PacBio

SMRT® Link user guide (v11.1)

Sequel® II and IIe systems

SMRT® Link user guide (v11.1)

Introduction

This document describes how to use PacBio's SMRT Link software. SMRT Link is the web-based end-to-end workflow manager for Sequel II systems and Sequel IIe systems. SMRT Link includes the following modules:

- **Sample Setup**: Calculate binding and annealing reactions for preparing DNA libraries for use on **all** Sequel II systems and Sequel IIe systems. (See "Sample Setup" on page 8 for details.)
- **Run Design**: Design sequencing runs and create and/or import sample sheets. (See "Run Design" on page 16 for details.)
- **Run QC**: Monitor run progress, status and quality metrics. (See "Run QC" on page 27 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 34 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 44 for details.)

Note: SMRT Link v11.1 is for use with **Sequel II** systems and **Sequel IIe** systems **only**. If you are using a Sequel system, use an **earlier** version of SMRT Link.

This document also describes:

- The data files generated by the Sequel II system and Sequel IIe systems for each cell transferred to network storage. (See "Sequel® II system and Sequel IIe system output files" on page 132 for details.)
- The data files generated by secondary analysis. (See "Secondary analysis output files" on page 137 for details.)
- Configuration and user management. (See "Configuration and user management" on page 140 for details.)
- SMRT Link client hardware/software requirements. (See "Hardware/ software requirements" on page 143 for details.)

Installation of SMRT Link **server** software is discussed in the document **SMRT Link software installation guide (v11.1)**.

New features, fixed issues and known issues are listed in the document **SMRT Link release notes (v11.1)**.

When you first start SMRT Link, you must specify which system you are using: **Sequel II**, or **Sequel IIe**. This choice affects some of the initial values used in the Sample Setup and Run Design modules. In those modules, you can switch between the two Sequel systems as needed.

Users with administrator access can configure SMRT Link to support **all** instrument types.

Contact information

For additional technical support, contact PacBio 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 131.

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 II system or Sequel IIe system, 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 jobs. (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 instruments (Sequel II system, Sequel IIe system) that SMRT Link will be used with, click **Instruments** and check the appropriate boxe(s).
  - **Admin users only**: Add/delete SMRT Link users and specify their roles. See "Adding and deleting SMRT Link users" on page 141 for details.
  - To specify how numbers are formatted, click **Number Formatting** and select **Period** or **Comma** as the decimal separator.
  - **(Sequel IIe system only)** To specify whether CCS analysis output includes kinetics information (used for epigenetics analysis), click **CCS Analysis Output** and select **Yes** or **No**. This is the default setting for **all** CCS analysis output, unless overwritten in individual Run Designs. **Note**: Adding kinetics information can increase the amount of storage used by the output BAM files by up to **5 times**.
- **About SMRT Link**
  - Displays software version information and available space on the server SMRT Link is connected to.
  - Click **Send** to send configuration information and/or analysis usage information to PacBio Technical Support for help in troubleshooting failed jobs.
  - **Admin users only**: 1) Update the SMRT Link **Chemistry Bundle**, which includes kit and DNA Control Complex names used in the Sample Setup and Run Design modules. 2) Update the SMRT Link **UI Bundle**, which includes changes and bug fixes to the SMRT Link Graphical User Interface or UI for a SMRT Link module.
- **Sign Out**
  - Logs you out and displays the initial login page.

Working with tables

- To sort table columns: Click a column title.
- To see additional columns: Click the > symbol next to a column title.

- To search within a table: Enter a unique search string into the Search field. (For details, see "Appendix B - Data search" on page 147.)

Sending information to Technical Support

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

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

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

Sample Setup

To prepare your samples for sequencing, use SMRT Link's **Sample Setup** module to generate a customized protocol for primer annealing and polymerase binding to SMRTbell® templates, with subsequent sample cleanup. 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. Select **High-Throughput** mode, which provides a more streamlined workflow to efficiently process multiple samples with similar library properties (such as mean insert size and DNA concentration) in parallel. You can also export the calculated values to a CSV file for laboratory automation.
    Note: **Classic** mode is provided for legacy support purposes **only,** and is described later in this document.
4. Click **+ New Calculation**.
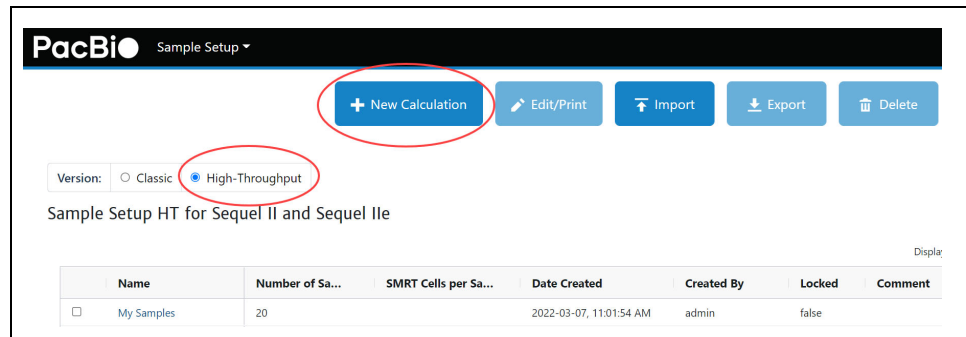
5. Enter the sample **name**.

Application-based calculations

6. Select a sequencing **application** for the sample. The following fields are **auto-populated** and display in green:
    – Binding Kit
    – Cleanup Anticipated Yield

7. Enter the **number of samples** for this calculation. Samples should be substantially equivalent to each other; all should have insert sizes and concentrations within +/- 15% of the specified values.
8. Enter the **number of SMRT® Cells** to bind per sample.
9. Enter the available **volume per sample**, in µL. When preparing multiple samples, this should be the **minimum** volume available for any sample.
10. Specify an **insert size,** in base pairs. The insert size is the length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters. This matches the mean insert size for the sample; the size range boundaries are described in the library preparation protocol. Enter the mean insert size of the sample(s).
11. Enter the sample **concentration(s)**, in ng/ul. Note that the acceptable range of input concentrations depends on insert size:

Insert size	Concentration range
10kb and up	20 - 60 ng/ul
3kb − 9999 bp	6 − 20 ng/ul
1.5kb − 2999 bp	3 − 10 ng/ul
500bp − 1499 bp	1 − 3 ng/ul

12. If necessary, edit the **Cleanup anticipated yield**. Adjust this percentage based on previous experience. (Cleanup removes excess primers/polymerase from bound complexes, which results in higher quality data.)
13. Specify the on-plate loading concentration (OPLC), in pM.
14. 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.
15. Optionally, do one of the following:
    − Click **Copy** to start a **new** sample group using the information entered. Then, edit specific fields for each sample group.
    − Click **Automate** to generate a CSV file. This exports the calculated values to a CSV file for lab automation.

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

Custom calculations

1. To accommodate new or unique sample types, choose **Application > Custom** and enter all settings manually.
2. Click **Set Custom Preset Values** to save any custom application settings you may have specified. The next time you select **Application > Custom**, those settings are retrieved.

Classic mode

Note: Classic mode is provided for legacy support purposes **only**. We **highly recommend** using High-Throughput mode even for single samples.

1. Select **Classic** mode.
2. Select **Sequel II** or **Sequel IIe**.
3. Click **+ New Calculation**.
4. Enter the sample **name**.
5. Select a sequencing **application** for the sample.
6. Enter the available sample **volume**, in µL.
7. Enter the sample **concentration**, in ng/ul.
8. Specify an **insert size,** in base pairs.
9. Select the **Internal Control** version to use for this run from the list, or type in a part number. PacBio **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 Internal Control for consumables to be eligible for reimbursement consideration.)
10. If necessary, edit the **Cleanup anticipated yield**. Adjust this percentage based on previous experience. (Cleanup removes excess primers/polymerase from bound complexes, which results in higher quality data.)
11. Specify the on-plate loading concentration (OPLC), in pM.
12. Enter the number of SMRT® Cells to bind, at the specified on-plate loading concentration.
13. Rather than leave a small amount of library behind, use the entire library volume available if desired by selecting **Prepare Entire Sample > Yes**. This generates annealing, binding and cleanup instructions for the **entire available sample volume**. The instructions for loading the sample plate will still follow the scale indicated by the specified number of SMRT Cells to run.
14. In the complex cleanup step, enter the pre- and post-cleanup sample DNA quantitation and volume measurement results.

	Sample 1	✓
Volume of Purified Complex (uL)	▮▮▮ uL	
Purified Complex Concentration (ng/uL)	▮▮▮ ng/uL	
Molar Concentration of Purified Complex (pM)	???	
Ampure Cleanup Yield (%)	???	

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

Final Loading Dilution

Reagent	Sample 1	✓	
Warning	Values measured in cleanup step must be entered		
Sequel® Complex Dilution Buffer	0.0 uL		
Prepared sample	0.0 uL		
Diluted Internal Control (Dilution 2)	0.0 uL		
DTT	0.0 uL		
Sequel Additive	0.0 uL		
Total Volume	0.0 uL		
# of SMRTCells requested	5		
Show values for a different number of cells	▮▮▮		
Show values for a different OPLC	▮▮▮		

Load 115 uL of sample per well and store at 4C for up to 24 hours before use.

16. Optionally, do one of the following:
    – Click **Copy** to start a **new** sample using the information entered. Then, edit specific fields for each sample.
    – Click **Remove** to delete the current calculation.
    – Click **Lock** to lock the calculation. This is **required** before samples can be imported into the Run Design module, and also sends a finalized version of the instructions to the server for use in Data Set reports. After locking, no further changes can be made to a calculation. (Click **View** to see the locked instructions.) Locking ensures that calculations are always synchronized with their run time state if a report is generated at a later date. (**Lock** is **only** available If there are one or more samples visible **and** most fields have values entered.)
    – Click the **New Sample** button at the top of the screen to start a new, empty sample.
17. Specify whether to display the **full** instructions, or only the **loading** instructions.

Advanced options

- Specify the **Minimum Pipetting Volume**, in uL. This allows you to set a lower limit on pipetting volumes to use in certain protocol steps, such as sample annealing and binding. We recommend setting this to 1 uL, though in some cases, for example if sample availability is very limited, it may be appropriate to set a value below 1 uL. Some protocol steps include fixed values of 1 uL that will **not** be affected by this setting.
- Specify the **% of Annealing Reaction to Use in Binding.** This accommodates pipetting underage: Due to pipetting issues, volumes may not add up to what they should; a value below 100% helps ensure there will be enough annealed sample for binding.

Editing or printing calculations

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

Deleting calculations

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

Importing/exporting calculations

Sample Setup supports importing and exporting calculations in CSV format.

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

Note: The content of the CSV file generated using the **Export** button in the Sample Setup home screen is **different** from the content of the CSV file generated using the High-Throughput mode's **Automate** button used for lab automation.

1. Access SMRT Link using the Chrome web browser.
2. Select **Sample Setup**.
3. Select **High-Throughput**.
4. Select an existing calculation.
5. Click **Export**, then click **Download**.
6. Edit the exported calculation in Excel (changing sample names, concentrations, and so on), then save it under a new name.

7. In Sample Setup, click **Import**.
8. Click **Browse**, then select the CSV file you previously modified in Step 6 and click **Open**. If everything is correct, click **Continue**. The imported calculation displays.

Note:

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

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

Field name	Required	Description
Sample Name	**Yes**	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods **only**.
System Name	**Yes**	**Must** be `Sequel II`, or `Sequel IIe`.
Application	**Yes**	Enter one of the following values: • `HiFi Reads` • `Microbial Assembly` • `Iso-Seq Method` • `MAS-Seq single cell` • `HiFiViral SARS-CoV-2` • `Adeno-Associated Virus` • `Full-Length 16S rRNA Sequencing` • `Shotgun Metagenomic Profiling or Assembly` • `<3kb Amplicons` • `>=3kb Amplicons` • `Custom`
Available Starting Sample Volume (uL)	**Yes**	Enter a positive integer. Units are in microliters.
Starting Sample Concentration (ng/uL)	**Yes**	Enter a positive integer. Units are in nanograms per microliter.
Insert Size (bp)	**Yes**	Enter a positive integer. Units are in base pairs.
Control Kit	**No**	**Must** be blank or `Lxxxxx101717600123199`.
Cleanup Anticipated Yield (%)	**No**	Enter a positive integer. **Note**: If **Application** is set to `Custom`, this field is **required**.
On Plate Loading Concentration (pM)	**Yes**	Enter a positive integer. Units are in parts per million.
Cells to Bind (cells)	**Yes**	Enter a positive integer.
Prepare Entire Sample	**Yes**	Enter a Boolean value: `true`, `t`, `yes`, `y`, `false`, `f`, `no`, or `n`. Boolean values are **not** case-sensitive.
Sequencing Primer	**Yes**	Enter one of the following values: • `Sequencing Primer v2` • `Sequencing Primer v4` • `Sequencing Primer v5`

Field name	Required	Description
Binding Kit	**Yes**	For Sequel II/IIe Binding Kits 2.0, 2.1, 2.2, 3.1 and 3.2: • `Lxxxxx101780500123199` (2.0) • `Lxxxxx101820500123199` (2.1) • `Lxxxxx101894200123199` (2.2) • `Lxxxxx102194200123199` (3.1) • `Lxxxxx102194100123199` (3.2)
Target Annealing Concentration (nM)	**No**	Enter a positive integer. Units are in nanomolar. **Note**: If **Application** is set to `Custom`, this field is **required**.
Target Binding Concentration (nM)	**No**	Enter a positive integer. Units are in nanomolar. **Note**: If **Application** is set to `Custom`, this field is **required**.
Target Polymerase Concentration (X)	**No**	Enter a positive integer. **Note**: If **Application** is set to `Custom`, this field is **required**.
Binding Time (hours)	**No**	Enter a positive integer. **Note**: If **Application** is set to `Custom`, this field is **required**.
Cleanup Bead Type	**No**	**Must** be `AMPure` or `ProNex`. **Note**: If **Application** is set to `Custom`, this field is **required**.
Cleanup Bead Concentration (X)	**No**	Enter a positive integer. **Note**: If **Application** is set to `Custom`, this field is **required**.
Minimum Pipetting Volume (uL)	**No**	Enter a positive integer. Units are in microliters.
Percent of Annealing Reaction To Use In Binding (%)	**No**	Enter a positive integer. **Note**: If **Application** is set to `Custom`, this field is **required**.
AMPure Diluted Bound Complex Volume (uL)	**No**	Enter a positive integer. Units are in microliters.
AMPure Diluted Bound Complex Concentration (ng/uL)	**No**	Enter a positive integer. Units are in nanograms per microliter.
AMPure Purified Complex Volume (uL)	**No**	Enter a positive integer. Units are in microliters.
AMPure Purified Complex Concentration (ng/uL)	**No**	Enter a positive integer. Units are in nanograms per microliter.
ProNex Diluted Bound Complex Volume (uL)	**No**	Enter a positive integer. Units are in microliters.
ProNex Diluted Bound Complex Concentration (ng/uL)	**No**	Enter a positive integer. Units are in nanograms per microliter.
ProNex Purified Complex Volume (uL)	**No**	Enter a positive integer. Units are in microliters.
ProNex Purified Complex Concentration (ng/uL)	**No**	Enter a positive integer. Units are in nanograms per microliter.
Requested Cells Alternate (cells)	**No**	Enter a positive integer.
Requested OPLC Alternate (pM)	**No**	Enter a positive integer. Units are in parts per million.

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`

Following are the fields contained in the CSV file generated by the **Automate** button in High-Throughput mode. This includes **all** the fields

that display in the Sample Setup page, with the volumes listed in each table easily accessible for liquid handling automation purposes.

Row	Field name
1	**Export Version** Version number of the file format specification. Allows for scripts to check version numbers to ensure compatibility through subsequent software releases.
2	**Instructions Version** Version number of SMRT Link, chemistry bundle, and parameters.
3	**Sample Group Name**
4	**Annealing Number of Samples**
5	**Annealing Sample Volume**
6	**Annealing Master Mix Volume**
7	**Annealing Incubation Temperature (C)**
8	**Annealing Incubation Time (minutes)**
9	**Polymerase Stock Volume**
10	**Sequel II Polymerase Dilution Buffer Volume**
11	**Binding Number of Samples**
12	**Binding Annealed Sample Volume**
13	**Binding Master Mix Volume**
14	**Binding Diluted Polymerase Volume**
15	**Binding Incubation Temperature (C)**
16	**Binding Incubation Time (minutes)**
17	**ICD1 Sequel Complex Dilution Buffer Volume**
18	**ICD1 Internal Control Stock Volume**
19	**ICD2 Sequel Complex Dilution Buffer Volume**
20	**ICD2 Diluted Internal Control (ICD1) Volume**
21	**ICD3 Sequel Complex Dilution Buffer Volume**
22	**ICD3 Diluted Internal Control (ICD2) Volume**
23	**Cleanup S2 Sample Input Volume**
24	**Cleanup S2 Diluent Volume**
25	**Cleanup S2 Binding Buffer**
26	**Cleanup S3 Bead Solution Volume**
27	**Cleanup S5 Elution Volume**
28	**Cleanup S5 Elution Buffer**
28	**Final Loading Number of Samples**
30	**Final Loading Prepared Sample Volume**
31	**Final Loading Diluted Internal Control (ICD3) Volume**
32	**Final Loading Volume (micro-liter)**

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 II systems and Sequel IIe systems linked to the same SMRT Link server.

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

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

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

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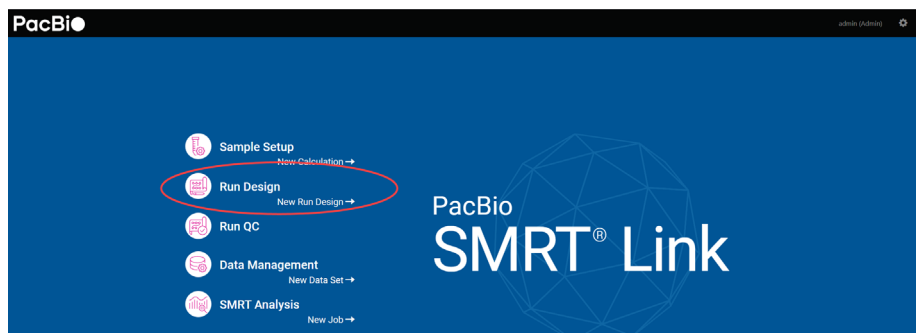
1. Access SMRT Link using the Chrome web browser.
2. Select **Run Design**.
3. Runs Designs can be sorted and searched for:
   - To sort Run Designs, click a **column title**.
   - To search for a Run Design, enter a unique search string into the **Search** field.
4. To initiate a new Run Design, click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

5. Specify if this Run Design is to be used with a Sequel II system or a Sequel IIe system. This affects the initial default values.
6. Enter a **Run Name**. (The software creates a new run name based on the current date and time; 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 are auto-populated as appropriate:
    – Sample Name
    – SMRTbell Adapter Design
    – Binding Kit
    – DNA Control Complex
    – Insert Size
    – On-Plate Loading Concentration

Application-based Run Designs

9. Select a sequencing **application** from the list. The following fields are auto-populated, and display in green:
   – Template Prep Kit
   – Binding Kit
   – Sequencing Kit
   – DNA Control Complex
   – Movie Time per SMRT Cell (hours)
   – Pre-Extension Time (hours)
10. Enter a **Well Sample Name**. (This is the name of the sequencing library loaded into one well. **Example**: `HG002_2019_11_02_10K`)
11. Enter a **Bio Sample Name**. (This is the name of the biological sample contained in the sequencing library, such as `HG002`. See "Working with barcoded data" on page 118 for details.)
12. (**Optional**) Enter **Sample Comments**.
13. Specify the **well position** used for this sample: Click the icon to the right of the entry field and choose a plate position.
14. Specify an **insert size** (500 base pairs minimum). The insert size is the length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters. This matches the average insert size for the sample; the size range boundaries are described in the library preparation protocol. **Note**: The default insert size for Subreads is 30,000; 10,000 for CCS reads.
15. Specify the **On-Plate loading concentration** (OPLC), in picomolarity.
16. (**Optional**) If you are using **barcoded samples**, see "Step 1: Specify the barcode setup and sample names in a Run Design" on page 119 for instructions. For details on secondary analysis of barcoded samples, see "Demultiplex Barcodes" on page 101.
17. Sample options:
    – Click **Copy**. This starts a new sample, using the values entered in the first sample.
    – Click **Delete**. This deletes the current sample.
    – Click **Add Sample**. This starts a new, empty sample.
18. 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.
19. 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.
20. (**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 126 for details.

If **MAS-Seq single cell** is selected as the application, the following additional **Auto Analysis** fields display:

- **Launch Read Segmentation**: Specify whether or not to split arrayed HiFi reads at MAS adapter positions, generating segmented reads (S-reads). If you specify **No**, you **cannot** launch Single-Cell Iso-Seq® Analysis.
- **Adapter Options**: Specify the MAS segmentation adapter choice to be **MAS16** (default) or **MAS12**.
- **Launch Single-Cell Iso-Seq Analysis**: Specify whether or not to enable analysis and functional characterization of full-length transcript isoforms with additional single-cell information, including single-cell barcodes and unique molecular identifiers (UMIs). To run this analysis, the **Read Segmentation** utility **must** be run first.
- **Sample Type**: Specify whether to use the **Human** or **Mouse** default reference genome and annotation set to align high-quality isoforms to, and to collapse isoforms mapped to the same genomic loci.
- **Analysis Name**: Specify the name of the analysis to be generated.

Custom Run Designs

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

- **SMRTbell Adapter Design, Template Prep Kit**, **Binding Kit**, or **Sequencing Kit**: Select one from the list, or type in a kit part number. If the barcode is invalid, "Invalid barcode" displays.
  Note: If the Sequencing or Binding kit selected is **incompatible**, an error message displays indicating the obsolete chemistry, and the run is **prevented** from proceeding.
- **DNA Control Complex**: PacBio **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 Internal Control for consumables to be eligible for reimbursement consideration.)
- **Movie time per SMRT Cell (hours)**: Enter a time between 0.5 and 30. **Note**: The **SMRT Cell 8M** part supports **all** movie times up to 30 hours.
- **Use Pre-Extension**: If selected, optionally specify the length of pre-extension time in hours. This initiates the sequencing reaction prior to data acquisition. After the specified time, the sequencing reagents are removed from the SMRT Cell and replenished with fresh reagents, and data acquisition starts. This feature is useful for short inserts (such as ≤15 kb) and provides a significant increase in read length.
- **Include 5mC Calls in CpG Motifs**: If selected, analyzes the kinetic signatures of cytosine bases in CpG motifs to identify the presence of 5mC.
- **Detect and Resolve Heteroduplex Reads**: Heteroduplexes are DNA molecules where the forward and reverse strands are not perfect reverse-complements. If the option is selected and heteroduplexes are detected, a consensus is called for **each** strand separately, and

the sequence of **both** strands is output.
Note: This option displays **only** if **Adeno-Associated Virus, Full-Length 16S rRNA Sequencing, <3kb Amplicons**, or **≥3kb Amplicons** are selected as the application.

Advanced options

- Specify whether to use **Adaptive Loading**. Adaptive Loading uses active monitoring of the ZMW loading process to predict a favorable loading end point. Certain steps (Cleanup and Sample Dilution) require a different buffer (Adaptive Loading Buffer) if this feature is used. **Note**: Adaptive Loading **requires** the use of Sequel® II binding kit 2.2. If you select **Yes**, fill in the following fields:
  - **Loading Target (P1 + P2)**: The fraction of ZMWs that the Adaptive Loading routine will aim to load with at least one sequencing complex. The default target for CCS applications is higher to accommodate loss of complexes during pre-extension, which is generally recommended for all CCS applications.
  - **Maximum Loading Time (hours)**: This defines the maximum time the system will allow loading to progress before proceeding to sequencing. (Loading time in Adaptive Loading is flexible.)
- Specify the length of time (1, 2 or 4 hours) for **immobilization** of SMRTbell templates. This is the length of time the SMRT Cell is at the Cell Prep Station to allow diffusion of SMRTbell templates into the ZMWs. This option is **not** available if **Adaptive Loading** is selected.
  - PacBio **highly recommends** using the default immobilization time of 2 hours.
- (**Sequel IIe systems only**) Specify, for this Run Design **only**, whether to include kinetics information (used for epigenetics analysis) in the CCS analysis output. This setting **overwrites** the global setting in **Gear > Configure > CCS Analysis Output**. **Note**: Adding kinetics information can increase the amount of storage used by the output BAM files by up to **5 times**.
- Specify, for this Run Design **only**, whether to include **low quality reads** (non-HiFi reads) in the CCS analysis output. Note that this option **disables** automatic demultiplexing, 5mC detection, and heteroduplex insert detection, if applicable.
- **Add Data to Project**: Specify that Data Sets generated by SMRT Cell(s) using this Run Design be associated with the selected Project. (This also applies to any Data Sets generated using Auto Analysis. By default, **all** Data Sets are assigned to **General Project**, which is accessible to all users.)

Editing or deleting Run Designs

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

Creating a Run Design by importing a CSV file

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

To obtain the sample CSV files

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

To update and import the CSV file

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

CSV file structure

- Each CSV file row represents **one sample**. (**Note**: The Run Design CSV has a limit of **15,000 samples**.)
- The first row contains run-level information such as Run Name, Run Comments, and so on.
- For demultiplexed samples **only**, **one additional row** per barcode/Bio Sample Name combination is added below the master sample row.

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

Outputting Subreads on the Sequel IIe system

The Sequel IIe system can be configured to output Subreads data in BAM format by using the Run Design CSV import mechanism. In addition to the other required columns, users can add the column **Emitted Subreads Percent** to the CSV file, with a value of 0-100 for a given collection. This results in the inclusion of Subreads from 0-100% of ZMWs in the Data Set transferred from the instrument, in a BAM file **separate** from the HiFi

reads. Note that this will **not** result in the inclusion of associated scraps data for each ZMW.

Run Design attribute	Required	Description
Experiment Name	No	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Defaults to Run Name. **Example**: Standard_Edna.1
Experiment Id	No	Enter a valid experiment ID. **Example**: 325/3250057 • Experiment IDs **cannot** contain the following characters: `<, >, :, ", \, |, ?, *,` or `)`. • Experiment IDs **cannot** start or end with a `/` and **cannot** have two adjacent `/` characters, such as `//`. • Experiment IDs **cannot** contain spaces. • Specifically, Experiment IDs **cannot** satisfy the regular expressions: `/[<>:"\\|?\*]/g, /(?:^\/)|\/\/|(?:\/$)/, / /g`
Experiment Description	No	Enter any ASCII string. Defaults to Run Comments. **Example**: 20170530_A6_VVnC_SampleSheet
Run Name	**Yes**	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. Run name must be entered for the first cell and will be applied to the remaining cells in the run. **Example**: 20170530_A6_VVnC_SampleSheet
System Name	No	Must be `Sequel II` or `Sequel IIe`.
Run Comments	No	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. **Example**: ecoliK12_March2021
Is Collection	No	Enter a Boolean value. (See Boolean details below.) Specifies whether the row designates a Collection (`TRUE`) or a barcoded sample (`FALSE`). • Collection lines should have the **Barcode Name** and **Bio Sample Name** fields blank. • Barcoded Sample lines only need to include the **Is Collection**, **Sample Name**, **Barcode Name**, and **Bio Sample Name** fields.
Sample Well	**Yes**	Must be specified in every row. Well number must start with a letter `A` through `H`, and end in a number `01` through `12`, i.e. `A01` through `H12`. It must satisfy the regular expression ``/^[A-H](?:0[1-9]|1[0-2])$/`` **Example**: A01
Well Sample Name	**Yes**	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. **Example**: A6_3230046_A01_SB_ChemKitv2_8rxnKit **Note**: The Sample Name **must** be unique within a run.
Movie Time per SMRT Cell (hours)	**Yes**	Enter a floating point number between 0.1 and 30. Time is in hours. **Example**: 5
Use Adaptive 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 alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. **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

Run Design attribute	Required	Description
Size Selection	No	Enter a Boolean value. (See Boolean details below.) Default is `FALSE`.
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 `diffusion` (not case-sensitive) or a custom script. (Sequel II systems do **not** support magbead loading.) A path can also be used, such as `/path/to/my/script/my_script.py`. The path will **not** be processed further, so if the full URI is required, it must be included in the CSV, such as `chemistry://path/to/my/script/my_script.py`.
Automation Parameters	No	To enable Pre-Extension time, enter the number of hours and set the boolean value to `TRUE`. **Example** 2 hours: `ExtensionTime=double:2|ExtendFirst=boolean:TRUE` (**Note**: Leave blank when not using Pre-Extension time, or set the boolean value to `FALSE`.)
Detect and Resolve Heteroduplex Reads	No	Enter a boolean value. (See Boolean details below.) Set to `TRUE` to allow for detection of heteroduplex reads. **Note**: Only applicable if **Application** is set to one of the following: • `Adeno-Associated Virus` • `<3kb Amplicons` • `>=3kb Amplicons` • `Custom`
Include 5mC Calls in CpG Motifs	No	Enter a boolean value. (See Boolean details below.) Set to `TRUE` to allow for 5mC calls in CpG motifs. **Note**: Only applicable if **Application** is set to `HiFi Reads` or `Custom`.
Sample is Barcoded	No	Enter a boolean value. (See Boolean details below.) Set to `TRUE` for a barcoded run.
Demultiplex Barcodes	No	Add any of the following values: `Do Not Generate`, `In SMRT Link`, or `On Instrument`. If left blank, the default is `Do Not Generate` for all systems. **Note**: This is available for all applications. The following values are recommended based upon your system: • **Sequel II system**: Enter one of the following values: `Do Not Generate` or `In SMRT Link`. • **Sequel IIe system**: Enter one of the following values: `Do Not Generate`, `In SMRT Link`, or `On Instrument`.
CCS Analysis Output - Include Low Quality Reads	No	Enter a boolean value. (See Boolean details below.) • Set to `TRUE` to allow for CCS analysis with `--all mode` activated and produce a `reads.bam` file • Set to `FALSE` to exclude all reads with `rq` < 0.99.

Run Design attribute	Required	Description
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 **Note**: This field is used for demultiplexed data.
Same Barcodes on Both Ends of Sequence	No	Enter a boolean value. (See Boolean details below.) Set to `TRUE` if symmetric, `FALSE` if asymmetric.
Barcode Name	No	Enter Barcode Names one per line. **Example**: `bc1001--bc1001` • Use double hyphens (`--`) to separate the 2 barcodes of each pair. • The barcode names **must** be contained within the specified Barcode Set. • A given barcode name **cannot** appear more than once in the spreadsheet. • A maximum of 15,000 barcodes is permitted per sample.
Bio Sample Name	**Yes**	Enter Bio Sample Names in the same row as their associated Barcode Names. Use alphanumeric characters, spaces (allowed but not recommended for compatibility with downstream software), hyphens, underscores, colons, or periods **only**. Bio Sample Names **cannot** be longer than 40 characters. **Example**: sample1 **Note**: This field is used for collections for non-multiplexed data, and for barcoded samples in multiplexed data.
Pipeline ID	No	**Note**: This field is **required** to create an Auto Analysis. • **5mC CpG Detection:** `cromwell.workflows.pb_detect_methyl` • **Demultiplex Barcodes:** `cromwell.workflows.pb_demux_ccs` • **Export Reads**: `cromwell.workflows.pb_export_ccs` • **Genome Assembly**: `cromwell.workflows.pb_assembly_hifi` • **HiFi Mapping:** `cromwell.workflows.pb_align_ccs` • **HiFiViral SARS CoV-2 Analysis**: `cromwell.workflows.pb_sars_cov2_kit` • **Iso-Seq Analysis**: `cromwell.workflows.pb_isoseq3_ccsonly` • **Mark PCR Duplicates**: `cromwell.workflows.pb_mark_duplicates` • **Microbial Genome Analysis**: `cromwell.workflows.pb_microbial_analysis` • **Read Segmentation:** `cromwell.workflows.pb_segment_reads` • **Read Segmentation and Single-Cell Iso-Seq Analysis**: `cromwell.workflows.pb_segment_reads_and_sc_isoseq` • **Single-Cell Iso-Seq Analysis**: `cromwell.workflows.pb_sc_isoseq` • **Structural Variant Calling**: `cromwell.workflows.pb_sv_ccs` • **Trim Ultra-Low Adapters:** `cromwell.workflows.pb_trim_adapters`
Analysis Name	No	Enter any ASCII string. See Auto Analysis Fields below for details. **Note**: This field is required for Auto Analysis, otherwise the name will be "". **Example**: sample 1 analysis

Run Design attribute	Required	Description

Entry Points	No	Entry Points only apply to Barcode Sets and Reference Sets. In addition, this field is **required** for Auto Analysis.
		
		Enter an ASCII string in the format `file_type;entry_id;uuid`, with parameters separated by | characters.
		  • To find the UUID: Click **Data Management > View Data > HiFi Reads or Subreads**. Click the Data Set of interest, then view the UUID.
		  • See the **SMRT® Tools reference guide** section **Appendix A - Application entry points and output files** to see the entry point names for each application.
		
		Example: PacBio.DataSet.BarcodeSet;eid_barcode;afe89e3f-17ca-e9b8-eae9-b701dbb1f02d|PacBio.DataSet.ReferenceSet;eid_ref_dataset;6b8db144-a601-4577-ab04-ba64cadc0548

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`. **Example**: pbmm2_align.task_options.minalnlength;integer;50 **Note**: This field is **optional** for Auto Analysis - any task options not specified will use pipeline defaults.
Application	No	• `HiFi Reads` • `Microbial Assembly` • `Iso-Seq Method` • `MAS-Seq single cell` • `HiFiViral SARS-CoV-2` • `Adeno-Associated Virus` • `Full-Length 16S rRNA Sequencing` • `Shotgun Metagenomic Profiling or Assembly` • `<3kb Amplicon Sequencing` • `≥3kb Amplicon Sequencing` • `Custom` If blank or contains invalid values, default is `Custom`.
CCS Analysis Output - Include Kinetics Information	No	Enter a boolean value. (See Boolean details below.) Set to `TRUE` to specify that CCS analysis output includes kinetics information (used for epigenetics analysis.) **Note**: Adding kinetics information can increase the amount of storage used by the output BAM files by up to **5 times**.

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

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

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

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

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

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

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

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

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

Auto Analysis fields

- The fields include Pipeline ID, Analysis Name, Entry Points, and Task Options.
- You can define **one** analysis for each Collection or Bio Sample. The Pipeline ID, Analysis Name and Entry Points fields are **required** to create an Auto Analysis.
- The analysis name is a concatenation of the values of the **Analysis Name** and **Bio Sample Name** fields.
- The **Task Options** field may be left blank; any task options not specified will use pipeline defaults.

Run QC

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

Metrics can be reviewed in the Run QC module. **All** Sequel II systems and Sequel IIe systems connected to SMRT Link can be reviewed using Run QC.

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

Accessing instrument status

1. Select **Instrument Status**. For **each** instrument connected to the instance of SMRT Link, this displays the instrument name and its current status, SMRT Cell status, when the run will be completed, any active alarms, and how many sequencing ZMWs are active.
   - A **red alarm** symbol displays next to the instrument status if any errors or warnings appear during a sequencing run.
   - If an instrument does **not** have a SMRT Cell tray loaded, the SMRT Cell Status field will **not** display any icons. The icons are:
     - **Fully green**: The SMRT Cell has completed sequencing.
     - **Half green**: The SMRT Cell is being prepared or is currently sequencing.
     - **White**: The SMRT Cell is in the queue for sequencing, but cell preparation has not started.

- **Run Completion**: Displays the estimated time remaining to complete sequencing run or the time elapsed since the sequencing run completed. Also displays the date (in YYYY-MM-DD format) when the last sequencing run was completed.
- **Sequencing ZMWs**: Displays a plot of how many ZMWs on a SMRT Cell are actively sequencing during a movie collection. For sequencing runs conducted with Binding kit 2.2 and 3.2, **only** the number of actively sequencing singly-loaded ZMWs (P1) displays. For sequencing runs conducted with Binding kit 2.1 and 3.1, the **total** number of actively sequencing ZMWs (P1 + P2) displays.

Note: Due to terminations, not **all** ZMWs are singly-loaded at the same time. Some ZMWs are singly-loaded only **at or near the end** of a movie collection, whereas others are singly-loaded only at the **beginning**. (**Singly-loaded** means that the ZMW contains only one active polymerase instead of two or more simultaneously active polymerases.) For runs conducted with Binding kit 2.2 and 3.2, the peak concurrent **Sequencing ZMWs** value shown in the plot will always be **less** than the final %P1 ZMW yield reported in the Run QC metrics table at the end of a movie collection. (For sequencing runs conducted with Binding kit 2.1 and 3.1, the peak concurrent Sequencing ZMWs value shown in the plot will always be **higher** than the final %P1 ZMW yield reported in Run QC.)
For a SMRT Cell that achieves ≥50% P1 loading and ≥10% P0, the ZMW Sequencing plot should typically display a peak value above 2,000,000.

See the figure below for an example comparison between the **Instrument Status report** (top) and **Run QC report** (bottom) for a WGS sample sequenced using Binding kit 3.2 with a 30-hour movie collection time. The Sequencing ZMWs plot in the Instrument Status report shows that the peak concurrent Sequencing ZMWs value for the last SMRT Cell in the run (Well D01) is approximately 3,000,000 ZMWs, whereas the final %P1 ZMW yield reported in the corresponding Run QC metrics table for Well D01 is 76.8% (or 6,144,000 P1 ZMWs.)
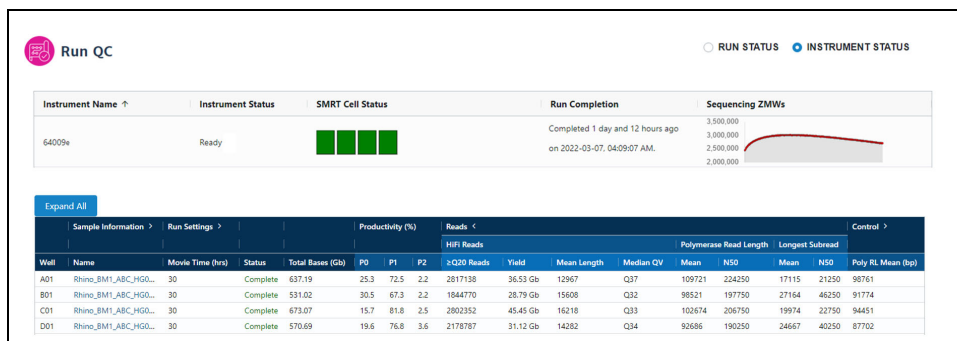
Accessing run information

1. Select **Run Status**.
2. Click **Sequel II** or **Sequel IIe** to view informations on runs for a specific instrument model.
3. Runs can be sorted and searched for:
   - To sort runs, click a **column title**.
   - To search for a run, enter a unique search string into the **Search** field.
4. Click the run to view information about that run.
5. To **export** Run QC data in CSV format: Select one or more runs in the table, then click **Export Selected**.

Table fields

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

- **Name**: A list of all runs for the instruments connected to SMRT Link. Click a run name to view more detailed information on the individual run page.
- **Summary**: A description of the run.
- **Dates**
  - **Run Date**: The date and time when the run was started.
  - **Completion Date**: The date and time the run was completed.
  - **Transferred Date**: The date and time the run results were transferred to the network.
- **Created By**: The name of the user who created the run.
- **Status**: The current status of the run. Can be one of the following: **Running**, **Complete**, **Failed**, **Terminated**, or **Unknown**.
- **Instrument Details**
  - **Instrument Name**: The name of the instrument.
  - **Instrument SN**: The serial number of the instrument.
  - **Instrument SW**: The versions of Sequel Instrument Control Software (ICS) installed on the instrument.
- **Cells**
  - **Total**: The total number of SMRT Cells used in the run.
  - **Completed**: The number of SMRT Cells that generated data for the run.

- **Failed**: The number of SMRT Cells that failed to generate data during the run.
- **Run ID**: An internally-generated ID number identifying the run.
- **Primary Analysis SW**: The version of Primary Analysis software installed on the instrument.
- **UUID**: Another internally-generated ID number identifying the run.

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

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

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

Run settings and metrics

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

- **Well**: The ID of an individual well used for this sample.
- **Sample Information**
  - **Name**: The sample name, as defined when creating the run. Clicking the name will take you to the corresponding entry in the **Data Management** module.
  - **Comment**: Sample comment, entered in Run Design.
- **Run Settings**
  - **Movie Time (hrs)**: The length of the movie associated with this SMRT Cell.
  - **Loading Concentration (pM):** The on-plate loading concentration, in picomolarity.
  - **Pre-extension Time (hrs)**: The pre-extension time used in the collection, if any.
  - **Workflow:** The instrument robotics workflow used for the run.
  - **Loading Time**: The time the system took for loading to progress before proceeding to sequencing.
- **Status**: The current collection status for the SMRT Cell. This can be one of the following: **Complete**, **Collecting**, **Aborted**, **Failed**, **In Progress**, or **Pending**.
- **Total Bases (Gb)**: Calculated by multiplying the number of **productive** (P1) ZMWs by the mean polymerase read length; displayed in Gigabases.
- **Unique Molecular Yield (Gb)**: The sum total length of unique single molecules that were sequenced. It is calculated as the sum of per-ZMW median subread lengths.
- **Productivity (%)**
  - **P0**: Empty ZMW; no signal detected.
  - **P1**: ZMW with a high quality read detected.
  - **P2**: Other, signal detected but no high quality read.
- **Reads**: Polymerase reads are trimmed to the high-quality region and include bases from adapters, as well as potentially multiple passes around a SMRTbell template.
  - **HiFi Reads ≥Q20 Reads**: The total number of CCS reads whose quality value is equal to or greater than 20.
  - **HiFi Reads Yield:** The total yield (in base pairs) of the CCS reads whose quality value is equal to or greater than 20.
  - **HiFi Reads Mean Length:** The mean read length of the CCS reads whose quality value is equal to or greater than 20.
  - **HiFi Reads Median QV:** The median number of CCS reads whose quality value is equal to or greater than 20.
  - **Polymerase Read Length Mean:** The mean high-quality read length of all polymerase reads. The value includes bases from adapters as well as multiple passes around a circular template.

- **Polymerase Read Length N50**: 50% of all read bases came from polymerase reads longer than this value.
- **Longest Subread Mean**: The mean subread length, considering only the longest subread from each ZMW.
- **Longest Subread N50:** 50% of all read bases came from subreads longer than this value when considering only the longest subread from each ZMW.
- **Control**
  - **Poly RL Mean (bp)**: The mean polymerase read length of the control reads.
  - **Total Reads**: The number of control reads obtained.
  - **Concordance Mean**: The average concordance (agreement) between the control raw reads and the control reference sequence.
  - **Concordance Mode**: The median concordance (agreement) between the control raw reads and the control reference sequence.
- **Local Base Rate**: The average base incorporation rate, excluding polymerase pausing events.
- **Template**
  - **Missing Adapter (%):** The percent of pre-filter ZMWs that are missing adapters.
  - **Adapter Dimer:** The percent of pre-filter ZMWs which have observed inserts of 0-10 bp. These are likely adapter dimers.
  - **Short Insert**: The percent of pre-filter ZMWs which have observed inserts of 11-100 bp. These are likely short fragment contamination.

8. 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.
- **Control Polymerase RL**: Displays the polymerase read length distribution of the control, if used.
- **Control Concordance**: Maps control reads against the known control reference and reports the concordance.
- **Base Yield Density**: Displays the number of bases sequenced in the collection, according to the length of the read in which they were observed. Values displayed are per unit of read length (i.e. the base yield density) and are averaged over 2000 bp windows to gently smooth the data. Regions of the graph corresponding to bases found in reads longer than the N50 and N95 values are shaded in medium and dark blue, respectively.
- **Read Length Density:** Displays a density plot of reads, hexagonally binned according to their high-quality read length and median subread length. For very large insert libraries, most reads consist of a single subread and will fall along the diagonal. For shorter inserts, subreads

will be shorter than the HQ read length, and will appear as horizontal features. This plot is useful for quickly visualizing aspects of library quality, including insert size distributions, reads terminating at adapters, and missing adapters.

- **HiFi Read Length Distribution**: Displays a histogram distribution of HiFi reads (QV ≥20), other CCS reads (three or more passes, but QV <20), and other reads, by read length.

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

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

- **5mC Detections**: If 5mC calling in CpG motifs was performed, this plot displays a reverse cumulative distribution of all detected CpG motifs according to their predicted probability of methylation.

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 a job. (See "Starting a job from a Data Set" on page 37 for details.)

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

For more information on Data Sets, click here.

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

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

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

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

Creating a Data Set

1. Access SMRT Link using the Chrome web browser.
2. Select **Data Management**.
3. Data Sets can be sorted and searched for:
   - To sort Data Sets, click a **column title**.
   - To search for a Data Set, use the Search function. See "Appendix B - Data search" on page 147 for details.
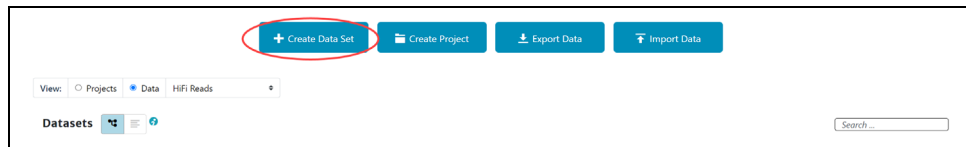
4. Click **+ Create Data Set**.
5. Enter a name for the new Data Set.
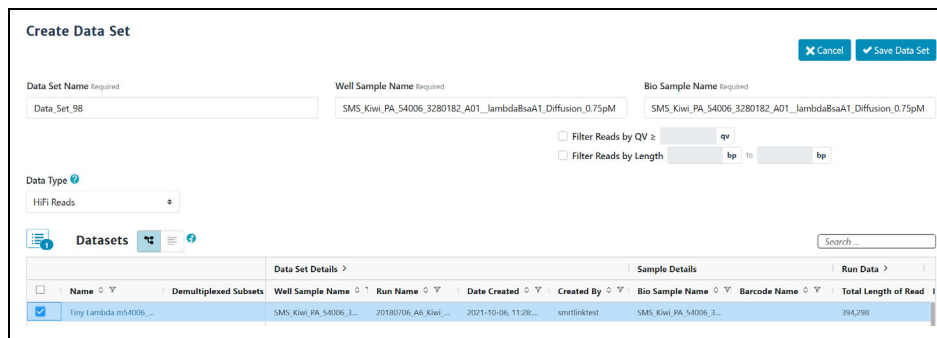
6. Select the type of data to include in the new Data Set:
   - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
   - **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.

   The Data Sets table displays the appropriate Data Sets available.
7. (**Optional**) Specify the **Project** that this new Data Set will be associated with using the **Projects** menu (located at the top-right of

the Data Management page.) **General Project:** This Data Set will be visible to **all** SMRT Link users. **All My Projects**: This Data Set will be visible **only** to users who have access to Projects that you are a member of.

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

8. In the **Data Sets** table, select one or more sets of sequence data.
9. (**Optional**) Choose how to **view** the Data Set table: 1) Tree Mode - A barcoded Data Set displays as **one row**. 2) Flat Mode - A barcoded Data Set and its demultiplexed subsets display as **separate rows**.
10. (**Optional**) Use the Search function to search for specific Data Sets. See "Appendix B - Data search" on page 147 for details.
11. (**Optional**) If you selected **one** Data Set **only**, click the **Filter Reads by QV≥** box above the Data Set list. Enter the minimum quality value to retain in the new Data Set.
12. (**Optional**) If you selected **one** Data Set **only**, click the **Filter Reads by Length** box above the Data Set list. Enter the minimum and/or maximum length to retain in the new Data Set.
13. Click **Save Data Set**. The new Data Set becomes available for starting analyses, viewing, or generating reports.
14. After the Data Set is created, click its name in the main Data Management screen to see reports, metrics, and charts describing the data included in the Data Set. See "Data Set QC reports" on page 37 for details.

Viewing Data Set information

1. On the home page, select **Data Management**.
2. Click **View > Data** and select the type of Data Set to view:
    - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
    - **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.

    The Data Sets table displays the appropriate Data Sets available.
3. (**Optional**) Use the Search function to search for Data Sets. See "Appendix B - Data search" on page 147 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:
    - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.

- **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.

  The Data Sets table displays the appropriate Data Sets available.
3. (**Optional**) Use the Search function to search for Data Sets. See for details.
4. Click the name of the Data Set to copy. The Data Set Reports page displays.
5. Click **Copy**. The main Data Management page displays; the new Data Set has (**copy**) appended to the name.

Deleting a Data Set

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

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

1. On the home page, select **Data Management**.
2. Click **View > Data** and select the type of data to delete:
   - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
   - **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.

   The Data Sets table displays the appropriate Data Sets available.
3. (**Optional**) Use the Search function to search for Data Sets. See for details.
4. Click the name of the Data Set to delete.
5. Click **Delete**. Note that this deletes the Data Set from the SMRT Link interface **only**; **not** from your server. To delete the Data Set from your server, **manually** delete it from the disk.
6. Click **Yes**. The Data Set is no longer available from SMRT Link.

Starting a job from a Data Set

From the Data Set reports page, a job can be started using the Data Set.

1. Click **Analyze...**, then name the job.
2. Select **Analysis** or **Utility** as the workflow type, then click **Next**.
3. Follow the instructions starting at Step 12 of .

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 reads and their total length in base pairs.
- The names of the run and instrument that generated the data.
- The biological sample name and well sample names of the sample used to generate the data.
- Path to the location on your cluster where the data is stored, which can be used for command-line navigation. For information on command-line usage, see **SMRT® Tools reference guide (v11.1)**.

Completed Analyses

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

Raw Data Report > Summary Metrics

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

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

What is a Project?

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

Projects include:

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

Data Sets and Projects

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

Creating a Project

1. Access SMRT Link using the Chrome web browser.
2. Select **Data Management**.

3. Click **+ Create Project**.
4. Enter a name for the new project.
5. (**Optional**) Enter a description for the project.
6. Click **Select Data Sets** and select one or more sets of sequence data to associate with the project.
   - (**Optional**) Use the Search function to search for Data Sets. See for details.
7. (**Optional**) Share the Project with other SMRT Link users. (**Note**: Unless a Project is shared, it is **only** visible to the owner.) There are two ways to specify who can access the new Project, using the controls in the **Members** section:
   - **Access for all SMRT Link Users: None** - No one can access the project other than the user who created it; **View** - Everyone can view the Project; **View/Edit**: Everyone can see and edit the Project.
   - **Access for Individual SMRT Link Users**: Enter a user name and click **Search By Name**. Choose **Owner**, **View**, or **View/Edit**, then click **Add Selected User**.
   - **Notes:** A) Projects can have **multiple** owners. B) If you enable **all** SMRT Link users to have **View/Edit** access, you cannot change an individual member's access to **View**.
8. Click **Save**. The new project becomes available for SMRT Link users who now have access.

Editing a Project

1. On the home page, select **Data Management**.
2. Click **View > Projects**.
3. Projects can be sorted and searched for:
   - To sort Projects: Click a **column title**.
   - To search for a Project, use the Search function. See for details.
4. Click the name of the project to edit.
   - (**Optional**) Edit the Project name or description.
   - (**Optional**) Delete a Data Set associated with the Project: Click **X**.
   - (**Optional**) Add one or more sets of sequence data to the Project: Click **Select Data Sets** and select one or more Data Sets to add.
   - (**Optional**) Delete members: Click **X** next to a Project member's name to delete that user from access to the Project.
   - (**Optional**) Add members to the Project: See Step 7 in **Creating a Project**.
5. Click **Save**. The modified Project is saved.

Deleting a Project

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

Viewing/deleting sequence, reference and barcode data

1. On the home page, select **Data Management**.
2. Click **View > Data**, then choose the type of data to view or delete:
   - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
   - **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.
   - **Barcodes**: Barcodes from barcoded samples.
   - **References**: Reference sequence FASTA files used when creating certain analyses.
3. (**Optional**) Use the Search function to search for specific Data Sets, barcode files or reference sequence files. See "Appendix B - Data search" on page 147 for details.
4. Click the name of the sequence, reference or barcode file of interest. Details for that sequence, reference sequence file or barcode file display.
5. (**Optional**) To delete the sequence data, reference sequence, or barcode file, click **Delete**.

Note: The **Copy** button is available for Subreads and HiFi reads, but **not** for Reference and Barcode data.
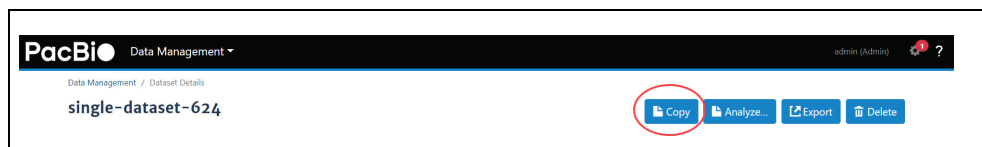
Importing sequence, reference and barcode data

Note: If your Sequel II system or Sequel IIe 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**.)
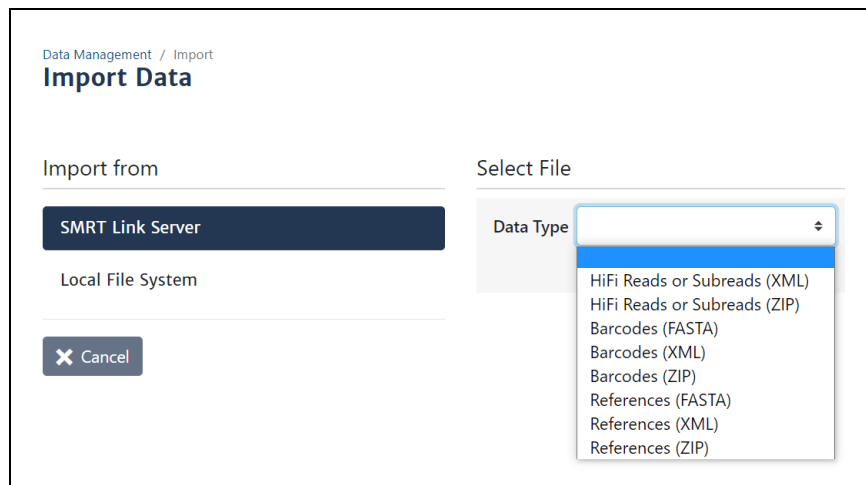
4. Select the data type to import:
   - **HiFi reads**: XML file (`.consensusreadset.xml`) or ZIP file containing information about HiFi reads (reads generated with CCS analysis whose quality value is equal to or greater than 20.) Use **only** ZIP files created by SMRT Link.
   - **Subreads**: XML file (`.subreadset.xml`) or ZIP file containing information about subreads from Sequel II systems, such as paths to the BAM files. Use **only** ZIP files created by SMRT Link.
   - **Barcodes**: FASTA (`.fa` or `.fasta`), XML (`.barcodeset.xml`), or ZIP files containing barcodes.
   - **References**: FASTA (`.fa` or `.fasta`), XML (`.referenceSet.xml`), or ZIP files containing a reference sequence for use in starting analyses. (**Note**: If importing from a **local system**, Reference files must be smaller than 15 MB.)
   - **Note**: FASTA files imported into SMRT Link must **not** contain empty lines or non-alphanumeric characters. The file name must **not** start with a number. For information about the file types listed here, click here.
5. Navigate to the appropriate file and click **Import**. The sequence data, reference, or barcodes are imported and becomes available in SMRT Link.

Exporting sequence, reference and barcode data

Two types of sequence data (HiFi reads and Subreads) can be exported, as well as barcode files and reference files.

1. On the home page, select **Data Management**.
2. Click **Export Data**.
3. Select the type of data to export:
   - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
   - **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.

- **Barcodes**: Files containing barcodes.
- **References**: Files containing a reference sequence for use in starting analyses.

4. (**Optional**) Use the Search function to search for Data Sets, barcode files, or reference files. See 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**.
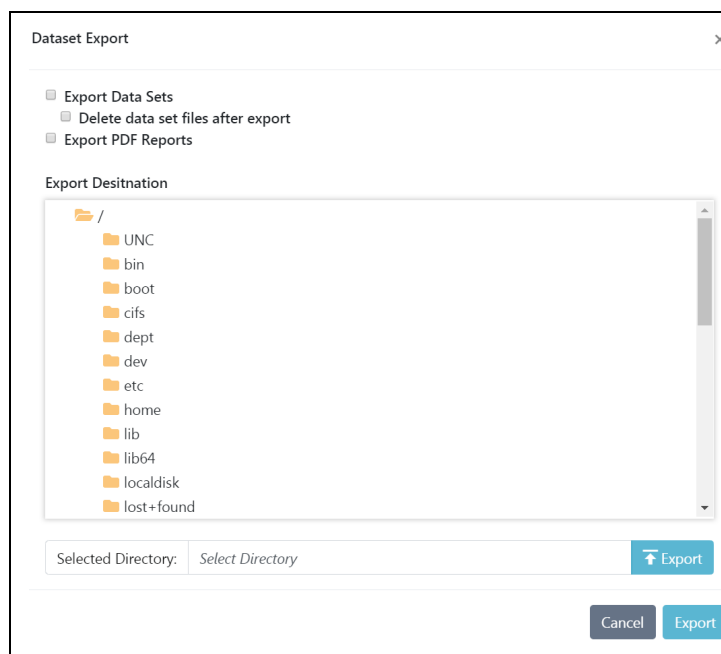
7. Navigate to the export destination directory.
8. (**Optional**) If exporting Data Sets, click **Delete data set files after export** to delete the Data Set(s) you selected from the SMRT Link installation. (Exporting, then deleting, Data Sets is useful for archiving Data Sets you no longer need.)
9. (**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 a job

1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
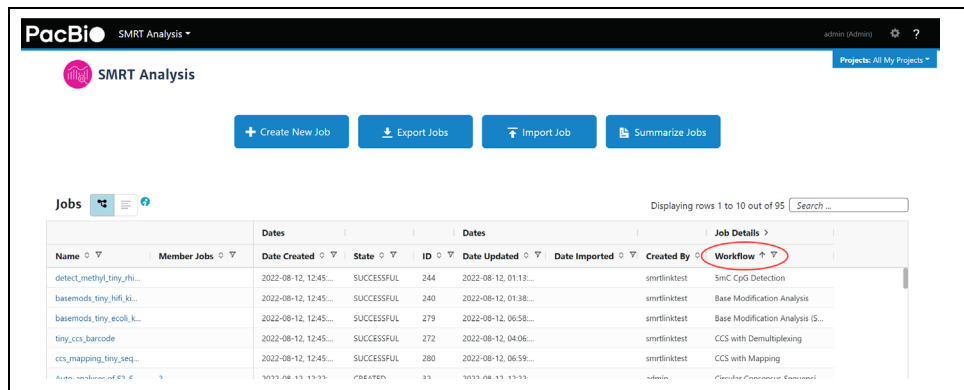
3. Jobs can be sorted, searched for, and filtered:
   – To **sort** jobs, click a **column title**.
   – To **search** for a job, use the Search function. See "Appendix B - Data search" on page 147 for details.)
   – To **filter** the list of jobs based on their **state**: Click the funnel in the **State** column header, then click one or more of the categories of interest: **Select All**, **Created**, **Running**, **Submitted**, **Terminated**, **Successful**, **Failed**, or **Aborted**.
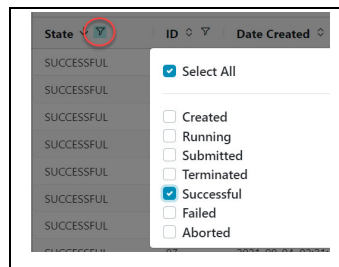
- To filter the list of jobs based on the **Project(s)** that they are associated with: Click the **Projects** menu (located at the top-right of the main SMRT Analysis page) and select a Project. See "What is a Project?" on page 39 for details.

4. Click **+ Create New Job**.
5. (**Optional**) Click **Copy From**..., choose a job whose settings you wish to reuse, then click **Select**. The job name and the Data Type are filled in. Go to Step 10 to select Data Set(s).
6. Enter a **name** for the job.
7. Specify the type of job to create:
   - **Analysis** - Uses applications designed to produce biologically-meaningful results. These applications **only** accept HiFi reads.
   - **Auto Analysis** - For information on the Auto Analysis feature, see "Automated analysis" on page 126 for details.
   - **Data Utility** - Data processing utilities used as **intermediate steps** to producing biologically-meaningful results.
8. If you selected **Data Utility**, select the type of data to use for the job:
   - **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
   - **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.
9. (**Optional**) Specify the **Project** that this job will be associated with using the **Projects** menu (located at the top-right of the SMRT Analysis page.) **General Project**: This job will be visible to **all** SMRT Link users. **All My Projects**: This job will be visible **only** to users who have access to Projects that you are a member of. To **restrict access** to a job, make sure to select a Project limited to the appropriate users **before** starting the job. **Note:** Selecting a Project **also** filters the Data Sets that you can use when **creating** the job.
10. In the **Data Sets** table, select one or more sets of data to be analyzed.
    - (**Optional**) Use the Search function to search for Data Sets. See "Appendix B - Data search" on page 147 for details.)
    - (**Optional**) Choose how to **view** the Data Set table: 1) Tree Mode - A barcoded Data Set displays as **one row**. 2) Flat Mode - A barcoded Data Set and its demultiplexed subsets display as **separate rows**.
    - (**Optional**) For Data Sets that include demultiplexed subsets, you can also select individual subsets as part of your selection. To do so:

      A) Click the Demultiplexed Subsets number link:
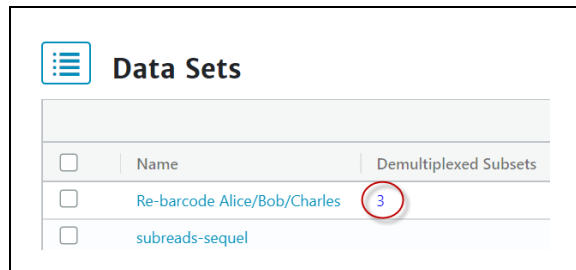
B) Select one or more subsets, then click **Back**:
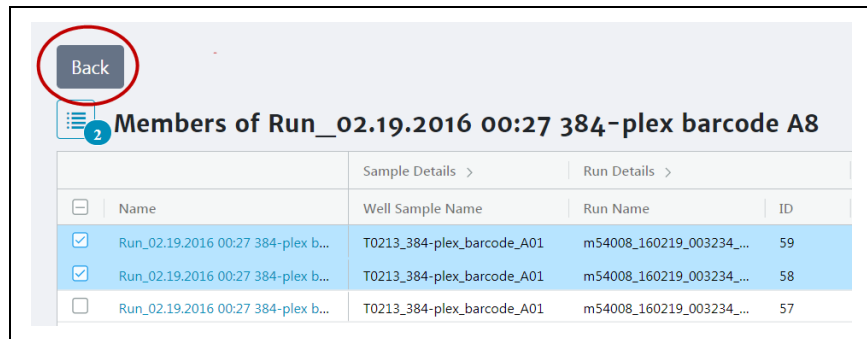
C) Click the list image to view or edit the full Data Set selection. (The small blue number specifies how many Data Sets and/or subsets were selected):
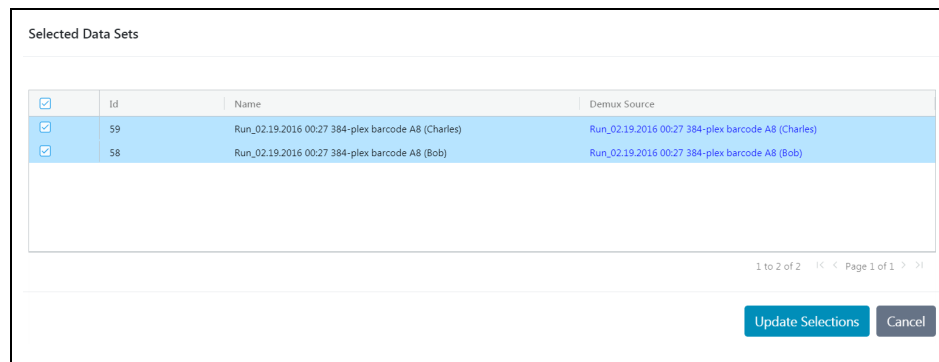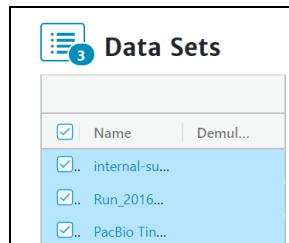
Note: For information on the Auto Analysis feature, see "Automated analysis" on page 126 for details.

11. If you selected **multiple** Data Sets as input for the job, additional options become available:

- **One Analysis for All Data Sets:** Runs one job using all the selected Data Sets as input, for a maximum of 30 Data Sets.
- **One Analysis per Data Set - Identical Parameters:** Runs one separate job for **each** of the selected Data Sets, using the **same** parameters, for a maximum of 10,000 Data Sets. Later in the process, optionally click **Advanced Parameters** and modify parameters.
- **One Analysis per Data Set - Custom Parameters:** Runs one separate job for **each** of the selected Data Sets, using **different** parameters for each Data Set, for a maximum of 16 Data Sets. Later in the process, click **Advanced Parameters** and modify parameters. Then click **Start and Create Next**. You can then specify parameters for **each** of the included Data Sets.
- **Note**: The number of Data Sets listed is based on testing using PacBio's suggested compute configuration, listed in **SMRT Link software installation guide (v11.1)**.

12. Click **Next**.
13. Select a secondary analysis application or data utility from the drop-down list. (Different choices display based on your initial choice of **Analysis** or **Data Utility** in Step 7. See"PacBio® secondary analysis applications" on page 54 or "PacBio® data utilities" on page 99 for details.)

- Each of the secondary analysis applications/data utilities has **required parameters** that are displayed. Review the default values shown.

Page 47

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

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

    The **Iso-Seq Analysis** application will be used as an example. This application characterizes full-length transcript isoforms.

15. 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 for details.)

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

    – To see information about parameters for **all** secondary analysis applications and data utilities provided by PacBio, see and .

Advanced Analysis Parameters

Min. CCS Predicted Accuracy (Phred Scale) ⓘ	Require and Trim Poly(A) Tail ⓘ	Minimum Mapped Length (bp) ⓘ
20	● ON ○ OFF	50

Minimum Gap-Compressed Identity (%) ⓘ	Minimum Mapped Coverage (%) ⓘ	Maximum Fuzzy Junction Difference (bp) ⓘ
95	99	5

Filters to Add to the Data Set ⓘ	Advanced pbmm2 Options ⓘ	Compute Settings ⓘ
		-- select -- ⇕

Ok Cancel

17. (**Optional**) Click **Export** to create a CSV file containing **all** the settings you specified for the application/data utility. You can then import this file when creating future jobs using the same application/data utility. You can also use this exported file as a template for use with later jobs.
18. (**Optional**) Click **Back** if you need to change any of the analysis attributes selected in Step 7.
19. Click **Start** to submit the job. (If you selected multiple Data Sets as input, click **Start Multiple Jobs** or **Start and Create Next**.)
20. Select **SMRT Analysis** from the Module Menu to navigate to the main SMRT Analysis screen. There, the status of the job displays. When the job has **completed**, click on its name - reports are available for the completed job.
21. (**Optional**) To **delete** the completed job: Click **Delete**, then click **Yes** in the confirmation dialog. The job is deleted from **both** the SMRT Link interface and from the server.
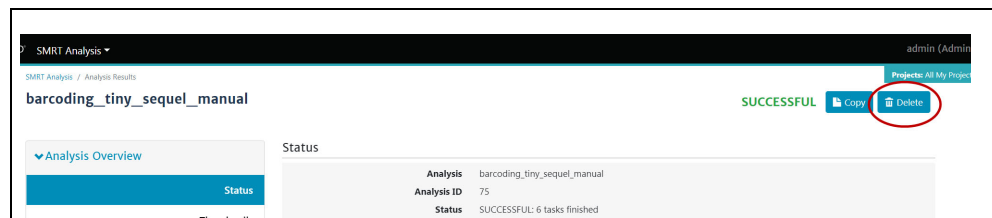
Starting a job after viewing sequence data

A job 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:
   – **HiFi reads:** Reads generated with CCS analysis whose quality value is equal to or greater than 20.
   – **Subreads**: Reads containing the sequence from one or more single passes of a polymerase on a single strand of an insert within a SMRTbell template.

   The Data Sets table displays the appropriate Data Sets available.
3. (**Optional**) Use the Search function to search for Data Sets. See "Appendix B - Data search" on page 147 for details.
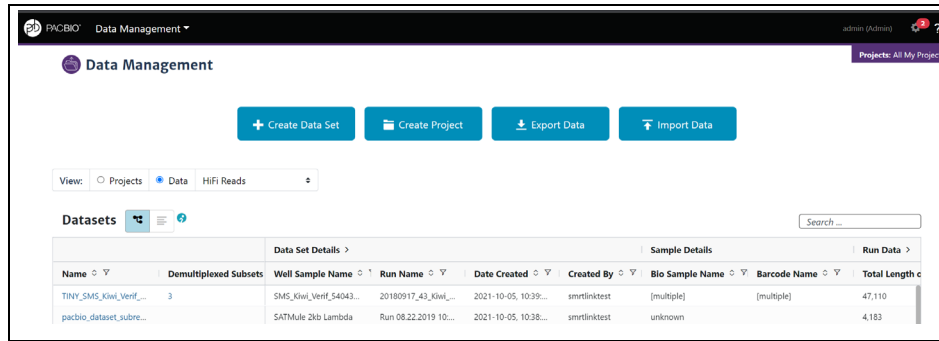
4. In the **Name** column, click the name of the sequence data of interest. Details for the selected sequence data display.
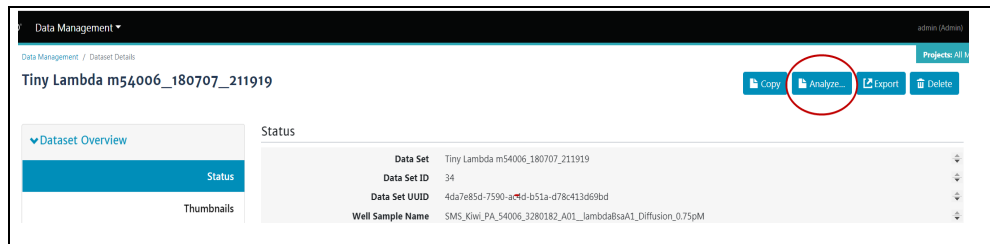
5. To **start** a job using this sequence data, click **Analyze...**, then name the job.
6. Select **Analysis** or **Utility** as the workflow type, then click **Next**.
7. Follow the instructions starting at Step 8 of "Creating and starting a job" on page 44.

Canceling a running job

1. On the home page, select **SMRT Analysis**.
2. Click the funnel in the **State** column header, then click **Running**. This displays **only** currently-running jobs.
3. Select a currently-running job to cancel.
4. Click **Cancel**.
5. Click **Yes** in the confirmation dialog. The canceled job displays as **Terminated**.

Restarting a failed job

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

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

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

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

1. On the home page, select **SMRT Analysis**.
2. Click the funnel in the **State** column header, then click **Failed**. This displays **only** failed jobs.
3. Select a failed job to restart.
4. Click **Restart**.

Viewing job results

1. On the home page, select **SMRT Analysis**. You see a list of **all** jobs.
2. (**Optional**) Click the funnel in the **State** column header, then click **Successful**. This displays **only** successfully-completed jobs.
3. (**Optional**) Use the Search function to search for specific jobs. See "Appendix B - Data search" on page 147 for details.
4. Click the job link of interest.
5. Click **Analysis Overview > Status** to see job information status, including which application/data utility was used for the job, and the inputs used.
6. Click Analysis **Overview > Thumbnails** or **Display All** to view thumbnails of the reports generated for the job. Click the link under a thumbnail to see a larger image.
7. Depending on the application/data utility used for the job, different job-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/data utilities, see "PacBio® secondary analysis applications" on page 54 or "PacBio® data utilities" on page 99.
8. To download data files created by SMRT Link: You can use these data files as input for further processing, pass on to collaborators, or upload to public genome sites. Click **Data > File Downloads**, then click

the appropriate file. The file is downloaded according to your browser settings.

9. (**Optional**) Specify prefixe(s) used in the names of files generated by the job. Example: **Run Name** can be included in the name of every file generated by the job. Click **Edit Output File Name Prefix**, check the type(s) of information to append to the file names, then click **Save**.

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

11. To visualize the secondary analysis results: See "Visualizing data using IGV" on page 129 for details.

Copying and running an existing job

If you run very similar jobs, you can **copy** an existing job, 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** jobs.

2. (**Optional**) Click the funnel in the **State** column header, then click **Successful**. This displays **only** successfully-completed jobs.

3. (**Optional**) Use the Search function to search for specific jobs. See "Appendix B - Data search" on page 147 for details.

4. Click the job link of interest.

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

6. Edit the name of the job.

7. Click **Next**.

8. (**Optional**) Edit any other parameters. See "PacBio® secondary analysis applications" on page 54 or "PacBio® data utilities" on page 99 for further details.

9. Click **Start**.

Exporting a job

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

1. On the home page, select **SMRT Analysis**.

2. Click **Export Job**.

3. (**Optional**) Use the Search function to search for specific analyses. See "Appendix B - Data search" on page 147 for details.

4. Select one or more jobs to export. This exports the entire contents of the job directory.

5. Click **Export Selected**.

6. Select the output directory for the job data and click **Export**.

Importing a job

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

1. On the home page, select **SMRT Analysis**.

2. Click **Import Job**.
3. Select a ZIP file containing the job to import.
4. Click **Import**. The job is imported and is available on the main SMRT Analysis page.

Summarizing Microbial Genome Analysis jobs

You can summarize information about **multiple** successfully-completed Microbial Genome Analysis jobs. This generates the following:

- A CSV report with the assembly statistics for each job on a separate row.
- A CSV report with the mapping statistics for each job on a separate row.
- A CSV report with the coverage statistics for each contig for each job.
- A text file containing a list of all the modified base motifs for each job.
- A ZIP file containing the assembly in FASTA format from each job. (Available in the **Data > File Downloads** section of the results page.)

1. On the home page, select **SMRT Analysis**.
2. Click **Summarize Jobs**. This displays all the successfully-completed Microbial Genome Analysis jobs.
3. (**Optional**) Use the Search function to search for specific analyses. See for details.
4. Select one or more Microbial Genome Analysis jobs to summarize, up to 384 jobs.
5. Click **Summarize Selected**. This creates a new job named **Job Summary.**
6. When the **Job Summary** job is successfully completed, click it to see summary information about the Microbial Genome Analysis jobs you selected in Step 4.

PacBio® secondary analysis applications

Following are the secondary analysis applications provided with SMRT Analysis v11.1. These applications are designed to produce biologically-meaningful results. Each application is described later in this document, including all analysis parameters, reports and output files generated by the application.

Note: These applications accept **only** HiFi reads as input.

Genome Assembly

- Generate *de novo* assemblies of genomes, using HiFi reads.
- See "Genome Assembly" on page 56 for details.

HiFi Mapping (was Mapping)

- Align (or map) reads to a user-provided reference sequence.
- See "HiFi Mapping" on page 59 for details.

HiFiViral SARS-CoV-2 Analysis

- Analyze multiplexed viral surveillance samples for SARS-CoV-2, using HiFi reads.
- See "HiFiViral SARS-CoV-2 Analysis" on page 63 for details.

Iso-Seq® Analysis

- Characterize full-length transcript isoforms, using HiFi reads.
- See "Iso-Seq® Analysis" on page 68 for details.

Microbial Genome Analysis

- **Note**: This combines and replaces the **Microbial Assembly** and **Base Modification Analysis** applications in the previous release.
- Generate *de novo* assemblies of small prokaryotic genomes between 1.9-10 Mb and companion plasmids between 2 – 220 kb, and identify methylated bases and associated nucleotide motifs.
- Optionally include identification of 6mA and 4mC modified bases and associated DNA sequence motifs.
- See "Microbial Genome Analysis" on page 75 for details.

Read Segmentation and Single-Cell Iso-Seq® Analysis

- Characterize full-length transcript isoforms with additional single-cell information, including single-cell barcodes and unique molecular identifiers (UMIs).
- The application is for use when using concatenation-based library preparations such as the MAS-Seq libraries.
- See "Read Segmentation and Single-Cell Iso-Seq® Analysis" on page 81 for details.

Single-Cell Iso-Seq® Analysis

- Characterize full-length transcript isoforms with additional single-cell information, including single-cell barcodes and unique molecular identifiers (UMIs).
- See for details.

Structural Variant Calling

- Identify structural variants (Default: ≥20 bp) in a sample or set of samples relative to a reference.
- See for details.

Genome Assembly

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

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

The application includes seven main steps:

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

Importing/exporting analysis settings

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

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

Parameters

Advanced parameters	Default value	Description
Genome Length	0	The approximate number of base pairs expected in the genome. This is used only for downsampling; if the value is ≤ 0, downsampling is disabled. Enter an integer, optionally followed by one of the metric suffixes: `k`, `M` or `G`. Example: `4500k` means "4,500 kilobases" or "4,500,000". `M` stands for Mega and `G` stands for Giga.
Downsampled coverage	0	The input Data Set can be downsampled to a desired coverage, provided that both the **Downsampled coverage** and **Genome Length** parameters are specified and >0.
		Downsampling applies to the **entire** assembly process, including polishing.
		This parameter selects reads randomly, using a fixed random seed for reproducibility.

Advanced parameters	Default value	Description
Run polishing	ON	Enables or disables the polishing stage of the workflow. Polishing can be disabled to perform fast draft assemblies.
Run phasing	ON	Enables or disables the phasing stage of the workflow. Phasing can be disabled to assemble haploid genomes, or to perform fast draft assemblies.
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** comma-separated) list of other filters to add to the Data Set.
Advanced Assembly Options	NONE	A semicolon-separated list of KEY=VALUE pairs. New line characters are not accepted.
Purge duplicate contigs from the assembly	ON	Enables or disables identification of "duplicate" alternate haplotype contigs which may be assembled in the primary contig file, and moves them to the associate contig (haplotig) file.
Cleanup intermediate files	ON	Removes intermediate files from the run directory to save space.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is `20` (QV 20), or 99% predicted accuracy.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Genome Assembly application generates the following reports:

Polished Assembly > Summary Metrics

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

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

Polished Assembly > Polished Contigs

- **Contig:** The name of the individual contig.
- **Length (bases):** The length of the contig, in bases.
- **Circular**: **Yes** if the contig is circular, **No** if it isn't.
- **Percent Polished**: The percent of contig bases that were polished.

- **Number of Polishing Reads**: The number of reads used to perform polishing on this contig.

Data > File Downloads

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

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

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

- Accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.
- Maps data to a provided reference sequence, and then identifies consensus and variants against this reference.
- Haploid variants and small indels, but **not** diploid variants, are called as a result to alignment to the reference sequence.

Importing/exporting analysis settings

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

Reference Set (Required)

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

Parameters

Advanced parameters	Default value	Description
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** 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.
Bio Sample Name of Aligned Dataset	NONE	Populates the Bio Sample Name (Read Group `SM` tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. **Note**: Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.
Minimum Gap-Compressed Identity (%)	70	The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is `20` (QV 20), or 99% predicted accuracy.
Advanced pbmm2 Options	NONE	Space-separated list of custom `pbmm2` options. Not all supported command-line options can be used, and HPC settings **cannot** be modified. See **SMRT® Tools reference guide v11.1** for details.
Target Regions (BED file)	NONE	(**Optional**) Specifies a BED file that defines regions for a **Target Regions** report showing coverage over those regions. See "Appendix C - BED file format for Target Regions report" on page 149 for details.
Compute Settings	Select	(**Optional**) Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The HiFi Mapping application generates the following reports:

Target Regions > Target Regions

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

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

Target Regions > Target Region Coverage

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

Mapping Report > Summary Metrics

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

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
- **Number of Alignments:** The number of alignments that mapped to the reference sequence.
- **Number of CCS reads (total):** The total number of CCS reads in the sequence.
- **Number of CCS reads (mapped)**: The number of CCS reads that mapped to the reference sequence.
- **Number of CCS reads (unmapped)**: The number of CCS reads not mapped to the reference sequence.
- **Percentage of CCS reads (mapped):** The percentage of CCS reads that mapped to the reference sequence.
- **Percentage of CCS reads (unmapped):** The percentage of CCS reads not mapped to the reference sequence.
- **Number of CCS Bases (mapped)**: The number of CCS bases that mapped to the reference sequence.
- **CCS Read Length Mean (mapped):** The mean read length of CCS reads that mapped to the reference sequence, starting from the first mapped base of the first mapped CCS read, and ending at the last mapped base of the last mapped CCS read.
- **CCS Read N50 (mapped)**: The read length at which 50% of the mapped bases are in CCS reads longer than, or equal to, this value.
- **CCS Read Length 95% (mapped)**: The 95th percentile of read 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 mapping statistics per movie.

- **Sample**: The sample name for which the following metrics apply.
- **Movie**: The movie name for which the following metrics apply.
- **Number of CCS Reads (mapped)**: The number of CCS reads that mapped to the reference sequence. This includes adapters.
- **CCS Read Length Mean (mapped)**: The mean read length of CCS reads that mapped to the reference sequence, starting from the first mapped base of the first mapped CCS read, and ending at the last mapped base of the last mapped CCS read.
- **CCS Read Length N50 (mapped)**: The read length at which 50% of the mapped bases are in CCS reads longer than, or equal to, this value.
- **Number of CCS Bases (mapped)**: The number of CCS bases that mapped to the reference sequence.
- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Mapping Report > Mapped CCS Read Length

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

Mapping Report > Mapped CCS Reads Concordance

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

Mapping Report > Mapped Concordance vs Read Length

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

Coverage > Summary Metrics

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

Coverage > Coverage Across Reference

- Maps coverage across the reference.

Coverage > Depth of Coverage

- Maps the reference regions against the percent coverage.

Coverage > Coverage vs. [GC] Content

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

Data > File Downloads

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

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

Data > IGV Visualization Files

The following files are used for visualization using IGV; see for details.

- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.

HiFiViral SARS-CoV-2 Analysis

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

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

Across **all** samples, this analysis provides:

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

Notes:

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

Auto Analysis in Run Design

Users may set the analysis to begin **automatically** after sequencing completes using Auto Analysis in Run Design. See for details.

HiFiViral SARS-CoV-2 application workflow

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

Preparing input data for the HiFiViral SARS-CoV-2 Analysis application

1. Run the **Demultiplex Barcodes** data utility, where the inputs are HiFi reads, and the primers are multiplexed barcode primers. (If HiFi reads have **not** been generated on the instrument, run CCS analysis first. See "Circular Consensus Sequencing (CCS)" on page 115 for details.)
   – The proper barcode sequences are provided by default:
     `Barcoded M13 Primer Plate`.
   – For the **Same Barcodes on Both Ends of Sequence** parameter, specify **No**; the barcode pairs are **asymmetric**.
   – Provide the correctly-formatted barcode pair-to-Bio Sample CSV file for the **Assign Bio Sample Names to Barcodes** option. (For details, see "Assign Bio Sample Names to Barcodes (Required)" on page 102.)

Running the HiFiViral SARS-CoV-2 Analysis application

1. **After** running the Demultiplex Barcode data utility, create a new job using **SMRT Analysis > Create New Job**.
2. Name the job.
3. Select all the demultiplex samples contained in the Data Set and choose **Analysis of Multiple Data Sets > One Analysis for All Data Sets**. Click **Next**.
4. Select **HiFiViral SARS-CoV-2 Analysis** from the Analysis Application list.
5. **SARS-CoV-2 Genome NC_045512.2** (the Wuhan reference genome) is automatically loaded; advanced users may select a different reference if desired.
6. To generate the optional Plate QC graphical summary, click **Advanced Parameters** and load a CSV file using the provided template (**assay-PlateQC_template_4by96.csv**) as a guide.
7. Click **OK**, then **Start**.

Importing/exporting analysis settings

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

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

Reference Genome (Required)

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

Parameters

Advanced parameters	Default value	Description
Plate QC CSV	NONE	(**Optional**) Specify a CSV file to generate the Plate QC report, which displays analysis results for each sample in the assay plate. The CSV file **must** contain barcode (asymmetric pairs), Bio Sample Name, assay plate IDs (can include 1-4 plates with unique names; avoid special characters), and assay plate well IDS in the format `A01, A02,…H12`. (To create a new file, click **Download Template**, edit, and then save the CSV file.) The plate and well information corresponds to the location of samples during the SARS-CoV-2 enrichment assay.
Probes FASTA	NONE	Specify probe sequences in FASTA format if using probes **other** than the standard probes shipped in the HiFiViral SARS-CoV-2 Kit.
Minimum Base Coverage	4	Specify the minimum read depth at each position to report either a variant or a reference base. Positions with **less** than this specified coverage will have an `N` base output in the consensus sequence FASTA file. Increasing the minimum base coverage may result in more `N`s and loss of variant detection. We do **not** recommend making this value lower than the default threshold of `4`, as it may increase the number of false positive variants called.
Minimum Variant Frequency	0.5	Specify that **only** variants whose frequency is **greater** than this value are reported. This frequency is determined based on the read depth (DP) and allele read count (AD) information in the VCF output file. We recommend using the default value to properly call the dominant alternative variant while also filtering out potential artifacts.
Advanced Processing Options	NONE	Additional options to pass to the `mimux` preprocessing tool for trimming and filtering reads by probe sequences. Options should be entered in space-separated format. See the HiFiViral SARS-CoV-2 Analysis section of **SMRT Tools reference guide (v11.1)** for details.
Minimum Barcode Score	80	A barcode score measures the alignment between a barcode attached to a read and an ideal barcode sequence, and is an indicator of how well the chosen barcode pair matches. It ranges between `0` (no match) and `100` (a perfect match). This parameter specifies that reads with barcode scores below this minimum value are **not** included in analysis.
Compute Settings	Select	(**Optional**) Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

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

Summary Report > Summary Metrics

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

Summary Report > Sample Summary

- **Bio Sample Name**: The name of the biological sample associated with the variants. (**Note**: Any spaces in the name are substituted by new line characters for consistency with output file names.)
- **Substitutions**: The count of all called substitutions in the consensus sequence for the sample.
- **Insertions:** The count of all called insertions in the consensus sequence for the sample.
- **Deletions:** The count of all called deletions in the consensus sequence for the sample.
- **Reads**: The total number of HiFi reads for the sample.
- **Read Coverage**: The mean number of mapped reads overlapping with each position in the reference genome.
- **On-Target Rate**: The mapping yield of reads; the number of unique mapped reads divided by the total number of reads.
- **Multiple Strains (Probability):** Samples are flagged as having multiple strains if the probability is at least 0.95. Samples may contain multiple strains due to sample contamination or presence of multiple strains in the RNA extract. To classify a sample as multi-strain, we tolerate error by using the binomial cumulative distribution function (with a fixed probability of 0.2). This feature is supported for samples with Ct < 26 with minor frequencies >20%. Samples **must** have >70% genome coverage to be called Multiple Strains.
- **Ns**: The number of bases in the consensus sequence that are Ns.
- **Genome Coverage**: The percentage of bases with at least four mapped reads overlapping their position by default. See the **Advanced Parameters** dialog to adjust minimum base coverage.

Summary Report > Genome Coverage

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

Summary Report > Plate QC

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

- **Blue** wells represent samples with at least 95% coverage.
- **Green** wells represent samples with at least 90% coverage.
- **Yellow** wells represent samples that passed the workflow but had genome coverage worse than 90%.
- **Red** wells represent samples that failed the workflow.

- **White** wells do **not** include a sample.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **All Samples, HiFi Reads FASTQ**: HiFi reads in FASTQ format for all samples.
- **All Samples, Consensus Sequence FASTA**: The full consensus genomic sequences; bases for which **no** consensus could be called are represented by Ns. See the **Advanced Parameters** dialog to adjust the minimum base coverage for outputting Ns.
- **All Samples, Genome Coverage Plots**: Plots for individual samples showing coverage depth across the genome.
- **All Samples, Variant Call VCF**: VCF file containing the final variant calls per sample.
- **All Samples, HiFi Reads Mapped BAM**: BAM file for each sample containing the HiFi reads aligned to the reference genome.
- **All Samples, Consensus Sequence Aligned BAM**: BAM file for each sample of consensus sequence aligned to the reference genome. The consensus sequence is split into fragments where there are Ns and each fragment is mapped.
- **All Samples, Raw Variant Calls VCF**: VCF file containing the intermediate variant calls per patient sample.
- **Sample Summary Table CSV**: CSV version of the data shown in the Sample Summary table.
- **All Samples, Probe Counts TSV**: Tab-delimited text file containing per-sample, per-probe counts. This file can be used to identify samples that are poorly sequenced or probes with high or low coverages.
- **Sample Inputs CSV**: CSV version of the Plate QC CSV, if supplied in the **Advanced Parameters** dialog.
- **Failed Sample Info**: CSV file containing information on samples for which the analysis was unable to generate a per-sample report due to an absence of usable data.
- **Failed Sample Info Analysis Logs**: Zipped log files describing failed samples.

Iso-Seq® Analysis

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

- The application accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis that have a quality value equal to or greater than Q20.

Notes on Multiplexed Data

There are two ways in which an Iso-Seq library can be multiplexed:

1. Barcoded adapter Iso-Seq libraries

- If using the SMRTbell Barcoded Adapter with the Iso-Seq Express protocol **on or after** April 21, 2022, demultiplex the Data Set **prior** to running the Iso-Seq application.
- To analyze the samples in a single Iso-Seq run, select **all** the demultiplexed Data Sets to combine and begin the Iso-Seq analysis.
- If following the standard Iso-Seq Express protocol, select **Iso-Seq cDNA Primers** as the Primer Set.

2. Barcoded cDNA primer Iso-Seq libraries

- If following multiplexing guidelines using the Iso-Seq Express protocol **on or prior to** April 21, 2022 and you ordered synthesized oligos listed in the **Appendix 3 - Recommended barcoded NEBNext single cell cDNA PCR primer and Iso-Seq Express cDNA PCR primer sequences** section of the document **Procedure & checklist - Preparing Iso-Seq® libraries using SMRTbell prep kit 3.0**, demultiplex your Data Set using the Iso-Seq application. In other words, do **not** run the Demultiplexing Barcodes utility first.
- See the **Primer Set Selection** column below for the correct choice of primer sequences.

Multiplexed method	Demultiplexed before Iso-Seq?	Primer set selection
Not multiplexed	NO	Iso-Seq cDNA Primers
Barcoded adapters	YES	Iso-Seq cDNA Primers
Barcoded cDNA primer	NO	Iso-Seq 12 Barcoded cDNA Primers Or Custom cDNA Primers

The application includes three main steps:

1. **Classify**: Identify and remove primers (which may be cDNA primers or barcoded cDNA primers). Identify full-length reads based on the asymmetry of 5' and 3' primers. Trim off polyA tails and remove artifactual concatemers.
2. **Cluster** (**Optional**): Perform *de novo* clustering and consensus calling. Output full-length consensus isoforms that are further separated into high-quality (HQ) and low-quality (LQ) based on estimated accuracies.

3. **Collapse** (**Optional**): When a reference genome is selected, map HQ isoforms to the genome, then collapse redundant isoforms into unique isoforms.

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

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

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

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

Importing/exporting analysis settings

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

Primer Set (Required)

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

Example 1: The Iso-Seq cDNA Primer primer set, included with the SMRT Link installation.

Users following the standard Iso-Seq Express protocol **without** multiplexing, or running a Data Set that has **already** been demultiplexed (either using Run Design or the SMRT Analysis application) should use this default option.

```
>IsoSeq_5p
GCAATGAAGTCGCAGGGTTGGG
>IsoSeq_3p
GTACTCTGCGTTGATACCACTGCTT
```

Example 2: The Iso-Seq 12 Barcoded cDNA Primers set, included with the SMRT Link installation.

Users using barcoded cDNA primers listed in the **Appendix 3 - Recommended barcoded NEBNext single cell cDNA PCR primer and Iso-Seq Express cDNA PCR primer sequences** section of the document **Procedure & checklist - Preparing Iso-Seq® libraries using SMRTbell prep kit 3.0**, should select this option.

```
>bc1001_5p
CACATATCAGAGTGCGGCAATGAAGTCGCAGGGTTGGGG
>bc1002_5p
ACACACAGACTGTGAGGCAATGAAGTCGCAGGGTTGGGG
…
```

(There are a total of 24 sequence records, representing 12 pairs of F/R barcoded cDNA primers.)

Example 3: An example of a custom cDNA primer set. 4 tissues were multiplexed using barcodes on the 3' end only.

```
>IsoSeq_5p
GCAATGAAGTCGCAGGGTTGGG
>dT_BC1001_3p
AAGCAGTGGTATCAACGCAGAGTACCACATATCAGAGTGCG
>dT_BC1002_3p
AAGCAGTGGTATCAACGCAGAGTACACACACAGACTGTGAG
>dT_BC1003_3p
AAGCAGTGGTATCAACGCAGAGTACACACATCTCGTGAGAG
>dT_BC1004_3p
AAGCAGTGGTATCAACGCAGAGTACCACGCACACGCGCG
```

Example 4: 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 genome to align High Quality isoforms to, and to collapse isoforms mapped to the same genomic loci.

Run Clustering (Default = ON)

- Specify **ON** to generate consensus isoforms.
- Specify **OFF** to classify reads **only** and not generate consensus isoforms. The Reference Set will also be ignored.

Cluster Barcoded Samples Separately (Default = OFF)

- Specify **OFF** if barcoded samples are from the **same** species, but different tissues, or samples of the same genes but different individuals. The samples are clustered with **all** barcodes pooled.
- Specify **ON** if barcoded samples are from **different** species. The samples are clustered separately by barcode.
- In either case, the samples on the results page are automatically named `BioSample_1` through `BioSample_N`.

Parameters

Advanced parameters	Default value	Description
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 Mapped Length (bp)	50	The minimum required mapped HQ isoform sequence length (in base pairs) for the Iso-Seq mapping-collapse step. **Note**: This is applicable **only** if a reference genome is provided.
Minimum Gap-Compressed Identity (%)	95	The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference. **Note**: This is applicable **only** if a reference genome is provided.
Minimum Mapped Coverage (%)	99	The minimum required HQ transcript isoform sequence alignment coverage (in percent) for the Iso-Seq mapping-collapse step. **Note**: This is applicable **only** if a reference genome is provided.
Maximum Fuzzy Junction Difference (bp)	5	The maximum junction difference between two mapped isoforms to be collapsed into a single isoform. If the junction differences are **all** less than the provided value, they will **all** be collapsed. Setting to `0` requires all junctions to be exact to be collapsed into a single isoform. Applicable **only** if a reference genome is provided.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for Iso-Seq Analysis is `20` (QV 20), or 99% predicted accuracy.
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** comma-separated) list of other filters to add to the Data Set.
Advanced pbmm2 Options	NONE	Space-separated list of custom `pbmm2` options. (`pbmm2` is already running with `--preset ISOSEQ`.) Not all supported command-line options can be used, and HPC settings **cannot** be modified. See **SMRT® Tools reference guide v11.1** for details.
Compute Settings	Select	**(Optional)** 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 Analysis Read Classification > Summary Metrics

- **Reads:** The total number of CCS reads.
- **Reads with 5' and 3' Primers:** The number of CCS reads with 5' and 3' cDNA primers detected.
- **Non-Concatemer Reads with 5' and 3' Primers:** The number of non-concatemer CCS reads with 5' and 3' primers detected.

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

CCS Analysis Read Classification > Primer Data

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

CCS Analysis Read Classification > Primer Read Statistics

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

CCS Analysis Read Classification > Primer Quality Scores

- Histogram of primer scores.

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

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

Transcript Clustering > Summary Metrics

- **Sample Name**: The sample name for which the following metrics apply.
- **Number of High-Quality Isoforms**: The number of consensus isoforms that have an estimated accuracy **above** the specified threshold.
- **Number of Low-Quality Isoforms**: The number of consensus isoforms that have an estimated accuracy **below** the specified threshold.

Transcript Clustering > Length of Consensus Isoforms

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

Transcript Mapping > Summary Metrics

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

Transcript Mapping > Length of Mapped Isoforms

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

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Primers Summary**: Text file listing how many ZMWs were filtered, how many ZMWs are the same or different, and how many reads were filtered.
- **Inferred Primers**: Inferred primers used in the analysis. The algorithm looks at the first 35,000 ZMWs, then selects primers with ≥10 counts and mean scores ≥45.
- **Full-Length Non-Concatemer 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**: Mapped, unique isoforms, in FASTQ format. This is the Mapping step output that is recommended output file to work with. When the input Data Set is a ConsensusReadSet, **only** a FASTA file is generated.
- **Collapsed Filtered Isoforms Groups**: Report of isoforms mapped into collapsed filtered isoforms.
- **Full-length Non-Concatemer Read Assignments**: Report of full-length read association with collapsed filtered isoforms, in text format.
- **Collapsed Filtered Isoform Counts**: Report of read count information for each collapsed filtered isoform.

Data > IGV Visualization Files

The following files are used for visualization using IGV; see "Visualizing data using IGV" on page 129 for details.

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

Note: For details on custom PacBio tags added to output BAM files by the Iso-Seq Application, see page 54 of **SMRT Tools reference guide (v11.1)**, or see here for details.

Microbial Genome Analysis

Use this application to generate *de novo* assemblies of small prokaryotic genomes between 1.9-10 Mb and companion plasmids between 2 – 220 kb. This application can optionally include analysis of 6mA and 4mC modified bases and associated DNA sequence motifs. (This requires kinetic information.)

Note: This combines and replaces the **Microbial Assembly** and **Base Modification Analysis** applications in the previous release.

The Microbial Genome Analysis application:

- Accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.
- Includes chromosomal- and plasmid-level *de novo* genome assembly, circularization, polishing, and rotation of the origin of replication for each circular contig.
- Performs base modification detection to identify 4mCm and 6mA and associated DNA sequence motifs. (This requires kinetic information.)
- Facilitates assembly of larger genomes (yeast) as well.

Importing/exporting analysis settings

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

Run Base Modification Analysis (Default = ON)

- Run Base Modification analysis on the final assembly. This **only** applies if the assembly is not empty, and the input data contains the correct kinetic tags.

Find Modified Base Motifs (Default = ON)

- Perform motif detection on the results of base modification analysis.

Parameters

Advanced parameters	Default value	Description
Downsampled coverage	0	The input Data Set can be downsampled to a desired coverage, provided that the **Downsampled coverage** value is >0. Downsampling applies to the **entire** assembly process, including polishing. This parameter selects reads randomly, using a fixed random seed for reproducibility.
Advanced Assembly Options for chromosomal stage	NONE	A semicolon-separated list of KEY=VALUE pairs. New line characters are not accepted. See **Appendix C** in **SMRT Tools reference guide (v11.1)** for details.
Advanced Assembly Options for plasmid stage	NONE	A semicolon-separated list of KEY=VALUE pairs. New line characters are not accepted. See **Appendix C** in **SMRT Tools reference guide (v11.1)** for details.
Maximum plasmid length, bp	300,000	Value that should be set higher than the maximum size of a plasmid in the input sample. The default value should work well in most cases.
Run secondary polish	ON	Specify that an additional polishing stage be run at the end of the workflow.
Base modifications to identify	m4C,m6A	Specify the base modifications to identify, in a comma-separated list.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is 20 (QV 20), or 99% predicted accuracy.
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** comma-separated) list of other filters to add to the Data Set.
Cleanup intermediate files	ON	Removes intermediate files from the run directory to save space.
Minimum Qmod Score	35	Specify the minimum Qmod score to use in motif-finding.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Microbial Genome Analysis application generates the following reports:

Mapping Report > Summary Metrics

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

- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
- **Number of Alignments:** The number of alignments that mapped to the reference sequence.
- **Number of CCS reads (total):** The total number of CCS reads in the sequence.
- **Number of CCS reads (mapped)**: The number of CCS reads that mapped to the reference sequence.
- **Number of CCS reads (unmapped)**: The number of CCS reads not mapped to the reference sequence.

- **Percentage of CCS reads (mapped):** The percentage of CCS reads that mapped to the reference sequence.
- **Percentage of CCS reads (unmapped):** The percentage of CCS reads not mapped to the reference sequence.
- **Number of CCS Bases (mapped)**: The number of CCS bases that mapped to the reference sequence.
- **CCS Read Length Mean (mapped):** The mean read length of CCS reads that mapped to the reference sequence, starting from the first mapped base of the first mapped CCS read, and ending at the last mapped base of the last mapped CCS read.
- **CCS Read Length N50 (mapped)**: The read length at which 50% of the mapped bases are in CCS reads longer than, or equal to, this value.
- **CCS Read Length 95% (mapped)**: The 95th percentile of read 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 mapping statistics per movie.

- **Sample**: The sample name for which the following metrics apply.
- **Movie**: The movie name for which the following metrics apply.
- **Number of CCS Reads (mapped)**: The number of CCS reads that mapped to the reference sequence. This includes adapters.
- **CCS Read Length Mean (mapped)**: The mean read length of CCS reads that mapped to the reference sequence, starting from the first mapped base of the first mapped CCS read, and ending at the last mapped base of the last mapped CCS read.
- **CCS Read Length N50 (mapped)**: The read length at which 50% of the mapped bases are in CCS reads longer than, or equal to, this value.
- **Number of CCS Bases (mapped)**: The number of CCS bases that mapped to the reference sequence.
- **Mean Concordance (mapped)**: The mean concordance of subreads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

Mapping Report > Mapped CCS Read Length

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

Mapping Report > Mapped CCS Reads Concordance

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

Mapping Report > Mapped Concordance vs Read Length

- Maps the percent concordance with the reference sequence against the CCS read 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 from Microbial Assembly HiFi

Displays a table of details about all assembled contigs.

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

Coverage > Summary Metrics

Displays depth of coverage across the *de novo*-assembled genome, as well as depth of coverage distribution.

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

Coverage > Coverage across Reference

- Displays coverage at each position of the draft genome assembly.

Coverage > Depth of Coverage

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

Coverage > Coverage vs. [GC] Content

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

Base Modifications > Kinetic Detections

- **Per-Base Kinetic Detections:** Maps the modification QV against per-strand coverage.
- **Kinetic Detections Histogram:** Histogram distribution of the number of bases by modification QV.

Modified Base Motifs > Modified Base Motifs

Displays statistics for the methyltransferase recognition motifs detected.

- **Motif**: The nucleotide sequence of the methyltransferase recognition motif, using the standard IUPAC nucleotide alphabet.
- **Modified Position**: The position within the motif that is modified. The first base is `0`. **Example**: The modified adenine in GATC is at position 2.
- **Modification Type**: The type of chemical modification most commonly identified at that motif. These are: 6mA, 4mC, or `modified_base` (modification not recognized by the software.)
- **% of Motifs Detected**: The percentage of times that this motif was detected as modified across the entire genome.
- **# of Motifs Detected**: The number of times that this motif was detected as modified across the entire genome.
- **# of Motifs In Genome**: The number of times this motif occurs in the genome.
- **Mean QV**: The mean modification QV for all instances where this motif was detected as modified.
- **Mean Coverage**: The mean coverage for all instances where this motif was detected as modified.
- **Partner Motif**: For motifs that are not self-palindromic, this is the complementary sequence.
- **Mean IPD Ratio**: The mean inter-pulse duration. An IPD ratio greater than 1 means that the sequencing polymerase slowed down at this base position, relative to the control. An IPD ratio less than 1 indicates speeding up.
- **Group Tag**: The motif group of which the motif is a member. Motifs are grouped if they are mutually or self reverse-complementary. If the motif isn't complementary to itself or another motif, the motif is given its own group.
- **Objective Score**: For a given motif, the objective score is defined as (`fraction methylated`) * (`sum of log-p values of matches`).

Modified Base Motifs > Modification QVs

- Maps motif sites against Modification QV for all genomic occurrences of a motif, for each reported motif, including "No Motif".

Modified Base Motifs > ModQV Versus Coverage by Motif

- Maps coverage against Modification QV for all genomic occurrences of a motif, for each reported motif.

Data > File Downloads

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

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

- **Final Polished Assembly**: The polished assembly before oriC rotation is applied, in FASTA format.
- **Final Polished Assembly Index**: The BAI index file for the polished assembly before oriC rotation is applied.
- **Final Polished Assembly for NCBI**: The final polished assembly with applied oriC rotation and header adjustment for NCBI submission, in FASTA format.
- **Coverage Summary**: Coverage summary for regions (bins) spanning the reference sequence.

Data > IGV Visualization Files

The following files are used for visualization using IGV; see for details.

- **Mapped BAM**: The BAM file of subread alignments to the draft contigs used for polishing.
- **Mapped BAM Index**: The BAI index file for the corresponding Mapped BAM file.
- **Final Polished Assembly**: The polished assembly before oriC rotation is applied, in FASTA format.
- **Final Polished Assembly Index**: The BAI index file for the polished assembly before oriC rotation is applied.
- **Per-Base IPDs for IGV**: BigWig file containing encoded per-base IPD ratios.

Read Segmentation and Single-Cell Iso-Seq® Analysis

The Read Segmentation and Single-Cell Iso-Seq Analysis application enables analysis and functional characterization of full-length transcript isoforms with additional single-cell information, including single-cell barcodes and unique molecular identifiers (UMIs), that were sequenced on PacBio instruments.

The application is for use when using concatenation-based library preparations such as the MAS-Seq libraries.

The Read Segmentation and Single-Cell Iso-Seq Analysis application:

- Accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.

Workflow

1. Split arrayed HiFi reads at adapter positions, generating **segmented reads** (S-reads) which are the comprising fragments. For each input HiFi read, the step creates multiple BAM records, one for each fragment. An arrayed HiFi read can contain many fragments.
2. Full-length reads are then identified by the presence of cDNA primers and polyA tails. Then, UMI and barcode information is extracted.
3. After barcode correction and UMI deduplication, the unique molecules are mapped to the reference genome and classified and filtered against reference annotation using `pigeon`, which is a transcript classification and filtering tool based on the SQANTI3 software.

Importing/exporting analysis settings

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

Segmentation Adapter Set

- Specify a FASTA file, provided by PacBio, containing segmentation adapters. If you need a **custom** segmentation adapter set, click **Advanced Parameters** and use a custom FASTA file formatted as described in the table below.

Primer Set (Required) (Default = 10x Chromium single cell 3' cDNA primers)

- Specify a primer sequence file in FASTA format to identify cDNA primers for removal. The primer sequence includes the 5' and 3' cDNA primers.

- 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 the example below.)
- Each primer sequence must be **unique**.

Example: The 10x Chromium single cell 3' cDNA primer set.

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGG
>3p
AGATCGGAAGAGCGTCGTGTAG
```

Reference Set (Required)

- Specify one of two default reference genome and annotation sets to align high quality isoforms to, and to collapse isoforms mapped to the same genomic loci. The default sets are `Human_hg38_Gencode_v39` and `Mouse_mm39_Gencode_vM28`.

Parameters

Advanced parameters	Default value	Description
10X Barcodes (text, gzipped)	3m-february-2018.txt.gz	A gzipped text file containing known 10X single-cell barcodes, one per line. This include list specifies the barcode-set to which raw cell barcodes are remapped by minimum edit distance.
Adapters FASTA	NONE	Specify a custom FASTA file containing segmentation adapters. If **not** specified, the adapters specified in the XML metadata are used. Adapters **must** be ordered in the expected order of adapters in the reads. There should be one entry per adapter (forward or reverse-complement orientation) with no overlapping adapter sequences. Duplicate names or sequences are not allowed. **Example:** >A AGCTTACTTGTGAAGA >B ACTTGTAAGCTGTCTA >C ACTCTGTCAGGTCCGA >D ACCTCCTCCTCCAGAA >E AACCGGACACACTTAG
Single cell barcode and UMI design	T-12U-16B	• `T` indicates the transcript position and is **mandatory**. It is the anchor that determines whether tags are located on the 3' or 5' side. • `U`, as in UMI, **must** be preceded by the length of the UMI. • `B`, as in cell barcode, **must** be preceded by the length of the cell barcode. **Example**: `T-12U-16B` indicates a 12bp UMI and 16bp cell barcode after the transcript (with polyA tail).

Advanced parameters	Default value	Description
Output prefix	isoseq	The output file name prefix for all non-matrix files. If **not** specified, the primary input classification file is used to determine the prefix. **Example:** `my_isoforms_classification.txt` inputs yields an output prefix of `my_isoforms`.
Cell Barcode Finding Method	knee	Select the `knee` or `percentile` method to determine the number of real cells based on the cell barcode ranking plot.
Cell Barcode Percentile Cutoff	99	An integer between 0-100 for the percentile cutoff if using the `percentile` cell barcode finding method. (This value is **ignored** if using the `knee` cell barcode finding method.)
Base task memory (MB)	512	Specifies the **minimum** memory per task when submitting a job.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Read Segmentation and Single-Cell Iso-Seq Analysis application generates the following reports:

Read Segmentation > Summary Metrics

- **Reads:** The number of input arrayed HiFi reads.
- **Segmented reads (S-reads):** The number of generated S-reads.
- **Mean length of S-reads:** The mean read length of the generated S-reads.
- **Percent of reads with full arrays:** The percentage of input reads containing all adapter sequences in the order listed in the segmentation adapter FASTA file.
- **Mean array size:** The mean number of fragments (or S-reads) found in the input reads.

Read Segmentation > Segmentation Statistics

- Histogram distribution of the number of S-reads per read.
- Heatmap of adapter ligations.

Read Segmentation > Length of Reads

- Histogram distribution of the number of HiFi reads by read length, in base pairs.

Read Segmentation > S-read Length Distribution

- Histogram distribution of the number of S-reads by the HiFi read length, in base pairs.

Read Statistics > Summary Metrics

- **Reads:** The total number of input reads.
- **Read Type:** The type of input reads - `CCS`, `SEGMENT`, or `mixed` if there are multiple input data sets with mixed data types.
- **Reads with 5' and 3' Primers with extracted UMIs and Barcodes:** The number of reads with 5' and 3' cDNA primers detected, and UMI/cell barcode information extracted. Also known as full-length tagged reads (FLT Reads).
- **Non-Concatemer Reads with 5' and 3' Primers and Poly-A Tail (FLNC Reads)**: The number of non-concatemer reads with 5' and 3' primers and polyA tails detected after UMI/cell barcode information has been extracted.

- **FLNC Reads with Valid Barcodes:** The number of full-length non-concatemer tagged reads that include valid barcodes, given a cell barcode whitelist.
- **FLNC Reads with Valid Barcodes, corrected:** The number of full-length non-concatemer tagged reads that include valid barcodes (given a cell barcode whitelist) after cell barcode correction.
- **Reads after Barcode Correction and UMI Deduplication:** The number of deduplicated reads, after barcode correction and deduplication.

Read Statistics > Length of Reads

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

Cell Statistics > Summary Metrics

- **Estimated Number of Cells:** The estimated number of cells.
- **Reads in Cells:** The percentage of reads in cells.
- **Mean Reads per Cell:** The mean number of reads per cell.
- **Median UMIs per Cell:** The median number of unique molecular identifiers (UMIs) per cell.

Cell Statistics > Barcode Rank Plot

- Displays the distribution of barcode counts and which barcodes were inferred to be associated with cells. The x-axis denotes barcodes ranked in decreasing order by UMI counts mapped to each barcode, and the y-axis denotes the UMI count for the x-th ranked barcode.

Transcript Statistics > Summary Metrics

- **FLNC reads mapped confidently to genome:** The number of FLNC reads mapped to the reference genome. This number is calculated first based on the number of deduplicated reads mapped to the genome, then expanded to account for duplicate FLNC reads for each unique molecule.
- **FLNC reads mapped confidently to transcriptome**: The number of FLNC reads mapped to the reference genome in which the read is later associated with a transcript that is classified as one of the following: `FSM`, `ISM`, `NIC`, or `NNC`.
- **Total unique genes**: The total number of unique genes across all cells.
- **Total unique genes, filtered**: The total number of unique genes, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique genes, known genes only**: The total number of unique genes across all cells in which the gene is annotated in the reference annotation.
- **Total unique genes, filtered, known genes only**: The total number of unique genes (genes annotated in the reference annotation) across all cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts**: The total number of unique transcripts across all cells.
- **Total unique transcripts, filtered**: The total number of unique transcripts across all cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts, known transcripts only**: The total number of unique transcripts across all cells in which the gene the transcript belongs to is annotated in the reference annotation.
- **Total unique transcripts, filtered, known transcripts only**: The total number of unique transcripts across all cells, after filtering out reads based on the SQANTI transcript filtering criteria. Only transcripts associated with known genes (genes annotated in the reference annotation) are included.

Transcript Statistics > Transcript Summary

- **Median genes per cell:** The median number of genes per cell.

- **Median genes per Cell, known genes only**: The median number of unique, known genes (genes annotated in the reference annotation) per input cell.
- **Median transcripts per cell:** The median number of transcripts per cell.
- **Median transcripts per cell, known transcripts only:** The median number of transcripts per cell. Only transcripts associated with known genes are included.
- **Total unique genes**: The total number of unique genes across all cells.
- **Total unique genes, known genes only**: The total number of unique, known genes (genes annotated in the reference annotation) across all cells.
- **Total unique transcripts**: The total number of unique transcripts across all cells.
- **Total unique transcripts, known transcripts only**: The total number of unique transcripts across all cells. Only transcripts associated with known genes are included.

Transcript Statistics > Transcript Summary, Filtered

- **Median genes per cell:** The median number of genes per cell, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Median genes per cell, known genes only:** The median number of unique known genes (genes annotated in the reference annotation) per cell, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Median transcripts per cell:** The median number of transcripts per cell, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Median transcripts per cell, known transcripts only:** The median number of transcripts per cell, after filtering out reads based on the SQANTI transcript filtering criteria. Only transcripts associated with known genes are included.
- **Total unique genes:** The total number of unique genes across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique genes, known genes only:** The total number of unique known genes (genes annotated in the reference annotation) across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts:** The total number of unique transcripts across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts, known transcripts only:** The total number of unique transcripts across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria. Only transcripts associated with known genes are included.

Transcript Statistics > Transcript Classification

- **Category:** Transcript classification assigned by the classification and filtering tool `pigeon`, based on the SQANTI3 software.

Category	Description
FSM (Full splice match)	The reference and query isoform have the **same** number of exons and each internal junction matches the positions of the reference. The exact 5' start and 3' end can differ within the first/last exons.
ISM (Incomplete splice match)	The query isoform has **fewer** external exons than the reference, but each internal junction matches the positions of the reference. The exact 5' start and 3' end can differ within the first/last exons.
NIC (Novel in catalog)	The query isoform does **not** have a FSM or ISM match, but is using a combination of known donor/acceptor sites.
NNC (Novel not in catalog)	The query isoform does **not** have a FSM or ISM match, and has at least **one** donor or acceptor site that is **not** annotated.
Antisense	The query isoform does **not** have overlap a same-strand reference gene but is anti-sense to an annotated gene.

Category	Description
Fusion	The query isoform overlaps two or more reference genes.
More junctions	The query isoform overlaps two or more reference genes, however some junctions are shared by multiple genes.
Genic intron	The query isoform is completely contained within a reference intron.
Genic genomic	The query isoform overlaps with introns and exons.
Intergenic	The query isoform is in the intergenic region.

- **Count:** The number of transcripts in a specific classification**.**
- **CAGE Detected:** The number of transcripts where the transcription start site falls within 50bp of an annotated CAGE (Cap Analysis of Gene Expression) peak site. (See here for more information.)
- **CAGE Detected (%):** The percentage of transcripts where the transcription start site falls within 50bp of an annotated CAGE peak site.
- **polyA Motif Detected:** The number of transcripts where a known polyA motif is detected upstream of the transcription end site.
- **polyA Motif Detected (%):** The percentage of transcripts where a known polyA motif is detected upstream of the transcription end site.

Transcript Statistics > Transcript Classification, filtered

- **Category:** Transcript classification assigned by the classification and filtering tool pigeon, based on the SQANTI3 software.
- **Count:** The number of transcripts, after filtering out reads based on the SQANTI filtering criteria, in a specific classification.
- **CAGE Detected:** The number of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where the transcription start site falls within 50bp of an annotated CAGE peak site.
- **CAGE Detected (%):** The percentage of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where the transcription start site falls within 50bp of an annotated CAGE peak site.
- **polyA Motif Detected:** The number of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where a known polyA motif is detected upstream of the transcription end site.
- **polyA Motif Detected (%):** The percentage of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where a known polyA motif is detected upstream of the transcription end site.

Transcript Statistics > Transcript Classification Plots

- **Isoform distributions across structural categories:**
  - Distribution of the percentage of transcripts by structural categories.
- **Structural categories by isoform lengths:**
  - Length distribution of transcripts in different structural categories.

Transcript Statistics > Transcript Classification Plots, Filtered

- **Isoform distributions across structural categories:**
  - Distribution of the percentage of isoforms by structural categories, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Structural categories by isoform lengths:**
  - Histogram display of the number of isoforms by their length in KB and their structural category, after filtering out reads based on the SQANTI transcript filtering criteria.

Transcript Statistics > Gene Saturation

Saturation plot showing the level of gene saturation based on the number of subsampled reads.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Report read_segmentation**: JSON report containing summary statistics.
- **Segmented Reads, passing, unaligned**: BAM file containing the generated S-reads that passed filtering.
- **Non-passing reads, unaligned:** BAM file containing HiFi reads that did **not** generate S-reads.
- **Transcript Classifications**: Text file containing detailed information on transcripts.
- **Transcript Exon Junctions**: Text file containing detailed information on the exon junctions of each transcript.
- **Mapped Transcripts BAM**: BAM file containing the transcripts that mapped to the reference genome.
- **Mapped Transcripts BAM Index**: BAM index file associated with the Mapped Transcript BAM file.
- **Cell barcode ranking and count information, after correction**: Tab-delimited CSV file containing single-cell barcode ranking and count information after cell barcode correction.
- **Unique mapped transcripts, GFF**: GFF file containing unique mapped transcripts.
- **Unique mapped transcripts, filtered, GFF**: GFF file containing unique mapped transcripts after filtering.
- **Unique mapped transcripts, classification TXT**: Text file containing unique mapped transcript classifications against annotations.
- **Unique mapped transcripts, filtered, classification TXT**: Text file containing unique mapped transcript classifications against annotations, after filtering.
- **Unique mapped transcripts, junctions TXT**: Text file containing information about unique mapped transcript junctions.
- **Unique mapped transcripts, filtered, junctions TXT**: Text file containing information about unique mapped transcript junctions, after filtering.
- **Deduplicated reads after cell barcode correction, unmapped, BAM**: BAM file containing unmapped reads after cell barcode correction and UMI deduplication.
- **Deduplicated reads after cell barcode correction, mapped, BAM**: BAM file containing mapped reads after cell barcode correction and UMI deduplication.
- **Deduplicated reads after cell barcode correction, mapped, BAM index**: BAM index file associated with the BAM file containing mapped reads after cell barcode correction and UMI deduplication.
- **Single-cell isoform and gene matrix, gzipped**: Gzipped file containing Seurat-compatible isoform and gene matrix files.
- **<Data Set> Segmented Reads**: Output Data Set, containing generated S-reads and supplementary files.

Single-Cell Iso-Seq® Analysis

The Single-Cell Iso-Seq application enables analysis and functional characterization of full-length transcript isoforms with additional single cell information, including single cell barcodes and UMIs, that were sequenced on PacBio instruments.

The Single-Cell Iso-Seq Analysis application:

- Accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.

Importing/exporting analysis settings

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

Primer Set (Required) (Default = 10x Chromium single cell 3' cDNA primers)

- 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).
- Each primer sequence must be **unique**.

Example 1: The Iso-Seq cDNA Primer primer set, included with the SMRT Link installation.

Users following the standard Iso-Seq Express protocol **without** multiplexing, or running a Data Set that has **already** been demultiplexed (either using Run Design or the SMRT Analysis application) should use this default option.

```
>IsoSeq_5p
GCAATGAAGTCGCAGGGTTGGG
>IsoSeq_3p
GTACTCTGCGTTGATACCACTGCTT
```

Example 2: The Iso-Seq 12 Barcoded cDNA Primers set, included with the SMRT Link installation.

Users using barcoded cDNA primers listed in the **Appendix 3 - Recommended barcoded NEBNext single cell cDNA PCR primer and Iso-Seq Express cDNA PCR primer sequences** section of the document

Procedure & checklist - Preparing Iso-Seq® libraries using SMRTbell prep kit 3.0, should select this option.

```
>bc1001_5p
CACATATCAGAGTGCGGCAATGAAGTCGCAGGGTTGGGG
>bc1002_5p
ACACACAGACTGTGAGGCAATGAAGTCGCAGGGTTGGGG
…
```

(There are a total of 24 sequence records, representing 12 pairs of F/R barcoded cDNA primers.)

Example 3: An example of a custom cDNA primer set. 4 tissues were multiplexed using barcodes on the 3' end only.

```
>IsoSeq_5p
GCAATGAAGTCGCAGGGTTGGG
>dT_BC1001_3p
AAGCAGTGGTATCAACGCAGAGTACCACATATCAGAGTGCG
>dT_BC1002_3p
AAGCAGTGGTATCAACGCAGAGTACACACACAGACTGTGAG
>dT_BC1003_3p
AAGCAGTGGTATCAACGCAGAGTACACACATCTCGTGAGAG
>dT_BC1004_3p
AAGCAGTGGTATCAACGCAGAGTACCACGCACACGCGCG
```

Example 4: 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 (Required)

- Specify one of two default reference genome sets to align High Quality isoforms to, and to collapse isoforms mapped to the same genomic loci. The default sets are `Human_hg38_Gencode_v39` and `Mouse_mm39_Gencode_vM28`.

Parameters

Advanced parameters	Default value	Description
10X Barcodes (text, gzipped)	3m-february-2018.txt.gz	Gzipped text file containing the allowed barcodes, one barcode per line.
Output prefix	isoseq	The output file name prefix for all non-matrix files. If **not** specified, the primary input classification file is used to determine the prefix. **Example:** `my_isoforms_classification.txt` inputs yields an output prefix of `my_isoforms`.

Advanced parameters	Default value	Description
Single cell barcode and UMI design	T-12U-16B	• `T` indicates the transcript position and is **mandatory**. It is the anchor that determines whether tags are located on the 3' or 5' side. • `U`, as in UMI, **must** be preceded by the length of the UMI. • `B`, as in cell barcode, **must** be preceded by the length of the cell barcode. **Example**: `T-12U-16B` indicates a 12bp UMI and 16bp cell barcode after the transcript (with polyA tail).
Cell Barcode Finding Method	knee	Select the `knee` or `percentile` method to determine the number of real cells based on the cell barcode ranking plot.
Cell Barcode Percentile Cutoff	99	An integer between 0-100 for the percentile cutoff if using the `percentile` cell barcode finding method. (This value is **ignored** if using the `knee` cell barcode finding method.)
Base task memory (MB)	512	Specifies the **minimum** memory per task when submitting a job.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Single-Cell Iso-Seq Analysis application generates the following reports:

Read Statistics > Summary Metrics

- **Reads:** The total number of input reads.
- **Read Type:** The type of input reads - `CCS`, `SEGMENT`, or `mixed` if there are multiple input data sets with mixed data types.
- **Reads with 5' and 3' Primers with extracted UMIs and Barcodes:** The number of reads with 5' and 3' cDNA primers detected, and UMI/cell barcode information extracted. Also known as full-length tagged reads (FLT Reads).
- **Non-Concatemer Reads with 5' and 3' Primers and Poly-A Tail (FLNC Reads)**: The number of non-concatemer reads with 5' and 3' primers and polyA tails detected after UMI/cell barcode information has been extracted.
- **FLNC Reads with Valid Barcodes:** The number of full-length non-concatemer tagged reads that include valid barcodes, given a cell barcode whitelist.
- **FLNC Reads with Valid Barcodes, corrected:** The number of full-length non-concatemer tagged reads that include valid barcodes (given a cell barcode whitelist) after cell barcode correction.
- **Reads after Barcode Correction and UMI Deduplication:** The number of deduplicated reads, after barcode correction and deduplication.

Read Statistics > Length of Reads

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

Cell Statistics > Summary Metrics

- **Estimated Number of Cells**: The estimated number of real cells, based on the cell barcode ranking plot.
- **Reads in Cells**: The number of reads in cells.
- **Mean Reads per Cell**: The mean number of reads per cell.
- **Median UMIs per Cell**: The median number of unique molecular identifiers (UMIs) per cell.

Cell Statistics > Barcode Rank Plot

- Displays the distribution of barcode counts and which barcodes were inferred to be associated with cells. The x-axis denotes barcodes ranked in decreasing order by UMI counts mapped to each barcode, and the y-axis denotes the UMI count for the x-th ranked barcode.

Transcript Statistics > Summary Metrics

- **FLNC reads mapped confidently to genome:** The number of FLNC reads mapped to the reference genome. This number is calculated first based on the number of deduplicated reads mapped to the genome, then expanded to account for duplicate FLNC reads for each unique molecule.
- **FLNC reads mapped confidently to transcriptome**: The number of FLNC reads mapped to the reference genome in which the read is later associated with a transcript that is classified as one of the following: FSM, ISM, NIC, or NNC.
- **Total unique genes**: The total number of unique genes across all cells.
- **Total unique genes, filtered**: The total number of unique genes, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique genes, known genes only**: The total number of unique genes across all cells in which the gene is annotated in the reference annotation.
- **Total unique genes, filtered, known genes only**: The total number of unique genes (genes annotated in the reference annotation) across all cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts**: The total number of unique transcripts across all cells.
- **Total unique transcripts, filtered**: The total number of unique transcripts across all cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts, known transcripts only**: The total number of unique transcripts across all cells in which the gene the transcript belongs to is annotated in the reference annotation.
- **Total unique transcripts, filtered, known transcripts only**: The total number of unique transcripts across all cells, after filtering out reads based on the SQANTI transcript filtering criteria. Only transcripts associated with known genes (genes annotated in the reference annotation) are included.

Transcript Statistics > Transcript Summary

- **Median genes per cell:** The median number of genes per cell.
- **Median genes per Cell, known genes only**: The median number of unique, known genes (genes annotated in the reference annotation) per input cell.
- **Median transcripts per cell:** The median number of transcripts per cell.
- **Median transcripts per cell, known transcripts only:** The median number of transcripts per cell. Only transcripts associated with known genes are included.
- **Total unique genes**: The total number of unique genes across all cells.
- **Total unique genes, known genes only**: The total number of unique, known genes (genes annotated in the reference annotation) across all cells.
- **Total unique transcripts**: The total number of unique transcripts across all cells.
- **Total unique transcripts, known transcripts only**: The total number of unique transcripts across all cells. Only transcripts associated with known genes are included.

Transcript Statistics > Transcript Summary, Filtered

- **Median genes per cell:** The median number of genes per cell, after filtering out reads based on the SQANTI transcript filtering criteria.

- **Median genes per cell, known genes only:** The median number of unique known genes (genes annotated in the reference annotation) per cell, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Median transcripts per cell:** The median number of transcripts per cell, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Median transcripts per cell, known transcripts only:** The median number of transcripts per cell, after filtering out reads based on the SQANTI transcript filtering criteria. Only transcripts associated with known genes are included.
- **Total unique genes:** The total number of unique genes across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique genes, known genes only:** The total number of unique known genes (genes annotated in the reference annotation) across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts:** The total number of unique transcripts across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Total unique transcripts, known transcripts only:** The total number of unique transcripts across all input cells, after filtering out reads based on the SQANTI transcript filtering criteria. Only transcripts associated with known genes are included.

Transcript Statistics > Transcript Classification

- **Category:** The type of transcript detected.

Category	Description
FSM (Full splice match)	The reference and query isoform have the **same** number of exons and each internal junction matches the positions of the reference. The exact 5' start and 3' end can differ within the first/last exons.
ISM (Incomplete splice match)	The query isoform has **fewer** external exons than the reference, but each internal junction matches the positions of the reference. The exact 5' start and 3' end can differ within the first/last exons.
NIC (Novel in catalog)	The query isoform does **not** have a FSM or ISM match, but is using a combination of known donor/acceptor sites.
NNC (Novel not in catalog)	The query isoform does **not** have a FSM or ISM match, and has at least **one** donor or acceptor site that is **not** annotated.
Antisense	The query isoform does **not** have overlap a same-strand reference gene but is anti-sense to an annotated gene.
Fusion	The query isoform overlaps two or more reference genes.
More junctions	The query isoform overlaps two or more reference genes, however some junctions are shared by multiple genes.
Genic intron	The query isoform is completely contained within a reference intron.
Genic genomic	The query isoform overlaps with introns and exons.
Intergenic	The query isoform is in the intergenic region.

- **Count:** The number of transcripts in a specific classification**.**
- **CAGE Detected:** The number of transcripts where the transcription start site falls within 50bp of an annotated CAGE (Cap Analysis of Gene Expression) peak site. (See here for more information.)
- **CAGE Detected (%):** The percentage of transcripts where the transcription start site falls within 50bp of an annotated CAGE peak site.
- **polyA Motif Detected:** The number of transcripts where a known polyA motif is detected upstream of the transcription end site.
- **polyA Motif Detected (%):** The percentage of transcripts where a known polyA motif is detected upstream of the transcription end site.

Transcript Statistics > Transcript Classification, filtered

- **Category:** Transcript classification assigned by the classification and filtering tool `pigeon`, based on the SQANTI3 software.
- **Count:** The number of transcripts, after filtering out reads based on the SQANTI filtering criteria, in a specific classification.
- **CAGE Detected:** The number of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where the transcription start site falls within 50bp of an annotated CAGE peak site.
- **CAGE Detected (%):** The percentage of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where the transcription start site falls within 50bp of an annotated CAGE peak site.
- **polyA Motif Detected:** The number of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where a known polyA motif is detected upstream of the transcription end site.
- **polyA Motif Detected (%):** The percentage of transcripts, after filtering out reads based on the SQANTI transcript filtering criteria, where a known polyA motif is detected upstream of the transcription end site.

Transcript Statistics > Transcript Classification Plots

- **Isoform distributions across structural categories:**
  - Distribution of the percentage of isoforms by structural categories.
- **Structural categories by isoform lengths:**
  - Histogram display of the number of isoforms by their length in KB and their structural category.

Transcript Statistics > Transcript Classification Plots, Filtered

- **Isoform distributions across structural categories:**
  - Distribution of the percentage of isoforms by structural categories, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Structural categories by isoform lengths:**
  - Histogram display of the number of isoforms by their length in KB and their structural category, after filtering out reads based on the SQANTI transcript filtering criteria.

Transcript Statistics > Gene Saturation

Saturation plot showing the level of gene saturation based on the number of subsampled reads.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Transcript Classifications**: Text file containing detailed information on transcripts.
- **Transcript Exon Junctions**: Text file containing detailed information on the exon junctions of each transcript.
- **Mapped Transcripts BAM**: BAM file containing the transcripts that mapped to the reference genome.
- **Mapped Transcripts BAM Index**: BAM index file associated with the Mapped Transcript BAM file.

- **Cell barcode ranking and count information, after correction**: Tab-delimited CSV file containing single-cell barcode ranking and count information after cell barcode correction.
- **Unique mapped transcripts, GFF**: GFF file containing unique mapped transcripts.
- **Unique mapped transcripts, filtered, GFF**: GFF file containing unique mapped transcripts after filtering.
- **Unique mapped transcripts, classification TXT**: Text file containing unique mapped transcript classifications against annotations.
- **Unique mapped transcripts, filtered, classification TXT**: Text file containing unique mapped transcript classifications against annotations, after filtering.
- **Unique mapped transcripts, junctions TXT**: Text file containing information about unique mapped transcript junctions.
- **Unique mapped transcripts, filtered, junctions TXT**: Text file containing information about unique mapped transcript junctions, after filtering.
- **Deduplicated reads after cell barcode correction, unmapped, BAM**: BAM file containing unmapped reads after cell barcode correction and UMI deduplication.
- **Deduplicated reads after cell barcode correction, mapped, BAM**: BAM file containing mapped reads after cell barcode correction and UMI deduplication.
- **Deduplicated reads after cell barcode correction, mapped, BAM index**: BAM index file associated with the BAM file containing mapped reads after cell barcode correction and UMI deduplication.
- **Single-cell isoform and gene matrix, gzipped**: Gzipped file containing Seurat-compatible isoform and gene matrix files.

Structural Variant Calling

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 **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.

Importing/exporting analysis settings

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

Reference Set (Required)

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

Parameters

Advanced parameters	Default value	Description
Minimum Length of Structural Variant (bp) (Required)	20	The minimum length of structural variants, in base pairs.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is `20` (QV 20), or 99% predicted accuracy.
Minimum % of Reads that Support Variant (any one sample) (Required)	10	Ignore calls supported by <P% of reads in every sample.
Minimum Reads that Support Variant (any one sample) (Required)	3	Ignore calls supported by <N reads in every sample.
Minimum Reads that Support Variant (total over all samples) (Required)	3	Ignore calls supported by <N reads total across samples.
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** 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 Gap-Compressed Identity (%)	70	The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.
Bio Sample Name of Aligned Dataset	NONE	Populates the Bio Sample Name (Read Group `SM` tag) in the aligned BAM file. If blank, uses the Bio Sample Name of the input file. **Note**: Avoid using spaces in Bio Sample Names as this may lead to third-party compatibility issues.

Advanced parameters	Default value	Description
Advanced pbmm2 Options	NONE	Space-separated list of custom `pbmm2` options. Not all supported command-line options can be used, and HPC settings **cannot** be modified. See **SMRT® Tools reference guide v11.1** for details.
Advanced pbsv Options	NONE	Additional `pbsv` command-line arguments. See **SMRT® Tools reference guide v11.1** for details.
Compute Settings	Select	**(Optional)** 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 Job**.
2. Enter a **name** for the analysis.
3. Ensure that **Analysis** is selected as the workflow type.
4. Select all the Data Sets for all the input samples.
5. In the **Analysis of Multiple Data Sets** list, select **One Analysis for All Data Sets**.
6. Click **Next**.
7. Select **Structural Variant Calling** from the Analysis Application list.

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

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

Reports and data files

The Structural Variant Calling application generates the following reports:

Report > Count by Sample (SV Type)

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

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

Report > Count by Sample (Genotype)

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

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

Report > Count by Annotation

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

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

Report > Length Histogram

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

Data > File Downloads

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

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

Data > IGV Visualization Files

The following files are used for visualization using IGV; see for details.

- **Aligned Reads (per sample)**: Aligned reads, in BAM format, separated by individual.
- **Index of Aligned Reads (per sample)**: BAM index files associated with the Aligned Reads BAM files.
- **Structural Variants**: All the structural variants, in VCF format. (See here for details.)

PacBio® data utilities

Following are data processing utilities provided with SMRT Analysis v11.1. These utilities are used as intermediate steps to producing biologically-meaningful results. Each utility is described later in this document, including all parameters, reports and output files generated by the utility.

Note: The following data utilities accept **only** HiFi reads as input.

5mC CpG Detection

- Analyze the kinetic signatures of cytosine bases in CpG motifs to identify the presence of 5mC.
- See for details.

Demultiplex Barcodes

- Separate reads by barcode.
- See for details.

Export Reads

- Export HiFi reads that pass filtering criteria as FASTA, FASTQ and BAM files.
- For **barcoded** runs, you must **first** run the **Demultiplex Barcodes** application to create BAM files **before** using this application.
- See for details.

Mark PCR Duplicates

- Remove duplicate reads from a HiFi reads Data Set created using an ultra-low DNA sequencing protocol.
- See for details.

Read Segmentation

- Splits arrayed HiFi reads at adapter positions, generating segmented reads (S-reads) comprised of multiple fragments.
- See for details.

Trim Ultra-Low Adapters (was Trim gDNA Amplification Adapters)

- Trim PCR Adapters from a HiFi reads Data Set created using an ultra-low DNA sequencing library.
- See for details.

Note: The following data utility accept **only** Subreads as input.

Circular Consensus Sequencing (CCS)

- Identify consensus sequences for single molecules.
- See for details.

5mC CpG Detection   Use this utility to analyze the kinetic signatures of cytosine bases in CpG motifs to identify the presence of 5mC.

- The utility accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20. The utility also requires kinetics information.

Importing/exporting analysis settings

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

Parameters

Advanced parameters	Default value	Description
Keep Kinetics in Output	OFF	If ON, specifies that the IPD and PulseWidth records are included in the output BAM file.
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** comma-separated) list of other filters to add to the Data Set.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is 20 (QV 20), or 99% predicted accuracy.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The 5mC CpG Detection utility generates the following reports:

5mC CpG Report > Methylation Probability

- **CpG Methylation in Reads**: The cumulative of percentage of CpG sites in the sample mapped against the predicted probability of methylation.
- **CpG Methylation in Reads (Histogram)**: Histogram displaying the percentage of CpG sites in the sample versus the predicted probability of methylation.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **HiFi reads with 5mC Calls**: BAM file containing all the HiFi reads in the sample that include 5mC calls.
- **<Input Data Set>(5mC)**: Output Data Set with the 5mC calls.

Demultiplex Barcodes   Use this utility to separate sequence reads by barcode. (See "Working with barcoded data" on page 118 for more details.)

- The utility accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.
- 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 of the known barcode sequences.

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

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

Notes on Iso-Seq Multiplexed Data

There are two ways in which an Iso-Seq library can be multiplexed:

1. **Barcoded adapter Iso-Seq libraries**
- If using the SMRTbell Barcoded Adapter with the Iso-Seq Express protocol **on or after** April 21, 2022, demultiplex the Data Set **prior** to running the Iso-Seq application.
- To analyze the samples in a **single** Iso-Seq run, select **all** the demultiplexed Data Sets to combine and begin the Iso-Seq analysis.

2. **Barcoded cDNA primer Iso-Seq libraries**
- If following multiplexing guidelines using the Iso-Seq Express protocol **prior to** April 21, 2022 and you ordered synthesized oligos listed in the **Appendix 2 - Recommended barcoded NEBNext single cell cDNA PCR primer and Iso-Seq Express cDNA PCR primer sequences** section, demultiplex your Data Set using the Iso-Seq application. In other words, do **not** run the Demultiplexing Barcodes utility first.
- See "Iso-Seq® Analysis" on page 68 for choices on Primer Sets to use.

Multiplexed method	Run Demultiplex Barcodes utility?
Not multiplexed	NO
Barcoded adapters	YES
Barcoded cDNA primer	NO

Importing/exporting analysis settings

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

Barcode Set (Required)

- Specify a barcode sequence file to separate the reads.

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

- Specify **Yes** to retain all the reads with the **same** barcodes on both ends of the insert sequence, such as symmetric and tailed designs. (See "Working with barcoded data" on page 118 for information on barcode designs.)
- Specify **No** to specify asymmetric designs where the barcodes are **different** on each end of the insert sequence.

Assign Bio Sample Names to Barcodes (Required)

SMRT Link automatically creates a CSV-format **Autofilled Barcoded Sample** file. The barcode name is populated based on your choice of barcode set, and if the barcodes are the same at both ends of the sequence. The file includes a column of automatically-generated Bio Sample Names 1 through N, corresponding to barcodes 1 through N, for the biological sample names. There are **two different ways** to specify which barcodes to use, and assign biological sample names to barcodes:

Interactively:

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

From a file:

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

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

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

Demultiplexed Output Data Set Name (Required)

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

Parameters

Advanced parameters	Default value	Description
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is 20 (QV 20), or 99% predicted accuracy.
Minimum Barcode Score	80	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 the analysis. This affects the output BAM file and the output demultiplexed Data Set XML file.
Advanced lima Options	NONE	Space-separated list of custom lima options. Not all supported command-line options can be used, and HPC settings **cannot** be modified. See the Demultiplex Barcodes section of the document **SMRT® Tools reference guide v11.1** for information on lima.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Demultiplex Barcodes utility generates the following reports:

Barcodes > Summary Metrics

- **Unique Barcodes**: The number of unique barcodes in the sequence data.
- **Barcoded HiFi Reads**: The number of correctly-barcoded reads in the HiFi sequence data.
- **Unbarcoded HiFi Reads**: The number of reads in the HiFi sequence data that do not contain barcodes.
- **Barcoded HiFi Read (%)**: The percentage of reads in the HiFi sequence data that contain barcodes.

- **Barcoded HiFi Yield (bp)**: The number of bases in HiFi sequence data reads that contain barcodes.
- **Unbarcoded HiFi Yield (bp)**: The number of bases in HiFi sequence data reads that do not contain barcodes.
- **Barcoded HiFi Yield (%)**: The percentage of bases in HiFi sequence data reads that contain barcodes.
- **Unbarcoded HiFi Yield (%)**: The percentage of bases in HiFi sequence data reads that do not contain barcodes.
- **Mean HiFi Reads per Barcode**: The mean number of HiFi reads per barcode combination.
- **Max. HiFi Reads per Barcode**: The maximum number of HiFi reads per barcode combination.
- **Min. HiFi Reads per Barcode**: The minimum number of HiFi reads per barcode combination.
- **Barcoded HiFi Read Length (mean, bp)**: The mean read length of HiFi reads per barcode combination, in base pairs.
- **Unbarcoded HiFi Read Length (mean, bp)**: The mean read length of HiFi reads not containing barcodes, in base pairs.

Barcodes > Barcode Data

- **Sample Name:** The name of the biological sample associated with the barcode combination.
- **Barcode**: A string containing the pair of barcode indices for which the following metrics apply.
- **Barcode Quality**: The barcode quality (QV) associated with the barcode combination.
- **HiFi Reads**: The number of HiFi reads associated with the barcode combination.
- **HiFi Read Length (mean, bp):** The mean read length of HiFi reads per barcode combination, in base pairs.
- **HiFi Read Quality (mean, QV)**: The mean barcode quality (QV) associated with the barcode combination.
- **HiFI Yield (bp):** The number of bases in HiFi sequence data reads that contain barcodes.
- **Polymerase Read Length (mean, bp)**: The mean read length of polymerase reads associated with the barcode combination, in base pairs.
- **Polymerase Yield (bp)**: The number of bases in polymerase reads associated with the barcode combination, in base pairs.

Barcodes > Inferred Barcodes

- **Barcode Name:** The barcode name.
- **Number of Reads:** The number of reads out of the first 35,000 that are inferred to be assigned to the barcode combination.
- **Mean Barcode Score:** The mean barcode score associated with the reads inferred to be associated with the barcode combination.
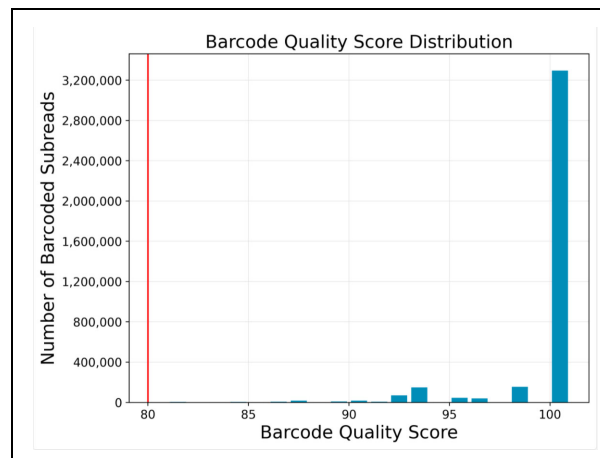
Barcodes > Barcoded Read Statistics

- **Number of Reads per Barcode**: Line graph displays the number of sorted reads per barcode.
  - **Good performance**: The Number of Reads per Barcode line (blue) should be mostly linear. Note that this depends on the choice of Y-axis scale. The mean Number of Reads per Barcode line (red) should be near the middle of the graph and should not be skewed by samples with too many or too few barcodes.

– **Questionable performance**: A sharp discontinuity in the blue line, followed by no yield, with the red line way far from the center. Check the output file **Inferred Barcodes**, note the correct barcodes used, and consider reanalyzing the multiplexed samples with the correct Bio Sample names for the barcodes actually used. If you reanalyze the data, ensure that the **Barcode Name** file includes **only** the correct barcodes used.

- **Barcode Frequency Distribution**: Histogram distribution of read counts per barcode.
  - **Good performance**: A uniform distribution, which is most often a fairly tight symmetric normal distribution, with few barcodes in the tails.
  - **Questionable performance**: A large peak at zero. This can indicate use of incorrect barcodes. Check the output file **Inferred Barcodes**, note the correct barcodes used, and consider reanalyzing the multiplexed samples with the correct Bio Sample names for the barcodes actually used. If you reanalyze the data, ensure that the **Barcode Name** file includes **only** the correct barcodes used.

- **Mean Read Length Distribution**: Histogram distribution of the mean polymerase read length for all samples.
  - **Good performance**: The distribution should be normal with a relatively tight range.
  - **Questionable performance**: A spread out distribution, with a mode towards the low end.

Barcodes > Barcode Quality Scores

- **Barcode Quality Score Distribution:** Histogram distribution of barcode quality scores. The scores range from 0-100, with 100 being a perfect match. Any significant modes or accumulation of scores <60 suggests issues with some of the barcode analyses. The red line is set at 80 – the minimum default barcode score.
  - **Good performance**: HiFi demultiplexing runs should have >90% of reads with barcode quality score ≥95.

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

Barcodes > Barcoded Read Binned Histograms

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

Data > File Downloads

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

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

Export Reads   Use this utility to export HiFi reads that pass filtering criteria as FASTA, FASTQ and BAM files.

- The utility accepts **HiFi reads** (BAM format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.
- For **barcoded** runs, you must **first** run the **Demultiplex Barcodes** utility to create BAM files **before** using this utility.
- This utility does **not** generate any reports.

Importing/exporting analysis settings

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

Output FASTA File (Default = ON)

- Outputs a single FASTA/FASTQ file containing all the reads that passed the filtering criteria.

Output BAM File (Default = OFF)

- Outputs a single BAM file containing all the reads that passed the filtering criteria.

Min. CCS Predicted Accuracy (Phred Scale) Default = 20

- Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is `20` (QV 20), or 99% predicted accuracy.

Parameters

Advanced parameters	Default value	Description
Filters to Add to the Data Set	NONE	A semicolon-separated (**not** comma-separated) list of other filters to add to the Data Set.
Compute Settings	Select	(**Optional**) 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 execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **hifi_reads.fasta.gz**: Sequence data that passed filtering criteria, converted to Gzipped FASTA format.

- **hifi_reads.fastq.gz**: Sequence data that passed filtering criteria, converted to Gzipped FASTQ format.
- **<Reads>.bam**: Sequence data that passed filtering criteria, in BAM format.

Mark PCR Duplicates

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

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

Note: If starting with a very low-input DNA sample using the **SMRTbell gDNA sample amplification kit**, you **must** run this utility (preceded by the **Trim Ultra-Low Adapters** utility) on the resulting Data Set **prior** to running any secondary analysis application.

Importing/exporting analysis settings

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

Parameters

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

Reports and data files

The Mark PCR Duplicates utility generates the following reports:

PCR Duplicates > Duplicate Rate (table)

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

PCR Duplicates > Duplicate Rate (chart)

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

Data > File Downloads

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

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

Read Segmentation	Use this utility to split arrayed HiFi reads at adapter positions, generating segmented reads (S-reads) which are the comprising fragments. For **each** input HiFi read, the utility creates multiple BAM records, one for each fragment. An arrayed HiFi read can contain many fragments.

- The utility accepts **HiFi reads** (BAM/Data Set format) as input. **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.

Importing/exporting analysis settings

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

Segmentation Adapter Set

- Specify a FASTA file, provided by PacBio, containing segmentation adapters. If you need a **custom** segmentation adapter set, click **Advanced Parameters** and use a custom FASTA file formatted as described below.

Parameters

Advanced parameters	Default value	Description
Adapters FASTA	NONE	Specify a custom FASTA file containing segmentation adapters. If **not** specified, the adapters specified in the XML metadata are used. Adapters **must** be ordered in the expected order of adapters in the reads. There should be one entry per adapter (forward or reverse-complement orientation) with no overlapping adapter sequences. Duplicate names or sequences are not allowed. **Example:** >A AGCTTACTTGTGAAGA >B ACTTGTAAGCTGTCTA >C ACTCTGTCAGGTCCGA >D ACCTCCTCCTCCAGAA >E AACCGGACACACTTAG
Base task memory (MB)	512	Specifies the **minimum** memory per task when submitting a job.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Read Segmentation utility generates the following reports:

Read Segmentation > Summary Metrics

- **Reads**: The number of input arrayed HiFi reads.
- **Segmented reads (S-reads)**: The number of generated S-reads.
- **Mean length of S-reads**: The mean read length of the generated S-reads.
- **Percent of reads with full arrays**: The percentage of input reads containing all adapter sequences in the order listed in the segmentation adapter FASTA file.
- **Mean array size**: The mean number of fragments (or S-reads) found in the input reads.

Read Segmentation > Segmentation Statistics

- Histogram distribution of the number of S-reads per read.
- Heatmap of adapter ligations.

Read Segmentation > Length of Reads

- Histogram distribution of the number of HiFi reads by read length, in base pairs.

Read Segmentation > S-read Length Distribution

- Histogram distribution of the number of S-reads by the HiFi read length, in base pairs.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Report read_segmentation**: JSON report containing summary statistics.
- **Segmented Reads, passing, unaligned**: BAM file containing the generated S-reads that passed filtering.
- **Non-passing reads, unaligned**: BAM file containing HiFi reads that did **not** generate S-reads.
- **<Data Set> Segmented Reads**: Output Data Set, containing generated S-reads and supplementary files.

Trim Ultra-Low Adapters

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

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

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

Importing/exporting analysis settings

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

PCR Adapters (Required)

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

Parameters

Advanced parameters	Default value	Description
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for **all** applications is 20 (QV 20), or 99% predicted accuracy.
Compute Settings	Select	**(Optional)** Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

Reports and data files

The Trim Ultra-Low Adapters utility generates the following reports:

PCR Adapters > Summary Metrics

- **Unique PCR Adapters**: The number of unique PCR adapters in the sequence data.
- **HiFi Reads With PCR Adapters**: The number of reads in the sequence data that contain PCR adapters.
- **HiFi Reads Without PCR Adapters**: The number of reads in the sequence data that do **not** contain PCR adapters.
- **Percent Reads with Adapters**: The percentage of reads in the sequence data that contain PCR adapters.
- **Percent Bases in Reads with Adapters**: The percentage of bases in reads in the sequence data that contain PCR adapters.
- **Mean HiFi Reads Per Adapter**: The mean number of reads per PCR adapter in the sequence data.

- **Max. HiFi Reads Per Adapter**: The maximum number of reads per PCR adapter in the sequence data.
- **Min. HiFi Reads Per Adapter**: The minimum number of reads per PCR adapter in the sequence data.
- **Mean HiFi Read Length**: The mean read length of reads per PCR adapter in the sequence data.

PCR Adapters > PCR Adapter Data

- **Bio Sample Name:** The name of the biological sample associated with the PCR adapters.
- **PCR Adapter Name**: A string containing the pair of PCR adapter indices for which the following metrics apply.
- **Mean PCR Adapter Quality:** The mean PCR adapter quality associated with the PCR adapter.
- **HiFi Reads**: The number of HiFi reads associated with the PCR adapter.
- **HiFi Read Length (mean, bp)**: The mean read length of HiFi reads associated with the PCR adapter.
- **HiFi Yield (bp)**: The total yield (in base pairs) of the HiFi reads associated with the PCR adapter.

PCR Adapters > PCR Adapter Read Statistics

- **Number of Reads Per PCR Adapter:** Histogram distribution of the mean number of reads per PCR adapter.
- **PCR Adapter Frequency Distribution:** Histogram distribution of reads with PCR adapter mapped to the number of barcoded samples.
- **Mean Read Length Distribution:** Maps the mean read length against the number of barcoded samples.

PCR Adapters > PCR Adapter Quality Scores

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

PCR Adapters > PCR Adapter Read Binned Histograms

- **Read Length Distribution By PCR Adapter**: Histogram distribution of the read length by PCR adapter. Each column of rectangles is similar to a read length histogram rotated vertically, seen from the top.
- **PCR Adapter Quality Distribution By Barcode**: Histogram distribution of the per-barcode version of the **Read Length Distribution by PCR Adapter** histogram.

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis execution.
- **SMRT Link Log**: Server-level analysis log information. (This file is displayed when you choose **Data > SMRT Link Log**.)
- **Reads Missing Adapters**: Reads Missing Adapters: BAM file containing the reads with missing PCR adapters from the input Data Set.
- **PCR Adapter Data CSV**: Includes the data displayed in the PCR Adapter Data table.
- **<Data Set> (trimmed)**: Output Data Set, with the PCR adapters removed.

Circular Consensus Sequencing (CCS)

Use this utility to identify consensus sequences for single molecules.

- The utility accepts **Subreads** (BAM format) as input.

Importing/exporting analysis settings

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

Detect 5mC Sites (Default = OFF)

- If set to **ON**, kinetics analysis to identify 5mC CpG sites will be performed.

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.
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 a Consensus for Each Strand	OFF	Generate a consensus for each strand. **Warning**: This is an experimental option for the CCS algorithm, and may **not** be compatible with all downstream applications. We recommend using command-line analysis for this feature.
Process All Reads	OFF	Specifies behavior identical to on-instrument CCS reads generation, overriding all other cutoffs. This setting writes a CCS read for **every** ZMW in the input Data Set. Set to OFF to specify more restrictive settings.
Include Kinetics information with CCS Analysis output	OFF	If ON, include kinetics per-base data required for methylation DNA analysis. **Note**: This results in a BAM file that is 3-4 times larger. This option applies only when **Process All Reads** is set to ON.
Advanced CCS Options	NONE	Space-separated list of additional command-line options to CCS analysis. Not all supported command-line options can be used, and HPC settings **cannot** be modified. See **SMRT® Tools reference guide v11.1** for details.
Minimum Predicted Accuracy (Deprecated)	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.) **Note**: This setting is **ignored** if the **Process All Reads** advanced parameter is set to ON.
Minimum Number of Passes (Deprecated)	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. **Note**: This setting is **ignored** if the **Process All Reads** advanced parameter is set to ON.
Detect And Split Heteroduplex Read	OFF	Specifies that any detected heteroduplexes are separated into separate reads.

Advanced parameters	Default value	Description
Compute Settings	Select	**(Optional)** 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) utility generates the following reports:

CCS Analysis Report > Summary Metrics

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

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

CCS Analysis Report > HiFi Read Length Summary

- **Read Length (bp):** The HiFi read length, ranging from ≥ 0 to ≥ 40,000 base pairs.
- **Reads:** The number of HiFi reads with the specified read length.
- **Reads (%)**: The percentage of HiFi reads with the specified read length.
- **Yield (bp):** The number of base pairs in the HiFi reads with the specified read length.
- **Yield (%):** The percentage of base pairs in the HiFi reads with the specified read length.

CCS Analysis Report > HiFi Read Quality Summary

- **Read Quality (Phred):** Phred-scale quality values, ranging from QV ≥20 to QV ≥50.
- **Reads:** The number of HiFi reads with the specified read quality.
- **Reads (%)**: The percentage of HiFi reads with the specified read quality.
- **Yield (bp):** The number of base pairs in the HiFi reads with the specified read quality.
- **Yield (%):** The percentage of base pairs in the HiFi reads with the specified read quality.

CCS Analysis Report > Read Length Distribution

- **HiFi Read Length Distribution**: Histogram distribution of HiFi reads by read length.
- **Yield by HiFi Read Length**: Histogram distribution of the cumulative yields of CCS reads by read length.
- **Read Length Distribution**: Histogram distribution of all reads by read length.

CCS Analysis Report > Number of Passes

- Histogram of the number of complete subreads in CCS reads, broken down by number of reads.

CCS Analysis Report > Read Quality Distribution

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

CCS Analysis Report > Predicted Accuracy vs. Read Length

- Heat map of CCS read lengths and predicted accuracies.

Data > File Downloads

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

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

Working with barcoded data

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

The canned data provided with SMRT Link v11.1 includes 9 barcode sets:

1. `gDNA Amplification Adapter`
2. `Iso-Seq 12 Barcoded cDNA Primers`
3. `Iso-Seq cDNA Primers`
4. `Barcoded Overhang Adapter Kits 8A and 8B`
5. `Barcoded M13 Primer Plate`
6. `SMRTbell Barcoded Adapter Plate 3.0`
7. `Mas-Seq Adapter v1(MAS16)`
8. `Mas-Seq Adapter v2(MAS12)`
9. `10x Chromium single cell 3' cDNA primers`

SMRT Link v11.1 supports sample traceability through the various modules in the application by using the **Bio Sample Name**.

Run Design in SMRT Link v11.1 contains a required **Bio Sample Name** field for both single and multiplexed samples.

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

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

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

Example Well Sample Names and Bio Sample Names

- **Non-barcoded Well Sample Name**: `HG002_2019_11_2_10K`
- **Non-barcoded Bio Sample Name**: `HG002`
- **Barcoded Well Sample Name**: `My Multiplexed Set of Bugs`
- **Barcoded Bio Sample Name:**
  `Unknown Microbe 1,...,Unknown Microbe N`

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

1. In SMRT Link, create a new Run Design as described in . **Before** you finish the new Run Design, perform the following steps.

2. Click **Barcoded Sample Options** and then click **Yes** for **Sample is Barcoded**. Additional fields related to barcoding display.
3. Specify a **Barcode Set** using the dropdown list.
   Note: You can specify up to 10,000 samples. Specifying **more** than 10,000 samples may cause a delay of several minutes in analysis submission.
4. Specify if the **same** barcodes are used on both ends of the sequences.
   – Selecting **Yes** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
   – Selecting **No** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such experiments retains any barcode pair combination identified in the Data Set.
5. SMRT Link automatically creates a CSV-format **Autofilled Barcode Name** file. The barcode name is populated based on your choice of barcode set, and if the barcodes are the same at both ends of the sequence. The file includes a column of automatically-generated Bio Sample Names 1 through N, corresponding to barcodes 1 through N, for the biological sample names. There are **two different ways** to specify which barcodes to use, and assign biological sample names

to barcodes. (**Note**: Bio Sample Names are hardcoded and can be traced through secondary analysis using SMRT Analysis.)

Interactively:

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

From a File:

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

6. Specify **if** and **where** to automatically generate **HiFi reads** (reads generated with CCS analysis whose quality value is equal to or greater than 20):

  – **On Instrument** (available **only** for the Sequel IIe system): HiFi reads are automatically generated on the instrument, **before** transfer to the compute cluster where SMRT Link is installed.
  – **In SMRT Link**: HiFi reads are automatically generated **after** transfer to the compute cluster where SMRT Link is installed.
  – **Do Not Generate**: HiFi reads are **not** generated for this run. Only subread data are transferred to the local compute cluster where SMRT Link is installed.

7. Click **Save**.

Step 2: Perform the sequencing run

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

Note: By default, `Demultiplex Barcodes (Auto)` creates **one** Data Set per autodetected barcode within the selected barcode set. It also applies

a Data Set filter of a minimum barcode score greater than 26 for optimal results in secondary analysis. If used, the analysis parameter **Filters to add to the Data Set** overrides other barcode filtering, even if the barcode score set with it is lower than `26`.

Step 3: (Optional) Run the Demultiplex Barcodes data utility

If instead you did **not** specify the barcode setup in the Run Design, or if you need to change any of the parameters used in the `Demultiplex Barcodes` analysis automatically launched from Run Design, run the **Demultiplex Barcodes** data utility. This 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 Job**.
2. Enter a **name** for the job.
3. Select **Data Utility** as the workflow type.
4. Select **HiFI reads** as the data type to use. The Data Sets table displays the appropriate Data Sets available for the job.
5. In the Data Sets table, select one or more Data Sets to be analyzed together.
6. Click **Next**.
7. Select **Demultiplex Barcodes** from the Applications list.
8. Specify a **Barcode Set** (barcode sequence file.)
   Note: You can specify up to 10,000 samples. Specifying **more** than 10,000 samples may cause a delay of several minutes in analysis sub-mission.
9. Specify if the **same** barcodes are used on both ends of the sequences.
   – Selecting **Yes** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
   – Selecting **No** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such data retains any barcode pair combination identified in the Data Set.
10. SMRT Link automatically creates a CSV-format **Autofilled Barcoded Sample** file. The barcode name is populated based on your choice of barcode set, and if the barcodes are the same at both ends of the sequence. The file includes a column of automatically-generated Bio Sample Names `1` through `N`, corresponding to barcodes `1` through `N`, for the biological sample names. There are **two different ways** to specify which barcodes to use, and assign biological sample names to barcodes:

Interactively:

• Click **Interactively**, then drag barcodes from the **Available Barcodes** column to the **Included Barcodes** column. (Use the check boxes to select multiple barcodes.)
• **(Optional)** Click a Bio Sample field to edit the Bio Sample Name associated with a barcode. **Note**: Avoid using spaces in Bio Sample Names as they may lead to third-party compatibility issues.

- (**Optional**) Click **Download as a file for later use**.
- Click **Submit** to save the edited barcodes/bio sample names. You see **Success** on the line below, assuming the file is formatted correctly.

From a File:

- Click **From a File**, then click **Download File.** Edit the file and enter the biological sample names associated with the barcodes in the second column, then save the file. Use alphanumeric characters, spaces (allowed but **not recommended** for compatibility with third-party downstream software), hyphens, underscores, colons, or periods **only** - other characters will be removed **automatically,** with a maximum of 40 characters. If you did **not** use all barcodes in the Autofilled Barcode Name file in the sequencing run, **delete** those rows.
- **Note**: Open the CSV file in a text editor and check that the columns are separated by **commas**, not semicolons or tabs.
- Select the **Barcoded Sample** file you just edited. You see **Success** on the line below, assuming the file is formatted correctly.
11. Specify the **name** for the new demultiplexed Data Set that will display in SMRT Link. The application creates a copy of the input Data Set, renames it to the name specified, and creates demultiplexed child Data Sets linked to it. The input Data Set remains separate and unmodified.
12. (**Optional**) Specify any advanced parameters.
13. 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 data utility, see "Demultiplex Barcodes" on page 99.

Step 4: Run applications using the demultiplexed data as input

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

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

1. Select the secondary analysis application/data utility to use.
2. Click the number in the **Demultiplexed Subsets** column, then select the demultiplexed Data Set to use as input:
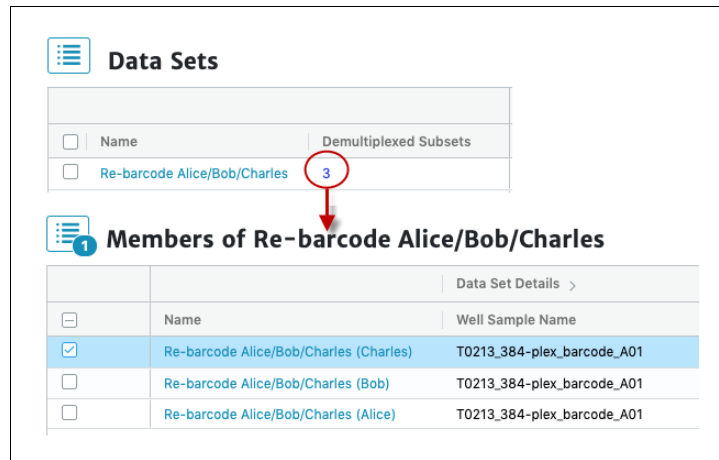
– You can select the **entire** Data Set as input, or one or more specific outputs from selected barcodes, to a maximum of 16 sub-Data Sets, 12 for Iso-Seq.

3. Additional **Analysis Type** options become available. You can select from the following options:
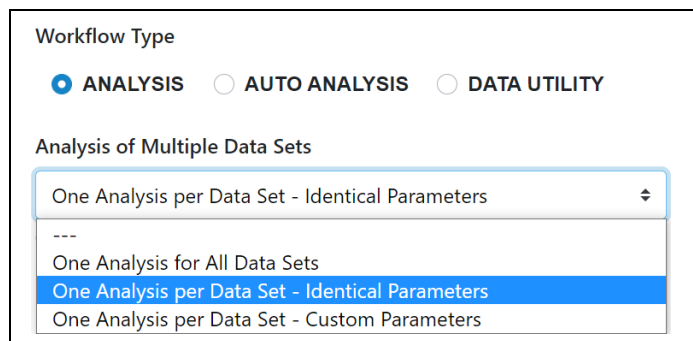
– **One Analysis for All Data Sets:** Runs one job using all the selected barcode Data Sets as input, for a maximum of 30 Data Sets.
– **One Analysis per Data Set - Identical Parameters:** Runs **one** separate job for **each** of the selected barcode Data Sets, using the **same** parameters, for a maximum of 10,000 Data Sets. Optionally click **Advanced Parameters** and modify parameters.
– **One Analysis per Data Set - Custom Parameters:** Runs **one** separate job for **each** of the selected barcode Data Sets, using **different** parameters for each Data Set, for a maximum of 16 Data Sets. Click **Advanced Parameters** and modify parameters. Then click **Start and Create Next**. You can then specify parameters for each of the included barcode Data Sets.
– **Note**: The number of Data Sets listed is based on testing using PacBio's suggested compute configuration, listed in **SMRT Link software installation guide (v11.1)**.

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

Demultiplex Barcodes details

The **Demultiplex Barcodes** data utility identifies barcode sequences in PacBio single-molecule sequencing data.

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

1. Barcoded target-specific primers
2. Barcoded universal primers
3. Barcoded overhang adapters
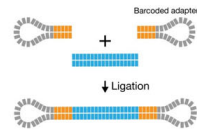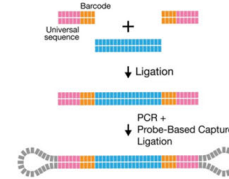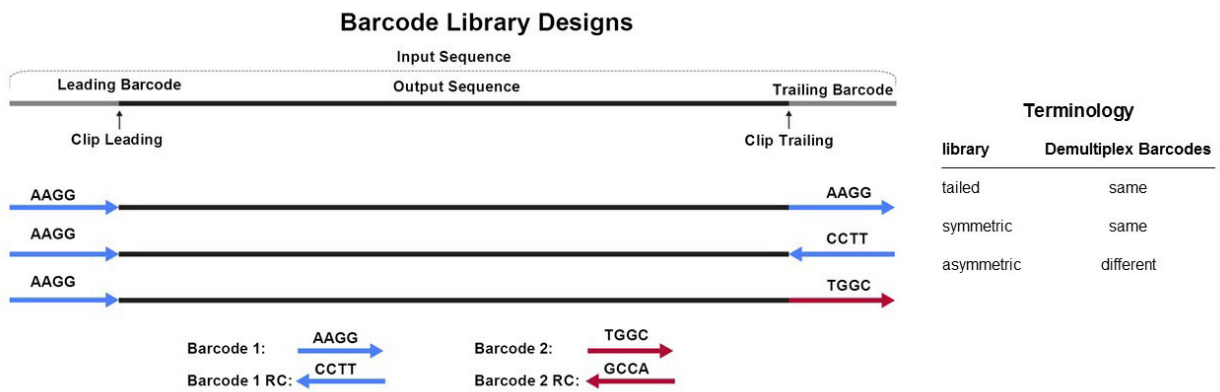4. Barcoded linear adapters (target capture)
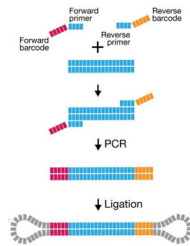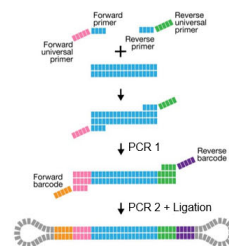
In addition, there are three different barcode library designs.

The **Demultiplex Barcodes** application in SMRT Link supports demultiplexing of subreads.

Demultiplexing of CCS reads is possible using the command line. (See **SMRT® Tools reference guide (v11.1)** for details.)

Symmetric mode

For **symmetric** and **tailed** library designs, the **same** barcode is attached to both sides of the insert sequence of interest. The only difference is the orientation of the trailing barcode. For barcode identification, one read with a single barcode region is sufficient. Symmetric barcoding is used for samples constructed using Barcoded overhang adapters, Barcoded universal primer and target enrichment (linear). This is also the default scoring mode in SMRT Link v10.2 and later.

Asymmetric mode

Barcode sequences are **different** on the ends of the SMRTbell template. Asymmetric mode is used with the M13 barcoding procedure. (See the document **Procedure & checklist - Preparing SMRTbell libraries using PacBio barcoded M13 primers for multiplex SMRT sequencing** for details.) PacBio using this mode **only** for small inserts (up to 5 kb) where both ends of the insert are expected to be sequenced. **Both** barcodes must be detected.

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

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

Mixed mode

Libraries with combined symmetric and asymmetric barcoding are **not** supported.

Automated analysis

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

- Auto Analysis can be set up in Run Design or SMRT Analysis **after** the Run Design is saved and **before** the run is loaded on the instrument.
- Auto Analysis can be run on HiFi reads, and includes **all** secondary analysis applications available.
- Auto Analysis works with **all** Sequel II systems and Sequel IIe systems.

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

- Read Type = **HiFi reads** and Generate HiFi reads = **On Instrument** or **In SMRT Link**.
- Read Type = **HiFi reads** and Sample is Barcoded = **Yes**.

Note: Pre Analysis works with **all** Sequel II systems and Sequel IIe systems.

Creating an Auto Analysis job from SMRT Analysis

This procedure includes only the basic steps - for more detailed information on creating jobs, see "Creating and starting a job" on page 44.

1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
3. Click **+ Create New Job**.
4. Enter a **name** for the job.
5. Specify **Auto Analysis** as the workflow type. The table displays available runs. (**Note**: Runs display here **only** if they are in the **Created** state - not if they are already running or have completed.)
6. Click a **Collections** link associated with a specific run in the table.
7. Select **one** sample, or drill down further and select from the barcoded samples of a single collection by clicking the **Barcoded Samples** link. You can select **multiple** barcoded samples at this level.
   Note: You **cannot** select a mix of samples and barcoded samples, or select barcoded samples from **multiple** samples. In addition, samples or barcoded samples will **not** be selectable if they have a Job Id.
8. Click **Next**.
9. Select a secondary analysis application from the dropdown list.
   Note: **Only** secondary analysis applications, **not** data utilities, are available for use with Auto Analysis.

10. **(Optional)** Click **Advanced Parameters** and specify the values of the parameters to change. Click **OK** when finished. (Different applications have different advanced parameters.)
    – To see information about parameters for **all** secondary analysis applications provided by PacBio, see "PacBio® secondary analysis applications" on page 54.
11. Click **Start** to submit the Auto Analysis job.

Creating Auto Analysis from a Run Design

1. Create a new Run Design (See "Creating a new Run Design" on page 16 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 job that will be associated with the collection.
3. Name the new job.
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 to change. Click **OK** when finished. To see information about parameters for **all** secondary analysis applications provided by PacBio, see "PacBio® secondary analysis applications" on page 54.
9. Click **Create**.

HiFiViral SARS-CoV-2: Creating Auto Analysis in Run Design

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

1. In Run Design, click **Create New Design**.
2. From the Application list, select **HiFiViral SARS-CoV-2**. Preloaded default values display in green.
3. Enter the **Well Sample Name**.
4. By default, **Auto Analysis** is set to **Yes** for **Automatic Launch of SARS-CoV-2 Analysis**.
5. Enter the **Analysis Name**.
6. By default, **Yes** is selected for **Sample Is Barcoded** and **No** for **Same Barcode on Both End of Sequence**.
7. By default, the barcode set **HiFiViral SARS-CoV-2 M13barcodes** is selected.
8. For **Assign Bio Sample Names to Barcodes**, select **From a File** and then click **Download File**.
9. Open the downloaded file (**assayPlateQC_template_4by96.csv**) in a text editor. See "HiFiViral SARS-CoV-2 Analysis" on page 54 for details on modifying this file for your samples.

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** jobs.
2. To filter the jobs, click the funnel in the **State** column header, then click **Created**. This displays **only** jobs in the **Created** state.
3. Click the job of interest.
4. 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. Scroll all the way to the right in the table. This displays information about the samples included in the run, including Pre Analysis and Auto Analysis details.

Visualizing data using IGV

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

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

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

- Iso-Seq Analysis
- HiFi Mapping
- Microbial Genome Analysis
- 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.

To visualize data using IGV

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

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

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

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

Using the PacBio® self-signed SSL certificate

SMRT Link v11.1 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.
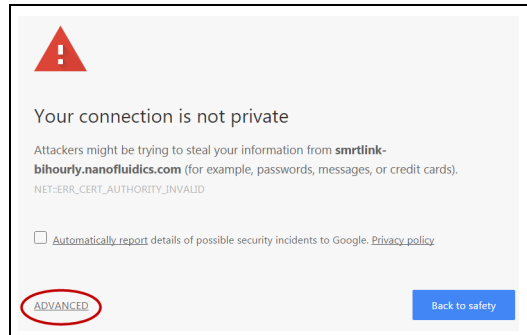
2. Click the **Proceed...** link. (You may need to scroll down.)
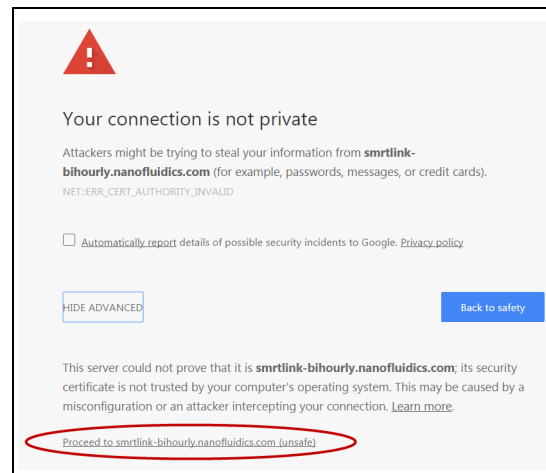
3. Close the window by clicking the **Close** box in the corner.

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® II system and Sequel IIe system output files

This section describes the data generated by the Sequel IIe system and Sequel II system for each SMRT Cell transferred to network storage.

Sequel IIe system output files

Following is a sample of the file and directory structure output by the **Sequel IIe** system (**not** including low-quality reads):

```
<your_specified_output_directory>/r64012_211206ee_183753/1_A01/
|--m64012ee_211206_183753.baz2bam_1.log
|--m64012ee_211206_183753.ccs.log
|--m64012ee_211206_183753.ccs_reports.json
|--m64012ee_211206_183753.ccs_reports.txt
|--m64012ee_211206_183753.consensusreadset.xml
|--m64012ee_211206_183753.hifi.reads.bam
|--m64012ee_211206_183753.hifi.reads.bam.pbi
|--m64012ee_211206_183753.sts.xml
|--m64012ee_211206_183753.zmw_metrics.json.gz
|--m64012ee_211206_183753.transferdone
```

If **5mC CpG Detection** is performed, the following additional files are output:

```
|-- m64012ee_211206_183753.5mc_report.json
|-- m64012ee_211206_183753.primrose.log
```

If **on-instrument demultiplexing** is performed, the following additional files are output. **Note**: The undemultiplexed `hifi_reads.bam` file is **not** transferred; it is partitioned into the file structure shown below:

```
|-- bc1001--bc1001/m64012e_211206_183753.bc1001--bc1001.consensusreadset.xml
|-- bc1001--bc1001/m64012e_211206_183753.hifi_reads.bc1001--bc1001.bam
|-- bc1001--bc1001/m64012e_211206_183753.hifi_reads.bc1001--bc1001.bam.pbi
|-- m64012e_211206_183753.barcodes.fasta
|-- m64012e_211206_183753.lima.log
|-- m64012e_211206_183753.lima_counts.txt
|-- m64012e_211206_183753.lima_guess.json
|-- m64012e_211206_183753.lima_guess.txt
|-- m64012e_211206_183753.lima_reports.txt
|-- m64012e_211206_183753.lima_summary.txt
|-- m64012e_211206_183753.unbarcoded.consensusreadset.xml
|-- m64012e_211206_183753.unbarcoded.hifi_reads.bam
|-- m64012e_211206_183753.unbarcoded.hifi_reads.bam.pbi
```

In these examples:

- `r64012ee_211206_183753` is a directory containing the output files associated with **one** run.
- `r64012ee` is the instrument ID number.

The run directory includes a subdirectory for each collection/cell associated with a sample well - in this case `1_A01`. The collection/cell subdirectory can include the following output files:

- `ccs.log`: Log file from the CCS analysis. Informative for debugging and performance tracking by PacBio.
- `ccs_reports.json, ccs_reports.txt`: Contains processing metrics summarizing how many ZMWs generated HiFi reads, and how many ZMWs failed to generate CCS reads. These files contain the same information, and are used internally by PacBio Technical Support.
- `hifi.reads.bam`: Contains the HiFi reads in BAM format.
  Note: If low-quality reads are included in the Run Design, the Sequel IIe system will output a `reads.bam` file, which contains **HiFi reads and non-HiFi** reads:
  - HiFi reads (QV 20 or higher)
  - Lower-quality but still polished consensus reads (QV 1 - QV 20)
  - Unpolished consensus reads (RQ=-1)
  - 0- or 1-pass subreads unaltered (RQ=-1)

  The `reads.bam` file should **not** be used by itself as input for non-SMRT Link tools that expect ≥QV 20. The BAM format is a binary, compressed, record-oriented container format for raw or aligned sequence reads. The associated SAM format is a text representation of the same data. The BAM specifications are maintained by the SAM/BAM Format Specification Working Group. BAM files produced by **all** Sequel II systems are **fully compatible** with the BAM specification. For more information on the BAM file format specifications, click here.
- `hifi.reads.bam.pbi`: Index file that allows for random access of HiFi reads in the BAM file.
- `sts.xml`: Contains summary statistics about the collection/cell and its post-processing.
- `zmw_metrics.json.gz`: Contains processing information used to generate RunQC plots.
- `5mc_report.json, primrose.log`: Contains information about 5mC CpG Detection analysis (using the `primrose` tool), if performed.
- `<Barcode Name>.consensusreadset.xml`: Contains reads associated with a specific barcode.
- `<Barcode Name>.bam`: Contains HiFi reads associated with a specific barcode, in .bam format.
- `<Barcode Name>.bam.pbi`: Index file that allows for random access of HiFi reads in the BAM file.
- `<Barcode Name>.fasta`: Contains reads associated with the specific barcode, in FASTA format.
- `lima.log`: Log file from the demultiplexing analysis, if performed. Informative for debugging and performance tracking by PacBio.
- `lima.counts.txt`: Contains the counts of each observed barcode pair. Only passing ZMWs are counted.

- `lima.guess.json`, `lima.guess.txt`: Describes the barcode subsetting process activated using the `--peek` and `--guess` options. These files contain the same information, and are used internally by PacBio Technical Support.
- `lima.reports.txt`: A tab-separated file describing each ZMW, unfiltered. This is useful information for investigating the demultiplexing process and the underlying data. A single row contains **all** reads from a single ZMW.
- `lima.summary.txt`: Lists how many ZMWs were filtered, how many ZMWs are the same or different, and how many reads were filtered.
- `unbarcoded.consensusreadset.xml`, `unbarcoded.hifi_reads.bam` `unbarcoded.hifi_reads.bam.pbi`: Contains information on HiFi reads **not** associated with any barcode.

Note: The Sequel IIe system runs CCS on-instrument by default and the `subreads.bam`, `subreads.bam.pbi`, `scraps.bam` and `scraps.bam.pbi` files are no longer generated and are **not** available. Even though the `subreads.bam` and `subreads.bam.pbi` files are **not** accessible by default, there is a mechanism available to enable their output. For detailed instructions on how to enable the output of these files, contact your Field Applications Support team members.

HiFi reads generation

A standard Run Design performs on-instrument CCS analysis **without** including low-quality reads and generates a `hifi_reads.bam` file and transfers it to the network server. If low-quality reads are included in the Run Design, the Sequel IIe system will produce a `reads.bam file`, which contains HiFi reads and non-HiFi reads, and should **not** be used unfiltered as input for tools that expect ≥QV 20. SMRT Link **automatically** launches an **Export Reads** job on the `reads.bam` to filter out the HiFi reads, and generates the following HiFi data files by default:

- `<Movie_Name>.hifi_reads.fastq.gz` - Gzipped FASTQ file containing HiFi reads.
- `<Movie_Name>.hifi_reads.fasta.gz` - Gzipped FASTA file containing HiFi reads.
- `<Movie_Name>.hifi_reads.bam` - BAM file containing HiFi reads.

If **not** using SMRT Link for subsequent analysis, use these three files as input with any third-party analysis tools.

Finding the hifi_reads files generated using On-Instrument CCS

1. In Run QC, click the desired run, then click the sample name to view the CCS Data Set.
2. Click **Analyses** in the left-side panel.
3. Click the **Export Reads** analysis.

4. To locate the directory containing the three `hifi_reads` files, append `/outputs` to the path shown.

Sequel II system output files

Following is a sample of the file and directory structure output by the **Sequel II** system:

```
<your_specified_output_directory>/r64008_20160116_003347/1_A01
|-- m64008_160116_003634.baz2bam_1.log
|-- m64008_160116_003634.scraps.bam
|-- m64008_160116_003634.scraps.bam.pbi
|-- m64008_160116_003634.subreads.bam
|-- m64008_160116_003634.subreads.bam.pbi
|-- m64008_160116_003634.subreadset.xml
|-- m64008_160116_003634.sts.xml
|-- m64008_160116_003634.transferdone
```

Files output by the Sequel II system include:

- `scraps.bam` and `scraps.bam.pbi`: These files contain sequence data outside of the high-quality region, rejected subreads, excised adapter and possible barcode sequences, as well as spike-in control sequences. (The basecaller marks regions of single molecule sequence activity as high-quality.) **Note**: This applies to files generated by Sequel Instrument Control Software (ICS) v3.1.0 or later.

- `subreads.bam`: The Sequel II system output **one** `subreads.bam` file per collection/cell, which contains unaligned base calls from high-quality regions. This file is transferred from the instrument to network storage, then is used as **input** for secondary analysis by PacBio's SMRT Analysis software. Data in a `subreads.bam` file is analysis-ready; all of the data present should be quality-filtered for analyses. Subreads that contain information such as double-adapter inserts or single-molecule artifacts are **not** used in secondary analysis, and are excluded from this file and placed in `scraps.bam`.

- `subreads.bam.pbi`: Provides backwards-compatibility with the APIs enabled for accessing the `cmp.h5` file.
- `subreadset.xml`: This file is needed to import data into SMRT Link.
- `sts.xml`: Contains summary statistics about the collection/cell and its post-processing.
- `transferdone`: Contains a list of files successfully transferred.

Frequently asked questions
What are the minimum files needed to analyze data on SMRT Link? • `.bam` file • `bam.pbi` file • `subreadset.xml` or `consensusreadset.xml` file
What is the average size of the file bundle for a 30-hour movie of HiFi reads? Approximately 50 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 Sequel II systems, while the `aligned.bam` file is created by SMRT Link after running mapping analysis applications.

Secondary analysis output files

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

- All files for a specific job reside in **one** directory named according to the job ID number.
- Every job 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 job.
  - `workflow.<UUID>.log`: Global log of each significant step in the job and snippets from a task's `stderr` output if the job 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 job settings and workflow diagrams.
  - – `datastore.json`: JSON file representing all output files imported by SMRT Link.
- `outputs/`: A directory containing symbolic links to all datastore files, which reside in the Cromwell execution directory. This is provided as a convenience and is **not** intended as a stable API; note that external resources from dataset XML and report JSON file are **not** included here. Demultiplexing outputs are nested in additional subdirectories.
- `pbscala-job.stderr`: Log collection of `stderr` output from the SMRT Link job manager.
- `pbscala-job.stdout`: Log collection of `stdout` output from the SMRT Link job manager. (**Note**: This is the file displayed as **Data > SMRT Link Log** on the analysis results page.)

A SMRT Link job generates several types of output files. You can use these data files as input for further processing, pass on to collaborators, or upload to public genome sites. Depending on the analysis application being used, the `output` directory contain files in the following formats:

- **BAM**: Binary version of the Sequence Alignment Map (SAM) format. (See here for details.)
- **BAI**: The `samtools` index file for a file generated in the BAM format.
- **BED**: Format that defines the data lines displayed in an annotation track. (See here for details.)
- **CSV**: Comma-Separated Values file. Can be viewed using Microsoft Excel or a text editor.
- **FASTA/FASTQ**: Sequence files that contains either nucleic acid sequence (such as DNA) or protein sequence information. FASTA/Q files store multiple sequences in a single file. FASTQ files also include per-base quality scores. (See here or here for details.)
- **GFF**: General Feature Format, used for describing genes and other features associated with DNA, RNA and protein sequences. (See here for details.)
- **PBI**: PacBio index file. (This is a PacBio-specific file type.)
- **VCF**: Variant Call Format, for use with the molecular visualization and analysis program VMD. (See here for details.)

To download data files created by SMRT Link:

1. On the home page, select **SMRT Analysis**. You see a list of **all** jobs.
2. Click the job 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. You can add new users **after** SMRT Link is integrated and configured to work with LDAP; you can also add new users using WSO2 API Manager or Keycloak **without** LDAP integration.

- For details on integrating LDAP and SMRT Link, see the document **SMRT Link software installation guide (v11.1)**.

SSL

SMRT Link requires 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 (WSO2 API Manager or Keycloak), 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: PacBio 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 guide (v11.1)**.

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

Adding and deleting SMRT Link users

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

Assigning user roles

SMRT Link supports three user roles: **Admin**, **Lab Tech**, and **Bioinformatician**. Roles define which SMRT Link modules a user can access. The following table lists the privileges associated with the three user roles:

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

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

Hardware/software requirements

Client hardware requirements

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

Client software requirements

- SMRT Link **requires** the Google® Chrome web browser, version 90 or later.

Note: SMRT Link **server** hardware and software requirement are listed in the document **SMRT Link software installation guide (v11.1)**.

Appendix A - PacBio terminology

General terminology

- **SMRT$^{®}$ Cell**: Consumable substrates comprising arrays of zero-mode waveguide nanostructures. SMRT Cells are used in conjunction with the DNA sequencing kit for on-instrument DNA sequencing.
- **SMRTbell$^{®}$ template**: A double-stranded DNA template capped by hairpin adapters (i.e., SMRTbell adapters) at both ends. A SMRTbell template is topologically circular and structurally linear, and is the library format created by the DNA template prep kit.
- **collection**: The set of data collected during real-time observation of the SMRT Cell; including spectral information and temporal information used to determine a read.
- **Zero-mode waveguide (ZMW)**: A nanophotonic device for confining light to a small observation volume. This can be, for example, a small hole in a conductive layer whose diameter is too small to permit the propagation of light in the wavelength range used for detection. Physically part of a SMRT Cell.
- **Run Design**: Specifies
  - The samples, reagents, and SMRT Cells to include in the sequencing run.
  - The run parameters such as movie time and loading to use for the sample.
- **adaptive loading**: Uses active monitoring of the ZMW loading process to predict a favorable loading end point.
- **unique molecular yield**: The sum total length of unique single molecules that were sequenced. It is calculated as the sum of per-ZMW median subread lengths.
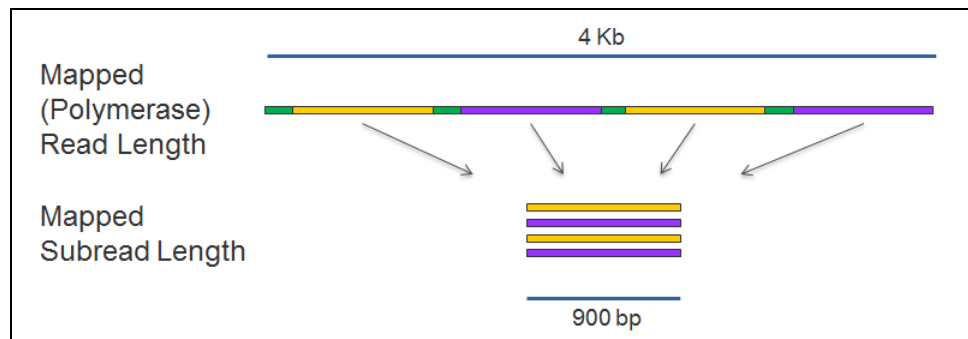
Read terminology

- **polymerase read**: A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell template. They can include sequences from adapters and from one or multiple passes around a circular template, which includes the insert of interest. Polymerase reads are most useful for quality control of the instrument run. Polymerase read metrics primarily reflect movie length and other run parameters rather than insert size distribution. Polymerase reads are trimmed to include only the high-quality region. **Note**: Sample quality is a major factor in polymerase read metrics.
- **subreads**: Each polymerase read is partitioned to form one or more subreads, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template and no adapter sequences. The subreads contain the full set of quality values and kinetic measurements. Subreads are useful for applications such as *de novo* assembly, base modification analysis, and so on.
- **longest subread length**: The mean of the maximum subread length per ZMW.

- **insert length**: The length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters.
- **circular consensus (CCS) reads**: The consensus sequence resulting from alignment between subreads taken from a single ZMW. Generating CCS reads does **not** include or require alignment against a reference sequence but **does** require at least two full-pass subreads from the insert. CCS reads are generated with CCS analysis. CCS reads with quality value equal to or greater than 20 are called **HiFi reads**.
- **HiFi reads:** Reads generated with CCS analysis whose quality value is **equal to or greater than 20**.
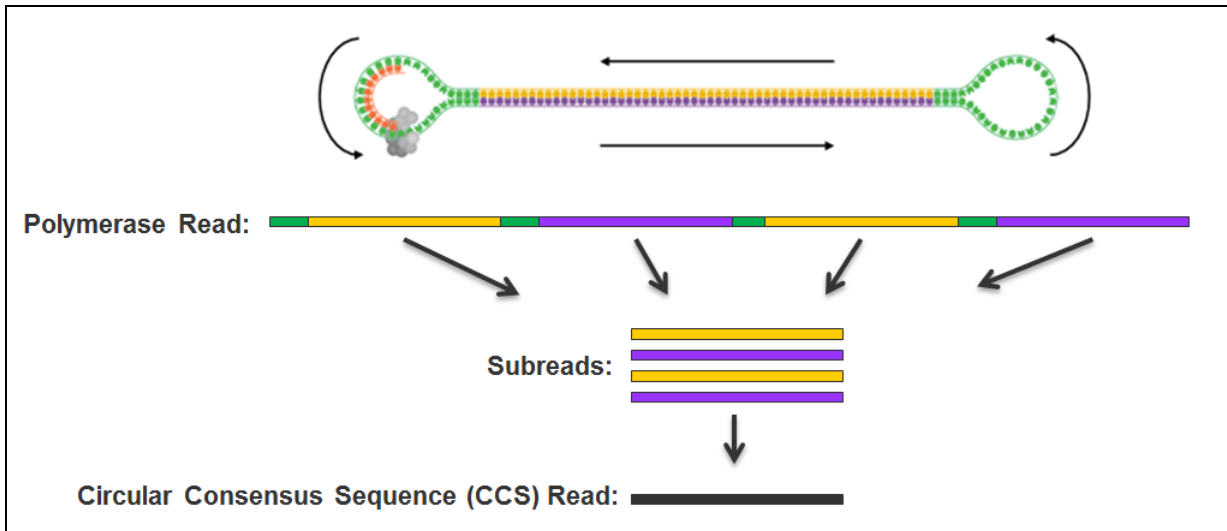
Read length terminology

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

Secondary analysis terminology

- **secondary analysis**: Follows primary analysis and uses basecalled data. It is application-specific, and may include:

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

- **secondary analysis application**: A secondary analysis workflow that may include multiple analysis steps. Examples include *de novo* assembly, RNA and epigenomics analysis.
- **consensus**: Generation of a consensus sequence from multiple-sequence alignment.
- **filtering**: Removes reads that do not meet the Read Length criteria set by the user.
- **mapping**: Local alignment of a read or subread to a reference sequence.
- **Auto Analysis**: Allows a specific analysis to be automatically run after a sequencing run has finished and the data is transferred to the SMRT Link server. The analysis can include demultiplexed outputs.
  - Auto Analysis works with **all** Sequel II systems and Sequel IIe systems.
- **Pre Analysis**: The process of CCS analysis and/or demultiplexing of Sequel basecalled data. Pre Analysis occurs **before** Auto Analysis.
  - Pre Analysis works with **all** Sequel II systems and Sequel IIe systems.

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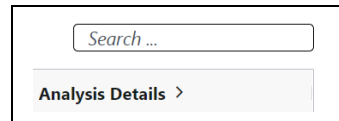
- **circular consensus accuracy**: Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.
- **consensus accuracy**: Accuracy based on aligning multiple sequencing reads or subreads together.
- **polymerase read quality**: A trained prediction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

Appendix B - Data search

Use this function to search for jobs, Data Sets, barcode files, or reference files.
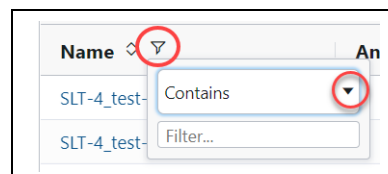
To search the entire table

1. Enter a text query into the Search box. This searches **every field** in the table, and displays **all** table rows containing the search characters.
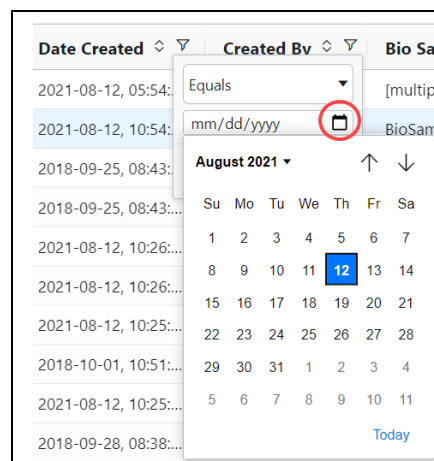

To search for a value within a column

1. Click the small filter icon at the right of the column name.
2. Enter a value; **all** table rows meeting the search criteria display. (To select a **different** search operator, click the droplist and select another search operator. Different search operators are available, based on the column's data type.)

- For the **Analysis State** column only, click one or more of the job states of interest: **Select All**, **Created**, **Running**, **Submitted**, **Terminated**, **Successful**, **Failed**, or **Aborted**.

- For **Date fields** only, click the small calendar and select a date.

Numeric field operators

- Equals, Not equal
- Greater than, Greater than or equals

- Less than, Less than or equals
- In range

Text field operators
- Contains, Not contains
- Equals, Not equal
- Starts with, Ends with

Date field operators
- Equals, Not equal
- Greater than, Less than
- In range

Appendix C - BED file format for Target Regions report

With the HiFi Mapping application, an optional **Target Regions** report can be generated that displays the number (and percentage) of reads and subreads that hit specified target regions.

The BED file required to generate the Target Regions report includes the following fields; with one entry per line:

1. `chrom`: The name of the chromosome (such as `chr3`, `chrY`, `chr2_random`) or scaffold (such as `scaffold10671`).
2. `chromStart`: The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered `0`.
3. `chromEnd`: The ending position of the feature in the chromosome or scaffold. The `chromEnd` base is **not** included in the display of the feature, however, the number in position format is represented. For example, the first 100 bases of chromosome 1 are defined as `chrom=1`, `chromStart=0`, `chromEnd=100`, and span the bases numbered 0-99 (**not** 0-100), but will represent the position notation `chr1:1-100`.
4. **(Optional)** Region Name.

Example: `lambda_NEB3011 15000 25000 Region2`

- Fields can be space- or tab-delimited.
- See here for details of the BED format.
- For details on the BED format's counting system, see here and here.

Appendix D - Additional information included in the CCS Data Set Export report

When you export a Data Set and select **Export PDF Reports**, a report is produced which includes additional fields, listed below.

- – See "Exporting sequence, reference and barcode data" on page 42 for details on exporting Data Sets.
- – The other fields and plots in this report are described in the appropriate Reports sections of "PacBio® secondary analysis applications" on page 54 and "PacBio® data utilities" on page 99.
- **ZMWs input:** The total number of ZMWs used as input in the Data Set.
- **ZMWs pass filters**: The number of ZMWs that passed **all** the filters.
- **ZMWs fail filters**: The number of ZMWs that failed **any** of the filters.
- **ZMWs shortcut filters**: The number of low-pass ZMWs skipped using the `--all` filter.
- **ZMWs with tandem repeats**: The number of ZMWs that did not generate CCS reads due to repeats larger than `--min-tandem-repeat-length`.
- **Below SNR threshold**: The number of ZMWs that did not generate CCS reads due to SNR below `--min-snr`.
- **Median length filter:** The number of ZMWs that did not generate CCS reads due to subreads that are <50% or >200% of the median subread length.
- **Lacking full passes**: The number of ZMWs that did not generate CCS reads due to having fewer than `--min-passes` full-length subreads.
- **Heteroduplex insertions**: The number of ZMWs that did not generate CCS reads due to single-strand artifacts.
- **Coverage drops**: The number of ZMWs that did not generate CCS reads due to coverage drops that would lead to unreliable polishing results.
- **Insufficient draft cov**: The number of ZMWs that did not generate CCS reads due to not having enough subreads aligned to the draft sequence end-to-end.
- **Draft too different**: The number of ZMWs that did not generate CCS reads due to having fewer than `--min-passes` full-length reads aligned to the draft sequence.
- **Draft generation error:** The number of ZMWs that did not generate CCS reads due to subreads that don't agree enough to generate a draft sequence.
- **Draft above --max-length**: The number of ZMWs that did not generate CCS reads due to a draft sequence longer than `--max-length`.
- **Draft below --min-length**: The number of ZMWs that did not generate CCS reads due to a draft sequence shorter than `--min-length`.
- **Reads failed polishing**: The number of ZMWs that did not generate CCS reads due to too many subreads dropped while polishing.

- **Empty coverage windows**: The number of ZMWs that did not generate CCS reads because at least one window had no coverage.
- **CCS did not converge**: The number of ZMWs that did not generate CCS reads because the draft sequence had too many errors that could not be polished in time.
- **CCS below minimum RQ**: The number of ZMWs that did not generate CCS reads because the predicted accuracy is below `--min-rq`.
- **Unknown error:** The number of ZMWs that did not generate CCS reads due to rare implementation errors.