</td>
<td>OFF</td>
<td>Specify whether to emit dq, iq, and sq “rich” quality tracks.</td>
</tr>
<tr>
<td>Filters to add to the Dataset</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

### Advanced Analysis Parameters - Iso-Seq

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allow Extra 5 Exon</td>
<td>OFF</td>
<td>If ON, collapse shorter 5' transcripts; if OFF, don't collapse them.</td>
</tr>
<tr>
<td>Bin by Primer (Iso-Seq application only)</td>
<td>OFF</td>
<td>Specify binning reads by Primers. This overwrites the Bin by Read Length Manually and Bin by Read Length in KB options. This option is removed in the Iso-Seq 2 application, where binning is done automatically.</td>
</tr>
<tr>
<td>Bin by Read Length Manually (Iso-Seq application only)</td>
<td>NONE</td>
<td>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. This option is removed in the Iso-Seq 2 application, where binning is done automatically.</td>
</tr>
<tr>
<td>Bin by Read Length in KB (Iso-Seq application only)</td>
<td>1</td>
<td>Specify binning reads by the read length, in kb. This option is removed in the Iso-Seq 2 application, where binning is done automatically.</td>
</tr>
</tbody>
</table>
### Advanced Analysis Parameters - Iso-Seq

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMAP nproc</td>
<td>24</td>
<td>The number of processing threads used to run the GMAP aligner. Adjust this value based on your processor power and reference genome size.</td>
</tr>
<tr>
<td>Minimum Quiver/Arrow Accuracy</td>
<td>0.99</td>
<td>Specify the minimum predicted consensus accuracy to classify an isoform as high-quality (HQ). All isoforms below this cutoff are considered LQ.</td>
</tr>
<tr>
<td>Maximum Fuzzy Junction</td>
<td>5</td>
<td>The maximum edit distance between mergeable fuzzy junctions.</td>
</tr>
<tr>
<td>Minimum FL Count</td>
<td>2</td>
<td>The minimum FL count to not filter a collapsed isoform.</td>
</tr>
<tr>
<td>Minimum GMAP aln coverage</td>
<td>0.99</td>
<td>The minimum query coverage to analyze a GMAP alignment.</td>
</tr>
<tr>
<td>Minimum GMAP aln identity</td>
<td>0.95</td>
<td>The minimum identity to analyze a GMAP alignment.</td>
</tr>
<tr>
<td>Minimum Sequence Length</td>
<td>50</td>
<td>The minimum sequence length to output.</td>
</tr>
<tr>
<td>Customer Primer Sequences</td>
<td>NONE</td>
<td>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 &quot;Appendix A - Barcoded Primers&quot; on page 115 for details.</td>
</tr>
<tr>
<td>Trim QVs 3'</td>
<td>30</td>
<td>Specify the number of bases whose Quality Value to ignore in the 3' end.</td>
</tr>
<tr>
<td>Trim QVs 5'</td>
<td>100</td>
<td>Specify the number of bases whose Quality Value to ignore in the 5' end.</td>
</tr>
<tr>
<td>Require PolyA</td>
<td>ON</td>
<td>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.</td>
</tr>
<tr>
<td>Sample Name</td>
<td>NONE</td>
<td>The name of the input sample. A random string is used when Sample Name is NONE.</td>
</tr>
</tbody>
</table>

### Reports and Data Files
The Iso-Seq/Iso-Seq 2 with Mapping 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.) For the Iso-Seq/Iso-Seq 2 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
- Maps CCS reads against 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).

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.

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.

**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 Details Log**: Log information for the analysis workflow.

• **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)

• **Consensus Isoforms**: Consensus Isoforms produced by clustering FLNC reads using the ICE algorithm.

• **Consensus Sequences (XML)**: Consensus sequences generated from CCS, in FASTA format.

• **Consensus Sequences**: Consensus sequences generated from CCS, in FASTQ format.

• **CSV Report (TXT)**: Summary of CCS performance and yield.

• **Draft Isoforms (XML)**: Intermediate Data Set used to get full-length reads.

• **Full-Length Non-Chimeric Reads (XML)**: Full-length non-chimeric subreads generated from pbtranscript classify.

• **High-Quality Isoforms**: Data Set of isoforms with high consensus accuracy.

• **High-Quality Isoforms (FASTQ)**: Data Set of isoforms with high consensus accuracy.

• **Low-Quality Isoforms**: Data Set of isoforms with low consensus accuracy.

• **Low-Quality Isoforms (FASTQ)**: Data Set 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 GFF**: Sequences and quality values of collapsed transcripts, in GFF format.

• **Collapsed Filtered Isoforms FASTQ**: Sequences and quality values of collapsed transcripts, in FASTQ format.

• **FLnFL Reads Status**: Status of FLNC and NFL reads associated with collapsed isoforms.

• **Gmap SAM Mapping HQ isoforms to Genome**: Displays SAM alignments mapping of High-Quality isoforms to the reference genome, using GMAP.
Use this application to determine phased consensus sequences for pooled amplicon data.

The LAA 2 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.

### Parameters

<table>
<thead>
<tr>
<th>Required Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Subread Length</td>
<td>3,000</td>
<td>The minimum length of subreads to use.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - LAA</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcode FASTA file</td>
<td>NONE</td>
<td>FASTA file name of the barcode sequences used, which overwrites any barcode names in the Data Set. <strong>Note</strong>: This is used only to find barcode names.</td>
</tr>
<tr>
<td>Chimera Filter</td>
<td>ON</td>
<td>Specify whether to activate the chimera filter and separate all consensus chimeric outputs.</td>
</tr>
<tr>
<td>Chimera Score Threshold</td>
<td>1</td>
<td>The minimum score to consider a sequence chimeric.</td>
</tr>
<tr>
<td>Cluster Inflation Factor</td>
<td>2</td>
<td>Markov clustering inflation parameter.</td>
</tr>
<tr>
<td>Cluster Loop Weight</td>
<td>0.001</td>
<td>Markov clustering loop weight parameter.</td>
</tr>
</tbody>
</table>
### Advanced Analysis Parameters - LAA

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>ON</td>
<td>Specify whether to activate the coarse clustering phase.</td>
</tr>
<tr>
<td>Barcode Name or Index Pairs</td>
<td>NONE</td>
<td>A comma-separated list of barcode pairs to analyze. This can be by name (&quot;lbc1--lbc1&quot;) or by Index (&quot;0--0&quot;).</td>
</tr>
<tr>
<td>Use Only Full Length Subreads</td>
<td>OFF</td>
<td>Specify whether to require that input reads have both flanking barcodes.</td>
</tr>
<tr>
<td>Disable Barcode Filtering</td>
<td>OFF</td>
<td>Specify that all data be treated as one sample.</td>
</tr>
<tr>
<td>Ignore End-Bases</td>
<td>0</td>
<td>When splitting, ignore N bases at the end. This prevents excessive splitting caused by degenerate primers.</td>
</tr>
<tr>
<td>Maximum Clustering Reads</td>
<td>500</td>
<td>The maximum number of input reads to cluster per barcode.</td>
</tr>
<tr>
<td>Maximum Subread Length</td>
<td>0</td>
<td>The maximum length of input reads to use. To disable, set to 0.</td>
</tr>
<tr>
<td>Maximum Phasing Reads</td>
<td>500</td>
<td>The maximum number of input reads to use for phasing and consensus.</td>
</tr>
<tr>
<td>Maximum Reads</td>
<td>2000</td>
<td>The maximum number of input reads to cluster per barcode.</td>
</tr>
<tr>
<td>Minimum Barcode Score</td>
<td>26</td>
<td>The minimum average barcode score required for subreads.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.95</td>
<td>The minimum predicted consensus accuracy below which a consensus is treated as noise.</td>
</tr>
<tr>
<td>Minimum Read Score</td>
<td>0.75</td>
<td>The minimum read score of input reads.</td>
</tr>
<tr>
<td>Minimum SNR</td>
<td>3.75</td>
<td>The minimum SNR of input reads.</td>
</tr>
<tr>
<td>Minimum Allele/Haplotype Read Fraction</td>
<td>0.10</td>
<td>The minimum fraction of reads favoring the minor phase required to split a haplotype.</td>
</tr>
<tr>
<td>Minimum Allele/Haplotype Reads</td>
<td>20</td>
<td>The minimum number of reads favoring the minor phase required to split a haplotype.</td>
</tr>
<tr>
<td>Minimum Allele/Haplotype Score</td>
<td>500</td>
<td>The global likelihood improvement required to split a haplotype.</td>
</tr>
<tr>
<td>Phasing</td>
<td>ON</td>
<td>Specify that the fine phasing step take place.</td>
</tr>
<tr>
<td>RNG Seed</td>
<td>42</td>
<td>Modulates the reservoir filtering of seeds.</td>
</tr>
<tr>
<td>Take Top N Sequences</td>
<td>0</td>
<td>Report only the top N consensus sequences for each barcode. To disable, use a number less than 1.</td>
</tr>
<tr>
<td>Trim Sequence Ends</td>
<td>0</td>
<td>Specify the number of bases to trim from each end of each consensus sequence.</td>
</tr>
</tbody>
</table>

### Reports and Data Files

The Long Amplicon Analysis (LAA2) 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 consensus sequences not categorized as Chimeric or Noise.
• **Good (%):** The percentage of consensus sequences not categorized as Chimeric or Noise.

• **Chimeric:** The number of consensus sequences flagged as likely coming from PCR cross-over events.

• **Chimeric (%):** The percentage of consensus sequences flagged as likely coming from PCR cross-over events.

• **Noise:** The number of consensus sequences that have 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 consensus sequences that have 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.

**Amplicon Consensus > Amplicon Consensus Summary**

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

• **Sequence Cluster:** A name given to the cluster of sequences roughly corresponding to one amplicon.

• **Sequence Phase:** A name 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 Details Log:** Log information for the analysis workflow.

• **SMRT Link Job Log:** Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)

• **Consensus Amplicons:** Consensus amplicons in FASTQ format.

• **Chimeric/Noise Sequences by Barcode:** Chimeric and noise sequences in FASTQ format.

• **Consensus Sequence Statistics:** Per-contig statistics.

• **Per-ZMW Predicted Accuracy:** Per-contig predicted accuracy, grouped by ZMW.
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, type 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

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

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - CCS</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Subread Length</td>
<td>21000</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Subread Length</td>
<td>10</td>
<td>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.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.99</td>
<td>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.</td>
</tr>
</tbody>
</table>
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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis 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.
- **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

<table>
<thead>
<tr>
<th>Config Version</th>
<th>Predefined v1.1, Pacific internal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Name</td>
<td>HIV HIV2</td>
</tr>
<tr>
<td>Reference Length</td>
<td>9719</td>
</tr>
<tr>
<td>Genes:</td>
<td></td>
</tr>
<tr>
<td>- 5LTR (1-634)</td>
<td></td>
</tr>
<tr>
<td>- p17 (700-1186)</td>
<td></td>
</tr>
<tr>
<td>- p24 (1186-1870)</td>
<td></td>
</tr>
<tr>
<td>- p2 (1879-1921)</td>
<td></td>
</tr>
<tr>
<td>- p7 (1921-2086)</td>
<td></td>
</tr>
<tr>
<td>- p1 (2086-2134)</td>
<td></td>
</tr>
<tr>
<td>- p6 (2134-2202)</td>
<td></td>
</tr>
<tr>
<td>- Protease (2253-2550)</td>
<td></td>
</tr>
<tr>
<td>- ATV/r: V122 L33F N46I M46L I47V G48V G49M I50L I54T I54A I54L I54K V82A V82T V82F V82S I84V N88S L90M</td>
<td></td>
</tr>
<tr>
<td>- DDC/r: V122 L33F I47V I50V I54L I54M I77V V8F I84V</td>
<td></td>
</tr>
<tr>
<td>- FV/r: V122 L33F N46I M46L I47V I50V I54L I54A I54L I54K I54M I54N I57V V82A V82T V82F V82S I84V L90M</td>
<td></td>
</tr>
<tr>
<td>- IDV/r: V122 M46I N46L I47V I54T I54A I54L I54M I77V I82A I82T I82F I82S I84V N88S L90M</td>
<td></td>
</tr>
<tr>
<td>- NVP: D33R L33F N46I M46L I47V G48V G48M I54T I54A I54L I54N I82A I82T I82F I82S I84V N88S L90M</td>
<td></td>
</tr>
<tr>
<td>- SQV/r: G48V G49M I54T I54A I54L I54N I82A I82T I82F I82S I84V N88S L90M</td>
<td></td>
</tr>
<tr>
<td>- TPV/r: V122 L33F N46I M46L I47V I54T I54A I54L I54K V82A V82T V82F V82S I84V</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

### 4. Drug Summaries

Summarizes the variants grouped by annotated drug mutations:
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.

- The row-wise variant calls are "transposed" onto per-column haplotypes. Each haplotype has an ID: `[A-Z]([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:

<table>
<thead>
<tr>
<th>Haplotype Category</th>
<th>#Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported</td>
<td>1735</td>
</tr>
<tr>
<td>Insufficient Coverage (unreported)</td>
<td>66</td>
</tr>
<tr>
<td>Overall Damaged (unreported)</td>
<td>3894</td>
</tr>
<tr>
<td>- Marginal Gaps</td>
<td>786</td>
</tr>
<tr>
<td>- Marginal Heteroduplexes</td>
<td>3709</td>
</tr>
<tr>
<td>- Marginal Partial</td>
<td>76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Haplotypes %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
</tr>
<tr>
<td>2.2</td>
</tr>
<tr>
<td>1.3</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>0.7</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

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 using the Arrow algorithm.

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

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Subread Length</td>
<td>0</td>
<td>The minimum length of subreads to use in the assembly.</td>
</tr>
<tr>
<td>Filters to Add to the Data Set</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Consensus</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Score</td>
<td>0</td>
<td>Specify the score to use in the display.</td>
</tr>
<tr>
<td>Minimum Confidence</td>
<td>40</td>
<td>The minimum confidence for a variant call to be output to the file variants.gff.</td>
</tr>
<tr>
<td>Track Description</td>
<td>NONE</td>
<td>Description to display in the header.</td>
</tr>
<tr>
<td>Track Name</td>
<td>variants</td>
<td>Name to display in the header.</td>
</tr>
<tr>
<td>Purpose</td>
<td>variants</td>
<td>Specify the run mode - Variants or Coverage.</td>
</tr>
<tr>
<td>Masking</td>
<td>ON</td>
<td>During the polish step, omit regions of reads that have low concordance with the template.</td>
</tr>
</tbody>
</table>
| 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. |
Reports and Data Files

The Resequencing application generates the following reports:

**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**

Maps the number of reads against the read length.

**Mapping Report > Mapped Subread Concordance**

• Maps the number of subreads against the percent concordance with the reference sequence.

**Mapping Report > Mapped Subread Length**

• Maps the number of subreads against 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.

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 ecoliK12_pbi_March2013
• Maps coverage of the ecoliK12_pbi_March2013 reference against the reference start position.

Coverage > Depth of Coverage Distribution
• Maps the reference regions against the percent coverage.

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).

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 ecoliK12_pbi_March2013**

• Maps the number of variants across the ecoliK12_pbi_March2013 reference against the reference start position.

**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 Details Log**: Log information for the analysis workflow.
• **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
• **Consensus Sequence**: Data Set containing consensus sequence.
• **Consensus Sequence**: Consensus sequence 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.
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

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Subread Length</td>
<td>0</td>
<td>The minimum length of subreads to use in the assembly.</td>
</tr>
<tr>
<td>Filters to Add to the Data Set</td>
<td>NONE</td>
<td>A comma-separated list of additional Data Set filters to use.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Consensus</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Score</td>
<td>0</td>
<td>Specify the score to use in the display.</td>
</tr>
<tr>
<td>Minimum Confidence</td>
<td>40</td>
<td>The minimum confidence for a variant call to be output to the file <code>variants.gff</code>.</td>
</tr>
<tr>
<td>Track Description</td>
<td>NONE</td>
<td>Description to display in the header.</td>
</tr>
<tr>
<td>Track Name</td>
<td>variants</td>
<td>Name to display in the header.</td>
</tr>
<tr>
<td>Purpose</td>
<td>variants</td>
<td>Specify the run mode - Variants or Coverage.</td>
</tr>
<tr>
<td>Masking</td>
<td>ON</td>
<td>During the polish step, omit regions of reads that have low concordance with the template.</td>
</tr>
</tbody>
</table>
| 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. |

<table>
<thead>
<tr>
<th>Advanced Analysis Parameters - Alignment</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of .bam files</td>
<td>1</td>
<td>Number of .bam files to create in consolidate mode.</td>
</tr>
<tr>
<td>Minimum Concordance</td>
<td>70</td>
<td>The minimum required alignment concordance, in percent.</td>
</tr>
</tbody>
</table>
Reports and Data Files

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

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.

### Advanced Analysis Parameters - Alignment

<table>
<thead>
<tr>
<th>Name</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consolidate .bam</td>
<td>OFF</td>
<td>Specify whether to merge chunked/gathered .bam files.</td>
</tr>
<tr>
<td>Concordant Alignment</td>
<td>ON</td>
<td>Specify whether to map subreads of a ZMW to the same genomic location.</td>
</tr>
<tr>
<td>Align Unsplit Polymerase Reads</td>
<td>OFF</td>
<td>Do not split reads into subreads even if subread regions are available.</td>
</tr>
<tr>
<td>Minimum Length</td>
<td>50</td>
<td>The minimum required alignment length, in base pairs.</td>
</tr>
<tr>
<td>Hit Policy</td>
<td>randombest</td>
<td>Specify how to treat multiple hits:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <strong>random</strong>: Selects a random hit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <strong>all</strong>: Selects all hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <strong>allbest</strong>: Selects all the best score hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <strong>randombest</strong>: Selects a random hit from all best score hits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- <strong>leftmost</strong>: Selects a hit which has the best score and the smallest mapping coordinate in any reference.</td>
</tr>
</tbody>
</table>

### Algorithm Options

List of space-separated arguments passed to BLASR. Default: -minMatch 12 -bestn 10 -minPctIdentity 70.0

### Advanced Analysis Parameters - Reports

<table>
<thead>
<tr>
<th>Name</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>1000</td>
<td>Specify the number of genome regions in the summary statistics. (This is used for guidance, and is not strict.)</td>
</tr>
<tr>
<td>Number of Variants</td>
<td>100</td>
<td>Specify the number of top variants to display.</td>
</tr>
<tr>
<td>Region Size</td>
<td>0</td>
<td>If specified, use a fixed region size.</td>
</tr>
<tr>
<td>Batch Sort Size</td>
<td>10,000</td>
<td>This is an intermediate sort size parameter.</td>
</tr>
<tr>
<td>Maximum Region Size</td>
<td>1,000,000</td>
<td>The upper limit for region size. This is ignored if Region Size is set explicitly.</td>
</tr>
<tr>
<td>Force the Number of Regions</td>
<td>OFF</td>
<td>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.</td>
</tr>
</tbody>
</table>
• **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 95\textsuperscript{th} 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 95\textsuperscript{th} 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**

• Maps the number of reads against the read length.

**Mapping Report > Mapped Subread Concordance**

• Maps the number of subreads against the percent concordance with the reference sequence.

**Mapping Report > Mapped Subread Length**

• Maps the number of subreads against 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.
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 lambda_NEB3011
• Maps coverage of the lambda_NEB3011 reference against the reference start position.

Coverage > Depth of Coverage Distribution
• Maps the reference regions against the percent coverage.

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).

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 ecoliK12_pbi_March2013

- Maps the number of variants across the ecoliK12_pbi_March2013 reference against the reference start position.

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.

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 Details Log**: Log information for the analysis workflow.
- **SMRT Link Job Log**: Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
- **BED file**: List of variants from the reference.
- **Consensus Contigs**: Data Set containing consensus contigs.
- **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.
Use this application to identify large structural variant insertions and deletions (default: ≥50 bp) in a sample or set of samples relative to a reference.

**Reference (Required):**
- Specify a reference genome against which to align the reads and call variants.

**Minimum Length of Structural Variant (bp) (Required, Default = 50):**
- Specify the minimum length of structural variants to be output, in base pairs.

**Minimum Reads That Support Variant (Count) (Required, Default = 2):**
- Specify that at least this number of reads summed across all samples must provide evidence of a structural variant for it to be output.
- Separate subreads from a ZMW count only once.

**Minimum Percentage of Reads That Support Variant (%) (Required, Default = 20):**
- Specify that at least this percentage of reads that span a structural variant locus must provide evidence of a variant in at least one sample for it to be output.
- Separate subreads from a ZMW count only once for support and overall coverage.

**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 Datasets** 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 23.
- 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**

This table describes the called variants broken down by 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.
- **Homozygous Variants:** The count of homozygous variants called in the sample.
- **Heterozygous Variants:** The count of heterozygous variants called in the sample.
- **Total Variants (total bp):** The count and total length (in base pairs) of all called insertions and deletions 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.

- **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.
- **Unannotated:** Variant sequence does not match any of the above patterns.
- **Total:** The sum of variants from all annotations.

- **Insertions (total bp):** The count and total length (in base pairs) of all called insertions with the annotation.
- **Deletions (total bp):** The count and total length (in base pairs) of all called deletions with the annotation.
- **Total Variants (total bp):** The count and total length (in base pairs) of all called insertions and deletions with the annotation.

**Report > Structural Variants > Length Histogram**

- The distribution of variant lengths, in base pairs, broken down by insertions and deletions. For each sample, separate distributions are provided for variants <1 kb (with bin size = 50 bp), and variants ≥1 kb (with bin size = 500 bp). 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 Details Log:** Log information for the analysis workflow.
- **SMRT Link Job Log:** Server-level analysis log information. (This file is displayed when you choose Data > Analysis Log.)
- **Structural Variants:** All the structural variants, in BED or VCF format.
• **Aligned Reads**: All reads aligned to the specified reference, in reference-coordinate sorted order, in BAM format. One BAM is available per sample.
Working with Barcoded Data

This section describes how to use SMRT Link to work with barcoded data. Note: There are major changes to barcoding functionality in this release.

The canned data provided with SMRT Link v5.1.0 includes 5 barcode sets:

- RSII_384_barcode
- RSII_96_barcode
- Sequel_RSII_16_barcode_v1
- Sequel_RSII_96_barcode_v1
- Sequel_RSII_384_barcode_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.

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 Yes 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 No 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 File, 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.

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 25 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 60.

### 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. Select the demultiplexed Data Set to use as input:

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

• **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. Click Start.

• **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.

Demultiplex Barcodes Application Details

The Demultiplex Barcodes application identifies barcode sequences in PacBio single-molecule sequencing data. It replaces pbbarcode and bam2bam for demultiplexing, starting with SMRT Analysis v5.1.0. The core alignment algorithm in the new Demultiplex Barcodes application is the same, but the algorithm to identify barcode pairs, as well as usability, are improved.

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 three different methods for barcoding samples with PacBio technology:

1. Sequence-specific primers
2. Barcoded universal primers
3. Barcoded adapters

Three Ways to Barcode Samples Using PacBio Technology

1. **Sequence-Specific Primers**

2. **Barcoded Universal Primers**

3. **Barcoded Adapters**
In addition, there are three different barcode library designs. As Demultiplex Barcodes supports raw subread and ccs read demultiplexing, the following terminology is based on the per (sub-) read view.

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 v5.1.0 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.
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.
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.
3. Click **Windows Offline (64-bit)**. This downloads a .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 macOS

**Note:** This requires macOS 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.
3. Click **Mac OS X**. This downloads a .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 v5.1.0 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.

![Advanced Link](image)

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

![Proceed Link](image)

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

![Close Box](image)

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.control.scraps.bam
|-- m54008_160116_003634.control.scraps.bam.pbi
|-- m54008_160116_003634.control.subreads.bam
|-- m54008_160116_003634.control.subreads.bam.pbi
|-- m54008_160116_003634.transferdone
```

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/3.0/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.
<table>
<thead>
<tr>
<th>Frequently Asked Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What are the minimum files needed to analyze data on SMRT Link?</strong></td>
</tr>
<tr>
<td>• .bam file</td>
</tr>
<tr>
<td>• bam.pbi file</td>
</tr>
<tr>
<td>• subreadset.xml file</td>
</tr>
<tr>
<td><strong>What is the average size of the file bundle for a 6-hour movie?</strong></td>
</tr>
<tr>
<td>Approximately 5 Gb.</td>
</tr>
<tr>
<td><strong>What is the difference between a regular .bam file and an aligned.bam file?</strong></td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>
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 (v5.1.0)].

**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 (v5.1.0)].
The following procedures are available only for SMRT Link users whose role is Admin.

Adding and Deleting SMRT Link Users
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 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. Roles define which SMRT Link modules a user can access. The following table lists the privileges associated with the three user roles:

<table>
<thead>
<tr>
<th>Tasks/Privileges</th>
<th>Admin</th>
<th>Lab Tech</th>
<th>Bioinformatician</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add/Delete SMRT Link Users</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Assign roles to SMRT Link users</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Update SMRT Link software</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Access Sample Setup Module</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Access Run Design Module</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Access Run QC Module</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Access Data Management Module</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Access SMRT Analysis Module</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

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 will run on the following:

Client Operating Systems
• Microsoft Windows 7 and later
• macOS 10.7 and later

Client Web Browser
• Google Chrome version 56 or later is required.

Note: SMRT Link Server hardware and software requirement are listed in the document SMRT Link Software Installation (v5.1.0).
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:

<table>
<thead>
<tr>
<th>Oligo</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>dT_BC1</td>
<td>AAGCAGTGGTATCAACGCAGAGTACtcagacgatgcgtcatTTTTTTTTTTTTTTTTTTTTTTTTTTTNV</td>
</tr>
<tr>
<td>dT_BC2</td>
<td>AAGCAGTGGTATCAACGCAGAGTAccatacatgactctgcTTTTTTTTTTTTTTTTTTTTTTTTTTTNV</td>
</tr>
</tbody>
</table>

Then the primer file should look like this:

> F0
AAGCAGTGGTATCAACGCAGAGTAC
>R0
tagacgcatcgcttagTACTCT6CGTTGATACCACTGCTT
>F1
AAGCAGTGGTATCAACGCAGAGTAC
>R1
gcagagctatgtagTACTCT6CGTTGATACCACTGCTT

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 Quality and Read Length parameters 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, average base QV, inter-pulse distance, and so on).