



SMRT Link User Guide

Sequel® System

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

P/N 101-039-100 Version 10 (January 2020)

© Copyright 2017- 2020, 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. BluePippin and SageELF are trademarks of Sage Science, Inc. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of 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



Introduction	.3
Sequel®/Sequel II System Users	.3
Contact Information	.3
Using SMRT® Link	.4
Module Menu Commands	.5
Gear Menu Commands	.5
Sample Setup	.6
Editing or Printing Calculations	.9
Deleting Calculations	.9
Run Design	.10
Creating a New Run Design	.10
Creating a Run Design by Importing a CSV File	.14
Editing or Deleting Run Designs	.17
Run QC	.18
Table Fields	.18
Run Settings and Metrics	.20
Data Management	.23
What is a Data Set?	.23
Creating a Data Set	.24
Viewing Data Set Information	.25
Copying a Data Set	.25
Deleting a Data Set	.25
Editing Data Set Information	.26
Starting an Analysis from a Data Set	.26
Data Set QC Reports	.26
What is a Project?	.27
Data Sets and Projects	.28
Creating a Project	.28
Editing a Project	.29
Deleting a Project	.29
Viewing/Deleting Sequence, Reference and Barcode Data	.29
Importing Sequence, Reference and Barcode Data	.30
Exporting Sequence, Reference and Barcode Data	.31
SMRT® Analysis	.33
Creating and Starting an Analysis	.33
Starting an Analysis After Viewing Sequence Data	.37
Canceling a Running Analysis	.38
Restarting a Failed Analysis	.38
Viewing Analysis Results	.39
Copying and Running an Existing Analysis	.40
Exporting an Analysis	.40

Importing an Analysis	41
PacBio® Secondary Analysis Applications	42
Assembly (HGAP 4) Application	44
Base Modification Analysis Application	49
CCS with Mapping Application	53
Circular Consensus Sequencing (CCS) Application	57
Convert BAM to FASTX Application	59
Demultiplex Barcodes Application	60
Iso-Seq® Analysis Application	64
Long Amplicon Analysis (LAA) Application	70
Mapping Application	73
Microbial Assembly Application	76
Minor Variants Analysis Application	80
Resequencing Application	86
Site Acceptance Test (SAT) Application	90
Structural Variant Calling Application	94
Working with Barcoded Data	97
Step 1: Specify the Barcode Setup and Sample Names in a Run Design	97
Step 2: Perform the Sequencing Run	98
Step 3: (Optional) Run the Demultiplex Barcodes Application	99
Step 4: Run Applications Using the Demultiplexed Data as Input	100
Demultiplex Barcodes Application Details	101
Automated Analysis	104
Creating Auto Analysis From a Run Design	104
Creating Auto Analysis Directly From SMRT Analysis	105
Getting Information About Analyses Created by Auto Analysis	105
Getting Information About Pre Analysis From SMRT Analysis	105
Getting Information About Pre Analysis From Run Design	106
Visualizing Data Using IGV	107
Using the PacBio® Self-Signed SSL Certificate	109
Sequel® System Output Files	110
Secondary Analysis Output Files	113
Configuration and User Management	116
LDAP	116
SSL	116
Adding and Deleting SMRT Link Users	117
Assigning User Roles	117
Hardware/Software Requirements	119
Appendix A - Barcoded Primers	120
Appendix B - Pacific Biosciences Terminology	121
Appendix C - Data Search	124

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/Sequel II Systems. SMRT Link includes the following modules:

- **Sample Setup:** Calculate binding and annealing reactions for preparing DNA samples. (See [“Sample Setup” on page 6](#) for details.)
- **Run Design:** Design sequencing runs and create and/or import sample sheets. (See [“Run Design” on page 10](#) for details.)
- **Run QC:** Monitor run progress, status and quality metrics. (See [“Run QC” on page 18](#) 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 23](#) 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 33](#) for details.)

This document also describes:

- The data files generated by the Sequel/Sequel II System for each cell that are transferred to network storage. (See [“Sequel® System Output Files” on page 110](#) for details.)
- The data files generated by secondary analysis. (See [“Secondary Analysis Output Files” on page 113](#) for details.)
- Configuration and user management. (See [“Configuration and User Management” on page 116](#) for details.)
- SMRT Link client hardware/software requirements. (See [“Hardware/Software Requirements” on page 119](#) for details.)

Installation of SMRT Link **Server** software is discussed in the document **SMRT Link Software Installation (v8.0)**.

New features, fixed issues and known issues are listed in the document **SMRT Link Release Notes (v8.0)**.

Sequel®/Sequel II System Users

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

Contact Information

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

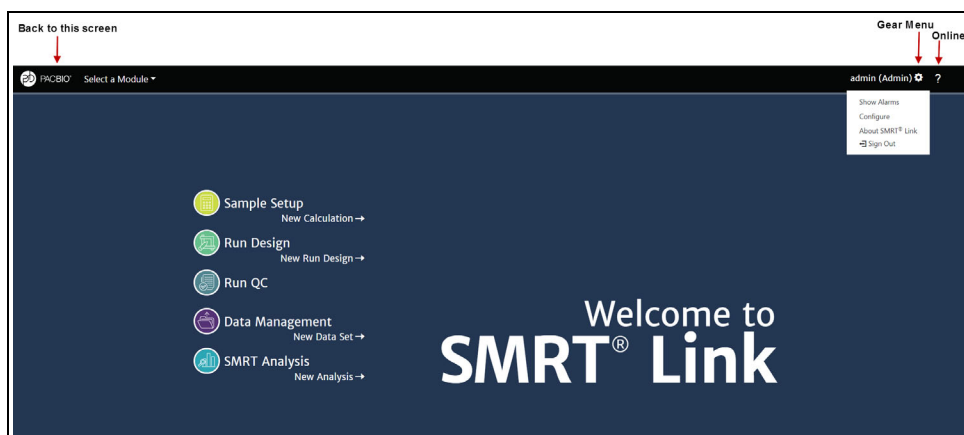
Using SMRT® Link

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 104](#).

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, 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 that SMRT Link will be used with Sequel and/or Sequel II Systems, click **Instrument** and check the appropriate boxes.
 - **Admin users only**: Add/delete SMRT Link users and specify their roles. See [“Adding and Deleting SMRT Link Users” on page 117](#) for details.
- **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**: Update the SMRT Link Chemistry Bundle, which includes kit and DNA Control Complex names used in the Sample Setup and Run Design modules.
- **Sign Out**
 - Logs you out and displays the initial login page.

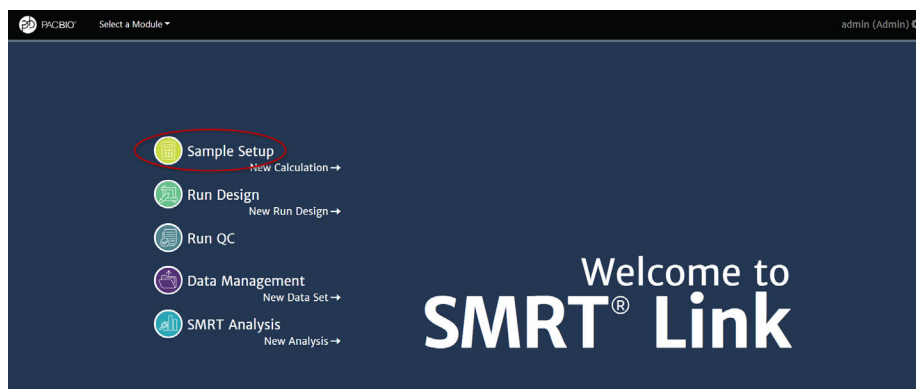
Working with Tables

- To **sort** table columns: Click a **column title**.
- To **search** within a table: Enter a unique search string into the **Search** field.

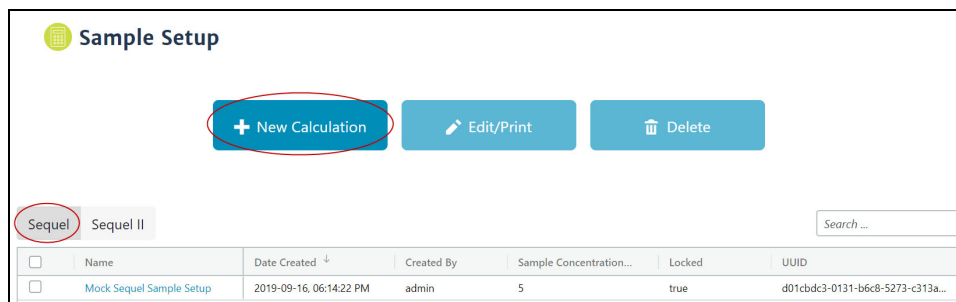
Name	Status	ID	Date Created ↑	Created By	Analysis Application
hgap4_tiny_synth5k	SUCCESSFUL	64	2019-08-19T09:16:19.962Z	smrtlinktest	Assembly (HGAP4)
assembly_microbial-fake-microbe	SUCCESSFUL	65	2019-08-19T09:16:19.981Z	smrtlinktest	Microbial Assembly

Sample Setup

Before setting up a run, use SMRT Link's **Sample Setup** module to generate a customized protocol for primer annealing and polymerase binding to SMRTbell® templates, with subsequent sample clean-up. 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. **(Optional)** Specify if this calculation is to be used with a Sequel or Sequel II System. This affects the initial default values.
4. Click **+ New Calculation**.



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

Sample 1	
Sample Name	Mock Sequel Sample Setup
Instrument Type	Sequel
Available Volume	20 uL
Concentration	5 ng/uL
Insert Size	1000 bp
Sequencing Primer	▼
Binding Kit	Sequel® Binding Kit 2.1
Sequencing Mode	CLR ▼
Iso-Seq experiment	No
Iso-Seq Version	None
Internal Control	+
Use Cleanup Step	No ▼
Cleanup Anticipated Yield	N/A
Cells to Bind	5 cells
Specify Concentration on Plate	6 pM
Number of SMRT Cells possible	261
▶ Advanced Options	
Warnings	
Actions	Copy Remove Lock

9. Select the **Sequencing Primer** to use for this run from the list, or type in a kit part number.
10. Select the **Binding Kit** to use for this run from the list, or type in a kit part number.
11. Select a **Sequencing Mode: CLR or CCS Reads**.
 - **Continuous Long Reads (CLR):** Generates 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. The CLR sequencing mode emphasizes the **longest** possible reads.
 - **CCS Reads:** Generates 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 advantageous for amplicon and RNA sequencing projects, and are **highly accurate** (>99% accuracy, Q_≥20).
12. **(CCS Reads Only)** Specify if this is an Iso-Seq experiment. If **Yes**, specify the Iso-Seq version.
13. **(Optional)** Select the **Internal Control** 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.)

-
14. Specify whether to **Clean up** the sample to remove excess primer/ polymerase. This results in higher quality data. If **Yes**, specify the anticipated yield.
 15. Enter the number of SMRT Cells to bind.
 16. Specify the on-plate loading concentration, in pM.
 17. (**Sequel II System Only**) Click **Advanced Options** and specify whether to use **Predictive Loading**. (This is a new feature under development, offered in a Beta version with limited support.) **Predictive 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 (Predictive Loading Buffer) if this feature is used.
 18. Click **Advanced Options** and 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.
 19. **Binding Volume Margin**: By default, a 10% volume excess is calculated for the annealing reaction, to ensure that any pipetting errors or losses do not compromise the following binding step. To use a different overage percentage, enter it here.
 20. Additional data entry is needed to generate the final steps of the protocol. The concentration and volume of the sample will need to be measured and their values input both immediately prior to and following the purification step.
 21. If desired, an alternative number of cells may be specified for the final sample dilution step. You may want to use this feature, for example, if you wish to prepare a large amount of sample, but only wish to load a portion of it initially.
 22. 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. 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 is just one sample **and** all fields have values entered.)
 - Click the **+ New Sample** button at the top of the screen to start a new, empty sample.
 23. To **print** the calculation(s) and instructions, use the browser's Print command (**Ctrl-P**).

Editing or Printing Calculations

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

Deleting Calculations

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

Run Design

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

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

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

Run Designs created in SMRT Link are accessible from **all** Sequel/Sequel II 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



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

1. Access SMRT Link using the Chrome web browser.
2. Select **Run Design**.
3. Run 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. **(Optional)** Specify if this Run Design is to be used with a Sequel or Sequel II System. This affects the initial default values.
 5. To initiate a new Run Design, click **Create New Design**.

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 **Select Sample** to import information from a previously-created Sample Setup entry. The following fields will be auto-populated as appropriate:
 - Sample Name
 - Binding Kit
 - DNA Control Complex
 - Sequencing Mode
 - Iso-Seq Experiment
 - Insert Size
 - On-Plate Loading Concentration
9. Enter a **Sample Name**.
10. **(Optional)** Enter **Sample Comments**.
11. Specify the well position used for this sample: Click **Select...** and choose a plate position.
12. Select a **Template Prep Kit** from the list, or type in a kit part number.
13. Select a **Binding Kit** from the list, or type in a kit part number.
14. Select a **Sequencing Kit** from the list, or type in a kit part number.
 - **Note:** If the Sequencing or Binding kit is **incompatible**, an error message displays indicating the obsolete chemistry, and the run is **prevented** from proceeding.
15. **(Optional)** Select a **DNA Control Complex** from the list. Pacific Biosciences **highly** recommends using the Internal Control to help distinguish between sample quality and instrument issues in the event

of suboptimal sequencing performance. (**Note:** PacBio **requires** the use of the Sequel Internal Control for consumables to be eligible for reimbursement consideration.)

16. Select a **Sequencing Mode: Continuous Long Reads or CCS Reads**.
 - **Continuous Long Reads (CLR):** Generates 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. The CLR sequencing mode emphasizes the **longest** possible reads.
 - **CCS Reads:** Generates 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 **highly accurate** (>99% accuracy, Q₂₀). Additional fields display when you select **CCS Reads**.
17. (**CCS Reads only**) Specify if this is an Iso-Seq experiment.
18. 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.

The screenshot displays the Sequel System software interface, divided into two main sections: Run Information and Sample Information.

Run Information:

- System Type:** SEQUEL II (selected).
- Run Name:** Run 09:16:2019 21:42
- Run Comments:** Test run
- Experiment Name:** Armadillo Sequencing
- Experiment ID:** (empty)
- Estimated Run Duration (hours):** 33
- Run Reagents / Consumables:**
 - 1 SMRT Cell
 - 1 sequencing reagent plate
 - 1 mineral oil tube
 - 3 boxes of tips
 - 1 mixing plate
 - 1 sample plate

Sample Information:

- On-Plate Loading Concentration (pM):** 0
- Movie Time per SMRT Cell (hours):** 30
- Use Pre-Extension:** NO (selected)
- Generate CCS Data:** YES (selected)
- Advanced Options:**
 - Use Predictive Loading (Beta):** NO (selected)
 - Immobilization Time (hrs):** 2
- Barcoded Sample Options:** (collapsed)
- Buttons:** Copy, Delete, Add Sample

19. Specify the **On-plate loading concentration**, in pM.
20. Specify the **Movie time (collection time)** per SMRT Cell (in hours): 0.5, 2, 4, 6, 8, 10, 15, 20, 24 or 30 hours. **Note:** For Sequel Systems, 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.
21. Specify whether to use **pre-extension**. If specified, 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.
22. **(CCS Reads only)** Specify whether to automatically generate **CCS Data**. (This is set to **Yes** by default if you selected the **CCS Reads Type** earlier.) CCS Reads:
 - Are generated from Sequel Data by CCS analysis.
 - Are advantageous due to their high accuracy and long read length.
 - Can be used as input for Structural Variant analysis, demultiplexing, and mapping applications in SMRT Link.
 23. **(Sequel II System Only, Optional)** Click **Advanced Options**, and specify whether to use **Predictive Loading**. (This is a new feature under development, offered in a Beta version with limited support.) **Predictive 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 (Predictive Loading Buffer) if this feature is used. If you select **Yes**, fill in the following fields:
 - **Loading Target (P1 + P2)**: The fraction of ZMWs that the Predictive 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 Predictive Loading is flexible.)
 24. **(Optional)** Click **Advanced Options**, then 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 Predictive Loading is selected.
 - PacBio **highly recommends** using the default immobilization time of 2 hours.
 25. **(Optional)** If you are using **barcoded samples**, see [“Step 1: Specify the Barcode Setup and Sample Names in a Run Design” on page 97](#) for instructions. For details on secondary analysis of barcoded samples, see [“Demultiplex Barcodes Application” on page 60](#).
 26. Sample options:
 - Click **Copy**. This starts a new sample, using the values entered in the first sample.
 - Click **Create**. This starts a new, empty sample.
 - Click **Delete**. This deletes the current sample.
 27. 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.

28. 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.
29. **(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 104](#) for details.

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 for Sequel and Sequel II

1. Go to the home page URL for your site’s SMRT Link installation and replace the last part of the URL (*welcome* or *home*) with `docs/xsd-datamodels/run-record/RunDesignTemplates_SL8.zip`.
2. Press the Enter key, and the ZIP file then downloads to your local machine.

To update and import the CSV file

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

Run Design Attribute	Required	Description
Experiment Name	No	Enter any ASCII string. Defaults to Run Name. Example: Standard_Edna.1
Experiment Id	No	Enter a valid experiment ID. Example: 325/3250057 <ul style="list-style-type: none"> • Experiment IDs cannot contain the following characters: <code><, >, :, ", \, , ?, *, or)</code>. • Experiment IDs cannot start or end with a <code>/</code> and cannot have two adjacent <code>/</code> characters, such as <code>//</code>. • Experiment IDs cannot contain spaces. • Specifically, Experiment IDs cannot satisfy the regular expressions: <code>/[<>:"\\ ?*\]/g</code>, <code>/(?:^\\/) \\\/ (?:\\\$)/, / /g</code>
Experiment Description	No	Enter any ASCII string. Defaults to Run Comments. Example: 20170530_A6_VVnC_SampleSheet
Run Name	Yes	Enter any ASCII string. Run name must be entered for the first cell and will be applied to the remaining cells in the run. Example: 20170530_A6_VVnC_SampleSheet
System Name	No	Must be either <i>Sequel</i> or <i>Sequel II</i> .

Run Design Attribute	Required	Description
Run Comments	No	Enter any ASCII string. Example: ecoliK12_March2019
Is Collection	No	Enter a Boolean value. (See Boolean details below.) Specifies whether the row designates a Collection (TRUE) or a barcoded sample (FALSE). <ul style="list-style-type: none"> Collection lines should have the Barcode Name and Bio Sample Name fields blank. Barcoded Sample lines only need to include the Is Collection, Sample Name, Barcode Name, and Bio Sample Name fields.
Sample Well	Yes	Well number must start with a letter A through H, and end in a number 01 through 12, i.e. A01 through H12. It must satisfy the regular expression ``^[A-H] (? : 0 [1-9] 1 [0-2]) \$/`` Example: A01
Sample Name	Yes	Enter any ASCII string. Example: A6_3230046_A01_SB_ChemKitv2_8rxnKit Note: The Sample Name must be unique within a run.
Sequencing Mode	No	Must be either CLR or CCS.
Movie Time per SMRT Cell (hours)	Yes	Enter a floating point number between 0.1 and 20 for Sequel; 0.1 and 30 for Sequel II. Time is in hours. Example: 5
Use Predictive Loading	No	Enter a Boolean value. (See Boolean details below.)
Loading Target (P1 + P2)	No	Enter a floating point number between 0.01 and 1. Example: 0.4
Maximum Loading Time (hours)	No	Enter a floating point number between 1 and 2. Time is in hours. Example: 1.2
Sample Comment	No	Enter any ASCII string. Example: A6_3230046_A01_SB_BindKit_ChemKit
Insert Size (bp)	Yes	Enter an integer ≥ 10. Units are in base pairs. Example: 2000
On Plate Loading Concentration (pM)	No	Enter a floating point number. Units are in parts per million. Example: 5
Size Selection	No	Enter a Boolean value. (See Boolean details below.) Default is <code>False</code> .
Template Prep Kit Box Barcode	Yes	Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM1117100259100111716
DNA Control Complex Box Barcode	No	Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM1234101084300123120
Binding Kit Box Barcode	Yes	Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM1117100862200111716
Sequencing Kit Box Barcode	Yes	Enter or scan a valid kit barcode. (See Kit Barcode Requirements details below.) Working example: DM0001100861800123120
Automation Name	No	Enter <code>diffusion</code> , <code>magbead</code> (not case-sensitive) or a custom script. (Sequel II does not support magbead loading.) A path can also be used, such as <code>/path/to/my/script/my_script.py</code> . The path will not be processed further, so if the full URI is required, it must be included in the CSV, such as <code>chemistry://path/to/my/script/my_script.py</code> .
Automation Parameters	No	To enable Pre-Extension time, enter the number of minutes. Example 60 minutes: <code>ExtensionTime=double:60 ExtendFirst=boolean:True</code> (Note: Leave blank when not using Pre-Extension time.)

Run Design Attribute	Required	Description
Generate CCS Data	No	Enter a boolean value. (See Boolean details below.)
Sample is Barcoded	No	Enter a boolean value. (See Boolean details below.) Set to 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 Data Management > View Data > Barcodes . Click the Barcode file of interest, then view the UUID. Example: dad4949d-f637-0979-b5d1-9777eff62008
Same Barcodes on Both Ends of Sequence	No	Enter a boolean value. (See Boolean details below.) Set to TRUE if symmetric, FALSE if asymmetric.
Barcode Name	No	Enter barcoded sample names that follow the guidelines below. Example: lbc1--lbc1;sample1 lbc2--lbc2;sample2 lbc3--lbc3;sample3 <ul style="list-style-type: none"> Put the entire mapping of barcode name-to-sample name into one spreadsheet cell. Use double hyphens (--) to separate the 2 barcodes of each pair, and semicolons to separate the barcode pair and sample name from the next ones. Barcoded sample names are included in a list separated by characters. Each item in the list follows the format [barcode name];[biosample name] The barcode names must be contained within the specified Barcode Set. A given barcode name cannot appear more than once in the list. The Bio Sample names can be any ASCII string but cannot contain the field separators and ;. The Bio Sample names cannot be longer than 40 characters. A maximum of 384 barcodes is permitted per sample.
Bio Sample Name	No	Enter any ASCII string. Example: sample1 Note: This is left blank in the Collection-level row.
Pipeline ID	No	Enter an ASCII string containing the ID of the secondary analysis application pipeline used. See Auto Analysis Fields below for details. Example: cromwell.workflows.pb_isoseq3
Analysis Name	No	Enter any ASCII string. See Auto Analysis Fields below for details. Example: sample 1 analysis
Entry Points	No	Enter an ASCII string, with parameters separated by characters: file_type;entry_id;uuid. To find the UUID: Click Data Management > View Data > CCS Data or Sequel Data . Click the Data Set of interest, then view the UUID. See Auto Analysis Fields below for details. Example: eid_ref_dataset=6b8db144-a601-4577-ab04-ba64cad0548
Task Options	No	Enter an ASCII string containing the options for the application referred to in the Pipeline ID field, with parameters separated by characters: task_id;value_type;value. See Auto Analysis Fields below for details. Example: genomic_consensus.task_options.min_confidence=45;genomic_consensus.task_options.track_name=variants

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

Editing or Deleting Run Designs

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

Run QC

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

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



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

The image shows the 'Run QC' module interface. At the top, there is a header with the 'Run QC' title and an 'Export Selected' button. Below the header, there is a filter bar with buttons for 'Complete', 'Running', 'Terminated', 'Aborted', 'Paused', and 'Ready'. A search field is also present. The main content is a table with columns for Name, Summary, Run Date, Created By, Status, Instrument Name, Total, Run ID, Primary Analysis SW, and UUID. Two rows of data are visible in the table.

Name	Summary	Run Date	Created By	Status	Instrument Name	Total	Run ID	Primary Analysis SW	UUID
PacBio Example SAT Run	54026_SAT_4pM_20181218	2018-12-18, 05:01:47 PM	unknown	Complete	54026	1	r54026_2018121...	6.0.0.47712	c427ff4-3a83-43...
20180917_43_Kwi_Verif	ecoliK12_pbi_March2013	2018-09-17, 02:08:48 PM	jziegle	Aborted	Flash43	4	r54043_2018091...	6.0.0.46934	d0f632ad-138f-4...

3. Runs can be sorted, searched for, and filtered:
 - To sort runs, click a **column title**.
 - To search for a run, enter a unique search string into the **Search** field.
 - To specify the status of the runs to display,
 - To filter the list of runs to display, click one or more of the following buttons: **Complete**, **Running**, **Terminated**, **Aborted**, **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

- **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 QC / Run Data

2018-07-27_64003_30kEcoli_30kSub COMPLETE Export Print

Overview

Run Start:
2018-07-27, 07:46:39 PM

Run Id:
r64003_20180728_024607

Instrument SN:
64003

Primary SW Version:
6.1.0.SNAPSHOT41861

Run Complete:
2018-07-28, 07:30:06 PM

Description:
Ecoli

Instrument Control SW Version:
6.1.0.SNAPSHOT41861

Transfer Complete:
2018-07-29, 06:57:22 AM

Instrument:
Sequel

Instrument Chemistry Bundle:
6.0.0.SNAPSHOT41529

Consumables


Well	Name	Movie Time (hrs)	Status	Total Bases (Gb)	Unique Molecular Y...	Productivity (%)			Read Length (bp)					
						P0	P1	P2	Polymerase	Longest Subread	Mean	N50	Mean	N50
A01	SMS_Spider_SP21_30kEcoli_7p5pMchip	10	Complete											
B01	SMS_Spider_SP21_30kSub_10pMchip	10	Complete	115.77	110.95	20.3	74.8	4.8	19002	32934	18597	31285		

Plots


▼ A01 Not yet imported: /pbi/collections/333/3330002/r64003_20180728_024607/1_A01/m64003_180728_025446.subreadset.xml

▼ B01


Polymerase Read Length



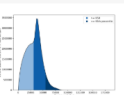
Longest Subread Length



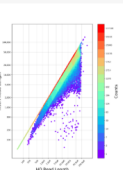
Loading Evaluation



Base Yield Density



Insert Read Length Density



Page 19

- **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

- **Well:** The ID of an individual well used for this sample.
- **Name:** The sample name, as defined when creating the run. Clicking the name will take you to the corresponding entry in the **Data Management** module.
- **Run Settings**
 - **Movie Time (hrs):** The length of the movie associated with this SMRT Cell.
 - **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.
- **Read Length (bp):** Polymerase reads are trimmed to the high quality region and include bases from adapters, as well as potentially multiple passes around a SMRTbell template.

-
- **Polymerase Mean:** The mean high-quality read length of all polymerase reads. The value includes bases from adapters as well as multiple passes around a circular template.
 - **Polymerase N50:** 50% of all read bases came from polymerase reads longer than this value.
 - **Longest Subread Mean:** The mean subread length, considering only the longest subread from each ZMW.
 - **Longest Subread N50:** 50% of all read bases came from subreads longer than this value when considering only the longest subread from each ZMW.
 - **Control**
 - **Poly RL Mean (bp):** The mean polymerase read length of the control reads.
 - **Total Reads (bp):** 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. Click the > arrow to expand rows to view plots for each SMRT Cell where data was successfully transferred. Clicking on an individual plot displays an expanded view. These plots include:
- **Polymerase Read Length:** Plots the number of reads against the polymerase read length.
 - **Longest Subread Length:** Plots the number of reads against the insert length.
 - **Control Polymerase RL:** Displays the Polymerase read length distribution of the control, if used.
 - **Control Concordance:** Maps control reads against the known control reference and reports the concordance.
 - **Loading Evaluation:** Displays the length distribution of unfiltered and filtered (polymerase) reads.
 - **Base Yield Density:** Displays the number of bases sequenced in the collection, according to the length of the read in which they were observed. Values displayed are per unit of read length (i.e. the base yield density) and are averaged over 2000 bp windows to gently smooth the data. Regions of the graph corresponding to bases found in reads longer than the N50 and N95 values are shaded in medium and dark blue, respectively.

-
- **Insert Read Length Density:** Displays a density plot of reads, hexagonally binned according to their HQ Read Length and median subread length. For very large insert libraries, most reads consist of a single subread and will fall along the diagonal. For shorter inserts, subreads will be shorter than the HQ read length, and will appear as horizontal features. This plot is useful for quickly visualizing aspects of library quality, including insert size distributions, reads terminating at adapters, and missing adapters.

Data Management

Use the **Data Management** module to:

- Create and manage Data Sets,
- View Data Set information,
- Create and manage Projects,
- View, import, 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 26](#) for details.)

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

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

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

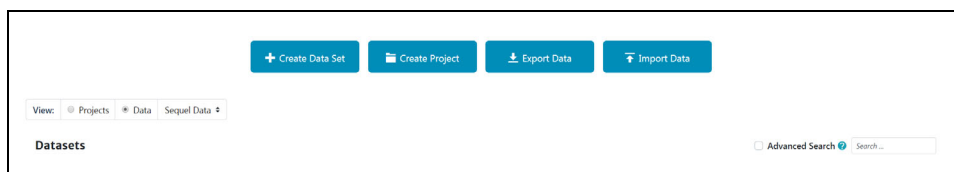
- **Basecalled data** Data Sets contain sequence data from one or multiple cells or collections.
- **Analyzed data** Data Sets contain data from previous analyse(s).

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

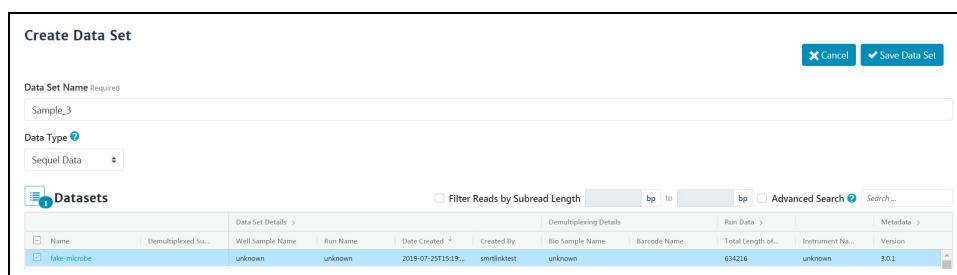
Creating a Data Set



1. Access SMRT Link using the Chrome web browser.
2. Select **Data Management**.
3. Data Sets can be sorted and searched for:
 - To sort Data Sets, click a **column title**.
 - To search for a Data Set **locally** or **remotely**, use the Search function. See [“Appendix C - Data Search” on page 124](#) for details.



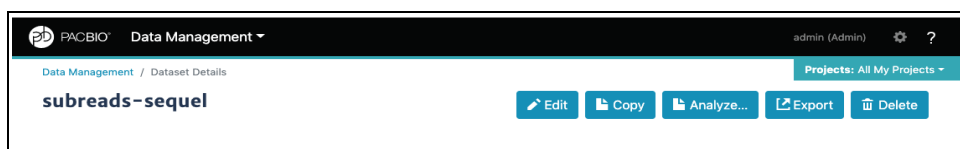
4. Click **+ Create Data Set**.
5. Enter a name for the new Data Set.



6. Click **View > Data** and select the type of data to include in the new Data Set:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.

The Data Sets table displays the appropriate Data Sets available.
7. In the **Data Sets** table, select one or more sets of sequence data.

8. **(Optional)** Use the Search function to search for specific Data Sets **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) for details.
9. **(Optional)** Click the **Filter reads by subread length** box above the Data Set list. Enter the minimum and/or maximum subread length to retain in the new Data Set.
10. Click **Save Data Set**. The new Data Set becomes available for starting analyses, viewing, or generating reports.
11. After the Data Set is created, click its name in the main Data Management screen to see reports, metrics, and charts describing the data included in the Data Set.



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:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.

The Data Sets table displays the appropriate Data Sets available.
3. **(Optional)** Use the Search function to search for Data Sets **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) 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:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.

The Data Sets table displays the appropriate Data Sets available.
3. **(Optional)** Use the Search function to search for Data Sets **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) for details.
4. Click the name of the Data Set to copy. The Data Set Reports page displays.
5. Click **Copy**. The main Data Management page displays; the new Data Set has **(copy)** appended to the name.

Deleting a Data Set

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

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

1. On the Home Page, select **Data Management**.
2. Click **View > Data** and select the type of data to delete:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.The Data Sets table displays the appropriate Data Sets available.
3. **(Optional)** Use the Search function to search for Data Sets **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) for details.
4. Click the name of the Data Set to delete.
5. Click **Delete**. Note that this **only** deletes the Data Set from the SMRT Link interface; **not** from your server. To delete the Data Set from your server, **manually** delete it from the disk.
6. Click **Yes**. The Data Set is no longer available from SMRT Link.

Editing Data Set Information

1. On the Home Page, select **Data Management**.
2. Click **View > Data** and select the type of Data Set to edit:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.The Data Sets table displays the appropriate Data Sets available.
3. **(Optional)** Use the Search function to search for Data Sets **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) for details.
4. Click the name of the Data Set to edit.
5. Click **Edit**.
6. Edit the Well Sample Name and/or the Biological Sample Name.
7. Click **Save**.

Starting an Analysis from a Data Set

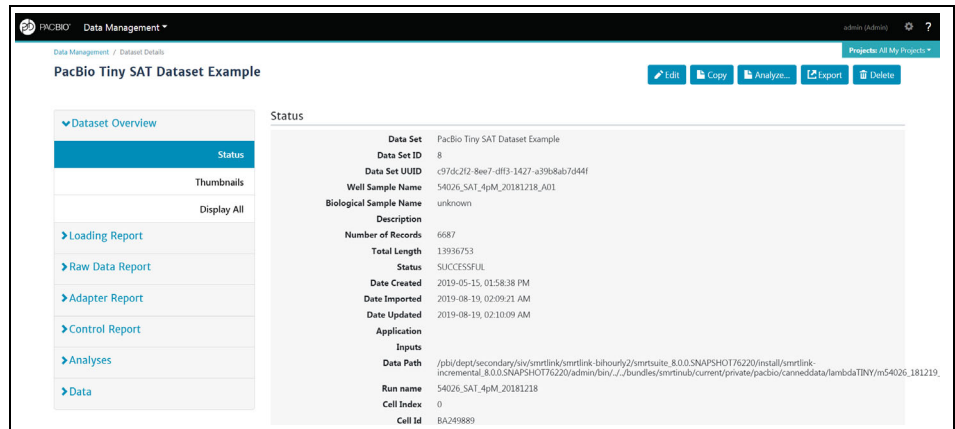
From the Data Set Reports page, an analysis can be started using the Data Set.

1. Click **Analyze...** then follow the instructions starting at Step 12 of [“Creating and Starting an Analysis” on page 33](#).

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 names of the well and sample used to generate the data.
- The name of the analysis application used to generate the data.

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.

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.

Creating a Project

The screenshot shows the 'Create Project' interface. It has a breadcrumb 'Data Management / Projects' and a 'Projects: All My Prj' indicator. The main form is titled 'Create Project' and includes a 'Cancel' and 'Save' button. The 'Project Name' field is required and contains 'QC_Group'. There is a 'Description' text area and an 'Associated Data Sets' section with a 'Select Data Sets' button. The 'Members' section allows setting access for all SMRT@ Link users (currently 'None') and for individual users. A search for 'QA' shows a table with one result: 'sappi-adm-qa' with email 'sappi-adm-qa@pacificbiosciences.com' and an 'Add Selected User' button.

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 **locally** or **remotely**. See [“Appendix C - Data Search”](#) on page 124 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:
 - For **all** SMRT Link Users: **None** - No one can access the project other than the user who created it; **View** - Everyone can view the Project; **View/Edit**: Everyone can see and edit the Project.
 - **To give individual users access to the Project:** Enter a user name and click **Search By Name**. Choose **Owner**, **View**, or **View/Edit**, then click **Add Selected User**.
 - **Notes:** A) Projects can have **multiple** owners. B) If you enable **all** SMRT Link users to have **View/Edit** access, you cannot change an individual member's access to **View**.
 8. Click **Save**. The new project becomes available for SMRT Link users who now have access.

Editing a Project

1. On the Home Page, select **Data Management**.
2. Click **View > Projects**.
3. Projects can be sorted and searched for:
 - To sort Projects: Click a **column title**.
 - To search for a Project: Enter a unique search string into the **Search** field.
4. Click the name of the project to edit.
 - **(Optional)** Edit the project name or description.
 - **(Optional)** Delete a Data Set associated with the project: 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**.
 - Click **Save**. The modified Project is saved.

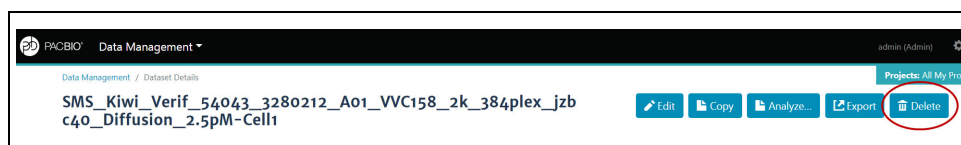
Deleting a Project

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:
 - **Sequel Data:** Subreads from Sequel or Sequel II Systems.
 - **CCS Data:** Single Molecule consensus reads generated from the CCS analysis.

- **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 files **locally** or **remotely**. See “[Appendix C - Data Search](#)” on page 124 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**.

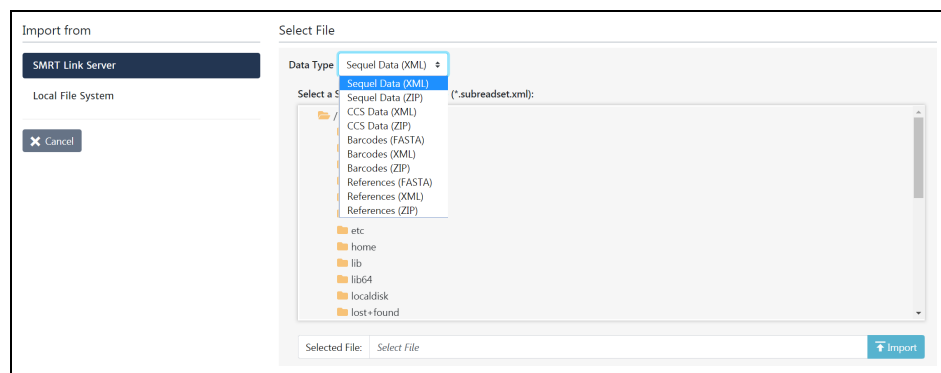


Importing Sequence, Reference and Barcode Data

Note: If your Sequel/Sequel II 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**.)



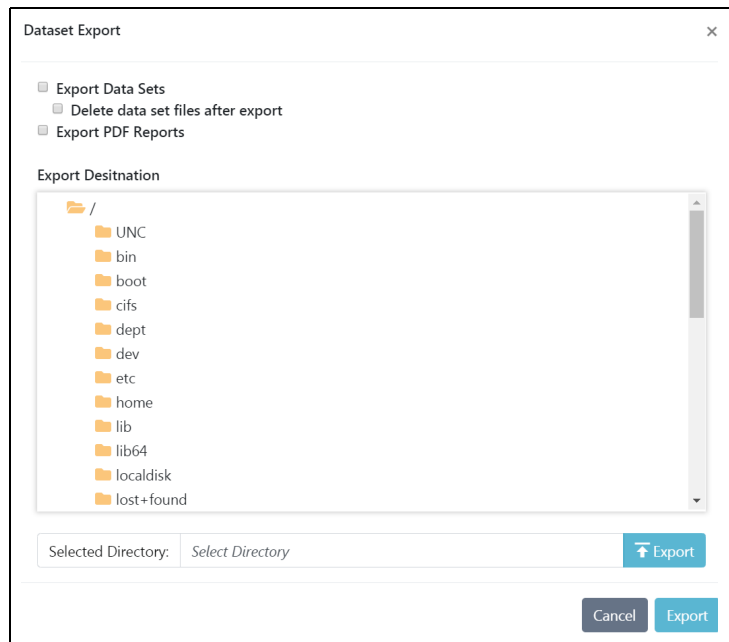
4. Select the data type to import:
 - **Sequel Data:** XML file (`.subreadset.xml`) or ZIP file containing information about Sequel/Sequel II sequence data, such as paths to the BAM files.
 - **CCS Data:** XML file (`.subreadset.xml`) or ZIP file containing information about CCS sequence data.

- **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.
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 (CCS Data and Sequel Data) can be exported, as well as barcode files and reference files.

1. On the Home Page, select **Data Management**.
2. Click **Export Data**.
3. Click **View** and select the type of data to export:
 - **Sequel Data:** Subreads from Sequel or Sequel II Systems.
 - **CCS Data:** Single Molecule consensus reads generated from the CCS analysis.
 - **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 **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) 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**.



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**.

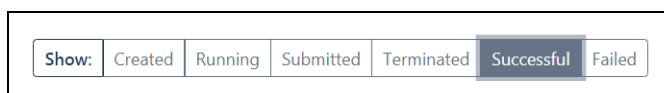
SMRT® Analysis

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, enter a unique search string into the **Search** field. (Use the Search function to search for analyses **locally** or **remotely**. See [“Appendix C - Data Search”](#) on page 124 for details.)
 - To filter the list of analyses to display, click one or more of the following buttons: **Created**, **Running**, **Submitted**, **Terminated**, **Successful**, and/or **Failed**. (Click **Show** to remove or select **all** the filters.)



4. Click **+ Create New Analysis**.
5. **(Optional)** Click **Copy From...** and choose an analysis whose settings you wish to reuse. The analysis name and the Data Type are filled in. Go to Step 8 to select Data Set(s).
6. Enter a **name** for the analysis.
7. Select the type of data to use for the analysis:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads. CCS Reads:
 - Are generated from Sequel Data by CCS analysis.
 - Can be automatically generated once the Sequel Data is transferred off the instrument (if the corresponding setting is selected in Run Design).

- Are advantageous for amplicon applications due to their high accuracy and long read length.
- Are **highly accurate** – by default, analysis parameters are set to >99% accuracy, Q≥20.

The Data Sets table displays the corresponding Data Sets available for the analysis.

- 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 **locally** or **remotely**. See “Appendix C - Data Search” on page 124 for details.)
 - **(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:

<input type="checkbox"/>	Name	Demultiplexed Subsets
<input type="checkbox"/>	Re-barcode Alice/Bob/Charles	3
<input type="checkbox"/>	subreads-sequel	

B) Select one or more subsets, then click **Back**:

<input type="checkbox"/>	Name	Sample Details >	Run Details >	ID
		Well Sample Name	Run Name	
<input checked="" type="checkbox"/>	Run_02.19.2016 00:27 384-plex b...	T0213_384-plex_barcode_A01	m54008_160219_003234_...	59
<input checked="" type="checkbox"/>	Run_02.19.2016 00:27 384-plex b...	T0213_384-plex_barcode_A01	m54008_160219_003234_...	58
<input type="checkbox"/>	Run_02.19.2016 00:27 384-plex b...	T0213_384-plex_barcode_A01	m54008_160219_003234_...	57

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):

<input checked="" type="checkbox"/>	Name	Demul...
<input checked="" type="checkbox"/>	internal-su...	
<input checked="" type="checkbox"/>	Run_2016...	
<input checked="" type="checkbox"/>	PacBio Tin...	

Selected Data Sets

<input checked="" type="checkbox"/>	Id	Name	Demux Source
<input checked="" type="checkbox"/>	59	Run_02.19.2016 00:27 384-plex barcode A8 (Charles)	Run_02.19.2016 00:27 384-plex barcode A8 (Charles)
<input checked="" type="checkbox"/>	58	Run_02.19.2016 00:27 384-plex barcode A8 (Bob)	Run_02.19.2016 00:27 384-plex barcode A8 (Bob)

1 to 2 of 2 << < Page 1 of 1 > >>

Note: For information on the Auto Analysis feature, see “Automated Analysis” on page 104 for details.

- If you selected **multiple** Data Sets as input for the analysis, additional options become available:

Analysis Type

AUTO ANALYSIS ANALYSIS

Analysis of Multiple Data Sets

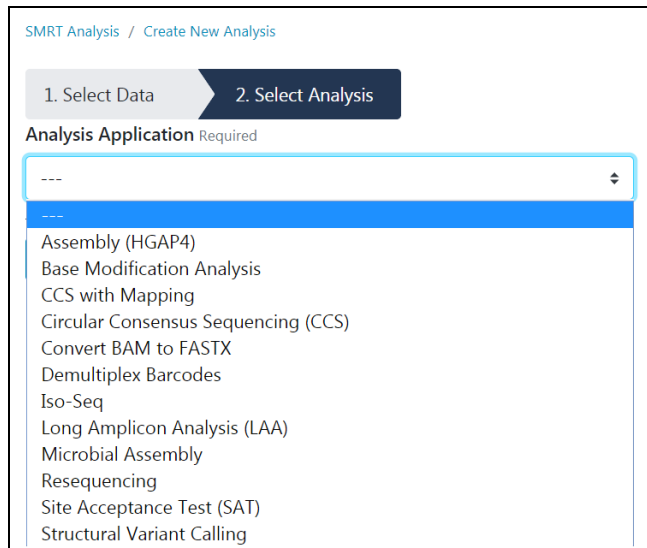
One Analysis per Data Set - Identical Parameters

One Analysis for All Data Sets

One Analysis per Data Set - Identical Parameters

One Analysis per Data Set - Custom Parameters

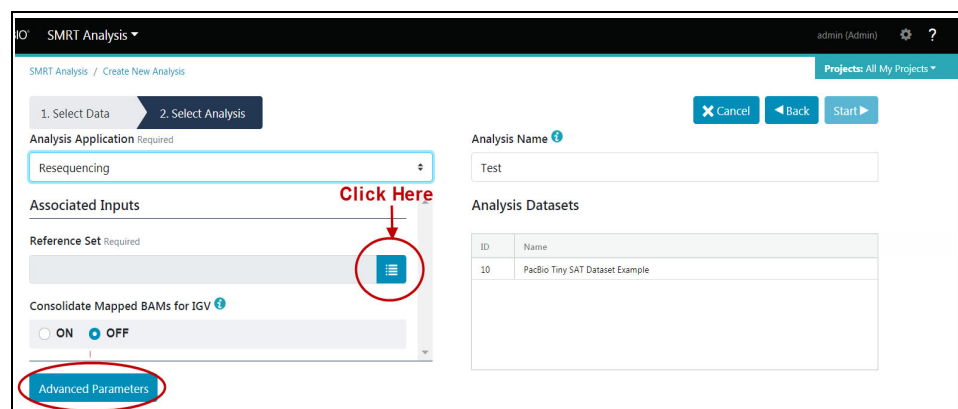
- **One Analysis for All Data Sets:** Runs **one** analysis using all the selected Data Sets as input.
 - **One Analysis per Data Set - Identical Parameters:** Runs one separate analysis for **each** of the selected Data Sets, using the **same** parameters. Later in the process, optionally click **Advanced 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. 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.
- (Optional) Specify the **Project** that this analysis will be associated with. **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. **Note:** To **restrict** access to an analysis, make sure to select a project limited to the appropriate users **before** starting the analysis.
 - Click **Next**.
 - Select a secondary analysis application from the dropdown list. (Different applications display based on your choice of Data Type in Step 7.)



- 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.

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

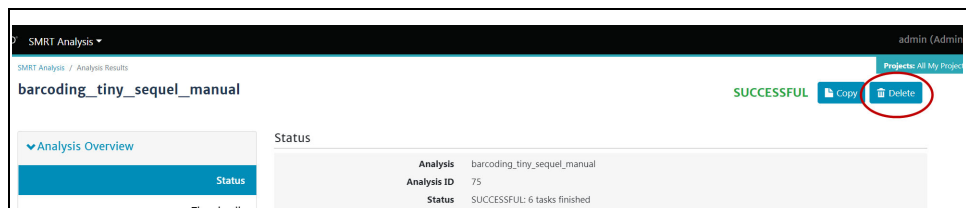
13. Click the **Reference Set** field and select a reference sequence from the dialog. (The reference sequences available in SMRT Link and displayed in the dialog were imported into SMRT Analysis. See [“Importing Sequence, Reference and Barcode Data”](#) on page 30 for details.)



14. **(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 42.

15. (Optional) Click **Back** if you need to change any of the analysis attributes selected in Step 7.
16. Click **Start** to submit the analysis. (If you selected multiple Data Sets as input, click **Start Multiple Jobs** or **Start and Create Next**.)
17. 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.
18. (Optional) To **delete** the completed analysis: Click **Delete**, then click **Yes** in the confirmation dialog. The analysis is deleted.

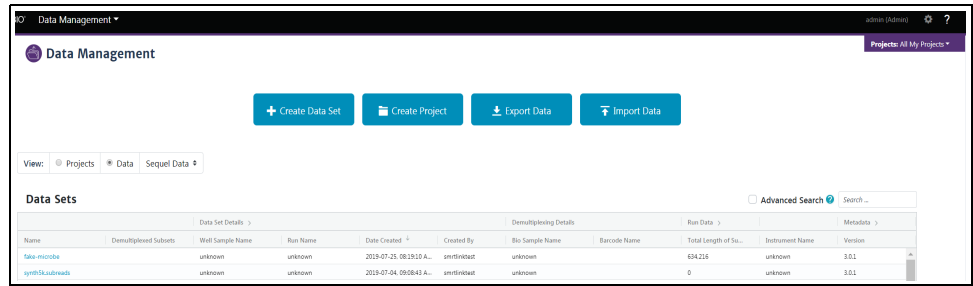


Starting an Analysis After Viewing Sequence Data

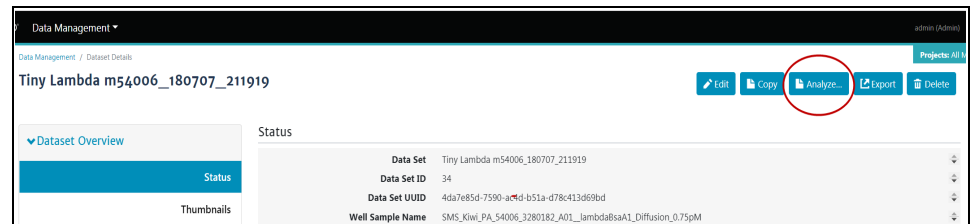
An analysis can be started by **first** viewing information about specific sequence data:

1. On the Home Page, select **Data Management**.
2. Click **View > Data** and select the type of Data Set to use:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.

The Data Sets table displays the appropriate Data Sets available.
3. (Optional) Use the Search function to search for Data Sets **locally** or **remotely**. See [“Appendix C - Data Search”](#) on page 124 for details.



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



- 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 33.

Canceling a Running Analysis

- On the Home Page, select **SMRT Analysis**.
- Click **Show** to remove all filters, then click the **Running** button. This displays only currently-running analyses.
- Select a currently-running analysis to cancel.
- Click **Cancel**.
- Click **Yes** in the confirmation dialog.

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**.

The screenshot shows the PacBio SMRT Analysis web interface. At the top, the user is logged in as 'mbudagyan (Admin)'. The main heading is 'Bsub_HGAP4 assembly'. A red 'FAILED' label is present, along with buttons for 'Send Log Files', 'Restart' (circled in red), 'Copy', and 'Delete'. On the left, there is a navigation menu with options like 'Analysis Overview', 'Status', 'Thumbnails', 'Display All', 'Coverage', 'Alignment to Draft Assembly', 'Preassembly', and 'Data'. The main content area shows the 'Status' of the analysis:

Analysis	Bsub_HGAP4 assembly
Analysis ID	101448
Status	FAILED: Task genomic_consensus.tasks.variantcaller-10 FAILED in 1496.24 sec
Created By	sdhillon
Date Created	2019-09-04, 10:32:25 AM
Date Updated	2019-09-04, 11:48:56 AM
Application	Assembly (HGAP 4)
SMRT Link Version	7.1.0.71715
Inputs	Sequel Data: Demux_32 plex_SD (bc1059--bc1059)
Path	/pbi/dept/secondary/siv/smrtlink/smrtlink-beta/smrtsuite_166987/userdata/jobs_root/101/101448
Error Message	Pbsmrtpipe job /pbi/dept/secondary/siv/smrtlink/smrtlink-beta/smrtsuite_166987/userdata/jobs_root/101/101448 failed with exit code 1. Task genomic_consensus.tasks.variantcaller-10 task genomic_consensus.tasks.variantcaller failed (exit-code 2) after 1496.24 sec Failed task genomic_consensus.tasks.variantcaller exit code 2 in 20.37 sec (See file '/pbi/dept/secondary/siv/smrtlink/smrtlink-beta/smrtsuite_166987/userdata/jobs_root/101/101448/tasks/genomic_consensus.task-10/stderr'.) Extracted from stderr Unidentifiable sequencing chemistry present in dataset. Check if your SMRTanalysis installation is out-of-date. Unidentifiable sequencing chemistry present in dataset. Check if your

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

1. On the Home Page, select **SMRT Analysis**.
2. Click **Show** to remove all filters, then click the **Failed** button. 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 **Show** to remove all filters, then click the **Successful** button. This displays **only** successfully-completed analyses.
3. **(Optional)** Use the Search function to search for specific analyses **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) 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 42](#).
8. To download data files created by SMRT Link: You can use these data files as input for further downstream processing, pass on to collabora-

-
- tors, 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 107](#) 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 **Show** to remove all filters, then click the **Successful** button. This displays **only** successfully-completed analyses.
3. **(Optional)** Use the Search function to search for specific analyses **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) 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 parameter. See [“PacBio® Secondary Analysis Applications” on page 42](#) 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 **locally** or **remotely**. See [“Appendix C - Data Search” on page 124](#) for details.
4. Select one or more analyses to export. This exports the entire contents of the analysis directory. To **also** export the input sequence data files associated with the analyses, select **Include Input Sequence Data**.
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 v8.0. Each application is described later, including all parameters and the reports and data files output by the application.

Assembly (HGAP 4)

- Generate *de novo* assemblies of genomes.
- See [“Assembly \(HGAP 4\) Application” on page 44](#) for details.

Base Modification Analysis

- Identify putative sites of base modification as well as common bacterial base modifications (6mA, 4mC).
- Optionally analyze the methyltransferase recognition motifs.
- See [“Base Modification Analysis Application” on page 49](#) 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 53](#) for details.

Circular Consensus Sequencing (CCS)

- Identify consensus sequences for single molecules.
- See [“Circular Consensus Sequencing \(CCS\) Application” on page 57](#) for details.

Convert BAM to FASTX

- Convert sequence data in BAM file format to the FASTX file format.
- For **barcoded** runs, you must **first** run the **Demultiplex Barcodes** application to create BAM files **before** using this application.
- See [“Convert BAM to FASTX Application” on page 59](#) for details.

Demultiplex Barcodes

- Separate reads by barcode.
- See [“Demultiplex Barcodes Application” on page 60](#) for details.

Iso-Seq® Analysis

- Characterize full-length transcripts. Optionally map the transcripts back to the reference genome if a reference genome is supplied.
- See [“Iso-Seq® Analysis Application” on page 64](#) for details.

Long Amplicon Analysis (LAA)

- Identify phased consensus sequences from a heterogeneous pool of amplicons.
- See [“Long Amplicon Analysis \(LAA\) Application” on page 70](#) for details.

Mapping

- Align (or map) CCS Reads to a user-provided reference sequence.
- See [“Mapping Application” on page 73](#) 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 76](#) for details.

Minor Variants Analysis

- Identify and phase minor single nucleotide substitution variants in complex populations.
- See [“Minor Variants Analysis Application” on page 80](#) for details.

Resequencing

- Map sequencing reads against a reference sequence and identify variants.
- See [“Resequencing Application” on page 86](#) for details.

Site Acceptance Test (SAT)

- Generate a report displaying instrument acceptance test metrics. (The application is designed **only** for analysis of Site Acceptance data.)
- See [“Site Acceptance Test \(SAT\) Application” on page 90](#) for details.

Structural Variant Calling

- Identify structural variants (Default: ≥ 20 bp) in a sample or set of samples relative to a reference.
- See [“Structural Variant Calling Application” on page 94](#) for details.

Assembly (HGAP 4) Application

Use this application (**Hierarchical Genome Assembly Process**) to generate high quality *de novo* assemblies of genomes, using PacBio data.

- HGAP 4 includes pre-assembly, *de novo* assembly and assembly polishing steps.
- HGAP 4 uses Falcon for *de novo* assembly and Arrow for polishing.
- The application accepts Sequel data (BAM format) as input.

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

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

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

Advanced Parameters	Default Value	Description
Aggressive Mode	OFF	If ON , allows more overlaps to be detected and reported, which creates longer preads that go into assembly. This can be useful when a Data Set assembles poorly using the defaults, possibly due to lower quality input subreads. The default is OFF as this is not as well tested as the default options and may cause side-effects on larger, more complex genomes.
Seed Length Cutoff	-1	Only reads as long as this value will be used as seeds in the draft assembly. -1 means this will be calculated automatically so that the total number of seed bases equals (Genome Length times Seed Coverage.)
Consensus Algorithm	best	<ul style="list-style-type: none"> • Best is the best algorithm based on the data provided. • Arrow is a more sophisticated algorithm that provides additional information about each read, allowing more accurate consensus calls. Arrow does not use the alignment provided by the mapper except for determining how to group reads together at the gross level. Arrow implicitly performs its own realignment, so it is highly sensitive to all variant types, including indels. • Plurality is a very simple variant-calling algorithm which does not perform any local realignment. It is heavily biased by the alignment produced by the mapper, and it is insensitive at detecting indels.
Downsampling Factor	0	If > 1, a filter is applied to the input Data Set to reduce the number of reads by this factor. Example: A downsampling factor of 10 means that only 1/10th of the reads will be used. (This is mostly used for internal debugging.)
Minimum Mapped Concordance (%)	70	The minimum required alignment concordance, in percent.
Seed Coverage	30	A target value for the total number of "raw" postprimary reads, divided by the total number of seed reads. Valid values are 20 to 100.

Advanced Parameters	Default Value	Description
FALCON cfg Overrides	NONE	Allows PacBio Support engineers to override the configuration file generated from other options. This is a semicolon-separated list of KEY=VALUE pairs. New line characters are accepted, but ignored.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Biological Sample Name	NONE	The Sample Name for all Read groups. If not specified, sample names from the Data Set are used, with the following precedence: A) SM Field in the Input Read Group B) Bio Sample Name C) Well Sample name D) "UnnamedSample".
Minimum Mapped Length (bp)	50	The minimum required mapped read length, in base pairs.
Override pbmm2 Options	NONE	Custom pbmm2 mapping options.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

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

Alignment to Draft Assembly > Summary Metrics

Displays statistics on reads that aligned to the draft assembly.

- **Percent Aligned Bases:** The number of subread bases that aligned to the draft assembly, divided by the total number of bases in the BAM file.
- **Mean Concordance (aligned):** The mean concordance of subreads that aligned to the draft assembly.
- **Number of Subread Bases (aligned):** The number of subread bases that aligned to the draft assembly.
- **Number of Subreads (aligned):** The number of subreads that aligned to the draft assembly.
- **Number of Alignments:** The number of alignments that mapped to the reference sequence.
- **Alignment Length Mean (aligned):** The mean length of alignments that aligned to the draft assembly.
- **Alignment Length N50 (aligned):** The alignment length at which 50% of the alignments are longer than, or equal to, this value.
- **Alignment Length 95% (aligned):** The 95th percentile of length of alignments that aligned to the draft assembly.
- **Alignment Length Max (aligned):** The maximum length of alignments that aligned to the draft assembly.
- **Number of Polymerase Reads (aligned):** The number of polymerase reads that aligned to the draft assembly. This includes adapters.
- **Polymerase Read Length Mean (aligned):** 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 (aligned):** 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.

-
- **Polymerase Read Length 95% (aligned)**: The 95th percentile of read length of polymerase reads that aligned to the draft assembly.
 - **Polymerase Read Length Max (aligned)**: The maximum length of polymerase reads that aligned to the draft assembly.

Alignment to Draft Assembly > Alignment Statistics Summary

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

- **Movie**: Movie name for which the following metrics apply.
- **Number of Polymerase Reads (aligned)**: The number of polymerase reads that aligned to the draft assembly. This includes adapters.
- **Polymerase Read Length Mean (aligned)**: 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 (aligned)**: 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 (aligned)**: The number of subreads that aligned to the draft assembly.
- **Number of Subread Bases (aligned)**: The number of subread bases that aligned to the draft assembly.
- **Subread Length Mean (aligned)**: The mean length of the mapped portion of subreads that aligned to the draft assembly.
- **Mean Concordance (aligned)**: The mean concordance of subreads that aligned to the draft assembly.

Alignment to Draft Assembly > Aligned Polymerase Read Length

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

Alignment to Draft Assembly > Aligned Subread Length

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

Alignment to Draft Assembly > Aligned Subread Concordance

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

Alignment to Draft Assembly > Aligned Concordance vs Read Length

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

Polished Assembly > Summary Metrics

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

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

Polished Assembly > Contig Coverage vs Confidence

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

Coverage > Summary Metrics

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

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

Coverage > Coverage across Reference

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

Coverage > Depth of coverage Distribution

- Histogram distribution of the reference regions by the coverage.

Preassembly > Summary Metrics

Displays statistics on the pre-assembly process.

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

-
- **Seed Coverage (bases/genome_size):** The number of seed bases divided by the number of base pairs expected in the genome.
 - **Number of Pre-Assembled Reads:** The number of reads output by the pre-assembler. Pre-assembled reads are very long, highly accurate reads that can be used as input to a *de novo* assembler.
 - **Pre-Assembled Read Length Mean:** The mean length of the pre-assembled reads.
 - **Pre-Assembled Read Length (N50):** 50% of the pre-assembled reads are longer than this value.
 - **Pre-Assembled Read Length 95%:** The 95th percentile of the length of the reads output by the pre-assembler.
 - **Pre-Assembled E-size (sum of squares/sum):** The expected contig size for a random base in the pre-assembled reads.
 - **Number of Pre-Assembled Bases (total):** The total number of bases output by the pre-assembler.
 - **Pre-Assembled Coverage (bases/genome_size):** The number of bases output by the pre-assembler divided by the number of base pairs expected in the genome.
 - **Pre-Assembled Yield (bases/seed_bases):** The percentage of seed read bases that were successfully aligned to generate pre-assembled reads.
 - **Average Number of Reads that Each Seed is Broken Into:** The average number of preliminary reads that each seed is broken into. (Preliminary reads are derived from seeds using error correction; some portions of seeds might be too "noisy" to use.)
 - **Average Number of Bases Lost from Each Seed:** The average number of bases from each seed that were completely discarded.

Data > File Downloads

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

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

Base Modification Analysis Application

Use this application to identify putative sites of base modification as well as 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 Sequel data (BAM format) as input.

Reference Set (Required):

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

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

Advanced Parameters	Default Value	Description
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Minimum Mapped Concordance (%)	70	The minimum required alignment concordance, in percent.
Minimum Mapped Length (bp)	50	The minimum required alignment length, in base pairs.
Compute Methyl Fraction (experimental)	OFF	When identifying specific modifications (6mA and/or 4mC), enabling this option will estimate the methylated fraction, along with 95% confidence interval bounds.
Minimum Methylated Fraction	0.3	The minimum methylated fraction to identify a motif.
Downsampling Factor	0	If > 1, a filter is applied to the input Data Set to reduce the number of reads by this factor. Example: A downsampling factor of 10 means that only 1/10th of the reads will be used. (This is mostly used for internal debugging.)
P-Value	0.001	The probability value cutoff.
Minimum Qmod Score	30	The minimum QMod score used to identify a motif.
Biological Sample Name	NONE	The Sample Name for all Read groups. If not specified, sample names from the Data Set are used, with the following precedence: A) SM Field in the Input Read Group B) Bio Sample Name C) Well Sample name D) "UnnamedSample".
Override pbmm2 Options	NONE	Custom pbmm2 mapping options.

Advanced Parameters	Default Value	Description
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The Base Modification Detection application generates the following reports:

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 of the reference against the reference start position.

Coverage > Depth of coverage Distribution

- Histogram distribution of the reference regions by the coverage.

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

Modified Base Motifs > Modification QVs

- Maps motif sites against Modification QV.

Base Modifications > Per-Base Kinetic Detections

- Maps the modification QV against per-strand coverage.

Base Modifications > Kinetic Detections Histogram

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

Mapping Report > Summary Metrics

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

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped)**: The number of subread bases that mapped to the reference sequence.
- **Number of Subreads (mapped)**: The number of subreads 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.

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

-
- **Number of Subreads (mapped):** The number of subreads that mapped to the reference sequence.
 - **Number of Subread Bases (mapped):** The number of subread bases that mapped to the reference sequence.
 - **Subread Length Mean (mapped):** The mean length of the mapped portion of subreads that mapped to the reference sequence.
 - **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the reference sequence.

Mapping Report > Mapped Polymerase Read Length

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

Mapping Report > 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.

Mapping Report > Mapped Concordance vs Read Length

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

Data > File Downloads

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

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

CCS with Mapping Application

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

The CCS with Mapping application:

- Generates consensus sequences from single molecules.
- Accepts Sequel data (BAM format) as input.
- 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 takes multiple subreads of the same SMRTbell template and combines them to produce one high-quality consensus sequence. The Circular Consensus Sequences are then mapped to a reference sequence.

Reference Set (Required):

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

Minimum Number of Passes: (Default = 3)

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

Minimum Predicted Accuracy: (Default = 0.99)

- 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.)

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

Advanced Parameters	Default Value	Description
Minimum CCS Read Length	10	The minimum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.

Advanced Parameters	Default Value	Description
Minimum Mapped Length (bp)	50	The minimum mapped read length, in base pairs.
Override pbmm2 Options	NONE	Custom pbmm2 mapping options.
Maximum CCS Read Length	50,000	The maximum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Downsampling Factor	0	If > 1, a filter is applied to the input Data Set to reduce the number of reads by this factor. Example: A downsampling factor of 10 means that only 1/10th of the reads will be used. (This is mostly used for internal debugging.)
Minimum Mapped Concordance (%)	70	The minimum required alignment concordance, in percent.
Biological Sample Name	NONE	The Sample Name for all Read groups. If not specified, sample names from the Data Set are used, with the following precedence: A) SM Field in the Input Read Group B) Bio Sample Name C) Well Sample name D) "UnnamedSample".
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The CCS with Mapping application generates the following reports:

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

Coverage > Depth of Coverage Distribution

- Maps the reference regions against the percent coverage.

CCS Report > Summary Metrics

- **≥Q20 Reads:** The total number of CCS reads whose quality value is equal to or greater than 20.
- **≥Q20 Yield (bp):** The total yield (in base pairs) of the CCS reads whose quality value is equal to or greater than 20.
- **≥Q20 Read Length (mean, bp):** The mean read length of the CCS reads whose quality value is equal to or greater than 20.
- **≥Q20 Read Quality (median):** The median number of CCS reads whose quality value is equal to or greater than 20.

CCS Report > Read Length Distribution

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

CCS Report > CCS Read Quality Distribution

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

Mapping Report > Summary Metrics

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

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

Mapping Report > CCS Mapping Statistics Summary

Displays CCS mapping statistics per movie.

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

Mapping Report > Mapped CCS Read Length

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

Mapping Report > Mapped CCS Read Concordance

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

Mapping Report > Mapped Concordance vs Read Length

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

Mapping Report > Mapped QV Calibration

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

Data > File Downloads

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

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

**Circular
Consensus
Sequencing
(CCS)
Application**

Use this application to identify consensus sequences for single molecules. The application accepts Sequel data (BAM format) as input.

Minimum Number of Passes: (Default = 3)

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

Minimum Predicted Accuracy: (Default = 0.99)

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

Parameters

Advanced Parameters	Default Value	Description
Minimum CCS Read Length	10	The minimum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Maximum CCS Read Length	50,000	The maximum length for the median size of insert reads to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Generate consensus for each strand	OFF	Generate a consensus for each strand.
Downsampling Factor	0	If > 1, a filter is applied to the input Data Set to reduce the number of reads by this factor. Example: A downsampling factor of 10 means that only 1/10th of the reads will be used. (This is mostly used for internal debugging.)
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

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

CCS Report > Summary Metrics

- **≥Q20 Reads:** The total number of CCS reads whose quality value is equal to or greater than 20.
- **≥Q20 Yield (bp):** The total yield (in base pairs) of the CCS reads whose quality value is equal to or greater than 20.
- **≥Q20 Read Length (mean, bp):** The mean read length of the CCS reads whose quality value is equal to or greater than 20.
- **≥Q20 Read Quality (median):** The median number of CCS reads whose quality value is equal to or greater than 20.

CCS Report > Read Length Distribution

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

CCS Report > Read Quality Distribution

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

Data > File Downloads

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

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

Convert BAM to FASTX Application

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

- For **barcoded** runs, you must **first** run the **Demultiplex Barcodes** application to create BAM files **before** using this application.
- This application does **not** generate any reports.
- The application accepts Sequel data (BAM format) as input.

Parameters

Advanced Parameters	Default Value	Description
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Downsampling Factor	0	If > 1, a filter is applied to the input Data Set to reduce the number of reads by this factor. Example: A downsampling factor of 10 means that only 1/10th of the reads will be used. (This is mostly used for internal debugging.)
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Data > File Downloads

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

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

Demultiplex Barcodes Application

Use this application to separate sequence reads by barcode. (See [“Working with Barcoded Data” on page 97](#) for more details.)

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

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

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

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

Barcode Set (Required):

- Specify a barcode sequence file to separate the reads.

New Data Set Name (Required):

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

Same Barcodes on Both Ends of Sequence (Default = ON)

- Specify **On** to retain all the reads with the **same** barcodes on both ends of the insert sequence, such as symmetric and tailed designs. (See [“Working with Barcoded Data” on page 97](#) for information on barcode designs.)
- Specify **Off** to specify asymmetric designs where the barcodes are **different** on each end of the insert.

Minimum Barcode Score (Default = 0)

- 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 downstream analysis.

Infer Barcodes Used (Default = ON)

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

Parameters

Advanced Parameters	Default Value	Description
Write Unbarcoded Reads	ON	If specified, write out a separate Data Set containing reads that are not barcoded.
Filter Minimum Barcode Quality	26	Reads with barcodes below this threshold are filtered out from the demultiplexed Data Set XML file.
Ignore BioSamples Records from Run Design	OFF	If ON, Demultiplex Barcodes will ignore BioSample elements in the input SubreadSet XML; this prevents exiting with an error if the annotated barcodes cannot be found.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The Demultiplex Barcodes application generates the following reports:

Barcodes > Summary Metrics

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

Barcodes > Barcode Data

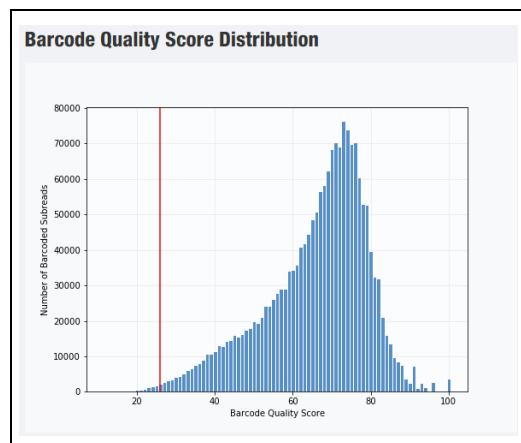
- **Bio Sample Name:** The name of the biological sample associated with the barcode.
- **Barcode Index:** The index number associated with the barcode.
- **Barcode Name:** A string containing the pair of barcode indices for which the following metrics apply.
- **Polymerase Reads:** The number of polymerase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

Barcodes > Barcoded Read Statistics

- **Number of Reads per Barcode:** Line graph displays the number of sorted reads per barcode.
 - **Good performance:** The Number of Reads per Barcode line (blue) should be mostly linear. Note that this depends on the choice of Y-axis scale. The mean Number of Reads per Barcode line (red) should be near the middle of the graph and should not be skewed by samples with too many or too few barcodes.
 - **Questionable performance:** A sharp discontinuity in the blue line, followed by no yield, with the red line way off center. This indicates that the user should allow the software to infer the barcodes.
- **Barcode Frequency Distribution:** Histogram distribution of read counts per barcode.
 - **Good performance:** A uniform distribution, which is most often a fairly tight symmetric normal distribution, with few barcodes in the tails.
 - **Questionable performance:** A large peak at zero indicates that the user should rerun the **Demultiplex Barcodes** application with the **Infer Barcodes** option set to **On**.
- **Mean Read Length Distribution:** Histogram distribution of the mean polymerase read length for all samples.
 - **Good performance:** The distribution should be normal with a relatively tight range.
 - **Questionable performance:** A spread out distribution, with a mode towards the low end.

Barcodes > Barcode Quality Scores

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



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

Barcodes > Barcoded Read Binned Histograms

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

Data > File Downloads

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

- **Analysis Log:** Log information for the analysis workflow.
- **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Barcode Files:** Barcoded subread Data Sets; one file per barcode.
- **Barcode Summary CSV:** Data displayed in the reports, in CSV format.
- **Barcodes Unassigned:** Barcodes not assigned to a specific isoform.

Note: You can get the demultiplexed BAM files using the Data Management module's "Export Data Sets" feature. In the demultiplexed BAM output file a tag `bc` is added for each read, indicating the assigned barcode. The `bc` tag is the zero-based index of the barcodes in the FASTA file. For example, when using the barcodes `RSII_96_barcodes`, a subread with barcode `1bc1` identified on both sides will have the tag `bc:B:S,0,0` in the BAM output file.

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

Iso-Seq® Analysis Application

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

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

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

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

The application accepts **either** Sequel/Sequel II SubreadSet or ConsensusReadSet files as input, but does **not** accept mixed input.

The application includes four main steps:

1. **CCS**: Build Circular Consensus Sequences (CCSs) from each sequencing ZMW. (The Iso-Seq application skips this step when the input Data Set is a ConsensusReadSet.)
2. **Classify**: Identify and remove primers (which includes cDNA primers and optionally barcodes). Identify strandedness based on the 5' and 3' primers.
3. **Cluster (Optional)**: Trim off polyA tails. Also remove artificial concatemers but do **not** remove PCR chimeras. Perform *de novo* clustering and consensus calling. Output full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on estimated accuracies.
4. **Collapse (Optional)**: When a reference genome is selected, the Iso-Seq application maps HQ isoforms to the selected reference genome, and then collapses isoforms which mapped to similar genomic loci into unique isoform groups.

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

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

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.

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 IsoSeq v2 primer set.

```
>NEB_5p
GCAATGAAGTCGCAGGGTTGGG
>Clontech_5p
AAGCAGTGGTATCAACGCAGAGTACATGGGG
>NEB_Clontech_3p
GTACTCTGCGTTGATACCACTGCTT
```

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

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGGG
>tissue1_3p
atgacgcatcgtctgaGTACTCTGCGTTGATACCACTGCTT
>tissue2_3p
gcagagtcagttagGTACTCTGCGTTGATACCACTGCTT
>tissue3_3p
gagtgtactctagtaGTACTCTGCGTTGATACCACTGCTT
>tissue4_3p
catgtactgatacacaGTACTCTGCGTTGATACCACTGCTT
```

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
TGGATTGATATGTAATACGACTCACTATAG
>TeloPrimeModified_3p
CGCCTGAGA
```

Reference Set (Optional):

- Optionally specify a reference sequence 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.

Parameters

Advanced Parameters	Default Value	Description
Polish CCS	OFF	Specify whether to polish CCS sequences using Arrow.
Minimum CCS Read Length	50	The minimum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Maximum CCS Read Length	15,000	The maximum length for the median size of subreads in a ZMW to generate a consensus sequence. If the targeted template is known to be a particular size range, this can filter out alternative DNA templates.
Require and trim Poly(A) Tail	ON	ON means that polyA tails are required for a sequence to be considered full length. OFF means sequences do not need polyA tails to be considered full length.
Minimum Accuracy for High Quality Isoforms	0.99	The minimum estimated accuracy for an isoform to be considered "High-Quality". Isoforms below the cutoff will be "Low-Quality". Note: This option is only available when the input is a SubreadSet.
Minimum Mapped Length (bp)	50	The minimum required mapped read length, in base pairs.
Minimum Mapped Concordance (%)	95	The minimum required alignment concordance, in percent.
Minimum Mapped Coverage (%)	99	The minimum required alignment read coverage, in percent.
Maximum Fuzzy Junction Difference (bp)	5	Specify the maximum edit distance, in base pairs, between mergeable fuzzy junctions.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The Iso-Seq application generates the following reports:

CCS Report > Summary Metrics

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

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

CCS Report > By Movie

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

CCS Report > CCS Read Length

- Histogram of the CCS read lengths.

CCS Report > Number of Passes

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

Transcript Clustering > Summary Metrics

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

Transcript Clustering > Read Length of Consensus Isoforms Reads

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

Transcript Clustering > Average Quality Value of HQ and LQ Isoforms

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

CCS Read Classification > Summary Metrics

- **Reads:** The total number of CCS reads.
- **Reads with 5' and 3' Primers:** The number of CCS reads with 5' and 3' primer detected.
- **Non-Concatemer Reads with 5' and 3' Primers:** The number of non-concatemer CCS reads with 5' and 3' primer detected.
- **Non-Concatemer Reads with 5' and 3' Primers and Poly-A Tail:** The number of non-concatemer CCS reads with 5' and 3' primer and polyA tail detected. This is usually the number for full-length, non-concatemer (FLNC) reads, unless polyA tails are not present in the sample.
- **Mean Length of Full-Length Non-Concatemer Reads:** The mean length of the non-concatemer CCS reads with 5' and 3' Primers and polyA Tail 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.

CCS Read Classification > Primer Data

- **Bio Sample Name:** The name of the biological sample associated with the primer.
- **Primer Index:** The index number associated with the primer.

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

CCS Read Classification > Primer Read Statistics > Number Of Reads Per Primer

- Maps the number of reads per primer, sorted by primer ranking.

CCS Read Classification > Primer Read Statistics > Primer Frequency Distribution

- Maps the number of samples with primers by the number of reads with primers.

CCS Read Classification > Primer Read Statistics > Mean Read Length Distribution

- Maps the read mean length against the number of samples with primers.

CCS Read Classification > Primer Quality Scores

- Histogram of primer scores.

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

- Heat map of read lengths, sorted by ranking.

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

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

CCS Read Classification > Read 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 Tail detected.

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 > Read 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 workflow.
- **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **CCS FASTQ:** Circular Consensus Sequences in FASTQ format.
- **Full-Length Non-Concatemer Reads:** 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:** 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 FASTQ:** 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 TXT:** Report of full-length read association with collapsed filtered isoforms.
- **Collapsed Filtered Isoform Counts:** Report of read count information for each collapsed filtered isoform.

Note: For details on custom PacBio tags added to output BAM files by the Iso-Seq Application, see page 39 of **SMRT Tools Reference Guide**, or <https://pacbiofileformats.readthedocs.io/>.

Long Amplicon Analysis (LAA) Application

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

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

The application includes five main steps:

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

Parameters

Advanced Parameters	Default Value	Description
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Chimera Filter	ON	Specify whether to activate the chimera filter and separate all consensus chimeric outputs.
Phasing	ON	Specify that the fine phasing step take place.
Minimum Subread Length	3,000	The minimum length of input reads to use. To disable , set to 0.
Maximum Subread Length	0	The maximum length of input reads to use. To disable , set to 0.
Minimum SNR	2.5	The minimum required signal-to-noise ratio (SNR) for any of the four channels. Data with SNR <2.5 is typically considered lower quality.
Minimum Barcode Score	26	The minimum average barcode score required for subreads.

Advanced Parameters	Default Value	Description
Downsampling Factor	0	If > 1, adds a filter to the input Data Set to sample a random selection of ZMWs instead of the full set. Example: A factor of 10 means that 10% of the ZMWs will be used.
Clustering	ON	Specify whether to activate the coarse clustering phase.
Filter Input Reads by Presence of Both Flanking Barcodes	OFF	Specify whether to filter the input reads if both flanking barcodes are present.
Maximum Reads	2,000	The maximum number of input reads to cluster per barcode.
Minimum Read Score	0.75	The minimum read score of input subreads.
Ignore N Bases At End	0	When splitting, ignore N bases at the end. This prevents excessive splitting caused by degenerate primers.
Maximum Phasing Reads	500	The maximum number of input reads to use for phasing and consensus.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

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

Amplicon Inputs > Amplicon Input Molecule Summary

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

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

Amplicon Consensus > Amplicon Consensus Summary

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

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

-
- **Sequence Cluster:** An identifying number given to the cluster of sequences from which this consensus sequence was generated, roughly corresponding to one locus or amplicon.
 - **Sequence Phase:** An identifying number given to each phased haplotype within a sequence cluster.
 - **Length (Bp):** The length of the consensus amplicon sequence.
 - **Estimated Accuracy:** The estimated accuracy of the consensus amplicon sequence.
 - **Subreads Coverage:** The number of subreads used to call consensus for this sequence.

Data > File Downloads

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

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

Mapping Application

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

- Are generated from Sequel data by CCS analysis.
- Can be automatically generated once the Sequel Data is transferred off the instrument (if the corresponding setting is selected in Run Design).
- Are advantageous for amplicon applications due to their high accuracy and long read length.
- Are **highly accurate** – by default, analysis parameters are set to >99% accuracy, Q \geq 20.

The Mapping application:

- Excludes the CCS generation from the **CCS with Mapping** application.
- Accepts CCS data (BAM format) as input.
- Maps CCS 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.

Reference Set (Required):

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

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

Advanced Parameters	Default Value	Description
Biological Sample Name	NONE	The Sample Name for all Read groups. If not specified, sample names from the Data Set are used, with the following precedence: A) SM Field in the Input Read Group B) Bio Sample Name C) Well Sample name D) "UnnamedSample".
Minimum Mapped Length (bp)	50	The minimum required mapped read length, in base pairs.
Override pbmm2 Options	NONE	Custom pbmm2 mapping options.
Minimum Mapped Concordance (%)	70	The minimum required alignment concordance, in percent.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The Mapping application generates the following reports:

Mapping Report > Summary Metrics

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

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

Mapping Report > CCS Mapping Statistics Summary

Displays CCS mapping statistics per movie.

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

Mapping Report > Mapped CCS Read Length

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

Mapping Report > Mapped CCS Read Concordance

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

Mapping Report > Mapped Concordance vs Read Length

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

Mapping Report > Mapped QV Calibration

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

Coverage > Summary Metrics

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

Coverage > Coverage Across Reference

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

Coverage > Depth of Coverage Distribution

- Maps the reference regions against the percent coverage.

Data > File Downloads

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

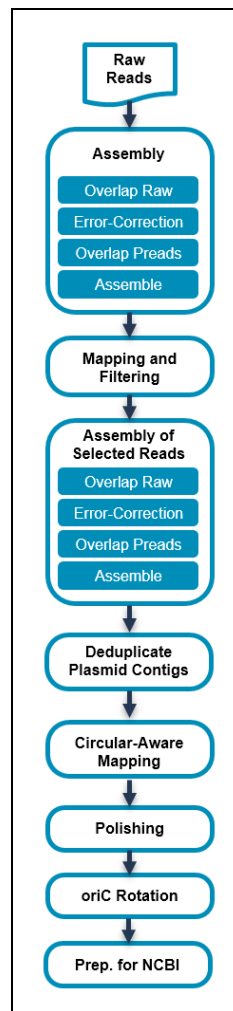
- **Analysis Log:** Log information for the analysis workflow.
- **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log.**)
- **Coverage Summary:** Coverage summary for regions (bins) spanning the reference.

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:

- 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.
- Accepts Sequel data (BAM format) as input.



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

Advanced Parameters	Default Value	Description
Seed Length Cutoff	-1	Only reads as long as this value will be used as seeds in the draft assembly. -1 means this will be calculated automatically so that the total number of seed bases equals (Genome Length x Coverage).
Coverage	30	A target value for the total amount of subread coverage used for assembly. This parameter is used, together with the genome size, to calculate the seed length cutoff.
Advanced Options	NONE	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.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The Microbial Assembly application generates the following reports:

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 Distribution

- Histogram distribution of the draft assembly regions by the coverage.

Alignment to Draft Assembly > Summary Metrics

Displays statistics on reads that aligned to the draft assembly.

- **Percent Aligned Bases:** The number of subread bases that aligned to the draft assembly, divided by the total number of bases in the BAM file.
- **Mean Concordance (aligned):** The mean concordance of subreads that aligned to the draft assembly.
- **Number of Subread Bases (aligned):** The number of subread bases that aligned to the draft assembly.

- **Number of Subreads (aligned):** The number of subreads that aligned to the draft assembly.
- **Number of Alignments:** The number of alignments that mapped to the reference sequence.
- **Alignment Length Mean (aligned):** The mean length of alignments that aligned to the draft assembly.
- **Alignment Length N50 (aligned):** The alignment length at which 50% of the alignments are longer than, or equal to, this value.
- **Alignment Length 95% (aligned):** The 95th percentile of length of alignments that aligned to the draft assembly.
- **Alignment Length Max (aligned):** The maximum length of alignments that aligned to the draft assembly.
- **Number of Polymerase Reads (aligned):** The number of polymerase reads that aligned to the draft assembly. This includes adapters.
- **Polymerase Read Length Mean (aligned):** 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 (aligned):** 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.
- **Polymerase Read Length 95% (aligned):** The 95th percentile of read length of polymerase reads that aligned to the draft assembly.
- **Polymerase Read Length Max (aligned):** The maximum length of polymerase reads that aligned to the draft assembly.

Alignment to Draft Assembly > Alignment Statistics Summary

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

- **Movie:** Movie name for which the following metrics apply.
- **Number of Polymerase Reads (aligned):** The number of polymerase reads that aligned to the draft assembly. This includes adapters.
- **Polymerase Read Length Mean (aligned):** 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 (aligned):** 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 (aligned):** The number of subreads that aligned to the draft assembly.
- **Number of Subread Bases (aligned):** The number of subread bases that aligned to the draft assembly.
- **Subread Length Mean (aligned):** The mean length of the mapped portion of subreads that aligned to the draft assembly.
- **Mean Concordance (aligned):** The mean concordance of subreads that aligned to the draft assembly.

Alignment to Draft Assembly > Aligned Polymerase Read Length

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

Alignment to Draft Assembly > Aligned Subread Length

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

Alignment to Draft Assembly > Aligned Subread Concordance

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

Alignment to Draft Assembly > Aligned Concordance vs Read Length

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

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 vs. Confidence

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

Data > File Downloads

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

- **Analysis Log:** Log information for the analysis workflow.
- **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Coverage Summary:** Coverage summary for regions (bins) spanning the reference.
- **Polished Assembly:** The polished assembly before oriC rotation is applied, in FASTA and FASTQ formats.
- **Final Assembly:** The final polished assembly with applied oriC rotation and header adjustment for NCBI submission, in FASTA format (.fsa extension).
- **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 applied.

Minor Variants Analysis Application

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

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

CCS Reads:

- Are generated from Sequel Data by CCS analysis.
- Can be automatically generated once the Sequel Data is transferred off the instrument (if the corresponding setting is selected in Run Design).
- Are advantageous for amplicon applications due to their high accuracy and long read length.
- Are **highly accurate** – by default, analysis parameters are set to $>99\%$ accuracy, $Q \geq 20$.

Reference Set (Required):

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

Target Config:

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

Parameters

Advanced Parameters - Minor Variants	Default Value	Description
Maximum Variant Frequency to Report (%)	100	Specify that only variants whose percentage of the population is less than this value be reported. Lowering this value helps to phase low-frequency variants when the highest frequency variant is different from the reference.
Minimum Variant Frequency to Report (%)	0.1	Specify that only variants whose percentage of the population is greater than this value be reported. Increasing this value helps to reduce PCR noise.
Phase Variants	ON	Specify whether to phase variants and cluster haplotypes.
Only Report Variants in Target Config	OFF	Specify whether to only report variants that confer drug resistance, as listed in the target configuration file.
Region of Interest	NONE	Specify genomic regions of interest; reads will be clipped to that region. If not specified, specifies all reads.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

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

1. Input Data

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

2. Target Config

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

▼ Target config

Config Version: Predefined v1.1, PacBio internal

Reference Name: HIV HXB2

Reference Length: 9719

Genes:

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

3. Variant Discovery

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

Each row represents a variant position. Each variant position consists of the reference codon, reference amino acid, relative amino acid position in

the gene, mutated codon, percentage, mutated amino acid, coverage, and possible affected drugs.

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

▼ Variant Discovery

HIV HXB2			Reverse Transcriptase					Affected Drugs*	
Codon	AA	Pos	AA	Codon	%	Coverage			
ATG	M	41	L	T T G	1	2793	ABC + DDI + TDF + D4T + ZDV		
AAA	K	65	R	A G A	1.1	2529	3TC + FTC + ABC + DDI + TDF + D4T		
			Pos	A	C	G	T	-	N
			-3	2947	0	0	0	0	51
			-2	2923	0	2	0	0	73
			-1	4	0	2952	0	0	42
			0	2606	0	0	0	339	53
			1	2905	0	29	0	0	64
			2	2938	0	0	0	0	60
			3	2938	0	0	0	0	60
			4	2942	0	0	0	0	56
			5	2751	0	0	0	0	247
TAT	Y	181	C	T G T	0.91	2946	NVP + EFV + ETR + RPV		
GGA	G	190	A	G C A	1	2947	NVP + EFV + ETR + RPV		
ACC	T	215	Y	T A C	0.93	2877	ABC + DDI + TDF + D4T + ZDV		

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

► Legend

4. Drug Summaries

Summarizes the variants grouped by annotated drug mutations:

▼ Drug Summaries

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

Phasing

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

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

Reverse Transcriptase										A	B	C	D	E	F	G	H	I
HXB2		Sample Variants					Haplotypes %											
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*	92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.8	0.7	
A T G	M	41	L	T T G	0.99	2903	ABC + DDI + TDF + D4T + ZDV											
A A A	K	65	R	A G A	1	2577	3TC + FTC + ABC + DDI + TDF + D4T											
G G G	G	99	G	G G T	0.72	2907												
T T A	L	100	F	T T T	0.85	2819	MGI											
T A T	Y	181	C	T G T	0.95	2939	NVP + EFV + ETR + RPV											
G G A	G	190	A	G C A	1	2941	MGI + NVP + EFV + ETR + RPV											
A C C	T	215	Y	T A C	0.88	2940	ABC + DDI + TDF + D4T + ZDV											

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

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

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

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

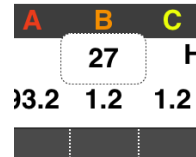
quality; as they are marginals, they are not exclusive and can overlap. The following image shows a sample with bad PCR conditions:

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

Haplotypes %

2.8 2.2 1.3 1 1 1 1 0.9 0.7

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



Resequencing Application

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

The Resequencing application:

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

Reference Set (Required):

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

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

Advanced Parameters	Default Value	Description
Consensus Algorithm	arrow	<ul style="list-style-type: none"> • Arrow is a more sophisticated algorithm that provides additional information about each read, allowing more accurate consensus calls. Arrow does not use the alignment provided by the mapper except for determining how to group reads together at the gross level. Arrow implicitly performs its own realignment, so it is highly sensitive to all variant types, including indels. • Plurality is a very simple variant-calling algorithm which does not perform any local realignment. It is heavily biased by the alignment produced by the mapper, and it is insensitive at detecting indels. • Best is the best algorithm based on the data provided.
Biological Sample Name	NONE	The Sample Name for all Read groups. If not specified, sample names from the Data Set are used, with the following precedence: A) SM Field in the Input Read Group B) Bio Sample Name C) Well Sample name D) "UnnamedSample".
Minimum Mapped Length (bp)	50	The minimum required mapped read length, in base pairs.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Minimum Mapped Concordance (%)	70	The minimum required alignment concordance, in percent.
Downsampling Factor	0	If > 1, adds a filter to the input Data Set to sample a random selection of ZMWs instead of the full set. Example: A factor of 10 means that 10% of the ZMWs will be used.

Advanced Parameters	Default Value	Description
Extract Unmapped Reads	OFF	Generates a BAM file containing all input reads that could not be mapped to the reference.
Override pbmm2 Options	NONE	Custom pbmm2 mapping options.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

The Resequencing application generates the following reports:

Mapping Report > Summary Metrics

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

- **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped):** The number of subread bases that mapped to the reference sequence.
- **Number of Subreads (mapped):** The number of subreads 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.

- **Movie:** Movie name for which the following metrics apply.
- **Number of Polymerase Reads (mapped):** The number of polymerase reads that mapped to the reference sequence. This includes adapters.
- **Polymerase Read Length Mean (mapped):** The mean read length of polymerase reads that mapped to the reference sequence, starting from the first mapped base of the first mapped subread, and ending at the last mapped base of the last mapped subread.

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

Mapping Report > Mapped Polymerase Read Length

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

Mapping Report > 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.

Mapping Report > Mapped Concordance vs Read Length

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

Consensus Variants > Summary Metrics

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

Consensus Variants > Consensus Calling Results

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

Consensus Variants > Observed Variants Across Reference

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

Top Variants > High-Confidence Variance Calls

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

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

Coverage > Summary Metrics

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

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

Coverage > Coverage Across Reference

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

Coverage > Depth of Coverage Distribution

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

Data > File Downloads

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

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

Site Acceptance Test (SAT) Application

Use this application to generate a report displaying site acceptance test metrics. This application is used to validate all new PacBio systems upon installation, and is designed to be run using specific lambda sequencing data included with the SMRT Link software.

The application accepts Sequel data (BAM format) as input.

Reference Set (Required):

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

Parameters

Advanced Parameters	Default Value	Description
Downsampling Factor	0	If > 1, adds a filter to the input Data Set to sample a random selection of ZMWs instead of the full set. Example: A factor of 10 means that 10% of the ZMWs will be used.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

Reports and Data Files

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

Site Acceptance Test Report > Summary Metrics

- **Instrument ID:** The ID number of the 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.

Top Variants > High-Confidence Variance Calls

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

- **Sequence:** The name of the reference sequence.
- **Position:** The position of the variant along the reference sequence.
- **Variant:** The variant position, type, and affected nucleotide.
- **Type:** The variant type: Insertion, Deletion, or Substitution.
- **Coverage:** The coverage at position.
- **Confidence:** The confidence of the variant call.

-
- **Genotype:** Includes the full number of chromosomes (diploid) or half the number (haploid).

Consensus Variants > Summary Metrics

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

Consensus Variants > Consensus Calling Results

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

Consensus Variants > Observed Variants Across Reference

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

Coverage > Summary Metrics

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

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

Coverage > Coverage across Reference

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

Coverage > Depth of Coverage Distribution

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

Mapping Report > Summary Metrics

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

- **Mean Concordance (mapped):** The mean concordance of subreads that mapped to the reference sequence.
- **Number of Subread Bases (mapped):** The number of subread bases that mapped to the reference sequence.
- **Number of Subreads (mapped):** The number of subreads 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.

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

Mapping Report > Mapped Polymerase Read Length

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

Mapping Report > 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.

Mapping Report > Mapped Concordance vs Read Length

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

Data > File Downloads

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

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

Structural Variant Calling Application

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

The application accepts Sequel data and CCS data (BAM format) as input.

Reference Set (Required):

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

Parameters

Advanced Parameters	Default Value	Description
Minimum Length of Structural Variant (bp)	20	Minimum length of Structural Variant, in base pairs.
Minimum Length of Copy Number Variant (bp)	1,000	Minimum length of a copy number variant, in base pairs.
Minimum % of Reads that Support Variant (any one sample)	20	Ignore calls supported by <P% of reads in every sample.
Minimum Reads that Support Variant (any one sample)	2	Ignore calls supported by <N reads in every sample.
Minimum Reads that Support Variant (total over all samples)	2	Ignore calls supported by <N reads total across samples.
Filters to Add to the Data Set	NONE	A semicolon or comma-separated list of other filters to add to the Data Set.
Minimum Mapped Length (bp)	50	The minimum required mapped read length, in base pairs.
Minimum Mapped Concordance (%)	70	The minimum required alignment concordance, in percent
Biological Sample Name	NONE	The Sample Name for all Read groups. If not specified, sample names from the Data Set are used, with the following precedence: A) SM Field in the Input Read Group B) Bio Sample Name C) Well Sample name D) "UnnamedSample".
Compute Settings	Select	Specify the distributed computing cluster settings configuration, if made available by the Site SMRT Link Administrator.

To Launch a Multi-Sample Analysis

1. Click + **Create New Analysis**.
2. Enter a **name** for the analysis.
3. Select the **type of data** (Sequel Data or CCS Data) 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.

- The Bio Sample name is **strongly recommended**. To add or edit this information, see [“Editing Data Set Information” on page 26](#).
- 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.

Report > Count by Sample (Genotype)

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

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

Report > Count by Annotation

This table describes the called variants broken down by a set of repeat annotations. Each variant is counted once (regardless of sample

genotypes) and assigned to exactly **one** annotation category. Only insertion and deletion variants are considered in this report.

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

Report > Structural Variants > Length Histogram

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

Data > File Downloads

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

- **Analysis Log:** Log information for the analysis workflow.
- **SMRT Link Log:** Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log.**)
- **Aligned Reads (per sample):** Aligned reads, in BAM format, separated by individual.
- **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.

Working with Barcoded Data

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

The canned data provided with SMRT Link v8.0 includes 7 barcode sets:

- `Sequel_16_barcodes_v1`
- `Sequel_16_Barcodes_v3`
- `Sequel_96_barcodes_v1`
- `Sequel_384_barcodes_v1`
- `RSII_96_barcodes`
- `RSII_384_barcodes`
- `IsoSeqPrimers_v2` (Includes the content of `IsoSeqPrimers` as well as support for NEB and Clontech primers.)

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

1. In SMRT Link, create a new run design as described in “[Creating a New Run Design](#)” on page 10. **Before** you finish the new Run Design, perform the following steps.

Barcoded Sample Options

Sample Is Barcoded YES NO

Barcode Set Required

Same Barcodes on Both Ends of Sequence YES NO

Autofilled Barcode Name File

Barcoded Sample Name File

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.
4. Specify if the **same** barcodes are used on both ends of the sequences.
 - Selecting **On** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
 - Selecting **Off** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such experiments retains any barcode pair combination identified in the Data Set.
5. SMRT Link **automatically** creates a CSV-format **Autofilled Barcode Name File**. The barcode name is populated based on your choice of barcode set, and if the barcodes are the same at both ends of the

sequence. The file includes a blank column for the biological sample names.

- **(Optional)** To specify the biological sample names corresponding to each barcode, click **Download Data**, enter the biological sample names associated with the barcodes (Maximum: 40 characters) in the second column, and save the file. If you did **not** use all barcodes in the Autofilled Barcode Name file in the sequencing run, either leave the biological sample name column blank for those barcodes, or delete those rows.
 - If you **don't** specify the biological sample name, it will automatically be set to the same value as the barcode name in SMRT Link.
 - **Note:** Open the CSV file in a text editor and check that the columns are separated by **commas**, not semicolons.
6. **(Optional)** Select the **Barcoded Sample Name File** you edited in **Step 5**. If you do **not** upload a Barcoded Sample Name File, the biological sample names for those barcodes will **automatically** be set to the barcode names.
 7. Click **Create**.

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

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

Example: bc1001--bc1001,biological sample name 1

Step 2: Perform the Sequencing Run

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

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

Step 3:
(Optional) Run
the Demultiplex
Barcodes
Application

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

1. Click **+ Create New Analysis**.
2. Enter a **name** for the analysis.
3. Select the type of data to use for the analysis:
 - **Sequel Data**: Subreads from Sequel or Sequel II Systems.
 - **CCS Data**: Single Molecule consensus reads generated from the CCS analysis.

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.

The screenshot shows the configuration interface for the 'Demultiplex Barcodes' application. It includes the following fields and options:

- Analysis Application** (Required): A dropdown menu with 'Demultiplex Barcodes' selected.
- Associated Inputs**: A section header.
- Barcode Set** (Required): A dropdown menu with 'Sequel_384_barcodes_v1' selected.
- New Dataset Name** (Required): A text input field containing 'Test_384_Dataset'. Below it, a note states: 'This is the name that the dataset will appear as in Data Management.'
- Same Barcodes on Both Ends of Sequence** (with a help icon): A toggle switch set to 'ON'.
- Minimum Barcode Score** (with a help icon): A text input field containing '26'.
- Infer Barcodes Used** (with a help icon): A toggle switch set to 'ON'.

7. Specify a Barcode Set (barcode sequence file.)
8. Specify the name for the new demultiplexed Data Set that will display in SMRT Link.
9. Specify if the **same** barcodes are used on both ends of the sequences.

- Selecting **On** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
 - Selecting **Off** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such data retains any barcode pair combination identified in the Data Set.
10. Specify the **Minimum Barcode Score**: Reads with barcode scores below the value are **not** included in downstream analysis. We recommend that you set this value to 26 for **all** applications.
 11. Specify if you want to infer which barcodes were used:
 - **On** infers which subset of barcodes from the selected barcode set were used, and outputs **one** data set for **each** of those inferred barcodes.
 - **Off** outputs **one** data set with **all** barcodes in the selected barcode set.
 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 61](#).

Step 4: Run Applications Using the Demultiplexed Data as Input

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

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

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

The screenshot shows a web interface for managing data sets. At the top, there's a 'Data Sets' section with a table. The table has two columns: 'Name' and 'Demultiplexed Subsets'. The first row is 'Re-barcode Alice/Bob/Charles' with a '3' in the 'Demultiplexed Subsets' column, which is circled in red. Below this is a section titled 'Members of Re-barcode Alice/Bob/Charles' with a sub-table. This sub-table has columns for 'Name' and 'Well Sample Name'. The first row is selected and highlighted in blue, showing 'Re-barcode Alice/Bob/Charles (Charles)' and 'T0213_384-plex_barcode_A01'.

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

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

Analysis Type

AUTO ANALYSIS ANALYSIS

Analysis of Multiple Data Sets

One Analysis per Data Set - Identical Parameters

One Analysis for All Data Sets

One Analysis per Data Set - Identical Parameters

One Analysis per Data Set - Custom Parameters

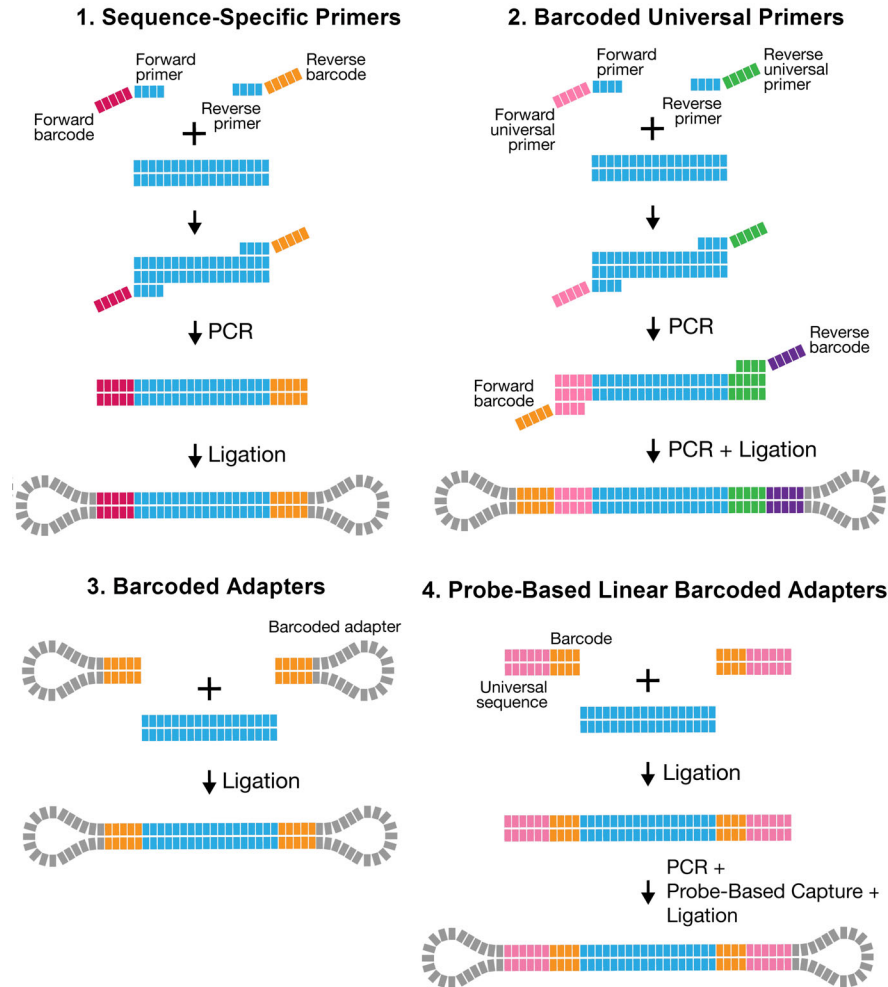
- **One Analysis on All Data Sets:** Runs **one** analysis using all the selected barcode Data Sets as input, for a maximum of 30 Data Sets.
 - **One Analysis per Data Set - Identical Parameters:** Runs **one** separate analysis for **each** of the selected barcode Data Sets, using the **same** parameters, for a maximum of 384 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.
4. Click **Start** to submit the analysis.

Demultiplex Barcodes Application Details

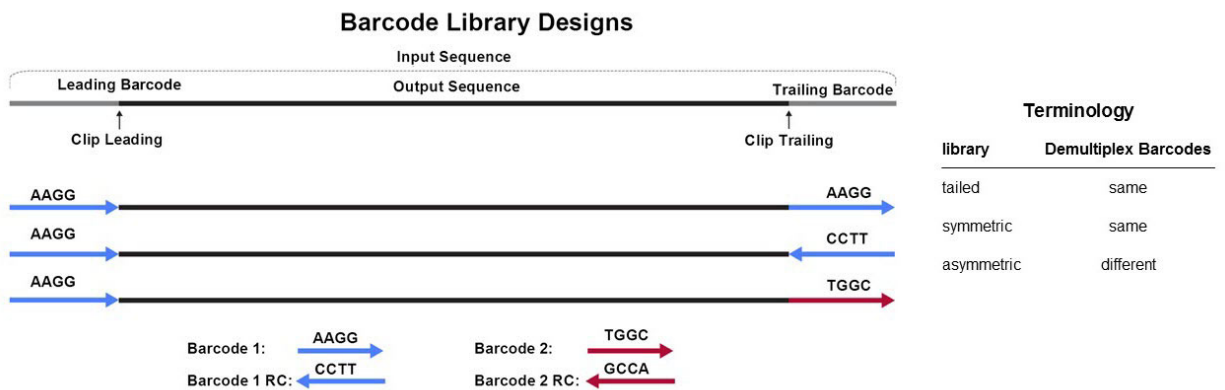
The **Demultiplex Barcodes** application identifies barcode sequences in PacBio single-molecule sequencing data. It **replaced** `pbbarcode` and `bam2bam` for demultiplexing, starting with SMRT Analysis v5.1.0.

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

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



In addition, there are three different barcode library designs.



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

Demultiplexing of CCS reads is possible on the command line.

Symmetric Mode

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

Asymmetric Mode

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

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

Mixed Mode

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

Workflow

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

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 CCS or CLR data, and includes **all** analysis applications available for the corresponding data type.
- Auto Analysis works with **both** Sequel and Sequel II 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 = **CCS Reads** and Generate CCS Data = **Yes** (Default)
- Read Type = **CCS Reads** and Sample is Barcoded = **Yes**.
- Read Type = **Continuous Long Reads** and Sample is Barcoded = **Yes**.

Note: Pre Analysis is available for **both** Sequel and Sequel II Systems.

Creating Auto Analysis From a Run Design

1. Create a new Run Design (See [“Creating a New Run Design” on page 10](#) 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 42](#).
9. Click **Create**.

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 <https://igv.org/> for further installation instruction and usage details.
- See <https://bit.ly/2NCTmAb> 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
- Mapping
- Microbial Assembly
- Resequencing
- 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**.

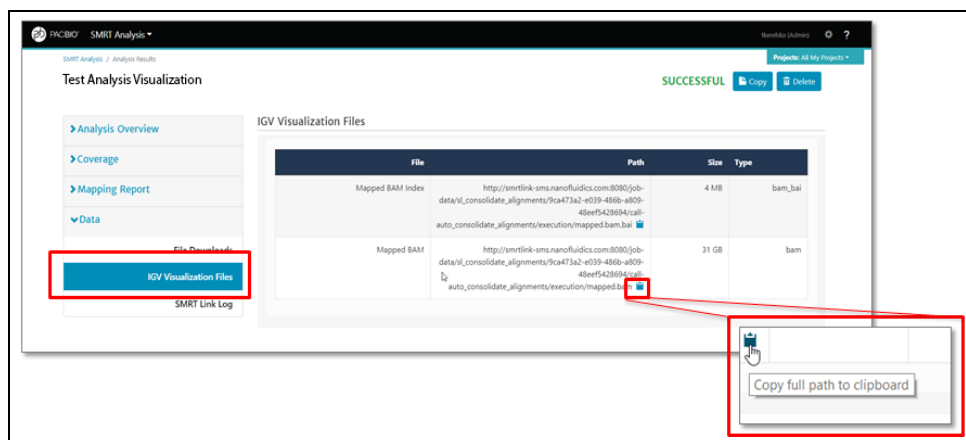
- When creating an analysis, you can specify that SMRT Link combines alignment BAM files for IGV visualization by setting the **Consolidate Mapped BAMs for IGV** option to **ON**.

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

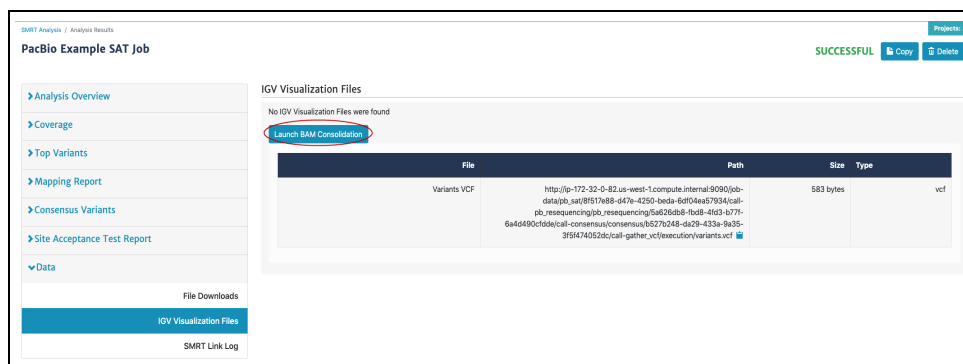
1. Create and run your analysis.
2. After the analysis has finished successfully, go to the **Data > IGV Visualization Files** section of the analysis report page.

- Open IGV and select the reference genome used for the analysis. (See <http://software.broadinstitute.org/software/igv/LoadGenome> for instructions on how to load a genome.)
- Copy a BAM file link from the **Data > IGV Visualization Files** section of the analysis report page.



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

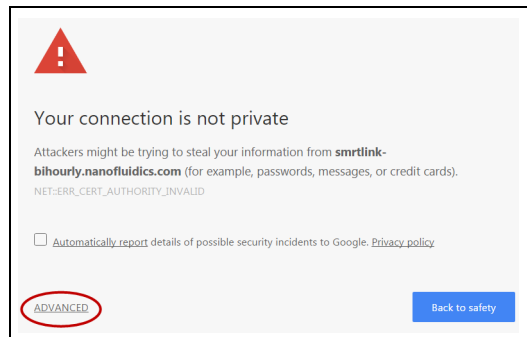
If you ran an analysis and there are **no Data > IGV Visualization Files** links, the analysis generated multiple alignment BAM files over 10 GB, but did **not** consolidate the files. Click the **Launch BAM Consolidation** button to consolidate them.



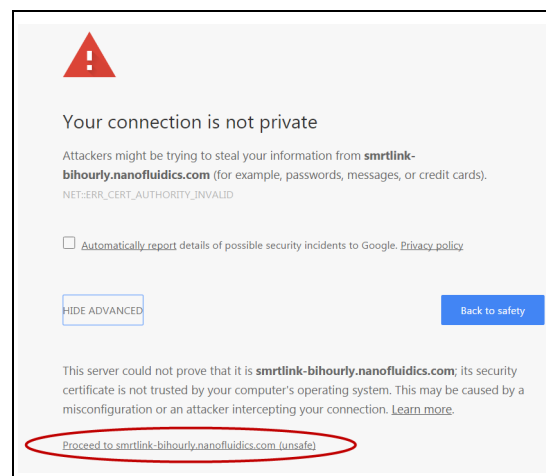
Using the PacBio® Self-Signed SSL Certificate

SMRT Link v8.0 ships with a PacBio Self-Signed SSL Certificate. If this is used at your site, security messages display when you try to login to SMRT Link for the **first time** using the Chrome browser. These messages may also display **other times** when accessing SMRT Link.

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



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



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



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

Sequel® System Output Files

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

File Structure

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

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

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

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

Subreads.BAM File

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

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

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

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

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

BAM.PBI File

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

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

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

For more information on the Pacific Biosciences BAM.PBI file format specifications, see <http://pacbiofileformats.readthedocs.io/en/5.1/PacBioBamIndex.html>.

Other Output Files

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

Note: This applies to files generated by Sequel Instrument Control Software (ICS) v3.1.0 or later.

- `.sts.xml`: Contains summary statistics about the collection/cell and its post-processing.
- `.control` files: Contains sequence data for spike-in-control reads.
- `.transferdone`: Contains a list of files successfully transferred.

Frequently Asked Questions

What are the minimum files needed to analyze data on SMRT Link?

- .bam file
- bam.pbi file
- subreadset.xml file

What is the average size of the file bundle for a 6-hour movie?

Approximately 5 Gb.

What is the difference between a regular .bam file and an aligned.bam file?

The `subreads.bam` file contains all the subreads sequences, while the `aligned.bam` file additionally contains the genomic coordinates of the reads mapped to a reference sequence.

The `subreads.bam` file is created by the PacBio Sequel/Sequel II System, while the `aligned.bam` file is created by SMRT Link after running Resequencing or Mapping analysis applications.

Secondary Analysis Output Files

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

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

```
$SMRT_ROOT/userdata/jobs_root/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.stderr
├─ pbscala-job.stdout
├─ 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.
- pbscala-job.stderr: Log collection of stderr output from pbscala.
- pbscala-job.stdout: Log collection of stdout output from pbscala. (**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 downstream 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 <http://genome.ucsc.edu/goldenPath/help/bam.html> 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 <http://genome.ucsc.edu/FAQ/FAQformat#format1> 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 http://en.wikipedia.org/wiki/FASTA_format or http://en.wikipedia.org/wiki/FASTQ_format for details.)
- **GFF:** General Feature Format, used for describing genes and other features associated with DNA, RNA and Protein sequences. (See <http://genome.ucsc.edu/FAQ/FAQformat#format3> for details.)
- **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 http://en.wikipedia.org/wiki/Variant_Call_Format 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 (v8.0)**.

SSL

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

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

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

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

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

- For details on updating SMRT Link to use an SSL Certificate, see the document **SMRT Link Software Installation (v8.0)**.

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

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

1. Choose **Gear > Configure**, then click **User Management**.
2. There are 2 ways to find users:
 - To display **all** SMRT Link users: Click **Display all Enabled Users**.
 - To find a specific user: Enter a user name, or partial name, and click **Search By Name**.
3. Click the desired user.
4. Click the **Role** field and select one of the three roles. (A **blank** role means that this user **cannot** access SMRT Link.)

-
- **Note:** There can be **multiple** users with the Admin role; but there **must** always be at least **one** Admin user.
5. Click **Save**.

Hardware/Software Requirements

Client Hardware Requirements

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

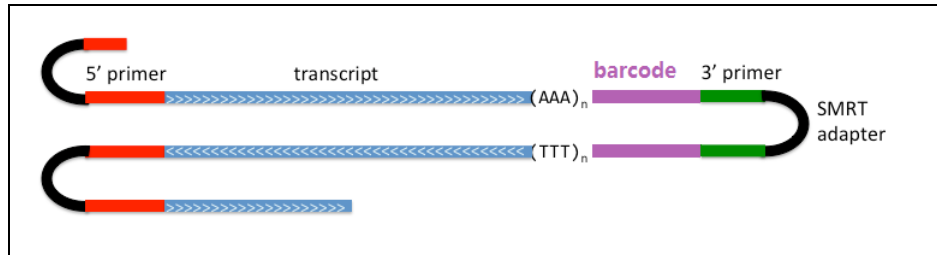
Client Software Requirements

- SMRT Link **requires** the Google® Chrome web browser, version 74 or later.

Note: SMRT Link **Server** hardware and software requirement are listed in the document **SMRT Link Software Installation (v8.0)**.

Appendix A - Barcoded Primers

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



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

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

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

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

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

Oligo	Sequence
dT_BC1	AAGCAGTGGTATCAACGCAGAGTACtcagacgatgctcatTTTVN
dT_BC2	AAGCAGTGGTATCAACGCAGAGTACctatacatgactctgcTTTVN

Then the primer file should look like this:

```
>F0
AAGCAGTGGTATCAACGCAGAGTAC
>R0
atgacgatgctctgaGTACTCTGCGTTGATACCACTGCTT
>F1
AAGCAGTGGTATCAACGCAGAGTAC
>R1
gcagagtcagtataagGTACTCTGCGTTGATACCACTGCTT
```

Appendix B - Pacific Biosciences Terminology

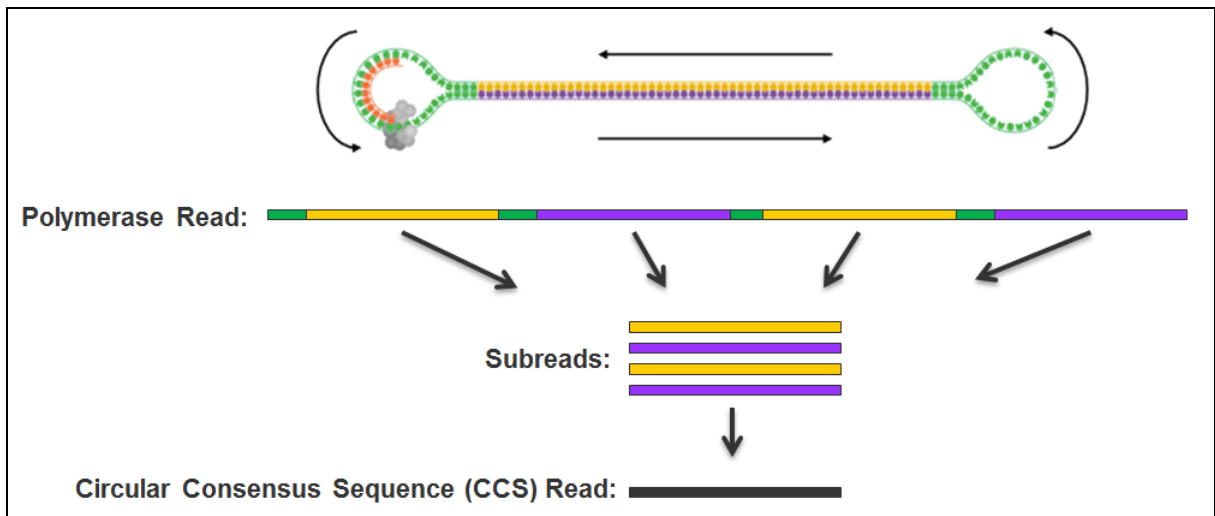
General Terminology

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

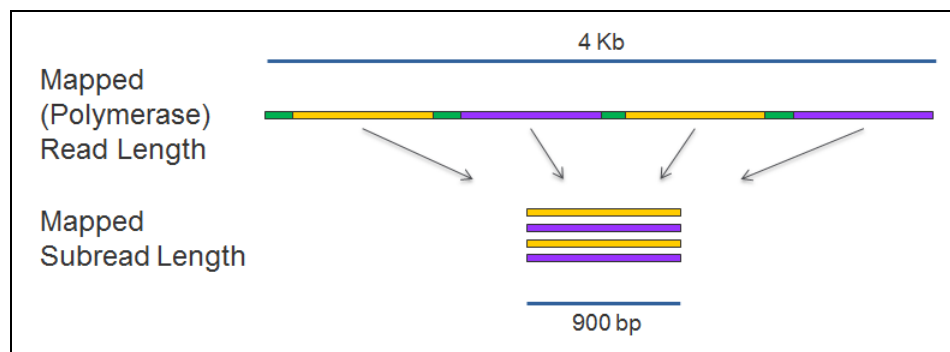
- **polymerase read:** A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell template. They can include sequences from adapters and from one or multiple passes around a circular template, which includes the insert of interest. Polymerase reads 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, resequencing, 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. CLR 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 advantageous for amplicon and RNA sequencing projects, and are **highly accurate** (>99% accuracy, Q₂₀).



Read Length Terminology

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



Secondary Analysis Terminology

- **secondary analysis:** Follows primary analysis and uses basecalled data. It is application-specific, and may include:
 - Filtering/selection of data that meets a desired criteria (such as quality, read length, and so on).
 - Comparison of reads to a reference or between each other for mapping and variant calling, consensus sequence determination, alignment and assembly (*de novo* or reference-based), variant identification, and so on.
 - Quality evaluations for a sequencing run, consensus sequence, assembly, and so on.
 - PacBio's SMRT Analysis contains a variety of secondary analysis applications including RNA and Epigenomics analysis tools.
- **secondary analysis application** (Formerly "Secondary analysis protocol"): A secondary analysis workflow that may include multiple analysis steps. Examples include *de novo* assembly, resequencing, RNA and epigenomics analysis.
- **consensus:** Generation of a consensus sequence from multiple-sequence alignment.
- **filtering:** Removes reads that do not meet the Read Length criteria set by the user.
- **mapping:** Local alignment of a read or subread to a reference sequence.
- **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.
 - **Sequel[®] II System:** Works with Sequel data, CCS data, and demultiplexed data.
 - **Sequel[®] System:** Can perform Auto Analysis or Pre Analysis, but **not** both.
- **Pre Analysis:** The process of CCS analysis and/or demultiplexing of Sequel basecalled data. Pre Analysis occurs **before** Auto Analysis.

Accuracy Terminology

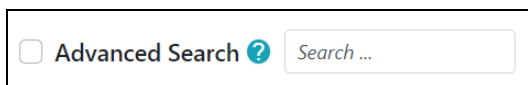
- **circular consensus accuracy:** Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.
- **consensus accuracy:** Accuracy based on aligning multiple sequencing reads or subreads together.
- **polymerase read quality:** A trained prediction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


Appendix C - Data Search

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

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

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

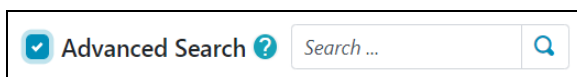




Advanced Search 

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

To perform an **advanced** search:

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



Advanced Search  

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

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

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

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

Example Data Set/barcode file/reference file search queries:

- name=TH_RC0_UnivABC_2pM_Diff-Cell11 (small)
- createdAt>2018-04-07
- numChildren>=2 AND isActive=true
- name=in:mito_mixtures_9plex,Duke_gDNA_SCNAcap_3

Example analysis search queries:

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

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

The tables below list allowed fields and their corresponding allowed operators and values.

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

Data Set, Barcode File and Reference File Fields

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

Analysis Fields

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

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