Welcome to SMRT® Link

SMRT Link User Guide
Sequel® Systems
For Research Use Only. Not for use in diagnostic procedures.

P/N 102-151-000 Version 01 (November 2021)

© Copyright 2017-2021, Pacific Biosciences of California, Inc. All rights reserved.

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

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

Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at https://www.pacb.com/legal-and-trademarks/terms-and-conditions-of-sale/.

Trademarks:
Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq and Sequel are trademarks of Pacific Biosciences. FEMTO Pulse and Fragment Analyzer are trademarks of Agilent Technologies Inc. All other trademarks are the sole property of their respective owners.

See https://github.com/broadinstitute/cromwell/blob/develop/LICENSE.txt for Cromwell redistribution information.

Pacific Biosciences of California, Inc.
1305 O'Brien Drive
Menlo Park, CA 94025
www.pacb.com
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>Sequel&lt;sup&gt;®&lt;/sup&gt; Systems Users</td>
<td>4</td>
</tr>
<tr>
<td>Contact Information</td>
<td>5</td>
</tr>
<tr>
<td>Using SMRT&lt;sup&gt;®&lt;/sup&gt; Link</td>
<td>5</td>
</tr>
<tr>
<td>Module Menu Commands</td>
<td>6</td>
</tr>
<tr>
<td>Gear Menu Commands</td>
<td>6</td>
</tr>
<tr>
<td>Sending Information to Technical Support</td>
<td>7</td>
</tr>
<tr>
<td>Sample Setup</td>
<td>8</td>
</tr>
<tr>
<td>Application-Based Calculations</td>
<td>8</td>
</tr>
<tr>
<td>Custom Calculations</td>
<td>11</td>
</tr>
<tr>
<td>Advanced Options</td>
<td>11</td>
</tr>
<tr>
<td>Editing or Printing Calculations</td>
<td>11</td>
</tr>
<tr>
<td>Deleting Calculations</td>
<td>11</td>
</tr>
<tr>
<td>Importing/Exporting Calculations</td>
<td>12</td>
</tr>
<tr>
<td>Run Design</td>
<td>15</td>
</tr>
<tr>
<td>Creating a New Run Design</td>
<td>15</td>
</tr>
<tr>
<td>Custom Run Designs</td>
<td>18</td>
</tr>
<tr>
<td>Advanced Options</td>
<td>19</td>
</tr>
<tr>
<td>Editing or Deleting Run Designs</td>
<td>19</td>
</tr>
<tr>
<td>Creating a Run Design by Importing a CSV File</td>
<td>19</td>
</tr>
<tr>
<td>Run QC</td>
<td>26</td>
</tr>
<tr>
<td>Table Fields</td>
<td>26</td>
</tr>
<tr>
<td>Run Settings and Metrics</td>
<td>29</td>
</tr>
<tr>
<td>Data Management</td>
<td>32</td>
</tr>
<tr>
<td>What is a Data Set?</td>
<td>32</td>
</tr>
<tr>
<td>Creating a Data Set</td>
<td>33</td>
</tr>
<tr>
<td>Viewing Data Set Information</td>
<td>34</td>
</tr>
<tr>
<td>Copying a Data Set</td>
<td>34</td>
</tr>
<tr>
<td>Deleting a Data Set</td>
<td>35</td>
</tr>
<tr>
<td>Starting an Analysis from a Data Set</td>
<td>35</td>
</tr>
<tr>
<td>Data Set QC Reports</td>
<td>35</td>
</tr>
<tr>
<td>What is a Project?</td>
<td>37</td>
</tr>
<tr>
<td>Data Sets and Projects</td>
<td>37</td>
</tr>
<tr>
<td>Creating a Project</td>
<td>37</td>
</tr>
<tr>
<td>Editing a Project</td>
<td>38</td>
</tr>
<tr>
<td>Deleting a Project</td>
<td>38</td>
</tr>
<tr>
<td>Viewing/Deleting Sequence, Reference and Barcode Data</td>
<td>39</td>
</tr>
<tr>
<td>Importing Sequence, Reference and Barcode Data</td>
<td>39</td>
</tr>
<tr>
<td>Exporting Sequence, Reference and Barcode Data</td>
<td>40</td>
</tr>
<tr>
<td>SMRT&lt;sup&gt;®&lt;/sup&gt; Analysis</td>
<td>42</td>
</tr>
<tr>
<td>Creating and Starting an Analysis</td>
<td>42</td>
</tr>
</tbody>
</table>
Starting an Analysis After Viewing Sequence Data 47
Canceling a Running Analysis 48
Restarting a Failed Analysis 48
Viewing Analysis Results 48
Copying and Running an Existing Analysis 49
Exporting an Analysis 50
Importing an Analysis 50

PacBio® Secondary Analysis Applications 51
Assembly (HGAP 4) Application 54
Base Modification Analysis Application 60
CCS with Demultiplexing Application 65
CCS with Mapping Application 71
Circular Consensus Sequencing (CCS) Application 76
Convert BAM to FASTX Application 79
Demultiplex Barcodes Application 80
Export Reads Application 85
Genome Assembly Application 87
HiFiViral SARS-CoV-2 Analysis Application 90
Iso-Seq® Analysis Application 96
Long Amplicon Analysis (LAA) Application 102
Mapping Application 105
Mark PCR Duplicates Application 110
Microbial Assembly Application 112
Minor Variants Analysis Application 119
Site Acceptance Test (SAT) Application 125
Structural Variant Calling Application 129
Trim gDNA Amplification Adapters Application 133

Working with Barcoded Data 135
Step 1: Specify Barcode Setup & Sample Names in a Run Design 136
Step 2: Perform the Sequencing Run 137
Step 3: (Optional) Run the Demultiplex Barcodes Application 138
Step 4: Run Applications Using the Demultiplexed Data as Input 139
Demultiplex Barcodes Application Details 141

Automated Analysis 143
Creating Auto Analysis From a Run Design 143
HiFiViral SARS-CoV-2: Creating Auto Analysis in Run Design 144
Creating Auto Analysis Directly From SMRT Analysis 144
Getting Information About Analyses Created by Auto Analysis 144
Getting Information About Pre Analysis From SMRT Analysis 145
Getting Information About Pre Analysis From Run Design 145

Visualizing Data Using IGV 146
Using the PacBio® Self-Signed SSL Certificate 148
Sequel® Systems Output Files 149
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequel Ile System Output Files</td>
<td>149</td>
</tr>
<tr>
<td>Sequel II and Sequel Systems Output Files</td>
<td>151</td>
</tr>
<tr>
<td><strong>Secondary Analysis Output Files</strong></td>
<td>153</td>
</tr>
<tr>
<td>Configuration and User Management</td>
<td>156</td>
</tr>
<tr>
<td>LDAP</td>
<td>156</td>
</tr>
<tr>
<td>SSL</td>
<td>156</td>
</tr>
<tr>
<td>Adding and Deleting SMRT Link Users</td>
<td>157</td>
</tr>
<tr>
<td>Assigning User Roles</td>
<td>157</td>
</tr>
<tr>
<td><strong>Hardware/Software Requirements</strong></td>
<td>159</td>
</tr>
<tr>
<td>Appendix A - Pacific Biosciences Terminology</td>
<td>160</td>
</tr>
<tr>
<td>Appendix B - Data Search</td>
<td>164</td>
</tr>
<tr>
<td>Appendix C - BED File Format for Target Regions Report</td>
<td>166</td>
</tr>
<tr>
<td>Appendix D - Additional Information: CCS Data Set Export Report</td>
<td>167</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. SMRT Link includes the following modules:

- **Sample Setup**: Calculate binding and annealing reactions for preparing DNA libraries for use on all Sequel Systems. (See “Sample Setup” on page 8 for details.)
- **Run Design**: Design sequencing runs and create and/or import sample sheets. (See “Run Design” on page 15 for details.)
- **Run QC**: Monitor run progress, status and quality metrics. (See “Run QC” on page 26 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 32 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 42 for details.)

This document also describes:

- The data files generated by the Sequel Systems for each cell transferred to network storage. (See “Sequel® Systems Output Files” on page 149 for details.)
- The data files generated by secondary analysis. (See “Secondary Analysis Output Files” on page 153 for details.)
- Configuration and user management. (See “Configuration and User Management” on page 156 for details.)
- SMRT Link client hardware/software requirements. (See “Hardware/Software Requirements” on page 159 for details.)

Installation of SMRT Link **Server** software is discussed in the document **SMRT Link Software Installation (v10.2)**.

New features, fixed issues and known issues are listed in the document **SMRT Link Release Notes (v10.2)**.

**Sequel® Systems Users**

When you first start SMRT Link, you must specify which System you are using: Sequel, Sequel II, or Sequel IIe. This choice affects some of the initial values used in the Sample Setup and Run Design modules. In those modules, you can switch between the three Sequel Systems as needed. Users with administrator access can configure SMRT Link to support all instrument types.
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 148.

After accessing SMRT Link, the Home page displays.

• Click the PacBio logo at the top left to navigate back to the SMRT Link Home page from within the application.
• Click the Gear menu to sign out, configure for the Sequel/Sequel II/Sequel IIe, 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. (A Select a Module menu displays next to the PacBio logo, allowing you to move between modules.)
• Click ? to view the SMRT Link Online help.
• Select Sign Out from the Gear menu to log out of SMRT Link.
Module Menu Commands

- **Sample Setup**: Displays the Sample Setup module.
- **Run Design**: Displays the Run Design module.
- **Run QC**: Displays the Run QC module.
- **Data Management**: Displays the Data Management module.
- **SMRT Analysis**: Displays the SMRT Analysis module.

Gear Menu Commands

- **Show Alarms**
  - Displays SMRT Link system-level alarms. To clear alarms, select and click **Clear Alarm** or **Clear All Alarms**.
- **Configure**
  - To specify the Sequel System(s) that SMRT Link will be used with, click **Instruments** and check the appropriate box(es).
  - **Admin users only**: Add/delete SMRT Link users and specify their roles. See “Adding and Deleting SMRT Link Users” on page 157 for details.
  - To specify how numbers are formatted, click **Number Formatting** and select **Period** or **Comma** as the decimal separator.
  - (Sequel IIe System only) To specify whether CCS Analysis output includes kinetics information (used for epigenetics analysis), click **CCS Analysis Output** and select **Yes** or **No**. This is the default setting for all CCS Analysis output, unless overwritten in individual Run Designs. **Note**: Adding kinetics information can increase the amount of storage used by the output BAM files by up to 5 times.
- **About SMRT Link**
  - Displays software version information and available space on the server SMRT Link is connected to.
  - Click **Send** to send configuration information and/or analysis usage information to Pacific Biosciences Technical Support for help in troubleshooting failed analyses.
  - **Admin users only**: 1) Update the SMRT Link **Chemistry Bundle**, which includes kit and DNA Control Complex names used in the Sample Setup and Run Design modules. 2) Update the SMRT Link **UI Bundle**, which includes changes and bug fixes to the SMRT Link Graphical User Interface or UI for a SMRT Link module.
- **Sign Out**
  - Logs you out and displays the initial login page.

Working with Tables

- To **sort** table columns: Click a **column title**.
- To see **additional** columns: Click the > symbol next to a column title.
- To **search** within a table: Enter a unique search string into the **Search** field. (For details, see “Appendix C - Data Search” on page 165.)
Sending Information to Technical Support

To open a case with Pacific Biosciences Technical Support, send an email to support@pacb.com.

Troubleshooting information can be sent to PacBio Technical Support two ways:

- From the SMRT Link menu: **About > Troubleshooting Information > Send**.
- From a SMRT Link “Failed” analysis Results page: Click **Send Log Files**.
Sample Setup

To prepare your samples for sequencing, 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. You can then print the instructions for use in the lab.

1. Access SMRT Link using the Chrome web browser.
2. Select **Sample Setup**.
3. Specify if this calculation is to be used with a Sequel, Sequel II or Sequel IIe System. This affects the initial default values.
4. Click **+ New Calculation**.
5. Enter the sample name.

**Application-Based Calculations**
6. Select a sequencing application for the sample. The following fields are auto-populated and display in green:
   - Sequencing Primer
   - Binding Kit

   **Note:** The following fields are located in **Advanced Options**:
   - Target Annealing Concentration
   - Target Binding Concentration
   - Target Polymerase Concentration (Relative)
   - Binding Time
   - Cleanup Bead Time
   - Cleanup Bead Concentration

7. Enter the available sample **volume**, in µL.
8. Enter the sample **concentration**, in ng/µL.
9. Specify an **Insert Size**, in 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 mean insert size for the sample; the size range boundaries are described in the library preparation protocol and in the **Quick Reference Card - Loading and Pre-Extension Recommendations for the Sequel/Sequel II System** documents. Enter the mean insert size of the sample.
10. Select the **Internal Control** version to use for this run from the list, or type in a part number. Pacific Biosciences **highly** recommends using the Internal Control to help distinguish between sample quality and instrument issues in the event of suboptimal sequencing performance.  
    **(Note:** PacBio requires the use of the Sequel Internal Control for consumables to be eligible for reimbursement consideration.)
11. If necessary, edit the **Cleanup anticipated yield**. Adjust this percentage based on previous experience. (Cleanup removes excess primers/polymerase from bound complexes, which results in higher quality data.)
12. Specify the on-plate loading concentration (OPLC), in pM.
13. Enter the number of SMRT® Cells to bind, at the specified on-plate loading concentration.

14. Rather than leave a small amount of library behind, use the entire library volume available if desired by selecting **Prepare Entire Sample > Yes**. This generates annealing, binding and cleanup instructions for the **entire available sample volume**. The instructions for loading the sample plate will still follow the scale indicated by the specified number of SMRT Cells to run.

15. In the complex cleanup step, enter the pre- and post-clean-up sample DNA quantitation and volume measurement results.

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume of Purified Complex (μL)</td>
<td>μL</td>
</tr>
<tr>
<td>Purified Complex Concentration (ng/μL)</td>
<td>ng/μL</td>
</tr>
<tr>
<td>Molar Concentration of Purified Complex (mM)</td>
<td>???</td>
</tr>
<tr>
<td>Ampure Cleanup Yield (%)</td>
<td>???</td>
</tr>
</tbody>
</table>

16. Optionally, specify an alternative number of cells or on-plate loading concentration (OPLC) for the final sample dilution step. Use this feature, for example, to initially set up a single-SMRT Cell run to test a specific loading concentration prior to conducting a multi-SMRT Cell sequencing run, or to set up a loading titration experiment to optimize the OPLC for your particular sample.

17. 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 **Lock** to lock the calculation. This is **required** before samples can be imported into the Run Design module, and also sends a finalized version of the instructions to the server for use in Data Set reports. After locking, no further changes can be made to a calculation. (Click **View** to see the locked instructions.) Locking ensures that calculations are always synchronized with their run time state if a report is generated at a later date. (**Lock** is **only** available if there are one or more samples visible and most fields have values entered.)

– Click the **New Sample** button at the top of the screen to start a new, empty sample.

18. Specify whether to display the **full** instructions, or only the **loading** instructions.

19. To **print** the calculation(s) and instructions, use the browser’s Print command (**Ctrl-P**).

**Custom Calculations**

1. To accommodate new or unique sample types, choose **Application > Custom** and enter all settings manually.

2. Click **Set Custom Preset Values** to save any custom application settings you may have specified. The next time you select **Application > Custom**, those settings are retrieved.

**Advanced Options**

• Specify the **Minimum Pipetting Volume**, in uL. This allows you to set a lower limit on pipetting volumes to use in certain protocol steps, such as sample annealing and binding. We recommend setting this to 1 uL, though in some cases, for example if sample availability is very limited, it may be appropriate to set a value below 1 uL. Some protocol steps include fixed values of 1 uL that will **not** be affected by this setting.

• Specify the **% of Annealing Reaction to Use in Binding**. This accommodates pipetting underage: Due to pipetting issues, volumes may not add up to what they should; a value below 100% helps ensure there will be enough annealed sample for binding.

**Editing or Printing Calculations**

1. On the **Sample Setup** screen, select one or more calculation names.

2. Click **Edit/Print**. (**Note:** If the samples use different versions of chemistry, a warning message displays.)

3. Edit the sample(s) as necessary.

4. Specify whether to display the **full** instructions, or only the **loading** instructions.

5. To print the calculation(s), use the browser’s **Print** command (**Ctrl-P**).

**Deleting Calculations**

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

2. Click **Delete**.
**Importing/Exporting Calculations**

Sample Setup supports importing and exporting calculations in CSV format.

To import a new calculation, first find (or create) a calculation similar to that you wish to import, then export it in CSV format. You can then customize the exported CSV file as needed, then import the modified CSV file.

1. Access SMRT Link using the Chrome web browser.
2. Select **Sample Setup**.
3. Select an existing calculation.
4. Click **Export**, then click **Download**.
5. Edit the exported calculation in Excel (changing sample names, concentrations, and so on), then save it under a new name.
6. In Sample Setup, click **Import**.
7. Click **Browse**, then select the CSV file you previously modified in Step 5 and click **Open**. If everything is correct, click **Continue**. The imported calculation displays.

**Note:** You can select multiple calculations to export to the same CSV file. You can also import multiple calculations by adding rows to the CSV file.

Following are the fields contained in the CSV-format Calculations file.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Required</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Name</td>
<td>Yes</td>
<td>Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only.</td>
</tr>
<tr>
<td>System Name</td>
<td>Yes</td>
<td>Must be Sequel, Sequel II, or Sequel IIE.</td>
</tr>
<tr>
<td>Application</td>
<td>Yes</td>
<td>Enter one of the following values:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• HiFi Reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Continuous Long Reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Low DNA Input</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Ultra-Low DNA Input</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Microbial Assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Variant Calling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Structural Variation Calling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• HiFiViral SARS-CoV-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Iso-Seq Method</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Full-Length 16S rRNA Sequencing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Shotgun Metagenomic Profiling or Assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• &lt;3kb Amplicons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• &gt;=3kb Amplicons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Custom</td>
</tr>
<tr>
<td>Available Starting Sample Volume (uL)</td>
<td>Yes</td>
<td>Enter a positive integer. Units are in microliters.</td>
</tr>
<tr>
<td>Starting Sample Concentration (ng/µL)</td>
<td>Yes</td>
<td>Enter a positive integer. Units are in nanograms per microliter.</td>
</tr>
<tr>
<td>Insert Size (bp)</td>
<td>Yes</td>
<td>Enter a positive integer. Units are in base pairs.</td>
</tr>
<tr>
<td>Field Name</td>
<td>Required</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Control Kit</td>
<td>No</td>
<td>Must be blank or Lxxxxx101717600123199.</td>
</tr>
<tr>
<td>Cleanup Anticipated Yield (%)</td>
<td>No</td>
<td>Enter a positive integer. Note: If Application is set to Custom, this field is required.</td>
</tr>
<tr>
<td>On Plate Loading Concentration (pM)</td>
<td>Yes</td>
<td>Enter a positive integer. Units are in parts per million.</td>
</tr>
<tr>
<td>Cells to Bind (cells)</td>
<td>Yes</td>
<td>Enter a positive integer.</td>
</tr>
<tr>
<td>Prepare Entire Sample</td>
<td>Yes</td>
<td>Enter a Boolean value: true, t, yes, y, false, f, no, or n. Boolean values are not case-sensitive.</td>
</tr>
</tbody>
</table>
| Sequencing Primer                               | Yes      | Enter one of the following values:  
  - Sequencing Primer v2  
  - Sequencing Primer v4  
  - Sequencing Primer v5 |
| Binding Kit                                     | Yes      | For Sequel II/IIe Binding Kits 1.0, 2.0, 2.1, and 2.2:  
  - Lxxxxx101717300123199  
  - Lxxxxx101789500123199  
  - Lxxxxx101820500123199  
  - Lxxxxx101894200123199  
For Sequel Binding Kits 2.0 and 3.1:  
  - Lxxxxx101365900123199  
  - Lxxxxx101500400123199 |
| Target Annealing Concentration (nM)             | No       | Enter a positive integer. Units are in nanomolar. Note: If Application is set to Custom, this field is required. |
| Target Binding Concentration (nM)               | No       | Enter a positive integer. Units are in nanomolar. Note: If Application is set to Custom, this field is required. |
| Target Polymerase Concentration (X)             | No       | Enter a positive integer. Note: If Application is set to Custom, this field is required. |
| Binding Time (hours)                            | No       | Enter a positive integer. Note: If Application is set to Custom, this field is required. |
| Cleanup Bead Type                               | No       | Must be AMPure or ProNex. Note: If Application is set to Custom, this field is required. |
| Cleanup Bead Concentration (X)                  | No       | Enter a positive integer. Note: If Application is set to Custom, this field is required. |
| Minimum Pipetting Volume (uL)                   | No       | Enter a positive integer. Units are in microliters.                       |
| Percent of Annealing Reaction To Use In Binding (%)| No       | Enter a positive integer. Note: If Application is set to Custom, this field is required. |
| AMPure Diluted Bound Complex Volume (uL)        | No       | Enter a positive integer. Units are in microliters.                       |
| AMPure Diluted Bound Complex Concentration (ng/uL) | No       | Enter a positive integer. Units are in nanograms per microliter.         |
| AMPure Purified Complex Volume (uL)             | No       | Enter a positive integer. Units are in microliters.                       |
| AMPure Purified Complex Concentration (ng/uL)   | No       | Enter a positive integer. Units are in nanograms per microliter.         |
| ProNex Diluted Bound Complex Volume (uL)        | No       | Enter a positive integer. Units are in microliters.                       |
| ProNex Diluted Bound Complex Concentration (ng/uL)| No  | Enter a positive integer. Units are in nanograms per microliter.         |
| ProNex Purified Complex Volume (uL)             | No       | Enter a positive integer. Units are in microliters.                       |
CSV File General Requirements

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

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

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

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

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

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

- Use SMRT Link’s Run Design module to create a new Run Design.
- 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

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 + Create New Design.
5. Specify if this Run Design is to be used with a Sequel, Sequel II or Sequel Ile System. This affects the initial default values.

6. Enter a **Run Name**. (The software creates a new run name based on the current date and time; you can edit the name as needed.)

7. **(Optional)** Enter **Run Comments**, **Experiment Name**, and **Experiment ID** as needed. (Note: Experiment ID must be alphanumeric.)

8. **(Optional)** Click **Import from Sample Setup** to import information from a previously-created Sample Setup entry. The following fields are auto-populated as appropriate:
   - Sample Name
   - Binding Kit
   - DNA Control Complex
   - Insert Size
   - On-Plate Loading Concentration

**Application-Based Run Designs**

9. Select a sequencing **application** from the list. The following fields are auto-populated, and display in green:
– Template Prep Kit
– Binding Kit
– Sequencing Kit
– DNA Control Complex
– Movie Time per SMRT Cell (hours)
– Pre-Extension Time (hours)

10. Enter a **Well Sample Name**. (This is the name of the sequencing library loaded into one well. **Example**: HG002_2019_11_02_10K)
11. Enter a **Bio Sample Name**. (This is the name of the biological sample contained in the sequencing library, such as HG002. See “Working with Barcoded Data” on page 135 for details.)
12. (Optional) Enter **Sample Comments**.
13. Specify the **well position** used for this sample: Click the icon to the right of the entry field and choose a plate position.
14. Specify an **Insert Size** (500 base pairs minimum). The Insert Size is the length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters. This matches the average insert size for the sample; the size range boundaries are described in the library preparation protocol and in the Quick Reference Card - Loading and Pre-Extension Recommendations for the Sequel System document. **Note**: The default Insert size for Continuous Long Reads is 30,000; 10,000 for CCS Reads.

15. Specify the **On-Plate loading concentration** (OPLC), in picomolarity.
16. Specify if and where to automatically generate **HiFi Reads** (reads generated with CCS Analysis whose quality value is equal to or greater than 20):
   - **On-Instrument** (available only for the Sequel IIe System): HiFi Reads are automatically generated on the instrument, **before** transfer to the compute cluster where SMRT Link is installed.
   - **In SMRT Link**: HiFi Reads are automatically generated **after** transfer to the compute cluster where SMRT Link is installed.
   - **Do Not Generate**: HiFi Reads are **not** generated for this run. Only subread data are transferred to the local compute cluster where SMRT Link is installed.

17. (Optional) If you are using **barcoded samples**, see “Step 1: Specify the Barcode Setup and Sample Names in a Run Design” on page
for instructions. For details on secondary analysis of barcoded samples, see “Demultiplex Barcodes Application” on page 80.

18. Sample options:
   – Click Copy. This starts a new sample, using the values entered in the first sample.
   – Click Delete. This deletes the current sample.
   – Click Add Sample. This starts a new, empty sample.

19. 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.

20. 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.

21. (Optional) Auto Analysis allows a specific analysis to be automatically run after a sequencing run has finished and the data transferred to the SMRT Link Server. See “Automated Analysis” on page 143 for details.

Custom Run Designs
To accommodate new or unique Run Designs, choose Application > Custom and enter all parameters manually. (See here for recommendations based on the analysis application used.)

– **Template Prep Kit, Binding Kit, or Sequencing Kit**: Select one from the list, or type in a kit part number. If the barcode is invalid, "Invalid barcode" displays.
   **Note**: If the Sequencing or Binding kit selected is incompatible, an error message displays indicating the obsolete chemistry, and the run is prevented from proceeding.

– **DNA Control Complex**: Pacific Biosciences highly recommends using the Internal Control to help distinguish between sample quality and instrument issues in the event of suboptimal sequencing performance. **(Note: PacBio requires the use of the Sequel Internal Control for consumables to be eligible for reimbursement consideration.)**

– **Movie time per SMRT Cell (hours)**: Enter a time between 0.5 and 20 (Sequel System) or between 0.5 and 30 (Sequel II Systems).
   **Note**: For the Sequel System, 15 and 20 hour movie times require the use of the SMRT Cell 1M LR part, and 20 hours is the maximum movie time. For Sequel II Systems, the SMRT Cell 8M part supports all movie times up to 30 hours.

– **Use Pre-Extension**: If selected, optionally specify the length of pre-extension time in hours. This initiates the sequencing reaction prior to data acquisition. After the specified time, the sequencing reagents are removed from the SMRT Cell and replenished with fresh reagents, and data acquisition starts. This feature is useful for short inserts (such as ≤15 kb) and provides a significant increase in read length.
Advanced Options

• (Sequel II and IIe Systems Only) Specify whether to use Adaptive Loading. Adaptive Loading uses active monitoring of the ZMW loading process to predict a favorable loading end point. Certain steps (Clean-up and Sample Dilution) require a different buffer (Adaptive Loading Buffer) if this feature is used. **Note:** Adaptive Loading requires the use of Sequel® II Binding Kit 2.2. If you select **Yes,** fill in the following fields:
  – **Loading Target (P1 + P2):** The fraction of ZMWs that the Adaptive Loading routine will aim to load with at least one sequencing complex. The default target for CCS applications is higher to accommodate loss of complexes during pre-extension, which is generally recommended for all CCS applications.
  – **Maximum Loading Time (hours):** This defines the maximum time the system will allow loading to progress before proceeding to sequencing. (Loading time in Adaptive Loading is flexible.)

• Specify the length of time (1, 2 or 4 hours) for **immobilization** of SMRTbell templates. This is the length of time the SMRT Cell is at the Cell Prep Station to allow diffusion of SMRTbell templates into the ZMWs. This option is **not** available if **Adaptive Loading** is selected.
  – PacBio **highly recommends** using the default immobilization time of 2 hours.

• (Sequel IIe Systems Only) Specify, for this Run Design only, whether to include kinetics information (used for epigenetics analysis) in the CCS Analysis output. This setting **overwrites** the global setting in Gear > Configure > CCS Analysis Output. **Note:** Adding kinetics information can increase the amount of storage used by the output BAM files by up to 5 times.

• **Add Data to Project:** Specify that Data Sets generated by SMRT Cell(s) using this Run Design be associated with the selected Project. (This also applies to any Data Sets generated using Auto Analysis. By default, all Data Sets are assigned to **General Project,** which is accessible to all users.)

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.

Creating a Run Design by Importing a CSV File

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

To obtain the sample CSV files

1. On the Home page, select Run Design.
2. Click **Import Run Design**.
3. Click **Download Template**. The ZIP file containing templates (one for Sequel Systems, and one for Sequel II Systems) downloads to your local machine.

**To update and import the CSV file**

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

**CSV File Structure**

- Each CSV file row represents **one sample**.
- The first row contains run-level information such as Run Name, Run Comments, and so on.
- For demultiplexed samples **only**, **one additional row** per barcode/Bio Sample Name combination is added below the master sample row.

**Note**: Specifying cluster settings configuration is **not yet** supported from the Run Design CSV.

<table>
<thead>
<tr>
<th>Run Design Attribute</th>
<th>Required</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment Name</td>
<td>No</td>
<td>Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Defaults to Run Name.</td>
</tr>
</tbody>
</table>
| Experiment Id          | No       | Enter a valid experiment ID. **Example**: 325/3250057  
  - Experiment IDs **cannot** contain the following characters: `<`, `>`, `:`, `\`, `,`, `?`, `*`, or `)`.  
  - Experiment IDs **cannot** start or end with a `/` and **cannot** have two adjacent `/` characters, such as `//`.  
  - Experiment IDs **cannot** contain spaces.  
  - Specifically, Experiment IDs **cannot** satisfy the regular expressions: `/^[<>:\"\\]\?\*$]/g, /(\?:\^\/)\/[\?\:\$]/, /\g` |
<p>| Experiment Description | No       | Enter any ASCII string. Defaults to Run Comments. | <strong>Example</strong>: 20170530_A6_VVnC_SampleSheet |
| Run Name               | Yes      | Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Run name <strong>must</strong> be entered for the first cell and will be applied to the remaining cells in the run. | <strong>Example</strong>: 20170530_A6_VVnC_SampleSheet |
| System Name            | No       | <strong>Must be</strong> Sequel, Sequel II, or Sequel IIe. |</p>
<table>
<thead>
<tr>
<th>Run Design Attribute</th>
<th>Required</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run Comments</td>
<td>No</td>
<td>Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Example: ecoliK12_March2019</td>
</tr>
<tr>
<td>Is Collection</td>
<td>No</td>
<td>Enter a Boolean value. (See Boolean details below.) Specifies whether the row designates a Collection (TRUE) or a barcoded sample (FALSE).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Collection lines should have the Barcode Name and Bio Sample Name fields blank.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Barcoded Sample lines only need to include the Is Collection, Sample Name, Barcode Name, and Bio Sample Name fields.</td>
</tr>
<tr>
<td>Sample Well</td>
<td>Yes</td>
<td>Must be specified in every row. Well number must start with a letter A through H and end in a number 01 through 12, i.e., A01 through H12.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>It must satisfy the regular expression `^[A-H](?:0[1-9]</td>
</tr>
<tr>
<td>Well Sample Name</td>
<td>Yes</td>
<td>Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Example: A6_3230046_A01_SB_ChemKitv2_8rxnKit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: The Sample Name must be unique within a run.</td>
</tr>
<tr>
<td>Movie Time per SMRT Cell (hours)</td>
<td>Yes</td>
<td>Enter a floating point number between 0.1 and 20 for Sequel; 0.1 and 30 for Sequel II Systems. Time is in hours. Example: 5</td>
</tr>
<tr>
<td>Use Adaptive Loading</td>
<td>No</td>
<td>Enter a Boolean value. (See Boolean details below.)</td>
</tr>
<tr>
<td>Loading Target (P1 + P2)</td>
<td>No</td>
<td>Enter a floating point number between 0.01 and 1. Example: 0.4</td>
</tr>
<tr>
<td>Maximum Loading Time (hours)</td>
<td>No</td>
<td>Enter a floating point number between 1 and 2. Time is in hours. Example: 1.2</td>
</tr>
<tr>
<td>Sample Comment</td>
<td>No</td>
<td>Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Example: A6_3230046_A01_SB_BindKit_ChemKit</td>
</tr>
<tr>
<td>Insert Size (bp)</td>
<td>Yes</td>
<td>Enter an integer ≥ 10. Units are in base pairs. Example: 2000</td>
</tr>
<tr>
<td>On Plate Loading Concentration (pM)</td>
<td>No</td>
<td>Enter a floating point number. Units are in parts per million. Example: 5</td>
</tr>
<tr>
<td>Size Selection</td>
<td>No</td>
<td>Enter a Boolean value. (See Boolean details below.) Default is FALSE.</td>
</tr>
<tr>
<td>Template Prep Kit Box Barcode</td>
<td>Yes</td>
<td>Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM1117100259100111716</td>
</tr>
<tr>
<td>DNA Control Complex Box Barcode</td>
<td>No</td>
<td>Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM1234101084300123120</td>
</tr>
<tr>
<td>Binding Kit Box Barcode</td>
<td>Yes</td>
<td>Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM1117100862200111716</td>
</tr>
<tr>
<td>Sequencing Kit Box Barcode</td>
<td>Yes</td>
<td>Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM0001100861800123120</td>
</tr>
<tr>
<td>Automation Name</td>
<td>No</td>
<td>Enter diffusion, magbead (not case-sensitive) or a custom script. (Sequel II Systems do not support magbead loading.) A path can also be used, such as /path/to/my/script/my_script.py. The path will not be processed further, so if the full URI is required, it must be included in the CSV, such as chemistry://path/to/my/script/my_script.py.</td>
</tr>
<tr>
<td>Run Design Attribute</td>
<td>Required</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>----------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Automation Parameters</td>
<td>No</td>
<td>To enable Pre-Extension time, enter the number of hours and set the boolean value to TRUE. Example 2 hours: ( \text{ExtensionTime}=\text{double}:2 \mid \text{ExtendFirst}=\text{boolean}:\text{TRUE} ) (Note: Leave blank when not using Pre-Extension time, or set the boolean value to \text{FALSE}.)</td>
</tr>
</tbody>
</table>
| Generate HiFi Reads                  | No       | **Sequel System, Sequel II System** Enter one of the following values: In \text{SMRT Link}, Do Not Generate.  
**Sequel Ile System** Enter one of the following values: In \text{SMRT Link}, \text{On Instrument}, Do Not Generate.  
If left blank, default is Do Not Generate for all Systems. |
| Sample is Barcoded                   | No       | Enter a boolean value. (See Boolean details below.) Set to \text{TRUE} for a barcoded run.                                                   |
| Barcode Set                          | No       | Must be a UUID for a Barcode Set present in the database.  
To find the UUID: Click \text{Data Management} > \text{View Data} > \text{Barcodes}.  
Click the Barcode file of interest, then view the UUID.  
Example: dad4949d-f637-0979-b5d1-9777eff62008  
Note: This field is used for demultiplexed data. |
| Same Barcodes on Both Ends of Sequence | No   | Enter a boolean value. (See Boolean details below.) Set to \text{TRUE} if symmetric, \text{FALSE} if asymmetric. |
| Barcode Name                         | No       | Enter Barcode Names one per line.  
Example: bc1001--bc1001  
• Use double hyphens (--) to separate the 2 barcodes of each pair.  
• The barcode names must be contained within the specified Barcode Set.  
• A given barcode name cannot appear more than once in the spreadsheet.  
• A maximum of 15,000 barcodes is permitted per sample. |
| Bio Sample Name                      | Yes      | Enter Bio Sample Names in the same row as their associated Barcode Names. Use alphanumeric characters, spaces (allowed but not recommended for compatibility with downstream software), hyphens, underscores, colons, or periods only. Bio Sample Names cannot be longer than 40 characters.  
Example: sample1  
Note: This field is used for collections for non-multiplexed data, and for barcoded samples in multiplexed data. |
<table>
<thead>
<tr>
<th>Run Design Attribute</th>
<th>Required</th>
<th>Description</th>
</tr>
</thead>
</table>
| Pipeline ID          | No       | Note: This field is required to create an Auto Analysis.  
• Assembly (HGAP4): cromwell.workflows.pb_hgap4  
• Base Modification Detection:  
cromwell.workflows.pb_basemods  
• CCS: cromwell.workflows.pb_ccs  
• CCS with Demultiplexing:  
cromwell.workflows.pb_ccs_demux  
• CCS w/Mapping: cromwell.workflows.pb_ccs_mapping  
• Convert BAM to FASTX:  
cromwell.workflows.pb_bam2fastx  
• Demultiplex Barcodes:  
cromwell.workflows.pb_demux_subreads  
• Demultiplex Barcodes (HiFi Reads Only):  
cromwell.workflows.pb_demux_ccs  
• Export Reads: cromwell.workflows.pb_export_ccs  
• Genome Assembly:  
cromwell.workflows.pb_assembly_hifi  
• HiFiViral SARS CoV-2 Analysis:  
cromwell.workflows.pb_sars_cov2_kit  
• Iso-Seq Analysis:  
cromwell.workflows.pb_isoseq3_ccs only  
• Long Amplicon Analysis: cromwell.workflows.pb_laa  
• Mapping: cromwell.workflows.pb_align_ccs  
• Mapping (HiFi Reads Only):  
cromwell.workflows.pb_ccs_subreads  
• Mark PCR Duplicates:  
cromwell.workflows.pb_mark_duplicates  
• Microbial Assembly:  
cromwell.workflows.pb_assembly_microbial  
• Microbial Assembly (HiFi Reads Only):  
cromwell.workflows.pb_assembly_hifi_microbial  
• Minor Variants Analysis: cromwell.workflows.pb_mv_ccs  
• Site Acceptance Test: cromwell.workflows.pb_sat  
• Structural Variant Calling:  
cromwell.workflows.pb_sv_clr  
• Structural Variant Calling (HiFi Reads Only):  
cromwell.workflows.pb_sv_ccs  
• Trim gDNA Amplification Adapters:  
cromwell.workflows.pb_trim_adapters |
| Analysis Name        | No       | Enter any ASCII string. See Auto Analysis Fields below for details.  
Note: This field is required for Auto Analysis, otherwise the name will be "".  
Example: sample 1 analysis |
CSV File General Requirements

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

Boolean Values

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

**Kit Barcode Requirements**

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

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

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

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

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

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

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

**Auto Analysis Fields**

• The fields include Pipeline ID, Analysis Name, Entry Points, and Task Options.

• You can define **one** analysis for each Collection or Bio Sample. The Pipeline ID, Analysis Name and Entry Points fields are **required** to create an Auto Analysis.

• The analysis name is a concatenation of the values of the Analysis Name and Bio Sample Name fields.

• The Task Options field may be left blank; any task options not specified will use pipeline defaults.
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.

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 filter the list of runs to display, click one or more of the following buttons: Complete, Running, Terminated, Aborted, Error, Paused, and/or Ready. (Click Show to remove or select all the filters.)
4. To export Run QC data in CSV format: Select one or more runs in the table, then click Export Selected.

Table Fields

Note: Not all table fields are shown by default. To see additional table fields, click the > symbol next to a column title.

- **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.
- **Summary**: A description of the run.
• Dates
  – Run Date: The date and time when the run was started.
  – Completion Date: The date and time the run was completed.
  – Transferred Date: The date and time the run results were transferred to the network.
• Created By: The name of the user who created the run.
• Status: The current status of the run. Can be one of the following: Running, Complete, Failed, Terminated, or Unknown.
• Instrument Details
  – Instrument Name: The name of the instrument.
  – Instrument SN: The serial number of the instrument.
  – Instrument SW: The versions of Sequel Instrument Control Software (ICS) installed on the instrument.
• Cells
  – Total: The total number of SMRT Cells used in the run.
  – Completed: The number of SMRT Cells that generated data for the run.
  – Failed: The number of SMRT Cells that failed to generate data during the run.
• Run ID: An internally-generated ID number identifying the run.
• Primary Analysis SW: The version of Primary Analysis software installed on the instrument.
• UUID: Another internally-generated ID number identifying the run.

5. 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.

6. Click the > arrow at the top of the **Consumables** table to see the sample wells used, consumable type, lot number, expiration date, and other information.
Run Settings and Metrics

**Note:** Click **Expand All** to expand all of the table columns. Click **Collapse All** to collapse the table columns.

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

- **Sample Information**
  - **Name:** The sample name, as defined when creating the run. Clicking the name will take you to the corresponding entry in the Data Management module.
  - **Comment:** Sample comment, entered in Run Design.

- **Run Settings**
  - **Movie Time (hrs):** The length of the movie associated with this SMRT Cell.
  - **Loading Concentration (pM):** The on-plate loading concentration, in picomolarity.
  - **Pre-extension Time (hrs):** The pre-extension time used in the collection, if any.
  - **Workflow:** The instrument robotics workflow used for 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.**

- **Total Bases (Gb):** Calculated by multiplying the number of productive (P1) ZMWs by the mean polymerase read length; displayed in Gigabases.

- **Unique Molecular Yield (Gb):** The sum total length of unique single molecules that were sequenced. It is calculated as the sum of per-ZMW median subread lengths.

- **Productivity (%)**
  - **P0:** Empty ZMW; no signal detected.
  - **P1:** ZMW with a high quality read detected.
  - **P2:** Other, signal detected but no high quality read.

- **Reads:** Polymerase reads are trimmed to the High Quality region and include bases from adapters, as well as potentially multiple passes around a SMRTbell template.
  - **HiFi Reads ≥Q20 Reads:** The total number of CCS Reads whose quality value is equal to or greater than 20.
  - **HiFi Reads Yield:** The total yield (in base pairs) of the CCS Reads whose quality value is equal to or greater than 20.
  - **HiFi Reads Mean Length:** The mean read length of the CCS Reads whose quality value is equal to or greater than 20.
  - **HiFi Reads Median QV:** The median number of CCS Reads whose quality value is equal to or greater than 20.
  - **Polymerase Read Length 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 Read Length N50:** 50% of all read bases came from polymerase reads longer than this value.
- **Longest Subread Mean**: The mean subread length, considering only the longest subread from each ZMW.
- **Longest Subread N50**: 50% of all read bases came from subreads longer than this value when considering only the longest subread from each ZMW.

**Control**
- **Poly RL Mean (bp)**: The mean polymerase read length of the control reads.
- **Total Reads**: The number of control reads obtained.
- **Concordance Mean**: The average concordance (agreement) between the control raw reads and the control reference sequence.
- **Concordance Mode**: The median concordance (agreement) between the control raw reads and the control reference sequence.

**Local Base Rate**: The average base incorporation rate, excluding polymerase pausing events.

**Template**
- **Adapter Dimer**: The % of pre-filter ZMWs which have observed inserts of 0-10 bp. These are likely adapter dimers.
- **Short Insert**: The % of pre-filter ZMWs which have observed inserts of 11-100 bp. These are likely short fragment contamination.

7. View plots for each SMRT Cell where data was successfully transferred. Clicking on an individual plot displays an expanded view. These plots include:

**Polymerase Read Length**: Plots the number of reads against the polymerase read length.

**Longest Subread Length**: Plots the number of reads against the insert length.

**Control Polymerase RL**: Displays the Polymerase read length distribution of the control, if used.

**Control Concordance**: Maps control reads against the known control reference and reports the concordance.

**Base Yield Density**: Displays the number of bases sequenced in the collection, according to the length of the read in which they were observed. Values displayed are per unit of read length (i.e. the base yield density) and are averaged over 2000 bp windows to gently smooth the data. Regions of the graph corresponding to bases found in reads longer than the N50 and N95 values are shaded in medium and dark blue, respectively.

**Read Length Density**: Displays a density plot of reads, hexagonally binned according to their HQ Read Length and median subread length. For very large insert libraries, most reads consist of a single subread and will fall along the diagonal. For shorter inserts, subreads will be shorter than the HQ read length, and will appear as horizontal features. This plot is useful for quickly visualizing aspects of library quality, including insert size distributions, reads terminating at adapters, and missing adapters.
• **HiFi Read Length Distribution**: Displays a histogram distribution of HiFi Reads (QV ≥20), other CCS Reads (three or more passes, but QV <20), and other reads, by read length.

• **Read Quality Distribution**: Displays a histogram distribution of HiFi Reads (QV ≥20) and other CCS Reads by read quality.

• **Read Length vs Predicted Accuracy**: Displays a heat map of CCS Read lengths and predicted accuracies. The boundary between HiFi Reads and other CCS Reads is shown as a dashed line at QV 20.
Data Management

Use the Data Management module to:

- Create and manage Data Sets,
- View Data Set information,
- Create and manage Projects,
- View, import, export, 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 35 for details.)

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

For more information on Data Sets, click here.

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, use the Search function. See “Appendix C - Data Search” on page 165 for details.

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

6. Select the type of data to include in the new Data Set:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
   - **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.

The Data Sets table displays the appropriate Data Sets available.
7. (Optional) Specify the Project that this new Data Set will be associated with using the Projects menu (located at the top-right of the Data Management page.) **General Project:** This Data Set will be visible to all SMRT Link users. **All My Projects:** This Data Set will be visible only to users who have access to Projects that you are a member of. **Note:** Selecting a Project also filters the Data Sets that you can use when creating the new Data Set.

8. In the Data Sets table, select one or more sets of sequence data.

9. (Optional) Choose how to view the Data Set table: 1) Tree Mode - A barcoded Data Set displays as one row. 2) Flat Mode - A barcoded Data Set and its demultiplexed subsets display as separate rows.

10. (Optional) Use the Search function to search for specific Data Sets. See “Appendix C - Data Search” on page 165 for details.

11. (Optional) If you selected one Data Set only, click the **Filter Reads by Length** box above the Data Set list. Enter the minimum and/or maximum length to retain in the new Data Set.

12. (Optional) If you selected one Data Set only, click the **Filter Reads by QV≥** box above the Data Set list. Enter the minimum quality value to retain in the new Data Set.

13. Click **Save Data Set**. The new Data Set becomes available for starting analyses, viewing, or generating reports.

14. 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. See “Data Set QC Reports” on page 35 for details.

---

**Viewing Data Set Information**

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

2. Click **View > Data** and select the type of Data Set to view:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
   - **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.

   The Data Sets table displays the appropriate Data Sets available.

3. (Optional) Use the Search function to search for Data Sets. See “Appendix C - Data Search” on page 165 for details.

4. Click the name of the Data Set to see information about the sequence data included in the Data Set, as well as QC reports.

---

**Copying a Data Set**

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

2. Click **View > Data** and select the type of data to copy:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
HiFi Reads: Reads generated with CCS Analysis whose quality value is equal to or greater than 20. The Data Sets table displays the appropriate Data Sets available.

3. (Optional) Use the Search function to search for Data Sets. See “Appendix C - Data Search” on page 165 for details.

4. Click the name of the Data Set to copy. The Data Set Reports page displays.

5. Click Copy. The main Data Management page displays; the new Data Set has (copy) appended to the name.

Deleting a Data Set

Note: SMRT Link’s Delete Data Set functionality deletes the Data Set from the SMRT Link interface only, 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. Click View > Data and select the type of data to delete:
   - Continuous Long Reads: Subreads from Sequel Systems.
   - HiFi Reads: Reads generated with CCS Analysis whose quality value is equal to or greater than 20. The Data Sets table displays the appropriate Data Sets available.

3. (Optional) Use the Search function to search for Data Sets. See “Appendix C - Data Search” on page 165 for details.

4. Click the name of the Data Set to delete.

5. Click Delete. Note that this deletes the Data Set from the SMRT Link interface only; 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.

Starting an Analysis from a Data Set

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

1. Click Analyze..., then name the analysis and click Next.

2. Follow the instructions starting at Step 12 of “Creating and Starting an Analysis” on page 42.

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:
Data Set Overview > Status

Displays the following information about the Data Set:

- The Data Set Name, ID, description, and when it was created and updated.
- The number of subreads and their total length in base pairs.
- The names of the run and instrument that generated the data.
- The biological sample name and well sample names of the sample used to generate the data.
- Path to the location on your cluster where the data is stored, which can be used for command-line navigation. For information on command-line usage, see SMRT® Tools Reference Guide (v10.2).

Completed Analyses

Lists all completed analyses that used the Data Set as input. To view details about a specific analysis, click its name.

Raw Data Report > Summary Metrics

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

Information on loading, control reads, and adapters is also displayed. Other information may display based on the Data Set type.
What is a Project?

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

Projects include:

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

Data Sets and Projects

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

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.
   – (Optional) Use the Search function to search for Data Sets. See “Appendix C - Data Search” on page 165 for details.
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, using the controls in the **Members** section:
   – **Access 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.
   – **Access for Individual SMRT Link Users**: Enter a user name and click **Search By Name**. Choose Owner, View, or View/Edit, then click **Add Selected User**.
   – **Notes**: A) Projects can have multiple owners. B) If you enable all SMRT Link users to have View/Edit access, you cannot change an individual member's access to View.
8. Click **Save**. The new project becomes available for SMRT Link users who now have access.

**Editing a Project**
1. On the Home Page, select **Data Management**.
2. Click **View > Projects**.
3. Projects can be sorted and searched for:
   – To sort Projects: Click a **column title**.
   – To search for a Project, use the Search function. See “Appendix C - Data Search” on page 165 for details.
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: Click **X**.
   – (Optional) Add one or more sets of sequence data to the project:
     Click **Select Data Sets** and select one or more Data Sets to add.
   – (Optional) Delete members: Click **X** next to a Project member's name to delete that user from access to the Project.
   – (Optional) Add members to the Project: See Step 7 in Creating a Project.
5. 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**

1. On the Home Page, select **Data Management**.
2. Click **View > Data**, then choose the type of data to view or delete:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
   - **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.
   - **Barcodes**: Barcodes from barcoded samples.
   - **References**: Reference sequence FASTA files used when creating certain analyses.
3. **(Optional)** Use the Search function to search for specific Data Sets, barcode files or reference sequence files. See “Appendix C - Data Search” on page 165 for details.
4. Click the name of the sequence, reference or barcode file of interest. Details for that sequence, reference sequence file or barcode file display.
5. **(Optional)** To delete the sequence data, reference sequence, or barcode file, click **Delete**.

**Note**: The **Copy** button is available for Continuous Long Reads and HiFi Reads, but **not** for Reference and Barcode data.

---

**Importing Sequence, Reference and Barcode Data**

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

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

1. On the Home Page, select **Data Management**.
2. Click **Import Data**.
3. Specify whether to import data from the **SMRT Link Server**, or from a **Local File System**. (**Note**: **Only** references and barcodes are available if you select **Local File System**.)
4. Select the data type to import:
   - **Continuous Long Reads**: XML file (.subreadset.xml) or ZIP file containing information about subreads from Sequel Systems, such as paths to the BAM files. Use only ZIP files created by SMRT Link.
   - **HiFi Reads**: XML file (.consensusreadset.xml) or ZIP file containing information about HiFi Reads (reads generated with CCS Analysis whose quality value is equal to or greater than 20.) Use only ZIP files created by SMRT Link.
   - **Barcodes**: FASTA (.fa or .fasta), XML (.barcodeset.xml), or ZIP files containing barcodes.
   - **References**: FASTA (.fa or .fasta), XML (.referenceSet.xml), or ZIP files containing a reference sequence for use in starting analyses. (Note: If importing from a local system, Reference files must be smaller than 15 MB.)
   - **Note**: FASTA files imported into SMRT Link must not contain empty lines or non-alphanumeric characters. The file name must not start with a number. For information about the file types listed here, click here.

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

**Exporting Sequence, Reference and Barcode Data**

Two types of sequence data (HiFi Reads and Continuous Long Reads) can be exported, as well as barcode files and reference files.

1. On the Home Page, select **Data Management**.
2. Click **Export Data**.
3. Select the type of data to export:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
   - **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.
   - **Barcodes**: Files containing barcodes.
   - **References**: Files containing a reference sequence for use in starting analyses.
4. (Optional) Use the Search function to search for Data Sets, barcode files, or reference files. See “Appendix C - Data Search” on page 165 for details.
5. Select one or more sets of data to export. (Multiple data files are combined as one ZIP file for export.)
6. Click **Export Selected**.

![Dataset Export](image)

7. Navigate to the export destination directory.
8. (Optional) If exporting Data Sets, 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. (Optional) If exporting Data Sets, click **Export PDF Reports** to create PDF files containing comprehensive information about the Data Set(s). Each PDF report contains extensive information about one Data Set, including loading statistics, run set up and QC information, analysis parameters and results including charts and histograms, and lists of the output files generated, all in one convenient document.
10. Click **Export**.
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, use the Search function. See “Appendix C - Data Search” on page 165 for details.
   – To filter the list of analyses based on their state: Click the funnel in the State column header, then click one or more of the categories of interest: Select All, Created, Running, Submitted, Terminated, Successful, Failed, or Aborted.
• To filter the list of analyses based on the Project(s) that they are associated with: Click the Projects menu (located at the top-right of the main SMRT Analysis page) and select a Project. See “What is a Project?” on page 37 for details.

4. Click + Create New Analysis.

5. (Optional) Click Copy From..., choose an analysis whose settings you wish to reuse, then click Select. The analysis name and the Data Type are filled in. Go to Step 9 to select Data Set(s).

6. Enter a name for the analysis.

7. Select the type of data to use for the analysis:
   – Continuous Long Reads: Subreads from Sequel Systems.
   – HiFi Reads: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.

8. (Optional) Specify the Project that this analysis will be associated with using the Projects menu (located at the top-right of the SMRT Analysis page.) General Project: 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. To restrict access to an analysis, make sure to select a project limited to the appropriate users before starting the analysis.

   Note: Selecting a Project also filters the Data Sets that you can use when creating the analysis.

9. In the Data Sets table, select one or more sets of data to be analyzed.
   – (Optional) Use the Search function to search for Data Sets. See “Appendix C - Data Search” on page 165 for details.)
   – (Optional) Choose how to view the Data Set table: 1) Tree Mode - A barcoded Data Set displays as one row. 2) Flat Mode - A barcoded Data Set and its demultiplexed subsets display as separate rows.
   – (Optional) For Data Sets that include demultiplexed subsets, you can also select individual subsets as part of your selection. To do so:

   A) Click the Demultiplexed Subsets number link:

   ![Data Sets Table]

   B) Select one or more subsets, then click Back:
C) Click the list image to view or edit the full Data Set selection. (The small blue number specifies how many Data Sets and/or subsets were selected):

**Note:** For information on the Auto Analysis feature, see “Automated Analysis” on page 143 for details.

10. If you selected **multiple** Data Sets as input for the analysis, additional options become available:
– **One Analysis for All Data Sets**: Runs one analysis using all the selected Data Sets as input, for a maximum of 30 Data Sets.

– **One Analysis per Data Set - Identical Parameters**: Runs one separate analysis for each of the selected Data Sets, using the same parameters, for a maximum of 10,000 Data Sets. Later in the process, optionally click **Advanced 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, for a maximum of 16 Data Sets. Later in the process, click **Advanced Parameters** and modify parameters. Then click **Start and Create Next**. You can then specify parameters for each of the included Data Sets.

– **Note**: The number of Data Sets listed is based on testing using PacBio's suggested compute configuration, listed in **SMRT Link Software Installation (v10.2)**.

11. Click **Next**.
12. Select a secondary analysis application from the dropdown list. (Different applications display based on your choice of Data Type in Step 7. See "PacBio® Secondary Analysis Applications" on page 51 for details.)

---

Each of the secondary analysis applications has **required parameters** that are displayed. Please review the default values shown.

Secondary analysis applications also have **advanced parameters**. These are set to default values, and need only be changed when analyzing data generated in non-standard experimental conditions.

13. (Optional) Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.

The **Iso-Seq** application will be used as an example. This application characterizes full-length transcript isoforms.
14. Click the **Reference Set** 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 Sequence, Reference and Barcode Data” on page 39 for details.)

15. **Optional** Click **Advanced Parameters** and specify the values of the parameters you would like to change. Click **OK** when finished. (Different applications have different advanced parameters.)

   - To see information about parameters for all secondary analysis applications provided by Pacific Biosciences, see “PacBio® Secondary Analysis Applications” on page 51.

16. **Optional** Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

17. **Optional** Click **Back** if you need to change any of the analysis attributes selected in Step 7.

18. Click **Start** to submit the analysis. (If you selected multiple Data Sets as input, click **Start Multiple Jobs** or **Start and Create Next**.)

19. Select **SMRT Analysis** from the Module Menu 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.

20. **(Optional)** To delete the completed analysis: Click **Delete**, then click **Yes** in the confirmation dialog. The analysis is deleted from both the SMRT Link interface and from the server.

---

**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 > Data** and select the type of Data Set to use:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
   - **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.

   The Data Sets table displays the appropriate Data Sets available.

3. **(Optional)** Use the Search function to search for Data Sets. See “Appendix C - Data Search” on page 165 for details.

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 12 of “Creating and Starting an Analysis” on page 42.
Canceling a Running Analysis

1. On the Home Page, select **SMRT Analysis**.
2. Click the funnel in the **State** column header, then click **Running**. This displays only currently-running analyses.
3. Select a currently-running analysis to cancel.
4. Click **Cancel**.
5. Click **Yes** in the confirmation dialog. The cancelled analysis displays as **Terminated**.

Restarting a Failed Analysis

You can **restart** a failed analysis; the execution speed from the start to the original point of failure is very fast, which can save time and computing resources. The restarted analysis may run to completion, depending on the source of failure.

- **Note**: As the restarted analysis uses information from the original failed analysis, do not delete the original analysis results.

If viewing the results page for the failed analysis: Click **Restart**.

If **not** viewing the results page for the failed analysis:

1. On the Home Page, select **SMRT Analysis**.
2. Click the funnel in the **State** column header, then click **Failed**. This displays only failed analyses.
3. Select a failed analysis to restart.
4. Click **Restart**.

Viewing Analysis Results

1. On the Home Page, select **SMRT Analysis**. You see a list of all analyses.
2. **(Optional)** Click the funnel in the **State** column header, then click **Successful**. This displays only successfully-completed analyses.

3. **(Optional)** Use the Search function to search for specific analyses. See “Appendix C - Data Search” on page 165 for details.

4. Click the analysis link of interest.

5. Click **Analysis Overview > Status** to see analysis information status, including which application was used for the analysis, and the inputs used.

6. Click Analysis **Overview > Thumbnails** or **Display All** to view thumbnails of the reports generated for the analysis. Click the link under 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 51.

8. To download data files created by SMRT Link: You can use these data files as input for further 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. **(Optional)** Specify prefix(es) used in the names of files generated by the analysis. Example: **Run Name** can be included in the name of every file generated by the analysis. Click **Edit Output File Name Prefix**, check the type(s) of information to append to the file names, then click **Save**.

10. To view analysis log details: Click **Data > SMRT Link Log**.

11. To visualize the secondary analysis results: See “Visualizing Data Using IGV” on page 146 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 funnel in the **State** column header, then click **Successful**. This displays only successfully-completed analyses.

3. **(Optional)** Use the Search function to search for specific analyses. See “Appendix C - Data Search” on page 165 for details.

4. Click the analysis link of interest.

5. Click **Copy** - this creates a copy of the analysis, named **Copy of <analysis name>**, using the **same** parameters.

6. Edit the name of the analysis.

7. Click **Next**.

8. **(Optional)** Edit any other parameters. See “PacBio® Secondary Analysis Applications” on page 51 for further details.
9. 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. **(Optional)** Use the Search function to search for specific analyses. See “Appendix C - Data Search” on page 165 for details.
4. Select one or more analyses to export. This exports the entire contents of the analysis directory.
5. Click **Export Selected Analyses**.
6. 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 v10.2. Each application is described later in this document, including all analysis parameters, reports and output files generated by the application.

**Note:** The Resequencing application is discontinued in this release; use the Mapping application instead. (See “Mapping Application” on page 105 for details.)

**Assembly (HGAP 4)**
- Generate *de novo* assemblies of genomes using Continuous Long Reads.
- See “Assembly (HGAP 4) Application” on page 54 for details.

**Base Modification Analysis**
- Identify common bacterial base modification (6mA, 4mC).
- Optionally analyze the methyltransferase recognition motifs.
- See “Base Modification Analysis Application” on page 60 for details.

**CCS with Demultiplexing**
- Generate consensus sequences from single molecules, then separate reads by barcodes.
- See “CCS with Demultiplexing Application” on page 65 for details.

**CCS with Mapping**
- Generate consensus sequences from single molecules, and map these consensus sequences to a user-provided reference sequence.
- See “CCS with Mapping Application” on page 71 for details.

**Circular Consensus Sequencing (CCS)**
- Identify consensus sequences for single molecules.
- See “Circular Consensus Sequencing (CCS) Application” on page 76 for details.

**Convert BAM to FASTX**
- Convert sequence data in BAM file format to the FASTA and FASTQ file formats.
- 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 79 for details.

**Demultiplex Barcodes**
- Separate reads by barcode.
- See “Demultiplex Barcodes Application” on page 80 for details.
Export Reads
- Export HiFi Reads that pass filtering criteria as FASTA, FASTQ and BAM files.
- For barcoded runs, you must first run the Demultiplex Barcodes application to create BAM files before using this application.
- See “Export Reads Application” on page 85 for details.

Genome Assembly
- Generate de novo assemblies of genomes, using HiFi Reads.
- See “Genome Assembly Application” on page 87 for details.

HiFiViral SARS-CoV-2 Analysis Application
- Analyze multiplexed viral surveillance samples for SARS-CoV-2, using HiFi Reads.
- See “HiFiViral SARS-CoV-2 Analysis Application” on page 90 for details.

Iso-Seq® Analysis
- Characterize full-length transcript isoforms, using HiFi Reads.
- See “Iso-Seq® Analysis Application” on page 96 for details.

Long Amplicon Analysis (LAA)
- Identify phased consensus sequences from a heterogeneous pool of amplicons.
- See “Long Amplicon Analysis (LAA) Application” on page 102 for details.

Mapping
- Align (or map) reads to a user-provided reference sequence.
- See “Mapping Application” on page 105 for details.

Mark PCR Duplicates
- Remove duplicate reads from a HiFi Reads Data Set created using an ultra-low DNA sequencing protocol.
- See “Mark PCR Duplicates Application” on page 110 for details.

Microbial Assembly
- Generate de novo assemblies of small prokaryotic genomes between 1.9-10 Mb and companion plasmids between 2 – 220 kb.
- See “Microbial Assembly Application” on page 112 for details.

Minor Variants Analysis
- Identify and phase minor single nucleotide substitution variants in complex populations.
- See “Minor Variants Analysis Application” on page 119 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 125 for details.

**Structural Variant Calling**

- Identify structural variants (Default: ≥20 bp) in a sample or set of samples relative to a reference.
- See “Structural Variant Calling Application” on page 129 for details.

**Trim gDNA Amplification Adapters**

- Trim PCR Adapters from a HiFi Reads Data Set created using an ultra-low DNA sequencing library.
- See “Trim gDNA Amplification Adapters Application” on page 133 for details.
Assembly (HGAP 4) Application

Use this application (Hierarchical Genome Assembly Process) to generate high quality \textit{de novo} assemblies of genomes, using Continuous Long Reads.

- The application accepts \textit{Continuous Long Reads} (BAM format) as input.
- HGAP 4 includes pre-assembly, \textit{de novo} assembly and assembly polishing steps.
- HGAP 4 uses Falcon for \textit{de novo} assembly and Arrow for polishing.

Importing/Exporting Analysis Settings

- Click \textbf{Import Analysis Settings} and select a previously-saved CSV file containing the desired settings (including \textbf{Advanced Parameters}) for the selected application. The imported application settings are set.
- Click \textbf{Export} to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

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

- The approximate number of base pairs expected in the genome, used to determine the coverage cutoff. Other parameters are set automatically based on this value.

Consolidate Mapped BAMs for IGV (Default = OFF)

- By default, SMRT Link consolidates chunked BAM files for viewing in IGV if the combined size is not more than 10 GB. Setting this option to \textbf{ON} ignores the file size cutoff and consolidates the BAM files.
- \textbf{Note}: This setting can double the amount of storage used by the BAM files, which can be considerable. Make sure to have enough disk space available. This setting may also result in longer run times.

Parameters

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive Mode</td>
<td>OFF</td>
<td>If \textbf{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>Seed Length Cutoff</td>
<td>-1</td>
<td>Only reads as long as this value will be used as seeds in the draft assembly. \texttt{-1} means this will be calculated automatically so that the total number of seed bases equals (Genome Length times Seed Coverage).</td>
</tr>
</tbody>
</table>
Reports and Data Files

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

**Polished Assembly > Summary Metrics**
Displays statistics on the contigs from the *de novo* assembly that were corrected by Arrow.

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

**Polished Assembly > Contig Coverage vs Confidence**

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

**Preassembly > Summary Metrics**

Displays statistics on the pre-assembly process.

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

• **Number of Pre-Assembled Bases (total):** The total number of bases output by the pre-assembler.

• **Pre-Assembled Coverage (bases/genome_size):** The number of bases output by the pre-assembler divided by the number of base pairs expected in the genome.

• **Pre-Assembled Yield (bases/seed_bases):** The percentage of seed read bases that were successfully aligned to generate pre-assembled reads.

• **Average Number of Reads that Each Seed is Broken Into:** The average number of preliminary reads that each seed is broken into. (Preliminary reads are derived from seeds using error correction; some portions of seeds might be too "noisy" to use.)

• **Average Number of Bases Lost from Each Seed:** The average number of bases from each seed that were completely discarded.

---

**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 that has zero coverage.

---

**Coverage > Coverage Across Reference**

• Maps coverage across the reference.

---

**Coverage > Depth of Coverage**

• Histogram distribution of the reference regions by the coverage.

---

**Coverage > Coverage vs. [GC] Content**

• Maps (as a percentage, over a 100 bp window) the number of Gs and Cs present across the coverage. The number of genomic windows with the corresponding % of Gs and Cs is displayed on top. Used to check that no coverage is lost over extremely biased base compositions.

---

**Alignment to Draft Assembly > Summary Metrics**

Displays statistics on reads that aligned to the draft assembly.

• **Percent Mapped Bases:** The percentage of bases that mapped to the draft assembly.

• **Number of Subreads (total):** The total number of subreads in the draft assembly.

• **Number of Subreads (mapped):** The number of subreads that mapped to the draft assembly.

• **Number of Subreads (unmapped):** The number of subreads not mapped to the draft assembly.

• **Percentage of Subreads (mapped):** The percentage of subreads that mapped to the draft assembly.

• **Percentage of Subreads (unmapped):** The percentage of subreads not mapped to the draft assembly.
• **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the draft assembly. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

• **Number of Subread Bases (mapped):** The number of subread bases that mapped to the draft assembly.

• **Number of Alignments:** The number of alignments that mapped to the draft assembly.

• **Alignment Length Mean (mapped):** The mean length of alignments that mapped to the draft assembly.

• **Alignment Length N50 (mapped):** The alignment length at which 50% of the alignments are longer than, or equal to, this value.

• **Alignment Length 95% (mapped):** The 95th percentile of length of alignments that mapped to the draft assembly.

• **Alignment Length Max (mapped):** The maximum length of alignments that mapped to the draft assembly.

• **Number of Polymerase Reads (mapped):** The number of polymerase reads that mapped to the draft assembly. This includes adapters.

• **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped 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 (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 draft assembly.

• **Polymerase Read Length Max (mapped):** The maximum length of polymerase reads that mapped to the draft assembly.

**Alignment to Draft Assembly > Alignment Statistics Summary**

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

• **Sample:** Sample name for which the following metrics apply.

• **Movie:** Movie name for which the following metrics apply.

• **Number of Polymerase Reads (mapped):** The number of polymerase reads that mapped to the draft assembly. This includes adapters.

• **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped 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 (mapped):** The read length at which 50% of the bases mapped to the draft assembly are in polymerase reads longer than, or equal to, this value.

• **Number of Subreads (mapped):** The number of subreads that mapped to the draft assembly.

• **Number of Subread Bases (mapped):** The number of subread bases that mapped to the draft assembly.

• **Subread Length Mean (mapped):** The mean length of the mapped portion of subreads that mapped to the draft assembly.

• **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the draft assembly. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

**Alignment to Draft Assembly > Mapped Polymerase Read Length**

• Histogram distribution of the number of reads by read length.
Alignment to Draft Assembly > Mapped Subread Length

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

Alignment to Draft Assembly > Mapped Subread Concordance

- Histogram distribution of the number of subreads against the percent concordance with the subreads that mapped to the draft assembly. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Alignment to Draft Assembly > Mapped Concordance vs Read Length

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

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **Mapped Reads**: All input reads that were mapped to the reference by the application.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference sequence.
- **Polished Assembly**: The final polished assembly, in Data Set, FASTA and FASTQ formats. This is the most important output file.

Data > IGV Visualization Files

The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.

- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.
Use this application to identify common bacterial base modifications (6mA, 4mC), and then optionally analyze the methyltransferase recognition motifs. Detection can use an in-silico control consisting of expected kinetic signals.

- The application accepts **Continuous Long Reads** (BAM format) as input.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Reference Set (Required)**

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

**Find Modified Base Motifs (Default = OFF)**

- Performs motif detection on the results of the Base Modification analysis.

**Consolidate Mapped BAMs for IGV (Default = OFF)**

- By default, SMRT Link consolidates chunked BAM files for viewing in IGV if the combined size is not more than 10 GB. Setting this option to **ON ignores** the file size cutoff and consolidates the BAM files.
- **Note**: This setting can **double** the amount of storage used by the BAM files, which can be considerable. Make sure to have enough disk space available. This setting may also result in longer run times.

### Parameters

<table>
<thead>
<tr>
<th>Advanced 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 semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Minimum Gap-Compressed Identity (%)</td>
<td>70</td>
<td>The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.</td>
</tr>
<tr>
<td>Minimum Mapped Length (bp)</td>
<td>50</td>
<td>The minimum required alignment length, in base pairs.</td>
</tr>
<tr>
<td>Compute Methyl Fraction (experimental)</td>
<td>OFF</td>
<td>When identifying specific modifications (6mA and/or 4mC), enabling this option estimates the methylated fraction, along with 95% confidence interval bounds.</td>
</tr>
<tr>
<td>Minimum Methylated Fraction</td>
<td>0.3</td>
<td>The minimum methylated fraction to identify a motif.</td>
</tr>
<tr>
<td>P-Value</td>
<td>0.001</td>
<td>The probability value cutoff for detecting base modifications.</td>
</tr>
<tr>
<td>Minimum Qmod Score</td>
<td>100</td>
<td>The minimum QMod score used to identify a motif.</td>
</tr>
</tbody>
</table>
Reports and Data Files

The Base Modification Detection application generates the following reports:

Base Modifications > Kinetic Detections

- **Per-Base Kinetic Detections**: Maps the modification QV against per-strand coverage.
- **Kinetic Detections Histogram**: Histogram distribution of the number of bases by 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 0. 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, 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 
  \[(\text{fraction methylated}) \times (\text{sum of log-p values of matches})\].

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced pbmm2 Options</td>
<td>NONE</td>
<td>Space-separated list of custom pbmm2 options. Not all supported command-line options can be used, and HPC settings cannot be modified. See SMRT® Tools Reference Guide v10.2 for details.</td>
</tr>
<tr>
<td>Bio Sample Name of Aligned Dataset</td>
<td>NONE</td>
<td>Populates the Bio Sample Name (Read Group SM tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. Note: Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>
Modified Base Motifs > Modification QVs
• Maps motif sites against Modification QV for all genomic occurrences of a motif, for each reported motif, including “No Motif”.

Modified Base Motifs > ModQV Versus Coverage by Motif
• Maps coverage against Modification QV for all genomic occurrences of a motif, for each reported motif.

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 that has zero coverage.

Coverage > Coverage across Reference
• Maps coverage across the reference.

Coverage > Depth of Coverage
• Histogram distribution of the reference regions by coverage.

Coverage > Coverage vs. [GC] Content
• Maps (as a percentage, over a 100 bp window) the number of Gs and Cs present across the coverage. The number of genomic windows with the corresponding % of Gs and Cs is displayed on top. Used to check that no coverage is lost over extremely biased base compositions.

Mapping Report > Summary Metrics
Mapping is local alignment of a read or subread to a reference sequence.

• **Percentage of Bases (mapped)**: The percentage of bases that mapped to the reference sequence.
• **Number of Subreads (total)**: The total number of subreads in the sequence.
• **Number of Subreads (mapped)**: The number of subreads that mapped to the reference sequence.
• **Number of Subreads (unmapped)**: The number of subreads not mapped to the reference sequence.
• **Percentage of Subreads (mapped)**: The percentage of subreads that mapped to the reference sequence.
• **Percentage of Subreads (unmapped)**: The percentage of subreads not mapped to the reference sequence.
• **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
• **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
• **Number of Alignments**: The number of alignments that mapped to the reference sequence.
• **Alignment Length Mean (mapped)**: The mean length of alignments that mapped to the reference sequence.
• **Alignment Length N50 (mapped)**: The alignment length at which 50% of the alignments are longer than, or equal to, this value.
• **Alignment Length 95% (mapped)**: The 95th percentile of length of alignments that mapped to the reference sequence.
• **Alignment Length Max (mapped)**: The maximum length of alignments 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.

• **Sample**: Sample name for which the following metrics apply.
• **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 the subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Mapping Report > Mapped Polymerase Read Length
• Histogram distribution of the number of reads by read length.

Mapping Report > Alignment Length
• Histogram distribution of the number of alignments by the alignment length.
Mapping Report > Alignment Concordance

• Histogram distribution of the number of alignments by the percent concordance with the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Mapping Report > Mapped Concordance vs Alignment Length

• Maps the percent concordance with the reference sequence against the alignment length, in base pairs.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **Mapped Reads**: All input reads that were mapped to the reference by the application.
- **Per-Base IPDs for IGV**: BigWig file containing encoded per-base IPD ratios.
- **Modifications**: Duplicate of the modification summary file.
- **Modified Base Motifs**: CSV file containing statistics for the methyltransferase recognition motifs detected.
- **Per-Base Kinetics**: CSV file containing per-base information.
- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.
- **Modified Bases**: GFF file listing every detected modified base in the genome.
- **Motif Annotations**: GFF file listing every modified nucleotide sequence motif in the genome.

Data > IGV Visualization Files

The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.

- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.
- **Per-Base IPDs for IGV**: BigWig file containing encoded per-base IPD ratios.
CCS with Demultiplexing Application

Use this application to first generate consensus sequences from single molecules, then to separate reads by barcodes.

- The application accepts **Continuous Long Reads** (BAM format) as input.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Barcode Set (Required)**

- Specify a barcode sequence file to separate the reads.

**Same Barcodes on Both Ends of Sequence (Default = Yes)**

- Specify **Yes** 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 135 for information on barcode designs.)
- Specify **No** to specify asymmetric designs where the barcodes are different on each end of the insert.

**Assign Bio Sample Names to Barcodes (Required)**

SMRT Link automatically creates a CSV-format **Autofilled Barcoded Sample 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 column of automatically-generated Bio Sample Names 1 through N, corresponding to barcodes 1 through N, for the biological sample names. There are **two different ways** to specify which barcodes to use, and assign biological sample names to barcodes:

**Interactively:**

1. Click **Interactively**, then drag barcodes from the **Available Barcodes** column to the **Included Barcodes** column. (Use the checkboxes to select multiple barcodes.)
2. (Optional) Click a Bio Sample field to edit the Bio Sample Name associated with a barcode. **Note:** Avoid using spaces in Bio Sample Names as they may lead to third-party compatibility issues.
3. (Optional) Click **Download as a file for later use**.
4. Click **Save** to save the edited barcodes/Bio Sample names. You see **Success** on the line below, assuming the file is formatted correctly.

**From a file:**

1. Click **From a File**, then click **Download File**. Edit the file and enter the biological sample names associated with the barcodes in the second
column, then save the file. Use alphanumeric characters, spaces (allowed but not recommended for compatibility with third-party downstream software), hyphens, underscores, colons, or periods only - other characters will be removed automatically, with a maximum of 40 characters. If you did not use all barcodes in the Autofilled Barcode Name file in the sequencing run, delete those rows.

- **Note**: Open the CSV file in a text editor and check that the columns are separated by commas, not semicolons or tabs.

2. Select the **Barcoded Sample File** you just edited. You see **Success** on the line below, assuming the file is formatted correctly.

**Demultiplexed Output Data Set Name**

- Specify the name for the new demultiplexed Data Set that will display in SMRT Link. The application creates a copy of the input Data Set, renames it to the name specified, and creates demultiplexed child Data Sets linked to it. The input Data Set remains separate and unmodified.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Min. CCS Predicted Accuracy (Phred Scale)</strong></td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
<tr>
<td><strong>Minimum CCS Read Length</strong></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><strong>Maximum CCS Read Length</strong></td>
<td>50,000</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><strong>Write Unbarcoded Reads</strong></td>
<td>ON</td>
<td>If specified, write out a separate Data Set containing reads that are not barcoded.</td>
</tr>
<tr>
<td><strong>Minimum Barcode Score</strong></td>
<td>80</td>
<td>A barcode score measures the alignment between a barcode attached to a read and an ideal barcode sequence, and is an indicator of how well the chosen barcode pair matches. It ranges between 0 (no match) and 100 (a perfect match). Specifies that reads with barcode scores below this minimum value are not included in analysis. This affects the output BAM file and the output demultiplexed Data Set XML file.</td>
</tr>
<tr>
<td><strong>Process All Reads</strong></td>
<td>ON</td>
<td>Specifies behavior identical to on-instrument CCS Reads generation, overriding all other cutoffs. This setting writes a CCS Read for every ZMW in the input Data Set. Set to OFF to specify more restrictive settings.</td>
</tr>
<tr>
<td><strong>Include Kinetics Information with CCS Analysis Output</strong></td>
<td>OFF</td>
<td>(Sequel IIe System only) If ON, include kinetics per-base data required for methylation DNA analysis. <strong>Note</strong>: This results in a BAM file that is 3-4 times larger. This option applies only when Process All Reads is set to ON.</td>
</tr>
<tr>
<td><strong>Minimum Predicted Accuracy (Deprecated)</strong></td>
<td>0</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.) <strong>Note</strong>: This setting is ignored if the Process All Reads advanced parameter is set to ON.</td>
</tr>
</tbody>
</table>
Reports and Data Files

The CCS with Demultiplexing application generates the following reports:

Barcodes > Summary Metrics

- **Unique Barcodes**: The number of unique barcodes in the sequence data.
- **Barcoded Reads**: The number of correctly-barcoded reads in the sequence data.
- **Mean Reads**: The mean number of reads per barcode combination.
- **Max. Reads**: The maximum number of reads per barcode combination.
- **Min. Reads**: The minimum number of reads per barcode combination.
- **Mean Read Length**: The mean read length of reads per barcode combination.
- **Unbarcoded Reads**: The number of reads without barcodes in the sequence data.
- **Percent Bases in Barcoded Reads**: The percentage of bases in reads in the sequence data that contain barcodes.
- **Percent Barcoded Reads**: The percentage of reads in the sequence data that contain barcodes.

Barcodes > Barcode Data

- **Bio Sample Name**: The name of the biological sample associated with the barcode combination.
- **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 combination.
- **Bases**: The number of bases associated with the barcode.
- **Mean Barcode Quality**: The mean barcode quality associated with the barcode combination.

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 far from the center. Check the output file Inferred Barcodes, note the correct barcodes used, and consider reanalyzing the multiplexed samples with the correct Bio Sample names for the barcodes actually used. If you reanalyze the data, ensure that the Barcode Name file includes only the correct barcodes used.
• **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. This can indicate use of incorrect barcodes. Check the output file Inferred Barcodes, note the correct barcodes used, and consider reanalyzing the multiplexed samples with the correct Bio Sample names for the barcodes actually used. If you reanalyze the data, ensure that the **Barcode Name** file includes only the correct barcodes used.

• **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 <60 suggests issues with some of the barcode analyses. The red line is set at 80 – the minimum default barcode score.
  
  – **Good performance**: HiFi demultiplexing runs should have >90% of reads with barcode quality score ≥95.
  
  – **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.

**CCS Analysis Report > Summary Metrics**

**Note:** CCS Reads with quality value equal to or greater than 20 are called HiFi Reads.

- **HiFi Reads:** The total number of CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Yield (bp):** The total yield (in base pairs) of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Read Length (mean, bp):** The mean read length of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Read Quality (median):** The median number of CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Number of Passes (mean):** The mean number of passes used to generate CCS Reads whose quality value is equal to or greater than 20.

**CCS Analysis Report > Read Length Distribution**

- Histogram distribution of HiFi Reads by read length.

**CCS Analysis Report > Number of Passes**

- Histogram of the number of complete subreads in CCS Reads, broken down by read type (HiFi Reads, other CCS Reads, other reads.)

**CCS Analysis Report > Read Quality Distribution**

- Histogram distribution of the CCS Reads by the read quality.

**CCS Analysis Report > Predicted Accuracy vs. Read Length**

- Heat map of CCS Read lengths and predicted accuracies.

**Data > File Downloads**

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

- **Analysis Log:** Log information for the analysis execution.
- **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **All Barcodes (FASTQ):** All barcoded reads, in FASTQ format.
- **CCS Analysis Per-Read Details:** Summary of CCS Analysis performance and yield.
- **Barcode Files:** Barcoded subread Data Sets; one file per barcode.
- **Barcode Summary CSV:** Data displayed in the reports, in CSV format. This includes Bio Sample Name.
- **Barcode Summary:** Text file listing how many ZMWs were filtered, how many ZMWs are the same or different, and how many reads were filtered.
- **Inferred Barcodes:** Inferred barcodes used in the analysis. The barcoding algorithm looks at the first 35,000 ZMWs, then selects barcodes with ≥10 counts and mean scores ≥45.
- **Unbarcoded Reads:** BAM file containing reads not associated with a barcode.
• **demultiplex.<barcode>.hifi.reads.fastq.gz**: Gzipped HiFi Reads in FASTQ format, one file per barcode.

• **All Reads (BAM)**: BAM file containing one CCS Read per ZMW, including the following types of reads:
  – HiFi Reads (Q20 or higher)
  – Lower-quality but still polished consensus reads (Q1-Q20)
  – Unpolished consensus reads (RQ=-1)
  – 0- or 1-pass subreads unaltered (RQ=-1)
Use this application to generate consensus sequences from single molecules, and map these consensus sequences to a user-provided reference sequence.

The CCS with Mapping application:

- Accepts Continuous Long Reads (BAM format) as input.
- 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 with Mapping uses 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.

**Note:** If the default CCS Analysis parameters are used, this pipeline will produce HiFi Reads and use them for mapping.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Reference Set (Required)**

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

**Consolidate Mapped BAMs for IGV (Default = OFF)**

- By default, SMRT Link consolidates chunked BAM files for viewing in IGV if the combined size is not more than 10 GB. Setting this option to ON ignores the file size cutoff and consolidates the BAM files.
- **Note:** This setting can double the amount of storage used by the BAM files, which can be considerable. Make sure to have enough disk space available. This setting may also result in longer run times.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
</tbody>
</table>
The CCS with Mapping application generates the following reports:

### Target Regions > Target Regions
Displays the number (and percentage) of reads that hit target regions specified by an input BED file. This is useful for targeted DNA sequencing applications. (This report displays only if a BED file is specified when creating the analysis.)

- **Coordinates**: The chromosome coordinates, as specified in the input BED file.
- **Region**: The name of the region, as specified in the input BED file.

---

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum CCS Read 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>Maximum CCS Read Length</td>
<td>50,000</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 Mapped Length (bp)</td>
<td>50</td>
<td>The minimum mapped read length, in base pairs.</td>
</tr>
<tr>
<td>Advanced pbmm2 Options</td>
<td>NONE</td>
<td>Space-separated list of custom pbmm2 options. Not all supported command-line options can be used, and HPC settings cannot be modified. See SMRT® Tools Reference Guide v10.2 for details.</td>
</tr>
<tr>
<td>Bio Sample Name of Aligned Dataset</td>
<td>NONE</td>
<td>Populates the Bio Sample Name (Read Group SM tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. <strong>Note</strong>: Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.</td>
</tr>
<tr>
<td>Minimum Gap-Compressed Identity (%)</td>
<td>70</td>
<td>The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.</td>
</tr>
<tr>
<td>Process All Reads</td>
<td>ON</td>
<td>Specifies behavior identical to on-instrument CCS Reads generation, overriding all other cutoffs. This setting writes a CCS Read for every ZMW in the input Data Set. Set to OFF to specify more restrictive settings.</td>
</tr>
<tr>
<td>Include Kinetics Information with CCS Analysis Output</td>
<td>OFF</td>
<td>(Sequel IIe System only) If ON, include kinetics per-base data required for methylation DNA analysis. <strong>Note</strong>: This results in a BAM file that is 3-4 times larger. This option applies only when Process All Reads is set to ON.</td>
</tr>
<tr>
<td>Advanced CCS Options</td>
<td>NONE</td>
<td>A space-separated list of additional command-line arguments to CCS Analysis. See SMRT® Tools Reference Guide v10.2 for details.</td>
</tr>
<tr>
<td>Target Regions (BED file)</td>
<td>NONE</td>
<td>(Optional) Specifies a BED file that defines regions for a Target Regions report showing coverage over those regions. See “Appendix D - BED File Format for Target Regions Report” on page 167 for details.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

---

### Reports and Data Files
The CCS with Mapping application generates the following reports:

---

Page 72
• **On-Target Reads**: The number (and percentage) of unique reads that map with any overlap to the target region.

**Target Regions > Target Region Coverage**

• Displays the number of hits per defined region of the chromosome.

**Mapping Report > Summary Metrics**

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

• **Mean Concordance (mapped)**: The mean concordance of the CCS Reads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

• **Number of Alignments**: The number of alignments that mapped to the reference sequence.

• **Number of CCS Reads (total)**: The total number of CCS Reads in the sequence,

• **Number of CCS Reads (mapped)**: The number of CCS Reads that mapped to the reference sequence.

• **Number of CCS Reads (unmapped)**: The number of CCS Reads that did not map to the reference sequence.

• **Percentage of CCS Reads (mapped)**: The percentage of CCS Reads that mapped to the reference sequence.

• **Percentage of CCS Reads (unmapped)**: The percentage of CCS Reads that did not map to the reference sequence.

• **Number of CCS Bases (mapped)**: The number of bases in the CSS Reads that mapped to the reference sequence.

• **CCS Read Length Mean (mapped)**: The mean length of CCS Reads that mapped to the reference sequence.

• **CCS Read Length N50 (mapped)**: The read length at which 50% of the bases are in CCS 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 Analysis mapping statistics per movie.

• **Sample**: Sample name for which the following metrics apply.

• **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 CCS 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.

• **Mean Concordance (mapped)**: The mean concordance of the CCS Reads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
Mapping Report > Mapped CCS Read Length

• Histogram distribution of the mapped CCS Reads by the read length.

Mapping Report > Mapped CCS Read Concordance

• Histogram distribution of the mapped CCS Reads by their concordance with the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Mapping Report > Mapped Concordance vs Read Length

• Maps the percent concordance with the reference sequence against CCS Read length.

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 Reference

• Maps coverage across the reference.

Coverage > Depth of Coverage Distribution

• Maps the reference regions against the percent coverage.

Coverage > Coverage vs. [GC] Content

• Maps (as a percentage, over a 100 bp window) the number of Gs and Cs present across the coverage. The number of genomic windows with the corresponding % of Gs and Cs is displayed on top. Used to check that no coverage is lost over extremely biased base compositions.

CCS Analysis Report > Summary Metrics

Note: CCS Reads with quality value equal to or greater than 20 are called HiFi Reads.

• HiFi Reads: The total number of CCS Reads whose quality value is equal to or greater than 20.

• HiFi Yield (bp): The total yield (in base pairs) of the CCS Reads whose quality value is equal to or greater than 20.

• HiFi Read Length (mean, bp): The mean read length of the CCS Reads whose quality value is equal to or greater than 20.

• HiFi Read Quality (median): The median number of CCS Reads whose quality value is equal to or greater than 20.

• HiFi Number of Passes (mean): The mean number of passes used to generate CCS Reads whose quality value is equal to or greater than 20.

CCS Analysis Report > Read Length Distribution

• HiFi Read Length Distribution: Histogram distribution of HiFi Reads by read length.

• Read Length Distribution: Histogram distribution of CCS Reads by read length, broken down by read type (HiFi Reads, other CCS Reads, other reads.)
CCS Analysis Report > Number of Passes
  • Histogram of the number of complete subreads in CCS Reads, broken down by read type (HiFi Reads, other CCS Reads, other reads.)

CCS Analysis Report > Read Quality Distribution
  • Histogram distribution of the CCS Reads by the read quality.

CCS Analysis Report > Predicted Accuracy vs. Read Length
  • Heat map of CCS Read lengths and predicted accuracies.

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

  • Analysis Log: Log information for the analysis execution.
  • SMRT Link Log: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
  • Mapped Reads: All input reads that were mapped to the reference by the application.
  • Mapped BAM: The BAM file of subread alignments to the draft contigs used for polishing.
  • Mapped BAM Index: The BAI index file for the corresponding Mapped BAM file.
  • CCS Analysis Per-Read Details: Summary of CCS Analysis performance and yield.
  • Coverage Summary: Coverage summary for regions (bins) spanning the reference sequence.
  • hifi_reads.fastq.gz: Gzipped HiFi Reads in FASTQ format.
  • All Reads (BAM): BAM file containing one CCS Read per ZMW, including the following types of reads:
    – HiFi Reads (Q20 or higher)
    – Lower-quality but still polished consensus reads (Q1-Q20)
    – Unpolished consensus reads (RQ=-1)
    – 0- or 1-pass subreads unaltered (RQ=-1)

Data > IGV Visualization Files
The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.

  • Mapped BAM: The BAM file of subread alignments to the draft contigs used for polishing.
  • Mapped BAM Index: The BAI index file for the corresponding Mapped BAM file.
Circular Consensus Sequencing (CCS) Application

Use this application to identify consensus sequences for single molecules.

- The application accepts **Continuous Long Reads** (BAM format) as input.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum CCS Read 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>Maximum CCS Read Length</td>
<td>50,000</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>Generate a Consensus for Each Strand</td>
<td>OFF</td>
<td>Generate a consensus for each strand. <strong>Warning</strong>: This is an experimental option for the CCS algorithm, and may not be compatible with all downstream applications. We recommend using command-line analysis for this feature.</td>
</tr>
<tr>
<td>Process All Reads</td>
<td>ON</td>
<td>Specifies behavior identical to on-instrument CCS Reads generation, overriding all other cutoffs. This setting writes a CCS read for <strong>every</strong> ZMW in the input Data Set. Set to OFF to specify more restrictive settings.</td>
</tr>
<tr>
<td>Include Kinetics Information with CCS Analysis Output</td>
<td>OFF</td>
<td>If <strong>ON</strong>, include kinetics per-base data required for methylation DNA analysis. <strong>Note</strong>: This results in a BAM file that is 3-4 times larger. This option applies only when <strong>Process All Reads</strong> is set to <strong>ON</strong>.</td>
</tr>
<tr>
<td>Advanced CCS Options</td>
<td>NONE</td>
<td>Space-separated list of additional command-line options to CCS Analysis. Not all supported command-line options can be used, and HPC settings <strong>cannot</strong> be modified. See <strong>SMRT® Tools Reference Guide v10.2</strong> for details.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy (Deprecated)</td>
<td>0</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.) <strong>Note</strong>: This setting is ignored if the <strong>Process All Reads</strong> advanced parameter is set to <strong>ON</strong>.</td>
</tr>
<tr>
<td>Minimum Number of Passes (Deprecated)</td>
<td>0</td>
<td>The minimum number of <strong>full</strong> passes for a ZMW to be used. Full passes <strong>must</strong> have an adapter hit before and after the insert sequence and so do not include any partial passes at the start and end of the sequencing reaction. <strong>Note</strong>: This setting is ignored if the <strong>Process All Reads</strong> advanced parameter is set to <strong>ON</strong>.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td><strong>(Optional)</strong> Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>
Reports and Data Files
The Circular Consensus Sequencing (CCS) application generates the following reports:

CCS Analysis Report > Summary Metrics

Note: CCS Reads with quality value equal to or greater than 20 are called HiFi Reads.

- **HiFi Reads**: The total number of CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Yield (bp)**: The total yield (in base pairs) of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Read Length (mean, bp)**: The mean read length of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Read Quality (median)**: The median number of CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Number of Passes (mean)**: The mean number of passes used to generate CCS Reads whose quality value is equal to or greater than 20.
- **<Q20 Reads**: The total number of CCS Reads whose quality value is less than 20.
- **<Q20 Yield (bp)**: The total yield (in base pairs) of the CCS Reads whose quality value is less than 20.
- **<Q20 Read Length (mean, bp)**: The mean read length of the CCS Reads whose quality value is less than 20.
- **<Q20 Read Quality (median)**: The median number of CCS Reads whose quality value is less than 20.

CCS Analysis Report > By Movie
Displays, per movie, statistics on HiFi Reads.

- **Movie**: Movie name for which the following metrics apply.
- **HiFi Reads**: The total number of CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Yield (bp)**: The total yield (in base pairs) of the CCS Reads whose quality value is equal to or greater than 20.
- **HiFi Read Length (mean, bp)**: The mean read length of the CCS Reads whose quality value is equal to or greater than 20.

CCS Analysis Report > Read Length Distribution

- **HiFi Read Length Distribution**: Histogram distribution of HiFi Reads by read length.
- **Read Length Distribution**: Histogram distribution of CCS Reads by read length, broken down by read type (HiFi Reads, other CCS Reads, other reads.)

CCS Analysis Report > Number of Passes

- Histogram of the number of complete subreads in CCS Reads, broken down by read type (HiFi Reads, other CCS Reads, other reads.)

CCS Analysis Report > Read Quality Distribution

- Histogram distribution of the CCS Reads by the read quality.
CCS Analysis Report > Predicted Accuracy vs. Read Length

- Heat map of CCS Read lengths and predicted accuracies.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **CCS Analysis Per-Read Details**: Summary of CCS Analysis performance and yield.
- **hifi_reads.fastq.gz**: Gzipped HiFi Reads in FASTQ format.
- **All Reads (BAM)**: BAM file containing one CCS Read per ZMW, including the following types of reads:
  - HiFi Reads (Q20 or higher)
  - Lower-quality but still polished consensus reads (Q1-Q20)
  - Unpolished consensus reads (RQ=-1)
  - 0- or 1-pass subreads unaltered (RQ=-1)
Use this application to convert sequence data in BAM file format to the FASTA and FASTQ file formats.

- The application accepts **Continuous Long Reads** (BAM format) as input.
- 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.

### Importing/Exporting Analysis Settings

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

### Parameters

<table>
<thead>
<tr>
<th>Advanced 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 semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td><em>(Optional)</em> Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</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 Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **FASTA file(s)**: Sequence data converted to FASTA format.
- **FASTQ file(s)**: Sequence data converted to FASTQ format.
Demultiplex Barcodes Application

Use this application to separate sequence reads by barcode. (See “Working with Barcoded Data” on page 135 for more details.)

**Note**: To demultiplex Iso-Seq samples in the SMRT Link interface, always choose the Iso-Seq Analysis application, not the Demultiplex Barcodes application.

- The application accepts **Continuous Long Reads** or **HiFi Reads** (BAM format) as input. **HiFi Reads** are reads generated with CCS Analysis whose quality value is equal to or greater than 20.
- 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 template 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.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Barcode Set (Required)**

- Specify a barcode sequence file to separate the reads.

**Same Barcodes on Both Ends of Sequence (Default = Yes)**

- Specify **Yes** 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 135 for information on barcode designs.)
- Specify **No** to specify asymmetric designs where the barcodes are **different** on each end of the insert sequence.
Assign Bio Sample Names to Barcodes (Required)

SMRT Link automatically creates a CSV-format Autofilled Barcoded Sample 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 column of automatically-generated Bio Sample Names \(1\) through \(N\), corresponding to barcodes \(1\) through \(N\), for the biological sample names. There are two different ways to specify which barcodes to use, and assign biological sample names to barcodes:

**Interactively:**
1. Click Interactively, then drag barcodes from the Available Barcodes column to the Included Barcodes column. (Use the checkboxes to select multiple barcodes.)
2. (Optional) Click a Bio Sample field to edit the Bio Sample Name associated with a barcode. Note: Avoid spaces in Bio Sample Names as they may lead to third-party compatibility issues.
3. (Optional) Click Download as a file for later use.
4. Click Save to save the edited barcodes/Bio Sample names. You see Success on the line below, assuming the file is formatted correctly.

**From a file:**
1. Click From a File, then click Download File. Edit the file and enter the biological sample names associated with the barcodes in the second column, then save the file. Use alphanumeric characters, spaces (allowed but not recommended for compatibility with third-party downstream software), hyphens, underscores, colons, or periods only - other characters will be removed automatically, with a maximum of 40 characters. If you did not use all barcodes in the Autofilled Barcode Name file in the sequencing run, delete those rows.
   – Note: Open the CSV file in a text editor and check that the columns are separated by commas, not semicolons or tabs.
2. Select the Barcoded Sample File you just edited. You see Success on the line below, assuming the file is formatted correctly.

Demultiplexed Output Data Set Name

- Specify the name for the new demultiplexed Data Set that will display in SMRT Link. The application creates a copy of the input Data Set, renames it to the name specified, and creates demultiplexed child Data Sets linked to it. The input data set remains separate and unmodified.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
<tr>
<td>Write Unbarcoded Reads</td>
<td>ON</td>
<td>If specified, write out a separate Data Set containing reads that are not barcoded.</td>
</tr>
</tbody>
</table>
**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 correctly-barcoded reads in the sequence data.
- **Mean Reads**: The mean number of reads per barcode combination.
- **Max. Reads**: The maximum number of reads per barcode combination.
- **Min. Reads**: The minimum number of reads per barcode combination.
- **Mean Read Length**: The mean read length of reads per barcode combination.
- **Mean Longest Subread Length**: The mean length of the longest subread in each barcoded sample.
- **Unbarcoded Reads**: The number of reads without barcodes in the sequence data.
- **Percent Bases in Barcoded Reads**: The percentage of bases in reads in the sequence data that contain barcodes.
- **Percent Barcoded Reads**: The percentage of reads in the sequence data that contain barcodes.

### Barcodes > Barcode Data

- **Bio Sample Name**: The name of the biological sample associated with the barcode combination.
- **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 combination.
- **Subreads**: The number of subreads associated with the barcode combination.
- **Bases**: The number of bases associated with the barcode combination.
- **Mean Read Length**: The mean read length of reads associated with the barcode combination.
- **Mean of Longest Subread Length**: The mean length of the longest subread associated with the barcode combination.
- **Mean Barcode Quality**: The mean barcode quality associated with the barcode combination.

---

**Advanced Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Barcode Score</td>
<td>80</td>
<td>A barcode score measures the alignment between a barcode attached to a read and an ideal barcode sequence, and is an indicator of how well the chosen barcode pair matches. It ranges between 0 (no match) and 100 (a perfect match). Specifies that reads with barcode scores below this minimum value are not included in analysis. This affects the output BAM file and the output demultiplexed Data Set XML file.</td>
</tr>
<tr>
<td>Advanced Lima Options</td>
<td>NONE</td>
<td>Space-separated list of custom lima options. Not all supported command-line options can be used, and HPC settings cannot be modified. See the Demultiplex Barcodes section of the document SMRT® Tools Reference Guide v10.2 for information on lima.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>(Optional)</td>
<td>Select the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>
• **Unique Molecular Yield**: The sum total length of unique single molecules associated with the barcode combination.

**Barcodes > Inferred Barcodes**

• **Barcode Name**: The barcode name.

• **Number of ZMWs**: The number of ZMWs out of the first 50,000 that are inferred to be assigned to the barcode combination.

• **Mean Barcode Score**: The mean barcode score associated with the reads inferred to be associated with the barcode combination.

• **Selected**: Yes if the number of ZMWs is at least 10, No otherwise.

**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 far from the center. Check the output file *Inferred Barcodes*, note the correct barcodes used, and consider reanalyzing the multiplexed samples with the correct Bio Sample names for the barcodes actually used. If you reanalyze the data, ensure that the **Barcode Name** file includes only the correct barcodes used.

• **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. This can indicate use of incorrect barcodes. Check the output file *Inferred Barcodes*, note the correct barcodes used, and consider reanalyzing the multiplexed samples with the correct Bio Sample names for the barcodes actually used. If you reanalyze the data, ensure that the **Barcode Name** file includes only the correct barcodes used.

• **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 <60 suggests issues with some of the barcode analyses. The red line is set at 80 – the minimum default barcode score.

  – **Good performance**: HiFi demultiplexing runs should have >90% of reads with barcode quality score ≥95.
Questionable performance: A bimodal distribution with a large second peak usually indicates that some barcodes that were sequenced were not included in the barcode scoring set.

Barcodes > Barcoded Read Binned Histograms

- **Read Length Distribution By Barcode**: Histogram distribution of the Polymerase read length by barcode. Each column of rectangles is similar to a read length histogram rotated vertically, seen from the top. Each sample should have similar Polymerase read length distribution. Non-smooth changes in the pattern looking from left to right might indicate suboptimal performance.

- **Barcode Quality Distribution By Barcode**: Histogram distribution of the per-barcode version of the Read Length Distribution by Barcode histogram. The histogram should contain a single cluster of hot spots in each column. All barcodes should also have similar profiles; significant differences in the pattern moving from left to right might indicate suboptimal performance.
  - **Good performance**: All columns show a single cluster of hot spots.
  - **Questionable performance**: A bimodal distribution would indicate missing barcodes in the scoring set.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **All Barcodes (FASTA)**: All barcoded reads, in FASTA format.
- **Barcode Files**: Barcoded subread Data Sets; one file per barcode.
- **Barcode Summary CSV**: Data displayed in the reports, in CSV format. This includes Bio Sample Name.
- **Barcode Summary**: Text file listing how many ZMWs were filtered, how many ZMWs are the same or different, and how many reads were filtered.
- **Inferred Barcodes**: Inferred barcodes used in the analysis. The barcoding algorithm looks at the first 35,000 ZMWs, then selects barcodes with ≥10 counts and mean scores ≥45.
- **Unbarcoded Reads**: BAM file containing reads not associated with a barcode.
- **demultiplex.<barcode>.hifi.reads.fastq.gz**: Gzipped HiFi Reads in FASTQ format, one file per barcode.
Export Reads Application
Use this application to export HiFi Reads that pass filtering criteria as FASTA, FASTQ and BAM files.

- The application accepts HiFi Reads (BAM format) as input. HiFi Reads are reads generated with CCS Analysis whose quality value is equal to or greater than 20.
- 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.

Importing/Exporting Analysis Settings
- Click Import Analysis Settings and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
- Click Export to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

Output FASTA File (Default = ON)
- Outputs a single FASTA/FASTQ file containing all the reads that passed the filtering criteria.

Output BAM File (Default = OFF)
- Outputs a single BAM file containing all the reads that passed the filtering criteria.

Min. CCS Predicted Accuracy (Phred Scale) Default = 20
- Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.

Parameters

<table>
<thead>
<tr>
<th>Advanced 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 semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td><em>(Optional)</em> Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</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 Log: Log information for the analysis execution.
- SMRT Link Log: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- hifi_reads.fasta.gz file: Sequence data that passed filtering criteria, converted to Gzipped FASTA format.
• **hifi_reads.fastq.gz** file: Sequence data that passed filtering criteria, converted to Gzipped FASTQ format.
• **<Reads>.bam** file: Sequence data that passed filtering criteria.
Genome Assembly Application

Use this application to generate high quality *de novo* assemblies of genomes, using HiFi Reads.

- The application accepts **HiFi Reads** (BAM format) as input. **HiFi Reads** are reads generated with CCS Analysis whose quality value is equal to or greater than 20.

The application includes seven main steps:

1. Convert input to a compressed database for fast retrieval.
2. Overlap reads using the **Pancake** tool.
3. Phase the overlapped reads using **Nighthawk**. Nighthawk also boosts contiguity of the assembly by removing overlaps between reads coming from different instances of a genomic repeat (such as segmental duplications.)
4. Remove chimeras and duplicate reads which do not span repeat regions. This improves contiguity and assembly quality.
5. Construct a string graph. Extract primary contigs and haplotigs. Haplotypes are represented by heterozygous bubbles.
6. Polish the contigs and haplotigs using phased reads. Phasing information is preserved. Polishing is done with **Racon**.
7. Identify potential haplotype duplications in the primary contig set using the **purge_dups** tool, and move them to the haplotig set. This final round of assembly processing is especially useful in high heterozygosity samples.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Length</td>
<td>0</td>
<td>The approximate number of base pairs expected in the genome. This is used only for downsampling; if the value is ( \leq 0 ), downsampling is disabled.</td>
</tr>
<tr>
<td>Downsampling Coverage</td>
<td>0</td>
<td>The input Data Set can be downsampled to a desired coverage, provided that both the <strong>Downsampling Coverage</strong> and <strong>Genome Length</strong> parameters are specified and ( &gt; 0 ). Downsampling applies to the entire assembly process, including polishing. This parameter selects reads randomly, using a fixed random seed for reproducibility.</td>
</tr>
<tr>
<td>Run Polishing</td>
<td>ON</td>
<td>Enables or disables the polishing stage of the workflow. Polishing can be disabled to perform fast draft assemblies.</td>
</tr>
</tbody>
</table>
Reports and Data Files

The Genome Assembly application generates the following reports:

**Polished Assembly > Summary Metrics**
Displays statistics on the contigs from the *de novo* assembly that were corrected by Racon.

- **Contig Type**: Primary or Haplotigs. Primary contigs represent pseudohaplotype assemblies, while haplotigs represent fully phased and assembled regions of the genome. Primary contigs are usually much longer than haplotigs due to allowed haplotype switching.
- **Polished Contigs**: The number of polished contigs.
- **Maximum Contig Length**: The length of the longest contig.
- **Mean Contig Length**: The mean length of the contigs.
- **Median Contig Length**: The median length of the contigs.
- **N50 Contig Length**: 50% of the contigs are longer than this value.
- **Sum of Contig Lengths**: Total length of all the contigs.
- **E-size (sum of squares/sum)**: The expected contig size for a random base in the polished contigs. Another interpretation: The area under the Nx curve (for x in range [0, 100]).
- **Number of Circular Contigs**: The number of assembled contigs that are circular.

**Polished Assembly > Polished Contigs**
- **Contig**: Name of the individual contig.
- **Length (bases)**: The length of the contig, in bases.
- **Circular**: Yes if the contig is circular, No if it isn’t.
- **Percent Polished**: The percent of contig bases that were polished.
- **Number of Polishing Reads**: The number of reads used to perform polishing on this contig.
Data > File Downloads

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

- Analysis Log: Log information for the analysis execution.
- SMRT Link Log: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- Haplotigs: The final polished haplotigs assembly, in FASTA format.
- Primary Contigs: The final polished primary contigs assembly, in FASTA format.
HiFiViral SARS-CoV-2 Analysis Application

Use this application to analyze multiplexed samples sequenced with the HiFiViral SARS-CoV-2 kit. For each sample, this analysis provides:

- Consensus sequence (FASTA)
- Variant calls (VCF)
- HiFi Reads aligned to the reference (BAM)
- Plot of HiFi Read coverage depth across the SARS-CoV-2 genome.

Across all samples, this analysis provides:

- Job summary table including passing sample count at 90 and 95% genome coverage.
- Sample summary table including, for each sample: Count of variable sites, genome coverage, read coverage, and probability of multiple strains, and other metrics.
- Plate QC graphical summary of performance across samples in assay plate layout.
- Plot of HiFi Read depth of coverage for all samples.

Notes:

- The application accepts HiFi Reads (BAM format) as input. HiFi Reads are reads generated with CCS Analysis that have a quality value equal to or greater than Phred-scaled Q20.
- This application is for SARS-CoV-2 analysis only and is not recommended for other viral studies. The Wuhan reference genome is needed to run the application, but advanced users may specify other reference genomes. We have not tested the application with reference genomes other than the Wuhan reference genome.
- The application is intended to identify variable sites and call a single consensus sequence per sample. The output consensus sequence is produced based on the dominant variant observed. Minor variant information that passes through a default threshold may be encoded in the raw VCF, but does not get propagated into the consensus sequence FASTA.
- The HiFiViral SARS-CoV-2 Analysis application can be run using the Auto Analysis feature available in Run Design. This feature allows users to complete all necessary analysis steps immediately after sequencing without manual intervention. The Auto Analysis workflow includes CCS, Demultiplex Barcodes, and HiFiViral SARS-CoV-2 Analysis.

Auto Analysis in Run Design

Users may set the analysis to begin automatically after sequencing completes using Auto Analysis in Run Design. See "HiFiViral SARS-CoV-2: Creating Auto Analysis in Run Design" on page 144 for details.
HiFiViral SARS-CoV-2 Application Workflow

1. Process the reads using the `mimux` tool to trim the probe arm sequences.
2. Align the reads to the reference genome using `pbmm2`.
3. Call and filter variants using `bcftools`, generating the raw variant calls in VCF file format. Filtering in this step removes low-quality calls (less than Q20), and normalizes indels.
4. Filter low-frequency variants using `vcfcons` and generate a consensus sequence by injecting variants into the reference genome. At each position, a variant is called only if both the base coverage exceeds the minimum base coverage threshold (Default = 4) and the fraction of reads that support this variant is above the minimum variant frequency threshold (Default = 0.5). See here for details.

Preparing Input Data for the HiFiViral SARS-CoV-2 Analysis Application

1. Run the `Demultiplex Barcodes` application, where the input to that application are HiFi Reads, and the primers are multiplexed barcode primers. (If HiFi Reads have not been generated on the instrument, run CCS Analysis first. See “Circular Consensus Sequencing (CCS) Application” on page 76 for details.)
   - The proper barcode sequences are provided by default: `HiFiViral_SARS-CoV-2_M13barcodes`.
   - For the `Same Barcodes on Both Ends of Sequence` parameter, specify No; the barcode pairs are asymmetric.
   - Provide the correctly formatted barcode pair-to-Bio Sample CSV file for the `Assign Bio Sample Names to Barcodes` option. (For details, see “Assign Bio Sample Names to Barcodes (Required)” on page 65.)

Running the HiFiViral SARS-CoV-2 Analysis Application

1. After running the Demultiplex Barcode application, create a new analysis using `SMRT Analysis > Create New Analysis`.
2. Name the analysis, then select `Data Types > HiFi Reads`.
3. Select all the demultiplex samples contained in the Data Set and choose `Analysis of Multiple Data Sets > One Analysis for All Data Sets`. Click Next.
4. Select `HiFiViral SARS-CoV-2 Analysis` from the Analysis Application list.
5. `SARS-CoV-2 Genome NC_045512.2` (the Wuhan reference genome) and `HiFiViral SARS-CoV-2 Enrichment Probes` are automatically loaded; advanced users may select a different reference or probe set if desired.
6. To generate the optional Plate QC graphical summary, click `Advanced Parameters` and load a CSV file using the provided template (`assayPlateQC_template_4by96.csv`) as a guide.
7. Click Start.
Importing/Exporting Analysis Settings

• Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.

• Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

Reference Genome (Required)

• Specify the full viral genome against which to align the reads and call variants. (The default is the Wuhan Reference genome.)

Probe Sequences (Required)

• Specify probe sequences in FASTA format. (The default probe sequence file is designed for the **HiFiViral SARS-CoV-2 Enrichment Probes**.)

Plate QC CSV (Optional)

• The file contains the following four columns: Barcode (asymmetric pairs), Bio Sample Name, assay plate ID (can include 1-4 plates with unique names; avoid special characters), and assay plate well (in the format A01, A02,…H12.) The plate and well information corresponds to the location of samples during the SARS-CoV-2 enrichment assay.
Parameters

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plate QC CSV</td>
<td>NONE</td>
<td>(Optional) Specify a CSV file to generate the Plate QC Report, which displays analysis results for each sample in the assay plate. The CSV file must contain barcode pairs, sample name, Plate IDs, and Well IDs. (To create a new file, click Download Template, edit, and then save the CSV file.)</td>
</tr>
<tr>
<td>Minimum Base Coverage</td>
<td>4</td>
<td>Specify the minimum read depth at each position to report either a variant or a reference base. Positions with less than this specified coverage will have an N base output in the consensus sequence FASTA file. Increasing the minimum base coverage may result in more Ns and loss of variant detection. We do not recommend making this value lower than the default threshold of 4, as it may increase the number of false positive variants called.</td>
</tr>
<tr>
<td>Minimum Variant Frequency</td>
<td>0.5</td>
<td>Specify that only variants whose frequency is greater than this value are reported. This frequency is determined based on the read depth (DP) and allele read count (AD) information in the VCF output file. We recommend using the default value to properly call the dominant alternative variant while also filtering out potential artifacts.</td>
</tr>
<tr>
<td>Advanced Processing Options</td>
<td>NONE</td>
<td>Additional options to pass to the mimux preprocessing tool for trimming and filtering reads by probe sequences. Options should be entered in space-separated format. See the HiFiViral SARS-CoV-2 Analysis section of SMRT Tools Reference Guide (v10.2) for details.</td>
</tr>
<tr>
<td>Minimum Barcode Score</td>
<td>80</td>
<td>A barcode score measures the alignment between a barcode attached to a read and an ideal barcode sequence, and is an indicator of how well the chosen barcode pair matches. It ranges between 0 (no match) and 100 (a perfect match). This parameter specifies that reads with barcode scores below this minimum value are not included in analysis.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

Reports and Data Files

The HiFiViral SARS-CoV-2 Analysis application generates the following reports:

Summary Report > Summary Metrics

- **Samples**: The count of all input samples, whether or not they passed analysis.
- **Samples with Genome Coverage > 90%**: The number of samples where at least 90% of bases have at least four mapped reads overlapping their position.
- **Samples with Genome Coverage > 95%**: The number of samples where at least 95% of bases have at least four mapped reads overlapping their position.
- **Samples Failing Workflow**: The number of samples for which the analysis was unable to generate a per-sample report due to an absence of usable data.

Summary Report > Sample Summary

- **Bio Sample Name**: The name of the biological sample associated with the variants. (Note: Any spaces in the name are substituted by new line characters for consistency with output file names.)
• **Substitutions**: The count of all called substitutions in the consensus sequence for the sample.

• **Insertions**: The count of all called insertions in the consensus sequence for the sample.

• **Deletions**: The count of all called deletions in the consensus sequence for the sample.

• **Reads**: The total number of HiFi Reads for the sample.

• **Read Coverage**: The mean number of mapped reads overlapping with each position in the reference genome.

• **On-Target Rate**: The mapping yield of reads; the number of unique mapped reads divided by the total number of reads.

• **Multiple Strains (Probability)**: Samples are flagged as having multiple strains if the probability is at least 0.95. Samples may contain multiple strains due to sample contamination or presence of multiple strains in the RNA extract. To classify a sample as multi-strain, we tolerate error by using the binomial cumulative distribution function (with a fixed probability of 0.2).

• **Ns**: The number of bases in the consensus sequence that are Ns.

• **Genome Coverage**: The percentage of bases with at least four mapped reads overlapping their position by default. See the Advanced Parameters dialog to adjust minimum base coverage.

**Summary Report > Genome Coverage**

- Coverage plot showing the per-sample mean read coverage within a window of 100 bp. The shaded region displays the 25th to 75th percentile in the range of coverage across all samples, and the darker solid line displays the median coverage across all samples.

**Summary Report > Plate QC**

Plot showing analysis results for each plate cell used. This plot is generated only if the user supplies a Plate QC CSV file mapping BioSample Names to Well IDs in Advanced Parameters.

- **Blue** wells represent samples with at least 95% coverage.
- **Green** wells represent samples with at least 90% coverage.
- **Yellow** wells represent samples that passed the workflow but had genome coverage worse than 90%.
- **Red** wells represent samples that failed the workflow.
- **White** wells do not include a sample.

**Data > File Downloads**

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **All Samples, HiFi Reads FASTQ**: HiFi Reads in FASTQ format for all samples.
- **All Samples, Consensus Sequence FASTA**: The full consensus genomic sequences; bases for which no consensus could be called are represented by Ns. See the Advanced Parameters dialog to adjust the minimum base coverage for outputting Ns.
• **All Samples, Genome Coverage Plots**: Plots for individual samples showing coverage depth across the genome.

• **All Samples, Variant Call VCF**: VCF file containing the final variant calls per sample.

• **All Samples, HiFi Reads Mapped BAM**: BAM file for each sample containing the HiFi Reads aligned to the reference genome.

• **All Samples, Consensus Sequence Aligned BAM**: BAM file for each sample of consensus sequence aligned to the reference genome. The consensus sequence is split into fragments where there are Ns and each fragment is mapped.

• **All Samples, Raw Variant Calls VCF**: VCF file containing the intermediate variant calls per patient sample.

• **Sample Summary Table CSV**: CSV version of the data shown in the Sample Summary table.

• **All Samples, Probe Counts TSV**: Tab-delimited text file containing per-sample, per-probe counts. This file can be used to identify samples that are poorly sequenced or probes with high or low coverages.
The Iso-Seq application enables analysis and functional characterization of full-length transcript isoforms for sequencing data generated on PacBio instruments.

- The application accepts Sequel Systems ConsensusReadSet (HiFi Reads) as input. HiFi Reads are reads generated with CCS Analysis whose quality value is equal to or greater than 20. By default, all reads are processed into CCS reads. QV filtering happens at the Iso-Seq Analysis step.
- Iso-Seq Analysis only supports CCS Reads input.
- Note: The default minimum CCS Analysis accuracy (Phred Scale) for Iso-Seq Analysis is QV 10. All other CCS-based workflows use HiFi Reads, which by default have QV 20.

**Note on Multiplexed Data**

- If you have multiplexed samples, provide the multiplexed barcodes as part of the primers. You should not run the Demultiplex Barcodes application first. See the Primer Set example below.

The application includes three main steps:

1. **Classify**: Identify and remove primers (which includes cDNA primers and optionally barcodes). Identify strandedness based on the 5’ and 3’ primers.
2. **Cluster (Optional)**: Trim off polyA tails and remove artificial concatemers. Perform de novo clustering and consensus calling. Output full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on estimated accuracies.
3. **Collapse (Optional)**: When a reference genome is selected, map HQ isoforms to the genome, then collapse isoforms into unique isoforms.

To obtain full-length non-concatemer (FLNC) reads and not complete the Cluster step: Ensure that the Run Clustering option is set to OFF.

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

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

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

**Importing/Exporting Analysis Settings**

- Click Import Analysis Settings and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
• Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Primer Set (Required)**

• Specify a primer sequence file in FASTA format to identify cDNA primers for removal. The primer sequence includes the 5’ and 3’ cDNA primers and (if applicable) barcodes.

• Primer IDs **must** be specified using the suffix _5p to indicate 5’ cDNA primers and the suffix _3p to indicate 3’ cDNA primers. The 3’ cDNA primer should **not** include the Ts and is written in reverse complement (see examples below).

• If barcodes were used, they should be included.

• Each primer sequence must be **unique**.

**Example 1:** The Iso-Seq v2 primer set (included with the SMRT Link installation).

>NEB_5p
GCAATGAAGTCGCAGGGTTGGG
>Clontech_5p
AAGCAGTGTTATCAACGCAGAGTACATGGGG
>NEB_Clontech_3p
GTACTCTGCGTTGATATCCTGCTT

**Note:** The Clontech kit is **unsupported**, and these primers will **not** be included in future SMRT Link releases.

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

>NEB_5p
GCAATGAAGTCGCAGGGTTGGG
>d_T_BC1001_3p
AAAGCAGTGTTATCAACGCAGAGTACACATATCAGAGTGCG
>d_T_BC1002_3p
AAAGCAGTGTTATCAACGCAGAGTACACACACAGACTGTGAG
>d_T_BC1003_3p
AAAGCAGTGTTATCAACGCAGAGTACACACATCTCGTGAGAG
>d_T_BC1004_3p
AAAGCAGTGTTATCAACGCAGAGTACACACGCACACACGCGCG

**Note:** NEB _5p is **not** an NEB primer sequence; it is PacBio’s Iso-Seq Express cDNA PCR primer sequence listed in the protocol.

**Special Handling for the TeloPrime cDNA Kit**

The Lexogen TeloPrime cDNA kit contains As in the 3’ primer that **cannot** be differentiated from the polyA tail. For best results, remove the As from the 3’ end as shown below:

>TeloPrimeModified_5p
TGGATGTGATATGATATACGACTCCTATAG
>TeloPrimeModified_3p
CGCCTGAGA


Reference Set (Optional)
• Optionally specify a reference genome to align High Quality isoforms to, and to collapse isoforms mapped to the same genomic loci.

Run Clustering (Default = ON)
• Specify ON to generate consensus isoforms.
• Specify OFF to classify reads only and not generate consensus isoforms. The Reference Set will also be ignored.

Cluster Barcoded Samples Separately (Default = OFF)
• Specify OFF if barcoded samples are from the same species, but different tissues, or samples of the same genes but different individuals. The samples are clustered with all barcodes pooled.
• Specify ON if barcoded samples are from different species. The samples are clustered separately by barcode.
• In either case, the samples on the results page are automatically named BioSample_1 through BioSample_N.

Parameters

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Require and trim Poly(A) Tail</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 Mapped Length (bp)</td>
<td>50</td>
<td>The minimum required mapped HQ isoform sequence length (in base pairs) for the Iso-Seq mapping-collapse step. Applicable only if a reference genome is provided.</td>
</tr>
<tr>
<td>Minimum Gap-Compressed Identity (%)</td>
<td>95</td>
<td>The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference. Note: Applicable only if a reference genome is provided.</td>
</tr>
<tr>
<td>Minimum Mapped Coverage (%)</td>
<td>99</td>
<td>The minimum required HQ transcript isoform sequence alignment coverage (in percent) for the Iso-Seq mapping-collapse step. Applicable only if a reference genome is provided.</td>
</tr>
<tr>
<td>Maximum Fuzzy Junction Difference (bp)</td>
<td>5</td>
<td>The maximum junction difference between two mapped isoforms to be collapsed into a single isoform. If the junction differences are all less than the provided value, they will all be collapsed. Setting to 0 requires all junctions to be exact to be collapsed into a single isoform. Applicable only if a reference genome is provided.</td>
</tr>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>10</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for Iso-Seq Analysis is 10 (QV 10), or 90% predicted accuracy. Note: Some of the automatically-generated CCS Analysis reports display information about QV 20 CCS Reads.</td>
</tr>
<tr>
<td>Filters to Add to the Data Set</td>
<td>NONE</td>
<td>A semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Advanced pbmm2 Options</td>
<td>NONE</td>
<td>Space-separated list of custom pbmm2 options. (pbmm2 is already running with --preset ISOSEQ) Not all supported command-line options can be used, and HPC settings cannot be modified. See SMRT® Tools Reference Guide v10.2 for details.</td>
</tr>
</tbody>
</table>
Reports and Data Files
The Iso-Seq application generates the following reports:

**CCS Analysis Read Classification > Summary Metrics**

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

**CCS Analysis Read Classification > Primer Data**

- **Bio Sample Name**: The name of the biological sample associated with the primer.
- **Primer Name**: A string containing the pair of primer indices associated with this biological sample.
- **CCS Reads**: The number of CCS Reads associated with this primer.
- **Mean Primer Quality**: The mean primer quality associated with the primer.
- **Reads with 5’ and 3’ Primers**: The number of CCS Reads with 5’ and 3’ cDNA primers detected.
- **Non-Concatemer Reads with 5’ and 3’ Primers**: The number of non-concatemer CCS Reads with 5’ and 3’ primers detected.
- **Non-Concatemer Reads with 5’ and 3’ Primers and Poly-A Tail**: The number of non-concatemer CCS Reads with 5’ and 3’ primers and polyA tails detected. This is usually the number for full-length, non-concatemer (FLNC) reads, unless polyA tails are not present in the sample.

**CCS Analysis Read Classification > Primer Read Statistics**

- **Number Of Reads Per Primer**: Maps the number of reads per primer, sorted by primer ranking.
- **Primer Frequency Distribution**: Maps the number of samples with primers by the number of reads with primers.
- **Mean Read Length Distribution**: Maps the read mean length against the number of samples with primers.
CCS Analysis Read Classification > Primer Quality Scores

- Histogram of primer scores.

CCS Analysis Read Classification > Primer Reads Binned Histograms

- Read Length Distribution By Primer: Heat map of read lengths, sorted by ranking.
- Primer Quality Distribution By Primer: Heat map of number of reads by primer scores, sorted by ranking.

CCS Analysis Read Classification > Length of Full-Length Non-Concatemer Reads

- Histogram of the read length distribution of non-concatemer CCS Reads with 5’ and 3’ primers and polyA tails detected.

Transcript Clustering > Summary Metrics

- Number of High-Quality Isoforms: The number of consensus isoforms that have an estimated accuracy above the specified threshold. (This is set by the Minimum Passes for High-Quality Isoforms option in the Advanced Parameters dialog.)
- Number of Low-Quality Isoforms: The number of consensus isoforms that have an estimated accuracy below the specified threshold. (This is set by the Minimum Passes for High-Quality Isoforms option in the Advanced Parameters dialog.)

Transcript Clustering > Length of Consensus Isoforms

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

Transcript Mapping > Summary Metrics

- Number of mapped unique isoforms: The number of unique isoforms, where each unique isoform is generated by collapsing redundant HQ isoforms (such as those have very minor differences from one to one another) to one isoform. Each unique isoform may be generated from one or multiple HQ isoforms.
- Number of mapped unique loci: The number of unique mapped genomic loci among all unique isoforms. Multiple unique isoforms may map to the same genomic location, indicating these unique isoforms are transcribed from the same gene family, but spliced differently.

Transcript Mapping > Length of Mapped Isoforms

- Histogram of mapped isoforms binned by read length and the distribution of mapped isoforms exceeding a read length cutoff.

Data > File Downloads

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

- Analysis Log: Log information for the analysis execution.
- SMRT Link Log: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- Primers Summary: Text file listing how many ZMWs were filtered, how many ZMWs are the same or different, and how many reads were filtered.
• **Inferred Primers**: Inferred primers used in the analysis. The algorithm looks at the first 35,000 ZMWs, then selects primers with ≥10 counts and mean scores ≥45.

• **Full-Length Non-Concatemer Read Assignments**: Full-length reads that have primers and polyA tails removed, in BAM format.

• **Full-Length Non-Concatemer Report**: Includes strand, 5’ primer length, 3’ primer length, polyA tail length, insertion length, and primer IDs for each full-length read that has primers and polyA tail, in CSV format.

• **Low-Quality Isoforms**: Isoforms with low consensus accuracy, in FASTQ and FASTA format. We recommend that you work only with High-Quality isoforms, unless there are specific reasons to analyze Low-Quality isoforms. When the input Data Set is a ConsensusReadSet, a FASTA file only is generated.

• **High-Quality Isoforms**: Isoforms with high consensus accuracy, in FASTQ and FASTA format. This is the recommended output file to work with. When the input Data Set is a ConsensusReadSet, a FASTA file only is generated.

• **Cluster Report**: Report of each full-length read into isoform clusters.

• **Isoform Counts by Barcode**: For each isoform, report supportive FLNC reads for each barcode.

• **Mapped High Quality Isoforms (BAM Index)**: Alignments mapping isoforms to the reference genome, in BAM and BAI (index) formats.

• **Collapsed Filtered Isoforms GFF**: Mapped, unique isoforms, in GFF format. This is the Mapping step output that is the recommended output file to work with.

• **Collapsed Filtered Isoforms**: Mapped, unique isoforms, in FASTQ format. This is the Mapping step output that is recommended output file to work with. When the input Data Set is a ConsensusReadSet, only a FASTA file is generated.

• **Collapsed Filtered Isoforms Groups**: Report of isoforms mapped into collapsed filtered isoforms.

• **Full-length Non-Concatemer Read Assignments**: Report of full-length read association with collapsed filtered isoforms, in text format.

• **Collapsed Filtered Isoform Counts**: Report of read count information for each collapsed filtered isoform.

**Data > IGV Visualization Files**

The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.

• **Mapped High Quality Isoforms**: Alignments mapping isoforms to the reference genome, in BAM and BAI (index) formats.

**Note**: For details on custom PacBio tags added to output BAM files by the Iso-Seq Application, see page 54 of **SMRT® Tools Reference Guide (v10.2)**, or click [here](#).
Use this application to determine phased consensus sequences for pooled amplicon data. The LAA application:

- Accepts **Continuous Long Reads** (BAM format) as input.
- 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.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced 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 semicolon-separated (<strong>not</strong> comma-separated) list of other filters to add to the Data Set.</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>Phasing</td>
<td>ON</td>
<td>Specify that the fine phasing step take place.</td>
</tr>
</tbody>
</table>
### Advanced Parameters

<table>
<thead>
<tr>
<th>Parameter</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 input reads to use. To disable, set to 0.</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>Minimum SNR</td>
<td>2.5</td>
<td>The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR &lt;2.5 is typically considered lower quality.</td>
</tr>
<tr>
<td>Minimum Barcode Score</td>
<td>26</td>
<td>The minimum average barcode score required for subreads.</td>
</tr>
<tr>
<td>Clustering</td>
<td>ON</td>
<td>Specify whether to activate the coarse clustering phase.</td>
</tr>
<tr>
<td>Maximum Clustering Reads</td>
<td>400</td>
<td>The maximum number of input reads to cluster per barcode.</td>
</tr>
<tr>
<td>Trim Sequence Ends</td>
<td>0</td>
<td>Trim this number of bases from each end of each consensus.</td>
</tr>
<tr>
<td>Advanced LAA Options</td>
<td>NONE</td>
<td>Space-separated list of custom laa options. Not all supported command-line options can be used, and HPC settings cannot be modified. See SMRT® Tools Reference Guide v10.2 for details.</td>
</tr>
<tr>
<td>Filter Input Reads by Presence of Both Flanking Barcodes</td>
<td>OFF</td>
<td>Specify whether to filter the input reads if both flanking barcodes are present.</td>
</tr>
<tr>
<td>Maximum Reads</td>
<td>2,000</td>
<td>The maximum number of input reads to cluster per barcode.</td>
</tr>
<tr>
<td>Minimum Read Score</td>
<td>0.75</td>
<td>The minimum read score of input subreads.</td>
</tr>
<tr>
<td>Ignore N Bases At End</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 Phasing Reads</td>
<td>500</td>
<td>The maximum number of input reads to use for phasing and consensus.</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 Read Fraction</td>
<td>0.1</td>
<td>The minimum fraction of reads favoring the minor phase required to split a haplotype.</td>
</tr>
<tr>
<td>Minimum Predicted Accuracy</td>
<td>0.95</td>
<td>The minimum predicted consensus accuracy below which a consensus is labeled as &quot;noise&quot;.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

### Reports and Data Files

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

**Amplicon Inputs > Amplicon Input Molecule Summary**

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

- **Barcode Name**: A string containing the pair of barcode names (or indices if not available) for which the following metrics apply.
- **Good**: The number of subreads used in a consensus sequence not categorized as Chimeric or Noise.
- **Good (%)**: The percentage of subreads used in a consensus sequence not categorized as Chimeric or Noise.
• **Chimeric**: The number of subreads used in a consensus sequence flagged as likely coming from PCR cross-over events.

• **Chimeric (%)**: The percentage of subreads used in a consensus sequence flagged as likely coming from PCR cross-over events.

• **Noise**: The number of subreads used in a consensus sequence that has a very low predicted accuracy (<95%) despite sufficient coverage (>20 reads and >10% of all sequences in the current bin) to be called a novel allele.

• **Noise (%)**: The percentage of subreads used in a consensus sequence that has a very low predicted accuracy (<95%) despite sufficient coverage (>20 reads and >10% of all sequences in the current bin) to be called a novel allele.

**Amplicon Consensus > Amplicon Consensus Summary**

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

• **Barcode Name**: A string containing the pair of barcode names (or indices if not available) for which the following metrics apply.

• **Sequence Cluster**: An identifying number given to the cluster of sequences from which this consensus sequence was generated, roughly corresponding to one locus or amplicon.

• **Sequence Phase**: An identifying number given to each phased haplotype within a sequence cluster.

• **Length (bp)**: The length of the consensus amplicon sequence.

• **Estimated Accuracy**: The estimated accuracy of the consensus amplicon sequence.

• **Subreads Coverage**: The number of subreads used to call consensus for this sequence.

**Data > File Downloads**

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

• **Analysis Log**: Log information for the analysis execution.

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

• **Consensus Sequences**: Consensus amplicons that passed all sequence quality filters, in FASTQ and zipped-FASTQ format.

• **Consensus Sequences Statistics (CSV)**: Statistics about consensus amplicons that passed all sequence quality filters, in CSV format.

• **Consensus Sequences by Barcode**: Consensus amplicons that passed all sequence quality filters, per barcode.

• **Chimeric/Noise Consensus Sequences**: Consensus amplicons that failed one or more sequence quality filters, in FASTQ and zipped-FASTQ format.

• **Chimeric/Noise Sequences by Barcodes**: Consensus amplicons that failed one or more sequence quality filters, per barcode.

• **Combined Results**: Consensus amplicons that passed all sequence quality filters (in FASTQ) format, plus CSV files summarizing the inputs and results.
Mapping Application

Use this application to align (or map) data to a user-provided reference sequence. The Mapping application:

- Accepts Continuous Long Reads or HiFi Reads (BAM format) as input. HiFi Reads are reads generated with CCS Analysis whose quality value is equal to or greater than 20.
- Excludes the CCS Analysis from the CCS with Mapping application.
- Maps data 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.

Importing/Exporting Analysis Settings

- Click Import Analysis Settings and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
- Click Export to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

Reference Set (Required)

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

Consolidate Mapped BAMs for IGV (Default = OFF)

- By default, SMRT Link consolidates chunked BAM files for viewing in IGV if the combined size is not more than 10 GB. Setting this option to ON ignores the file size cutoff and consolidates the BAM files.
- Note: This setting can double the amount of storage used by the BAM files, which can be considerable. Make sure to have enough disk space available. This setting may also result in longer run times.

Parameters

<table>
<thead>
<tr>
<th>Advanced 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 semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Minimum Mapped Length (bp)</td>
<td>50</td>
<td>The minimum required mapped read length, in base pairs.</td>
</tr>
<tr>
<td>Bio Sample Name of Aligned Dataset</td>
<td>NONE</td>
<td>Populates the Bio Sample Name (Read Group SM tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. Note: Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.</td>
</tr>
<tr>
<td>Minimum Gap-Compressed Identity (%)</td>
<td>70</td>
<td>The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.</td>
</tr>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
</tbody>
</table>
Reports and Data Files

The Mapping application generates the following reports:

**Target Regions > Target Regions**
Displays the number (and percentage) of reads that hit target regions specified by an input BED file. This is useful for targeted DNA sequencing applications. (This report displays only if a BED file is specified when creating the analysis.)

- **Coordinates:** The chromosome coordinates, as specified in the input BED file.
- **Region:** The name of the region, as specified in the input BED file.
- **On-Target Reads:** The number (and percentage) of unique reads that map with any overlap to the target region.

**Target Regions > Target Region Coverage**
- Displays the number of hits per defined region of the chromosome.

**Mapping Report > Summary Metrics**
Mapping is local alignment of a read or subread to a reference sequence.

- **Percentage of Bases (mapped):** The percentage of bases that mapped to the reference sequence.
- **Number of Subreads (total):** The total number of subreads in the sequence.
- **Number of Subreads (mapped):** The number of subreads that mapped to the reference sequence.
- **Number of Subreads (unmapped):** The number of subreads not mapped to the reference sequence.
- **Percentage of Subreads (mapped):** The percentage of subreads that mapped to the reference sequence.
- **Percentage of Subreads (unmapped):** The percentage of subreads not mapped to the reference sequence.
- **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
- **Number of Subread Bases (mapped):** The number of subread bases that mapped to the reference sequence.
• **Number of Alignments**: The number of alignments that mapped to the reference sequence.
• **Alignment Length Mean (mapped)**: The mean length of alignments that mapped to the reference sequence.
• **Alignment Length N50 (mapped)**: The alignment length at which 50% of the alignments are longer than, or equal to, this value.
• **Alignment Length 95% (mapped)**: The 95th percentile of length of alignments that mapped to the reference sequence.
• **Alignment Length Max (mapped)**: The maximum length of alignments 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.

• **Sample**: Sample name for which the following metrics apply.
• **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. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

**Mapping Report > Mapped Polymerase Read Length**

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

**Mapping Report > Alignment Length**

• Histogram distribution of the number of alignments by the alignment length.
Mapping Report > Alignment Concordance

- Histogram distribution of the number of alignments by the percent concordance with the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Mapping Report > Mapped Concordance vs Alignment Length

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

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 Reference

- Maps coverage across the reference.

Coverage > Depth of Coverage

- Maps the reference regions against the percent coverage.

Coverage > Coverage vs. [GC] Content

- Maps (as a percentage, over a 100 bp window) the number of Gs and Cs present across the coverage. The number of genomic windows with the corresponding % of Gs and Cs is displayed on top. Used to check that no coverage is lost over extremely biased base compositions.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
- **Mapped Reads**: All input reads that were mapped to the reference by the application.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference sequence.
- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.

Data > IGV Visualization Files

The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.
• **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.

• **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.
Mark PCR Duplicates Application

Use this application to remove duplicate reads from a HiFi Reads Data Set created using an ultra-low DNA sequencing protocol.

- The application accepts HiFi Reads (BAM format) as input. HiFi Reads are reads generated with CCS Analysis whose quality value is equal to or greater than 20.

Note: If starting with a very low-input DNA sample using the SMRTbell gDNA Sample Amplification Kit, you must run this application (preceded by the Trim gDNA Amplification Adapters application) on the resulting Data Set prior to running the any other secondary analysis application.

Importing/Exporting Analysis Settings

- Click Import Analysis Settings and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
- Click Export to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

Parameters

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify Duplicates Across Sequencing Libraries</td>
<td>ON</td>
<td>Duplicate reads are identified per sequencing library. The library is specified in the BAM read group LB tag, which is set using the Well Sample Name field in Run Design. By convention, different LB tags correspond to different library preparations. Use this option when the LB tag does not follow this convention to treat all reads as from the same sequencing library.</td>
</tr>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

Reports and Data Files

The Mark PCR Duplicates application generates the following reports:

PCR Duplicates > Duplicate Rate (table)

- Library: The name of the library containing duplicate molecules.
- Unique Molecules: The number of unique molecules in the library.
- Unique Molecules (%): The percentage of unique molecules in the library.
- Duplicate Reads: The number of duplicate reads in the library.
- Duplicate Reads (%): The percentage of duplicate reads in the library.

PCR Duplicates > Duplicate Rate (chart)

- Duplicate Rate: Displays the percentage of duplicate reads per library.
• **Duplicate Reads per Molecule:** Displays the percentage of duplicated molecules per library; broken down by the number of reads per duplicated molecule.

**Data > File Downloads**

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

• **Analysis Log:** Log information for the analysis execution.
• **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose Data > SMRT Link Log.)
• **PCR Duplicates:** BAM file containing duplicate reads with PCR adapters.
• **<Data Set> (deduplicated):** Output Data Set, with duplicate reads with PCR adapters removed.
Microbial Assembly Application

Use this application to generate de novo assemblies of small prokaryotic genomes between 1.9-10 Mb and companion plasmids between 2 – 220 kb.

The Microbial Assembly application:

- Accepts Continuous Long Reads or HiFi Reads (BAM format) as input. HiFi Reads are reads generated with CCS Analysis whose quality value is equal to or greater than 20.
- Includes chromosomal- and plasmid-level de novo genome assembly, circularization, polishing, and rotation of the origin of replication for each circular contig.
- Facilitates assembly of larger genomes (yeast) as well.

Importing/Exporting Analysis Settings

- Click Import Analysis Settings and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
• Click Export to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

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

• The approximate number of base pairs expected in the genome, used to determine the coverage cutoff. **Note:** It is better to slightly overestimate rather than underestimate the genome length to ensure good coverage across the genome.

**Consolidate Mapped BAMs for IGV (Default = OFF)**

• By default, SMRT Link consolidates chunked BAM files for viewing in IGV if the combined size is not more than 10 GB. Setting this option to ON ignores the file size cutoff and consolidates the BAM files. **Note:** This setting can double the amount of storage used by the BAM files, which can be considerable. Make sure to have enough disk space available. This setting may also result in longer run times.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed Length Cutoff</td>
<td>-1</td>
<td>Only reads as long as this value are used as seeds in the draft assembly. -1 means this will be calculated automatically so that the total number of seed bases equals ((\text{Genome Length} \times \text{Coverage})).</td>
</tr>
<tr>
<td>Coverage</td>
<td>30</td>
<td>A target value for the total amount of unique molecular coverage used for assembly. This parameter is used together with the genome size to generate a minimum length cutoff, retaining only subreads longer than this cutoff to include in assembly.</td>
</tr>
<tr>
<td>Advanced Assembly Options</td>
<td>NONE</td>
<td>Allows PacBio Support engineers to override the configuration file generated from other options. This is a semicolon-separated list of KEY=VALUE pairs. New line characters are not accepted.</td>
</tr>
<tr>
<td>Downsamplled Coverage</td>
<td>100</td>
<td>Randomly downsamples to the specified coverage value, then assembles the same coverage. (The full Data Set is still used for alignment and polishing.) This improves contiguity for high-coverage Data Sets.</td>
</tr>
<tr>
<td>Bio Sample Name of Aligned Dataset</td>
<td>NONE</td>
<td>Populates the Bio Sample Name (Read Group SM tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. <strong>Note:</strong> Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

**Reports and Data Files**

The Microbial Assembly application generates the following reports:

**Polished Assembly > Summary Metrics**

Displays statistics on the contigs from the *de novo* assembly that were corrected by Arrow.
• **Polished Contigs**: The number of polished contigs.
• **Maximum Contig Length**: The length of the longest contig.
• **N50 Contig Length**: 50% of the contigs are longer than this value.
• **Sum of Contig Lengths**: Total length of all the contigs.
• **E-size (sum of squares/sum)**: The expected contig size for a random base in the polished contigs.

**Polished Assembly > Polished Contigs**

Displays a table of details about all assembled contigs.

• **Contig**: Contig name.
• **Length (bases)**: The length of the contig, in base pairs, after polishing.
• **Circular**: Marks whether circularity of the contig was detected. Output values are yes and no.
• **Coverage**: The average coverage across the contig, calculated by the sum of coverage of all bases in the contig divided by the number of bases.
• **Mean QV**: The mean QV across the contig.

**Polished Assembly > Contig Confidence vs Coverage**

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

**Alignment to Draft Assembly > Summary Metrics**

Displays statistics on reads that aligned to the draft assembly.

• **Percent Mapped Bases**: The percentage of bases that mapped to the draft assembly.
• **Number of Subreads (total)**: The total number of subreads in the draft assembly.
• **Number of Subreads (mapped)**: The number of subreads that mapped to the draft assembly.
• **Number of Subreads (unmapped)**: The number of subreads not mapped to the draft assembly.
• **Percentage of Subreads (mapped)**: The percentage of subreads that mapped to the draft assembly.
• **Percentage of Subreads (unmapped)**: The percentage of subreads not mapped to the draft assembly.
• **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the draft assembly. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
• **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the draft assembly.
• **Number of Alignments**: The number of alignments that mapped to the draft assembly.
• **Alignment Length Mean (mapped)**: The mean length of alignments that mapped to the draft assembly.
• **Alignment Length N50 (mapped)**: The alignment length at which 50% of the alignments are longer than, or equal to, this value.
• **Alignment Length 95% (mapped)**: The 95th percentile of length of alignments that mapped to the draft assembly.
• **Alignment Length Max (mapped)**: The maximum length of alignments that mapped to the draft assembly.
• **Number of Polymerase Reads (mapped)**: The number of polymerase reads that mapped to the draft assembly. This includes adapters.
• **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped 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 (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 draft assembly.

• **Polymerase Read Length Max (mapped):** The maximum length of polymerase reads that mapped to the draft assembly.

Alignment to Draft Assembly > Alignment Statistics Summary

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

• **Sample:** Sample name for which the following metrics apply.

• **Movie:** Movie name for which the following metrics apply.

• **Number of Polymerase Reads (mapped):** The number of polymerase reads that aligned to the draft assembly. This includes adapters.

• **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that aligned 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 (mapped):** The read length at which 50% of the bases aligned to the draft assembly are in polymerase reads longer than, or equal to, this value.

• **Number of Subreads (mapped):** The number of subreads that aligned to the draft assembly.

• **Number of Subread Bases (mapped):** The number of subread bases that aligned to the draft assembly.

• **Subread Length Mean (mapped):** The mean length of the mapped portion of subreads that aligned to the draft assembly.

• **Mean Concordance (mapped):** The mean concordance of subreads that aligned to the draft assembly. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Alignment to Draft Assembly > Mapped Polymerase Read Length

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

Alignment to Draft Assembly > Mapped Subread Length

• Histogram distribution of the number of alignments by the alignment length.

Alignment to Draft Assembly > Mapped Subread Concordance

• Histogram distribution of the number of subreads against the percent concordance with the subreads that aligned to the draft assembly. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Alignment to Draft Assembly > 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 the *de novo*-assembled genome, as well as depth of coverage distribution.

- **Mean Coverage**: The mean depth of coverage across the assembled genome sequence.
- **Missing Bases**: The percentage of the genome's sequence that have zero depth of coverage.

Coverage > Coverage across Reference
- Displays coverage at each position of the draft genome assembly.

Coverage > Depth of Coverage
- Histogram distribution of the draft assembly regions by the coverage.

Coverage > Coverage vs. [GC] Content
- Maps (as a percentage, over a 100 bp window) the number of Gs and Cs present across the coverage. The number of genomic windows with the corresponding % of Gs and Cs is displayed on top. Used to check that no coverage is lost over extremely biased base compositions.

Preassembly > Summary Metrics
Displays statistics on the pre-assembly process.

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

• **Seed Coverage (bases/genome_size):** The number of seed bases divided by the number of base pairs expected in the genome.

• **Number of Pre-Assembled Reads:** The number of reads output by the pre-assembler. Pre-assembled reads are very long, highly accurate reads that can be used as input to a *de novo* assembler.

• **Pre-Assembled Read Length Mean:** The mean length of the pre-assembled reads.

• **Pre-Assembled Read Length (N50):** 50% of the pre-assembled reads are longer than this value.

• **Pre-Assembled Read Length 95%:** The 95th percentile of the length of the reads output by the pre-assembler.

• **Pre-Assembled E-size (sum of squares/sum):** The expected contig size for a random base in the pre-assembled reads.

• **Number of Pre-Assembled Bases (total):** The total number of bases output by the pre-assembler.

• **Pre-Assembled Coverage (bases/genome_size):** The number of bases output by the pre-assembler divided by the number of base pairs expected in the genome.

• **Pre-Assembled Yield (bases/seed_bases):** The percentage of seed read bases that were successfully aligned to generate pre-assembled reads.

• **Average Number of Reads that Each Seed is Broken Into:** The average number of preliminary reads that each seed is broken into. (Preliminary reads are derived from seeds using error correction; some portions of seeds might be too "noisy" to use.)

• **Average Number of Bases Lost from Each Seed:** The average number of bases from each seed that were completely discarded.

**Data > File Downloads**

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

• **Analysis Log:** Log information for the analysis execution.

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

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

• **Polished Assembly:** The polished assembly before oriC rotation is applied, in FASTA and FASTQ formats.

• **Final Polished Assembly:** The final polished assembly with applied oriC rotation and header adjustment for NCBI submission, in FASTA format.

• **Mapped BAM:** The BAM file of subread alignments to the draft contigs used for polishing.

• **Mapped BAM Index:** The BAI index file for the corresponding Mapped BAM file.

• **Polished Contigs After oriC Rotation:** Polished contigs with oriC rotation applied, before the NCBI adjustment process is used. This is the main output file, and can be used as a reference for downstream analysis.

• **Draft Assembly:** FASTA file containing draft assembly contigs before polishing.

• **Draft Assembly Index:** The FAI index file for the corresponding draft assembly FASTA file.
Data > IGV Visualization Files

The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.

- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.
- **Draft Assembly**: FASTA file containing draft assembly contigs before polishing.
- **Draft Assembly Index**: The FAI index file for the corresponding draft assembly FASTA file.
Use this application to identify and phase minor single nucleotide substitution variants in complex populations. This application is powered by the \textit{juliet} algorithm:

- Accepts \textbf{HiFi Reads} (BAM format) as input. \textbf{HiFi Reads} are reads generated with CCS Analysis whose quality value is equal to or greater than 20.
- Includes reference-based codon amino acid-calling (indel variants not called) in amplicons $\leq$4kb, fully spanned by long reads.
- Includes extensive application reports for the HIV pol coding region, including drug resistance annotation from publicly-available databases.
- Includes reliable 1\% minor variant detection with 6000 high-quality CCS Reads with predicted accuracy of $\geq$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.

**Importing/Exporting Analysis Settings**

- Click \textbf{Import Analysis Settings} and select a previously-saved CSV file containing the desired settings (including \textbf{Advanced Parameters}) for the selected application. The imported application settings are set.

- Click \textbf{Export} to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Reference Set (Required)**

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

**Target Config (Required)**

- Defines genes of interest within the reference and, optionally, drug resistance mutations for specific variants. Minor Variants Analysis contains one predefined target configuration for HIV HXB2. To specify this target configuration, enter \texttt{HIV\_HXB2} into the \textbf{Target Config} field. To specify a \textbf{custom} target configuration for any organism or gene other than HIV HXB2: Enter \textbf{either} the path to the target configuration JSON file on the SMRT Link server, \textbf{or} the entire content of the JSON file.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced 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 \textbf{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>
</tbody>
</table>
Reports and Data Files

The Minor Variants Analysis 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.

**Data > File Downloads**

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

• **Analysis Log**: Log information for the analysis execution.

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

• **Variants Summary**: Data from the Minor Variants Details report, in CSV format.

• **Mapped Reads**: All input reads that were mapped to the reference by the application.

• **Detailed Reports**: Minor variants report information generated, as a ZIP-compressed HTML file. This includes the full report, in human-readable format, and contains four sections:

1. **Input Data**

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

2. **Target Config**

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

```
▼ Target config

Config Version:  Predefined v1.1, PacBio internal
Reference Name: HIV HXB2
Reference Length: 9719
Genes:
  • 5'LTR (1-634)
  • p17 (790-1186)
  • p24 (1186-1879)
  • p2 (1879-1921)
  • p7 (1921-2086)
  • p1 (2086-2134)
  • p6 (2134-2292)
  • **Protease** (2293-2550)
    o AZT/3T: V32I L33F M463 N466 L147V G48V Q48K I50L I54V I54T I54A I54L I54I V82A V82T V82F V82S I84V N88S L90M
    o D47V/7: V32I L33F M463 N466 L147V I47A I50V I54V I54T I54A I54L I54M L76V V82A V82T V82F V82S I84V L90M
    o IDV/R: V32I M461 N461 L147V I54V I54T I54A I54L I54W L76V V82A V82T V82F V82S I84V N88S L90M
    o NVP: D30N L33F M461 N461 L147V G48V Q48K I54V I54T I54A I54L I54I V82A V82T V82F V82S I84V N88S L90M
    o SQV/R: G48V Q48K I54V I54T I54A I54L I54M V82A V82T I84V N88S L90M
    o TTV/R: V32I L33F M463 N466 L147V I47A I54V I54A I54W V82T V82L I84V
```

3. **Variant Discovery**
For each gene/open reading frame, there is one overview table.

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

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

<table>
<thead>
<tr>
<th>HIV</th>
<th>HXB2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codon</td>
<td>AA</td>
</tr>
<tr>
<td>A T G</td>
<td>M</td>
</tr>
<tr>
<td>A A A</td>
<td>K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pos</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td>2947</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-2</td>
<td>2933</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>4</td>
<td>0</td>
<td>2952</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2606</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>339</td>
</tr>
<tr>
<td>1</td>
<td>2905</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2938</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2938</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2942</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2751</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

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

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

<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>

The second type of tooltip is for each haplotype percentage and shows the number of reads contributing to this haplotype:
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 included with the SMRT Link software.

- The application accepts Continuous Long Reads (BAM format) as input.

Importing/Exporting Analysis Settings

- Click Import Analysis Settings and select a previously-saved CSV file containing the desired settings (including Advanced Parameters) for the selected application. The imported application settings are set.
- Click Export to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

Reference Set (Required)

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

Parameters

<table>
<thead>
<tr>
<th>Advanced 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 semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

Reports and Data Files

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

Top Variants > High-Confidence Variant Calls

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

- **Sequence**: The name of the reference sequence.
- **Position**: The position of the variant along the reference sequence.
- **Variant**: The variant position, type, and affected nucleotide.
- **Type**: The variant type: Insertion, Deletion, or Substitution.
- **Coverage**: The coverage at position.
- **Confidence**: The confidence of the variant call.
- **Genotype**: Includes the full number of chromosomes (diploid) or half the number (haploid).
Coverage > Summary Metrics
Displays depth of coverage across references, as well as depth of coverage distribution.

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

Coverage > Coverage across Reference
- Maps coverage across the reference.

Coverage > Depth of Coverage
- Histogram distribution of the percent coverage for reference regions.

Coverage > Coverage vs. [GC] Content
- Maps (as a percentage, over a 100 bp window) the number of Gs and Cs present across the coverage. The number of genomic windows with the corresponding % of Gs and Cs is displayed on top. Used to check that no coverage is lost over extremely biased base compositions.

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 > Variants Across Reference
- Maps the number of variants across the user-selected reference against the reference start position.

Mapping Report > Summary Metrics
Mapping is local alignment of a read or subread to a reference sequence.

- **Percentage of Bases (mapped)**: The percentage of bases that mapped to the reference sequence.
• **Number of Subreads (total):** The total number of subreads in the sequence.
• **Number of Subreads (mapped):** The number of subreads that mapped to the reference sequence.
• **Number of Subreads (unmapped):** The number of subreads not mapped to the reference sequence.
• **Percentage of Subreads (mapped):** The percentage of subreads that mapped to the reference sequence.
• **Percentage of Subreads (unmapped):** The percentage of subreads not mapped to the reference sequence.
• **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
• **Number of Subread Bases (mapped):** The number of subread bases that mapped to the reference sequence.
• **Number of Alignments:** The number of alignments that mapped to the reference sequence.
• **Alignment Length Mean (mapped):** The mean length of alignments that mapped to the reference sequence.
• **Alignment Length N50 (mapped):** The alignment length at which 50% of the alignments are longer than, or equal to, this value.
• **Alignment Length 95% (mapped):** The 95th percentile of length of alignments that mapped to the reference sequence.
• **Alignment Length Max (mapped):** The maximum length of alignments 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.

• **Sample:** Sample name for which the following metrics apply.
• **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.
• **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.
• **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. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

**Mapping Report > Mapped Polymerase Read Length**

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

**Mapping Report > Alignment Length**

• Histogram distribution of the number of alignments by the alignment length.

**Mapping Report > Alignment Concordance**

• Histogram distribution of the number of alignments by the percent concordance with the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

**Mapping Report > Mapped Concordance vs Alignment Length**

• Maps the percent concordance with the reference sequence against the alignment length, in base pairs.

**Site Acceptance Test Report > Summary Metrics**

• **Instrument ID**: The ID number of the PacBio instrument System 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 Log**: Log information for the analysis execution.

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

• **Mapped Reads**: All input reads that were mapped to the reference by the application.

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

• **Consensus FASTA/FASTQ**: Consensus sequences, in FASTA/FASTQ format.

• **Variants**: List of variants from the reference, in BED, GFF or VCF format.
Use this application to identify structural variants (Default: ≥20 bp) in a sample or set of samples relative to a reference. Variant types identified are insertions, deletions, duplications, copy number variants (CNVs), inversions, and translocations.

- The application accepts **Continuous Long Reads** or **HiFi Reads** (BAM format) as input. **HiFi Reads** are reads generated with CCS Analysis whose quality value is equal to or greater than 20.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.

- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**Reference Set (Required)**

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

### Parameters

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Length of Structural Variant (bp)</td>
<td>20</td>
<td>The minimum length of Structural Variant, in base pairs.</td>
</tr>
<tr>
<td>Minimum Length of Copy Number Variant (bp)</td>
<td>1,000</td>
<td>The minimum length of a copy number variant, in base pairs.</td>
</tr>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
<tr>
<td>Minimum % of Reads that Support Variant (any one sample)</td>
<td>20</td>
<td>Ignore calls supported by &lt;P% of reads in every sample.</td>
</tr>
<tr>
<td>Minimum Reads that Support Variant (any one sample)</td>
<td>2</td>
<td>Ignore calls supported by &lt;N reads in every sample.</td>
</tr>
<tr>
<td>Minimum Reads that Support Variant (total over all samples)</td>
<td>2</td>
<td>Ignore calls supported by &lt;N reads total across samples.</td>
</tr>
<tr>
<td>Filters to Add to the Data Set</td>
<td>NONE</td>
<td>A semicolon-separated (not comma-separated) list of other filters to add to the Data Set.</td>
</tr>
<tr>
<td>Minimum Mapped Length (bp)</td>
<td>50</td>
<td>The minimum required mapped read length, in base pairs.</td>
</tr>
<tr>
<td>Minimum Gap-Compressed Identity (%)</td>
<td>70</td>
<td>The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.</td>
</tr>
<tr>
<td>Bio Sample Name of Aligned Dataset</td>
<td>NONE</td>
<td>Populates the Bio Sample Name (Read Group SM tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. <strong>Note:</strong> Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.</td>
</tr>
</tbody>
</table>
To Launch a Multi-Sample Analysis

1. Click + Create New Analysis.
2. Enter a name for the analysis.
3. Select the type of data (Continuous Long Reads or HiFi Reads) to use for the analysis.
4. Select all the Data Sets for all the input samples.
5. In the Analysis of Multiple Data Sets list, select One Analysis for All Data Sets.
6. Click Next.
7. Select Structural Variant Calling from the Analysis Application list.

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

- If multiple Data Sets with the same Bio Sample Name are selected and submitted, the Structural Variant Calling application merges those Data Sets as belonging to the same sample.
- If any input Data Sets do not have a Bio Sample Name specified, they are merged (if there are multiple such Data Sets) and their Bio Sample Name is set to UnnamedSample in the analysis results.

Reports and Data Files

The Structural Variant Calling application generates the following reports:

Report > Count by Sample (SV Type)

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

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

Advanced Parameters

<table>
<thead>
<tr>
<th>Advanced pbmm2 Options</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NONE</td>
<td>Space-separated list of custom pbmm2 options. Not all supported command-line options can be used, and HPC settings cannot be modified. See SMRT® Tools Reference Guide v10.2 for details.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advanced pbsv Options</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
</table>
Report > Count by Sample (Genotype)

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

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

Report > Count by Annotation

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

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

Report > Length Histogram

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

Data > File Downloads

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

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

Data > IGV Visualization Files

The following files are used for visualization using IGV; see “Visualizing Data Using IGV” on page 146 for details.
- **Aligned Reads (per sample):** Aligned reads, in BAM format, separated by individual.
- **Index of Aligned Reads (per sample):** BAM index files associated with the Aligned Reads BAM files.
- **Structural Variants:** All the structural variants, in VCF format. (See [here](#) for details.)
Trim gDNA Amplification Adapters Application

Use this application to trim PCR Adapters from a HiFi Reads Data Set created using an ultra-low DNA sequencing library.

- The application accepts **HiFi Reads** (BAM format) as input. **HiFi Reads** are reads generated with CCS Analysis whose quality value is equal to or greater than 20.

**Note**: If starting with a very low-input DNA sample using the **SMRTbell gDNA Sample Amplification Kit**, you must run this application (followed by the Mark PCR Duplicates application) on the resulting Data Set prior to running the any other secondary analysis application.

**Importing/Exporting Analysis Settings**

- Click **Import Analysis Settings** and select a previously-saved CSV file containing the desired settings (including **Advanced Parameters**) for the selected application. The imported application settings are set.
- Click **Export** to create a CSV file containing all the settings you specified for the application. You can then import this file when creating future analyses using the same application. You can also use this exported file as a template for use with later analyses.

**PCR Adapters (Required)**

- Specify the file of PCR adapters used during library preparation of an ultra-low DNA sequencing library to be trimmed from the sequenced data.

**Parameters**

<table>
<thead>
<tr>
<th>Advanced Parameters</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. CCS Predicted Accuracy (Phred Scale)</td>
<td>20</td>
<td>Phred-scale integer QV cutoff for filtering HiFi Reads. The default for all applications (except Iso-Seq Analysis) is 20 (QV 20), or 99% predicted accuracy.</td>
</tr>
<tr>
<td>Compute Settings</td>
<td>Select</td>
<td>(Optional) Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.</td>
</tr>
</tbody>
</table>

**Reports and Data Files**

The Trim gDNA Amplification Adapters application generates the following reports:

**PCR Adapters > Summary Metrics**

- **Unique PCR Adapters**: The number of unique PCR adapters in the sequence data.
- **Reads with PCR Adapters**: The number of reads in the sequence data that contain PCR adapters.
- **Mean Reads Per Adapter**: The mean number of reads per PCR adapter in the sequence data.
- **Max. Reads Per Adapter**: The maximum number of reads per PCR adapter in the sequence data.
- **Min. Reads Per Adapter**: The minimum number of reads per PCR adapter in the sequence data.
• **Mean Read Length**: The mean read length of reads per PCR adapter in the sequence data.

• **Reads Without PCR Adapters**: The number of reads without PCR adapters in the sequence data.

• **Percent Bases in Reads with Adapters**: The percentage of bases in reads in the sequence data that contain PCR adapters.

• **Percent Reads with Adapters**: The percentage of reads in the sequence data that contain PCR adapters.

**PCR Adapters > PCR Adapter Data**

• **Bio Sample Name**: The name of the biological sample associated with the PCR adapters.

• **PCR Adapter Name**: A string containing the pair of PCR adapter indices for which the following metrics apply.

• **Polymerase Reads**: The number of polymerase reads associated with the PCR adapter.

• **Bases**: The number of bases associated with the PCR adapter.

• **Mean PCR Adapter Quality**: The mean PCR adapter quality associated with the PCR adapter.

**PCR Adapters > PCR Adapter Quality Scores**

• Histogram distribution of PCR adapter Quality scores. The scores range from 0-100, with 100 being a perfect match.

**Data > File Downloads**

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

• **Analysis Log**: Log information for the analysis execution.

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

• **Reads Missing Adapters**: Reads Missing Adapters: BAM file containing the reads with missing PCR adapters from the input Data Set.

• **PCR Adapter CSV**: Includes the data displayed in the PCR Adapter Data table.

• **Diagnostic Log**: Lists how many ZMWs were filtered, how many ZMWs are the same or different, and how many reads were filtered.

• **<Data Set> (trimmed)**: Output Data Set, with the PCR adapters removed.
Working with Barcoded Data

This section describes how to use SMRT Link to work with barcoded data. Demultiplex Barcodes analysis is powered by the Lima SMRT Analysis tool.

The canned data provided with SMRT Link v10.2 includes 10 barcode sets:

- gDNA_Sample_Amplification_Adapter
- IsoSeq_Primers_12_Barcodes_v1
- IsoSeqPrimers_v2 (Includes the content of IsoSeqPrimers as well as support for NEB and Clontech primers.)
- Sequel_16_barcodes_v1
- Sequel_16_barcodes_v3
- Sequel_96_barcodes_v1
- Sequel_384_barcodes_v1
- Sequel_64_M13barcodes_v1
- HiFiViral_SARS-CoV-2_M13barcodes
- SMRTbell Barcoded Adapter Plate 3.0 (bc2001-bc2096)

SMRT Link v10.2 supports sample traceability through the various modules in the application by using the Bio Sample Name.

Run Design in SMRT Link v10.2 contains a required Bio Sample Name field for both single and multiplexed samples.

- For multiplexed experiments, SMRT Link provides default names for one Bio Sample Name per barcode, which can be edited as needed in the Barcoded Sample Names File.
- For multiplexed Iso-Seq Analysis only, Bio Sample Names are not required.

Well Sample Name and Bio Sample Name entered in Sequel II Systems, and in Sequel System Run Designs for multiplexed runs:

- Display as column values in the Data Management and SMRT Analysis modules.
- Display as Data Set attributes in the Data Set details page in Data Management.
- Populate the LB and SM tags in read group headers of BAM files containing basecalled data.

Example Well Sample Names and Bio Sample Names

- Non-Barcoded Well Sample Name: HG002_2019_11_2_10K
- Non-Barcoded Bio Sample Name: HG002
- Barcoded Well Sample Name: My Multiplexed Set of Bugs
- Barcoded Bio Sample Name: Unknown Microbe 1,...,Unknown Microbe N
Step 1: Specify the Barcode Setup and Sample Names in a Run Design

**Note:** If you specified the barcode setup in Run design, the demultiplexing is performed **automatically** after the data is transferred to the SMRT link server. You can also specify the barcode setup **manually** by selecting **SMRT Analysis > Create New Analysis** and then selecting the Demultiplex Barcodes application.

1. In SMRT Link, create a new Run Design as described in “Creating a New Run Design” on page 15. **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 display.

3. Specify a **Barcode Set** using the dropdown list. **Note:** You can specify up to 10,000 samples. Specifying more than 10,000 samples may cause a delay of several minutes in analysis submission.

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 column of automatically-generated Bio Sample Names 1 through N, corresponding to barcodes 1 through N, for the biological sample names. There are **two different ways** to specify which barcodes to use, and assign biological sample names to
barcodes. (Note: Bio Sample Names are hardcoded and can be traced through secondary analysis using SMRT Analysis.)

Interactively:

• Click Interactively, then drag barcodes from the Available Barcodes column to the Included Barcodes column. (Use the check boxes to select multiple barcodes.)

• (Optional) Click a Bio Sample field to edit the Bio Sample Name associated with a barcode. Note: Avoid using spaces in Bio Sample Names as they may lead to third-party compatibility issues.

• (Optional) Click Download as a file for later use.
• Click Save to save the edited barcodes/Bio Sample names. You see Success on the line below, assuming the file is formatted correctly.

From a File:

• Click From a File, then click Download File. Edit the file and enter the biological sample names associated with the barcodes in the second column, then save the file. Use alphanumeric characters, spaces (allowed but not recommended for compatibility with third-party downstream software), hyphens, underscores, colons, or periods only - other characters will be removed automatically, with a maximum of 40 characters. If you did not use all barcodes in the Autofilled Barcode Name file in the sequencing run, delete those rows.

• Note: Open the CSV file in a text editor and check that the columns are separated by commas, not semicolons or tabs.

• Select the Barcoded Sample File you just edited. You see Success on the line below, assuming the file is formatted correctly.

6. Click Save.

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 Systems. 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) 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 analysis. If used, the analysis parameter Filters to add to the Data Set overrides other barcode filtering, even if the barcode score set with it is lower than 26.
Step 3: (Optional) Run the Demultiplex Barcodes Application

If instead 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 the type of data to use for the analysis:
   - **Continuous Long Reads**: Subreads from Sequel Systems.
   - **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.

The Data Sets box displays the appropriate Data Sets available for the analysis.

4. In the Data Sets box, select one or more Data Sets to be analyzed together.
5. Click Next.
6. Select Demultiplex Barcodes from the Applications list.
7. Specify a Barcode Set (barcode sequence file.)
   - **Note**: You can specify up to 10,000 samples. Specifying more than 10,000 samples may cause a delay of several minutes in analysis submission.
8. 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 data retains any barcode pair combination identified in the Data Set.
9. SMRT Link automatically creates a CSV-format Autofilled Barcoded Sample 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 column of automatically-generated Bio Sample Names 1 through N, corresponding to barcodes 1 through N, for the biological sample names. There are two different ways to specify which barcodes to use, and assign biological sample names to barcodes:

   - **Interactively**:
     - Click Interactively, then drag barcodes from the Available Barcodes column to the Included Barcodes column. (Use the check boxes to select multiple barcodes.)
     - **(Optional)** Click a Bio Sample field to edit the Bio Sample Name associated with a barcode. **Note**: Avoid using spaces in Bio Sample Names as they may lead to third-party compatibility issues.
     - **(Optional)** Click Download as a file for later use.
• Click **Submit** to save the edited barcodes/bio sample names. You see **Success** on the line below, assuming the file is formatted correctly.

**From a File:**

• Click **From a File**, then click **Download File**. Edit the file and enter the biological sample names associated with the barcodes in the second column, then save the file. Use alphanumeric characters, spaces (allowed but **not recommended** for compatibility with third-party downstream software), hyphens, underscores, colons, or periods **only** - other characters will be removed **automatically**, with a maximum of 40 characters. If you did **not** use all barcodes in the Autofilled Barcode Name file in the sequencing run, **delete** those rows.

• **Note:** Open the CSV file in a text editor and check that the columns are separated by **commas**, not semicolons or tabs.

• Select the **Barcoded Sample File** you just edited. You see **Success** on the line below, assuming the file is formatted correctly.

10. Specify the **name** for the new demultiplexed Data Set that will display in SMRT Link. The application creates a copy of the input Data Set, renames it to the name specified, and creates demultiplexed child Data Sets linked to it. The input Data Set remains separate and unmodified.

11. **(Optional)** Specify any advanced parameters.

12. 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 82.

---

**Step 4: Run Applications Using the Demultiplexed Data as Input**

All secondary analysis applications except **Demultiplex Barcodes** can use demultiplexed Data Sets as input.

**Note:** For **Iso-Seq** analysis with barcoded samples, use the Iso-Seq application **instead** of the Demultiplex Barcodes application, as the Iso-Seq application **already** includes the demultiplexing step as part of the pipeline. When performing multiplexed Iso-Seq analysis, ensure that the **Run Design Sample Is Barcoded** option is set to **No** (the default setting). Then, in SMRT Analysis, go straight to the Iso-Seq application and, in the parameters section, select a Primer Set containing multiple primers, such as *IsoSeq_Primers_12_Barcodes_v1*.

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

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

– One Analysis for All Data Sets: Runs one analysis using all the selected barcode Data Sets as input, for a maximum of 30 Data Sets.

– One Analysis per Data Set - Identical Parameters: Runs one separate analysis for each of the selected barcode Data Sets, using the same parameters, for a maximum of 10,000 Data Sets. Optionally click Advanced Parameters and modify parameters.

– One Analysis per Data Set - Custom Parameters: Runs one 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 Parameters and modify parameters. Then click Start and Create Next. You can then specify parameters for each of the included barcode Data Sets.

– Note: The number of Data Sets listed is based on testing using PacBio's suggested compute configuration, listed in SMRT Link Software Installation (v10.2).

4. Click Start to submit the analysis.
The Demultiplex Barcodes application identifies barcode sequences in PacBio single-molecule sequencing data. It replaced pbbarcode and bam2bam for demultiplexing, starting with SMRT Analysis v5.1.0.

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

1. Barcoded target-specific primers
2. Barcoded universal primers
3. Barcoded overhang adapters
4. Barcoded linear adapters (target capture)

In addition, there are three different barcode library designs.
The **Demultiplex Barcodes** application in SMRT Link supports demultiplexing of subreads. The following terminology is based on the per (sub-) read view.


**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. Symmetric barcoding is used for samples constructed using Barcoded Overhang Adapters, Barcoded Universal Primer and target enrichment (linear). This is also the default scoring mode in SMRT Link v10.2 and later.

**Asymmetric Mode**

Barcode sequences are **different** on the ends of the SMRTbell template. Asymmetric mode is used with the M13 barcoding procedure. (See the document [Procedure & Checklist - Preparing SMRTbell Libraries using PacBio Barcoded M13 Primers for Multiplex SMRT Sequencing](https://pacificbiomes.org/images/Tools/SMRT/ProcedureChecklist_SMRTbell_multiplex_M13.pdf) for details.) Pacific Biosciences recommends using this mode **only** for small inserts (up to 5 kb) where both ends of the insert are expected to be sequenced. **Both** barcodes must be detected.

**Note:** For both Symmetric and Asymmetric modes, the limit for unique individual barcode sequences is 768, and the limit for the number of different barcode pairs is 10,000.

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 with combined symmetric and asymmetric barcoding 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.
Automated Analysis

**Auto Analysis** and **Pre Analysis** allow a specific analysis to be automatically run after a sequencing run has finished and the data is transferred to the SMRT Link Server. The analysis can include demultiplexed output.

- Auto Analysis can be set up in Run Design or SMRT Analysis after the Run Design is saved and before the run is loaded on the instrument.
- Auto Analysis can be run on HiFi Reads or Continuous Long Reads, and includes all analysis applications available for the corresponding data type.
- Auto Analysis works with all Sequel Systems.

**Pre Analysis** is the process of CCS Analysis and/or demultiplexing of Sequel basecalled data. Pre Analysis occurs before Auto Analysis, and is defined when you create a Run Design and specify one or more of the following:

- Read Type = HiFi Reads and Generate HiFi Reads = On Instrument or In SMRT Link.
- Read Type = HiFi Reads and Sample is Barcoded = Yes.
- Read Type = Continuous Long Reads and Sample is Barcoded = Yes.

**Note:** Pre Analysis works with all Sequel Systems.

**Creating Auto Analysis From a Run Design**

1. Create a new Run Design (See “Creating a New Run Design” on page 15 for details) and save it. The Auto Analysis button is enabled only after you save the Run Design.
2. Click Auto Analysis. This takes you into SMRT Analysis, where you create the new analysis that will be associated with the collection.
3. Name the new analysis.
4. Click the numbered Collections link (Column 2 of the Runs table) associated with the run that you defined in Step 1. (Note: Runs display here only if they are in the Created state - not if they are already running or have completed.)
5. Select a collection for analysis.
6. Click Next.
7. Select a secondary analysis application to use for the analysis.
8. (Optional) Click Advanced 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 51.
9. Click Create.
HiFiViral SARS-CoV-2: Creating Auto Analysis in Run Design

The HiFiViral SARS-CoV-2 Application includes a streamlined version of Auto Analysis as part of creating a Run Design.

1. In Run Design, click Create New Design.
2. From the Application list, select HiFiViral SARS-CoV-2. Preloaded default values display in green.
3. Enter the Well Sample Name.
4. By default, Auto Analysis is set to Yes for Automatic Launch of SARS-CoV-2 Analysis.
5. Enter the Analysis Name.
6. By default, Yes is selected for Sample Is Barcoded and No for Same Barcode on Both End of Sequence.
7. By default, the barcode set HiFiViral SARS-CoV-2 M13barcodes is selected.
8. For Assign Bio Sample Names to Barcodes, select From a File and then click Download File.
9. Open the downloaded file (assayPlateQC_template_4by96.csv) in a text editor. See “Plate QC CSV (Optional)” on page 92 for details on modifying this file for your samples.

Creating Auto Analysis Directly From SMRT Analysis

1. Select SMRT Analysis. Click + Create New Analysis.
2. Enter a name for the analysis.
3. Click Auto Analysis. The table displays all runs available for use with Auto Analysis.
4. Follow the procedure Creating Auto Analysis from a Run Design, starting at Step 4.

Getting Information About Analyses Created by Auto Analysis

There are several ways to obtain information on the state of an analysis created using the Auto Analysis feature.

From SMRT Analysis:

1. On the Home Page, select SMRT Analysis. You see a list of all analyses.
2. To filter the analyses, click Show to remove all filters, then click the Created button. This displays only analyses in the Created state.
3. Click the analysis of interest.
4. Click the From Multi-Job link.
5. Click Analysis Overview > Status of Individual Analyses. This displays information about the analysis, including the application used.
From Run Design:

1. On the Home Page, select Run Design.
2. Click the Run Design of interest.
3. Click the From Multi-Job link.
4. Scroll all the way to the right in the table. This displays information about the samples included in the run.
5. Click the Auto Analysis ID link for a sample. This displays information about the analysis, including the application used.

Getting Information About Pre Analysis From SMRT Analysis

1. On the Home Page, select SMRT Analysis. You see a list of all analyses.
2. To filter the analyses, click Show to remove all filters, then click the Created button. This displays only analyses in the Created state.
3. Click the analysis of interest.
4. Click the Pre Analysis link.
5. Click Analysis Overview > Status of Individual Analyses. This displays information about the Pre Analysis, including the application used.

Getting Information About Pre Analysis From Run Design

1. On the Home Page, select Run Design.
2. Click the Run Design of interest.
3. On the left side (above the consumables list), click the Pre Analysis ID link. This displays information about the Pre Analysis, including the application used.
Visualizing Data Using IGV

Once an analysis has successfully completed, visualize the results using the Integrative Genomics Viewer (IGV).

- See here for further installation instruction and usage details.
- See here for PacBio-specific settings and visualizations.

You can visualize data generated by the following secondary analysis applications:

- Assembly (HGAP4)
- Base Modification Analysis
- CCS with Mapping
- Genome Assembly
- Iso-Seq
- Mapping
- Microbial Assembly
- Structural Variant Calling

IGV requires the following files for visualization:

- One consolidated alignment BAM file
- BAM index file
- Genome reference file

If an analysis generates multiple alignment BAM files, those files must first be combined into one consolidated alignment BAM file for visualization with IGV.

SMRT Link defaults to combining chunked alignment BAM files if the combined file sizes are 10 GB or less.

- When creating an analysis, you can specify that SMRT Link combines alignment BAM files for IGV visualization by setting the Consolidate Mapped BAMs for IGV option to ON.

  Note: This setting doubles the amount of storage used by the BAM files, which can be considerable. Make sure to have enough disk space available. This setting may also result in longer run times.

To visualize data using IGV

1. Create and run your analysis.
2. After the analysis has finished successfully, go to the Data > IGV Visualization Files section of the analysis report page.
3. Open IGV and select the reference genome used for the analysis. (See here for instructions on how to load a genome.)
4. Copy a BAM file link from the **Data > IGV Visualization Files** section of the analysis report page.

**Note:** If you are performing *de novo* assembly, you **must** use links to the **draft** assembly BAM files, which are clearly labeled.

5. In IGV, choose **File > Load from URL...** and paste the link into the File URL input field. Click **OK**.
6. Repeat for the remaining links.

If you ran an analysis and there are **no Data > IGV Visualization Files** links, the analysis generated multiple alignment BAM files over 10 GB, but did **not** consolidate the files. Click the **Launch BAM Consolidation** button to consolidate them.
Using the PacBio® Self-Signed SSL Certificate

SMRT Link v10.2 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 SSL Certificate Alert](image)

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

![Advanced SSL Certificate Alert](image)

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

![Closed SSL Certificate Alert](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® Systems Output Files

This section describes the data generated by the Sequel Ile, Sequel II, and Sequel Systems for each SMRT Cell transferred to network storage.

Sequel Ile System Output Files

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

```
<your_specified_output_directory>/r64009_20200825_221039/1_A01/
|-- m64009_200825_222052.baz2bam_1.log
|-- m64009_200825_222052.ccs.log
|-- m64009_200825_222052.ccs_reports.json
|-- m64009_200825_222052.ccs_reports.txt
|-- m64009_200825_222052.consensusreadset.xml
|-- m64009_200825_222052.reads.bam
|-- m64009_200825_222052.reads.bam.pbi
|-- m64009_200825_222052.sts.xml
|-- m64009_200825_222052.transferdone
|-- m64009_200825_222052.zmw_metrics.json.gz
```

In this example:

- `/r64009_20200825_221039` is a directory containing the output files associated with one run.
- `r64009` is the instrument ID number.
- `20200825_222052` 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 includes the following output files:

- `baz2bam_1.log`: Log file for post-primary analysis processing.
- `ccs_reports.json, ccs_reports.txt`: Contains processing metrics summarizing how many ZMWs generated HiFi Reads, and how many ZMWs failed CCS Reads generation. These files contain the same information, and are used internally by PacBio Technical support.
- `consensusreadset.xml`: This file is needed to import data into SMRT Link.
- `sts.xml`: Contains summary statistics about the collection/cell and its post-processing.
- `transferdone`: Contains a list of files successfully transferred.
- `zmw_metrics.json.gz`: Contains processing information used to generate RunQC plots.
- `reads.bam.pbi`: Provides backwards-compatibility with the APIs enabled for accessing the `cmp.h5` file.
reads.bam: The Sequel IIe System outputs one reads.bam file per collection/cell, containing one read per productive ZMW. The file includes:
- HiFi Reads (QV 20 or higher)
- Lower-quality but still polished consensus reads (QV 1 - QV 20)
- Unpolished consensus reads (RQ=-1)
- 0- or 1-pass subreads unaltered (RQ=-1)

Note: reads.bam contains HiFi Reads and non-HiFi Reads, and should not be used by itself as input for non-SMRT Link tools that expect ≥ QV 20.

- 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 all Sequel Systems are fully compatible with the BAM specification. For more information on the BAM file format specifications, click here.

Note: If CCS Analysis is run on the Sequel IIe System, the subreads.bam, scraps.bam and scraps.bam.pbi files are no longer generated and available. If CCS Analysis is run in SMRT Link, Sequel IIe instrument output includes the subreads.bam, scraps.bam and scraps.bam.pbi files.

HiFi Reads Generation

An on-instrument CCS analysis generates a reads.bam file and transfers it to the network server. The reads.bam file contains HiFi Reads and non-HiFi Reads, and should not be used unfiltered as input for tools that expect ≥ QV 20. SMRT Link automatically launches an Export Reads analysis on the reads.bam to filter out the HiFi Reads, and generates the following HiFi data files by default:

- <Movie_Name>.hifi_reads.fastq.gz - Gzipped FASTQ file containing HiFi Reads.
- <Movie_Name>.hifi_reads.fasta.gz - Gzipped FASTA file containing HiFi Reads.
- <Movie_Name>.hifi_reads.bam - BAM file containing HiFi Reads.

If not using SMRT Link for subsequent analysis, please use these three files as input with any third-party analysis tools.

Finding the hifi_reads Files Generated Using On-Instrument CCS

1. In Run QC, click the desired run, then click the sample name to view the CCS Data Set.
2. Click Analyses in the left-side panel.
3. Click the Export Reads analysis.
4. To locate the directory containing the three `hifi_reads` files, append `/outputs` to the path shown.

Sequel II and Sequel Systems Output Files

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

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

Files output by the Sequel II and Sequel Systems include:

- `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.

- `subreads.bam`: The Sequel II and Sequel Systems output 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 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.

- subreads.bam.pbi: Provides backwards-compatibility with the APIs enabled for accessing the cmp.h5 file.
- subreadset.xml: This file is needed to import data into SMRT Link.
- sts.xml: Contains summary statistics about the collection/cell and its post-processing.
- transferdone: Contains a list of files successfully transferred.

### Frequently Asked Questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
</table>
| **What are the minimum files needed to analyze data on SMRT Link?**       | • .bam file  
• bam.pbi file  
• subreadset.xml file                                                                                                                                 |
| **What is the average size of the file bundle for a 6-hour movie?**       | Approximately 5 Gb.                                                                                                                                              |
| **What is the difference between a regular .bam file and an aligned.bam file?** | The subreads.bam file contains all the subreads sequences, while the aligned.bam file additionally contains the genomic coordinates of the reads mapped to a reference sequence.  

The subreads.bam file is created by the PacBio Sequel Systems, while the aligned.bam file is created by SMRT Link after running Mapping analysis applications.
Secondary Analysis Output Files

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

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

```bash
$SMRT_ROOT/userdata/jobs_root/0000/0000000/0000000002/
├── cromwell-job -> $SMRT_ROOT/userdata/jobs-root/cromwell-executions/
│   └── pb_demux_subreads_auto/24e691c8-8d0d-4670-9db3-c7cb1126e8f8
│       └── entry-points
│           └── ae6f1c2c-b4a2-41cc-8e44-98b494f12a57.subreadset.xml
│       └── logs
│           ├── pb_simple_mapping
│           │   └── 24e691c8-8d0d-4670-9db3-c7cb1126e8f8
│           │       └── call-mapping
│           │           └── execution
│           │               └── stderr
│           │               └── stdout
│           │       └── workflow.24e691c8-8d0d-4670-9db3-c7cb1126e8f8.log
│           └── outputs
│               └── mapping.report.json -> $SMRT_ROOT/userdata/jobs-root/cromwell-executions/
│                   └── pb_simple_mapping/24e691c8-8d0d-4670-9db3-c7cb1126e8f8/call-mapping/execution/
│                       └── mapping.report.json
│               └── mapped.bam -> $SMRT_ROOT/userdata/jobs-root/cromwell-executions/
│                   └── pb_simple_mapping/24e691c8-8d0d-4670-9db3-c7cb1126e8f8/call-mapping/execution/
│                       └── mapped.bam
│   └── pbscala-job.stdout
│   └── pbscala-job.stderr
│       └── workflow
│           └── analysis-options.json
│           └── datastore.json
│           └── engine-options.json
│           └── inputs.json
│           └── metadata.json
│           └── metadata-summary.json
│           └── task-timings.metadata.json
│           └── timing-diagram.html
```

- logs/: Contains log files for the analysis job.
  - workflow.<UUID>.log: Global log of each significant step in the analysis and snippets from a task’s stderr output if the analysis failed.
  - The same directory contains stdout and stderr for individual tasks.
- cromwell-job/: Symbolic link to the actual Cromwell execution directory, which resides in another part of the jobs-root directory. Contains subdirectories for each workflow task, along with executable scripts, output files, and stderr/stdout for the task.
  - call-tool_name/execution/: Example of an individual task directory (This is replaced with <task_id> below.)
  - <task_id>/stdout: General task stdout log collection.
– `<task_id>/stderr`: General task stderr log collection.
– `<task_id>/script`: The SMRT® Tools command for the given analysis task.
– `<task_id>/script.submit`: The JMS submission script wrapping run.sh.
– `<task_id>/stdout.submit`: The stdout collection for the script.submit script.
– `<task_id>/stderr.submit`: The stderr collection for the script.submit script.

• `workflow/`: Contains JSON files for analysis settings and workflow diagrams.
– `datastore.json`: JSON file representing all output files imported by SMRT Link.

• `outputs/`: A directory containing symbolic links to all datastore files, which reside in the Cromwell execution directory. This is provided as a convenience and is not intended as a stable API; note that external resources from dataset XML and report JSON file are not included here. Demultiplexing outputs are nested in additional subdirectories.
– `pbscala-job.stdout`: Log collection of stdout output from the SMRT Link job manager.
– `pbscala-job.stderr`: Log collection of stderr output from the SMRT Link job manager. (Note: This is the file displayed as Data > SMRT Link Log on the Analysis Results page.)

A SMRT Link Analysis job generates several types of output files. You can use these data files as input for further processing, pass on to collaborators, or upload to public genome sites. Depending on the analysis application being used, the output directory contain files in the following formats:

• **BAM**: Binary version of the Sequence Alignment Map (SAM) format. (See here for details.)
• **BAI**: The samtools index file for a file generated in the BAM format.
• **BED**: Format that defines the data lines displayed in an annotation track. (See here for details.)
• **CSV**: Comma-Separated Values file. Can be viewed using Microsoft Excel or a text editor.
• **FASTA/FASTQ**: Sequence files that contains either nucleic acid sequence (such as DNA) or protein sequence information. FASTA/Q files store multiple sequences in a single file. FASTQ files also include per-base quality scores. (See here or here for details.)
• **GFF**: General Feature Format, used for describing genes and other features associated with DNA, RNA and Protein sequences. (See here for details.)
• **PBI**: PacBio index file. (This is a PacBio-specific file type.)
• **VCF**: Variant Call Format, for use with the molecular visualization and analysis program VMD. (See here for details.)
To Download Data Files Created by SMRT Link:

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

**LDAP**

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

- For details on integrating LDAP and SMRT Link, see the document `SMRT Link Software Installation (v10.2)`.

**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 (v10.2)`. 
The following procedures are available only for SMRT Link users whose role is **Admin**.

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

1. Choose **Gear > Configure**, then click **User Management**.
2. There are two ways to find users:
   - To display all SMRT Link users: Click **Display all Enabled Users**.
   - To find a specific user: Enter a user name, or partial name, and click **Search By Name**.
3. Click the desired user.
4. Click the **Role** field and select one of the three roles. (A blank role means that this user cannot access SMRT Link.)
• **Note:** There can be **multiple** users with the Admin role; but there **must** always be at least **one** Admin user.

5. Click **Save**.
Hardware/Software Requirements

**Client Hardware Requirements**
SMRT Link requires a minimum screen resolution of 1600 by 900 pixels.

**Client Software Requirements**
- SMRT Link requires the Google® Chrome web browser, version 90 or later.

**Note:** SMRT Link Server hardware and software requirement are listed in the document SMRT Link Software Installation (v10.2).
Appendix A - 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.

• **adaptive loading**: Uses active monitoring of the ZMW loading process to predict a favorable loading end point.

• **unique molecular yield**: The sum total length of unique single molecules that were sequenced. It is calculated as the sum of per-ZMW median subread lengths.

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, base modification analysis, and so on.

• **longest subread length**: The mean of the maximum subread length per ZMW.
• **insert length**: The length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters.

• **Continuous Long Reads (CLR)**: Reads with a subread length approximately equivalent to the polymerase read length indicating that the sequence is generated from a single continuous template from start to finish. Continuous Long Reads are the longest possible reads.

• **circular consensus (CCS) reads**: The consensus sequence resulting from alignment between subreads taken from a single ZMW. Generating CCS Reads does not include or require alignment against a reference sequence but does require at least two full-pass subreads from the insert. CCS Reads are generated with CCS Analysis. CCS Reads with quality value equal to or greater than 20 are called HiFi Reads.

• **HiFi Reads**: Reads generated with CCS Analysis whose quality value is equal to or greater than 20.

---

**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, RNA and epigenomics analysis.

• **consensus**: Generation of a consensus sequence from multiple-sequence alignment.

• **filtering**: Removes reads that do not meet the Read Length criteria set by the user.

• **mapping**: Local alignment of a read or subread to a reference sequence.

• **Auto Analysis**: Allows a specific analysis to be automatically run after a sequencing run has finished and the data is transferred to the SMRT Link Server. The analysis can include demultiplexed outputs.
  - Auto Analysis works with all Sequel Systems.

• **Pre Analysis**: The process of CCS Analysis and/or demultiplexing of Sequel basecalled data. Pre Analysis occurs **before** Auto Analysis.
  - Pre Analysis works with all Sequel Systems.

Accuracy Terminology

• **circular consensus accuracy**: Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.

• **consensus accuracy**: Accuracy based on aligning multiple sequencing reads or subreads together.
• **polymerase read quality**: A trained prediction of a read’s mapped accuracy based on its pulse and base file characteristics (peak signal-to-noise ratio, inter-pulse distance, and so on).
Appendix B - Data Search

Use this function to search for analyses, Data Sets, barcode files, or reference files.

To Search the Entire Table
1. Enter a text query into the Search box. This searches every field in the table, and displays all table rows containing the search characters.

To Search for a Value Within a Column
1. Click the small filter icon at the right of the column name.
2. Enter a value; all table rows meeting the search criteria display. (To select a different search operator, click the droplist and select another search operator. Different search operators are available, based on the column’s data type.)

- For the Analysis State column only, click one or more of the analysis states of interest: Select All, Created, Running, Submitted, Terminated, Successful, Failed, or Aborted.
- For Date fields only, click the small calendar and select a date.

Numeric Field Operators
- Equals, Not equal
- Greater than, Greater than or equals
– Less than, Less than or equals
– In range

Text Field Operators
– Contains, Not contains
– Equals, Not equal
– Starts with, Ends with

Date Field Operators
– Equals, Not equal
– Greater than, Less than
– In range
Appendix C - BED File Format for Target Regions Report

With the CCS with Mapping and Mapping applications, an optional Target Regions report can be generated that displays the number (and percentage) of reads and subreads that hit specified target regions.

The BED file required to generate the Target Regions report includes the following fields; with one entry per line:

1. **chrom**: The name of the chromosome (such as chr3, chrY, chr2_random) or scaffold (such as scaffold10671).
2. **chromStart**: The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd**: The ending position of the feature in the chromosome or scaffold. The chromEnd base is not included in the display of the feature, however, the number in position format is represented. For example, the first 100 bases of chromosome 1 are defined as chrom=1, chromStart=0, chromEnd=100, and span the bases numbered 0-99 (not 0-100), but will represent the position notation chr1:1-100.
4. **(Optional) Region Name**.

**Example**: lambda_NEB3011 15000 25000 Region2

- Fields can be space- or tab-delimited.
- See [here](#) for details of the BED format.
- For details on the BED format's counting system, see [here](#) and [here](#).
Appendix D - Additional Information Included in the CCS Data Set

Export Report

When you export a Data Set and select Export PDF Reports, a report is produced which includes additional fields, listed below.

- See “Exporting Sequence, Reference and Barcode Data” on page 40 for details on exporting Data Sets.
- The other fields and plots in this report are described in the appropriate Reports sections of “PacBio® Secondary Analysis Applications” on page 51.

• **ZMWs input**: The total number of ZMWs used as input in the Data Set.
• **ZMWs pass filters**: The number of ZMWs that passed all the filters.
• **ZMWs fail filters**: The number of ZMWs that failed any of the filters.
• **ZMWs shortcut filters**: The number of low-pass ZMWs skipped using the --all filter.
• **ZMWs with tandem repeats**: The number of ZMWs that did not generate CCS Reads due to repeats larger than --min-tandem-repeat-length.
• **Below SNR threshold**: The number of ZMWs that did not generate CCS Reads due to SNR below --min-snr.
• **Median length filter**: The number of ZMWs that did not generate CCS Reads due to subreads that are <50% or >200% of the median subread length.
• **Lacking full passes**: The number of ZMWs that did not generate CCS Reads due to having fewer than --min-passes full-length subreads.
• **Heteroduplex insertions**: The number of ZMWs that did not generate CCS Reads due to single-strand artifacts.
• **Coverage drops**: The number of ZMWs that did not generate CCS Reads due to coverage drops that would lead to unreliable polishing results.
• **Insufficient draft cov**: The number of ZMWs that did not generate CCS Reads due to not having enough subreads aligned to the draft sequence end-to-end.
• **Draft too different**: The number of ZMWs that did not generate CCS Reads due to having fewer than --min-passes full-length reads aligned to the draft sequence.
• **Draft generation error**: The number of ZMWs that did not generate CCS Reads due to subreads that don't agree enough to generate a draft sequence.
• **Draft above --max-length**: The number of ZMWs that did not generate CCS Reads due to a draft sequence longer than --max-length.
• **Draft below --min-length**: The number of ZMWs that did not generate CCS Reads due to a draft sequence shorter than --min-length.
• **Reads failed polishing**: The number of ZMWs that did not generate CCS Reads due to too many subreads dropped while polishing.

• **Empty coverage windows**: The number of ZMWs that did not generate CCS Reads because at least one window had no coverage.

• **CCS did not converge**: The number of ZMWs that did not generate CCS Reads because the draft sequence had too many errors that could not be polished in time.

• **CCS below minimum RQ**: The number of ZMWs that did not generate CCS Reads because the predicted accuracy is below `--min-rq`.

• **Unknown error**: The number of ZMWs that did not generate CCS Reads due to rare implementation errors.