



# SMRT<sup>®</sup> Link user guide (v25.2)

Revio<sup>®</sup> and Vega<sup>™</sup> systems



Research use only. Not for use in diagnostic procedures.

PN 103-651-300 Version 01 (March 2025)

© 2025 Pacific Biosciences of California, Inc. ("PacBio")

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

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

Certain notices, terms, conditions and/or use restrictions may pertain to your use of PacBio products and/or third party products. Refer to the applicable PacBio terms and conditions of sale and to the applicable license terms at <https://www.pacb.com/legal-and-trademarks/product-license-and-use-restrictions/>.

Trademarks:

Pacific Biosciences, the PacBio logo, PacBio, Circulomics, Omniome, SMRT, SMRTbell, Iso-Seq, Sequel, Nanobind, SBB, Revio, Onso, Apton, Kinnex, PureTarget, SPRQ, and Vega are trademarks of PacBio.

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

1305 O'Brien Drive Menlo Park, CA 94025 [www.pacb.com](http://www.pacb.com)

Introduction .....	5
Contact information .....	6
Using SMRT Link .....	6
Settings menu .....	7
Notifications menu.....	8
Working with tables.....	8
Sending information to Technical Support .....	8
<b>Instruments</b> .....	9
Viewing instrument information .....	9
<b>Sample Setup</b> .....	10
Creating a new calculation .....	10
<b>Runs</b> .....	14
Viewing information about runs.....	14
Run Design.....	19
Creating run designs for the Revio system .....	19
Creating run designs for the Vega™ system.....	24
Editing or deleting run designs.....	26
Duplicating run designs or creating run designs by importing a CSV file.....	27
<b>Data Management</b> .....	37
Understanding Data Sets.....	37
Understanding <b>Projects</b> .....	44
Importing and exporting data.....	46
Exporting sequence, reference, and barcode data .....	47
<b>SMRT® Analysis</b> .....	49
Creating and starting a job .....	49
Starting a job after viewing sequence data .....	53
Copying and running an existing job.....	55
Summarizing Microbial Genome Analysis jobs .....	56
SMRT Analysis applications.....	57
Target Enrichment.....	63
Iso-Seq® Analysis .....	67
Microbial Genome Analysis.....	75
Read Segmentation and Iso-Seq® <b>Analysis</b> .....	85
Read Segmentation and Single-Cell <b>Iso-Seq® Analysis</b> .....	94
Single-Cell Iso-Seq® Analysis .....	103
Demultiplex Barcodes .....	117
<b>Read Segmentation</b> .....	125
Trim Ultra-Low Adapters.....	127
Undo Demultiplexing.....	129
<b>Secondary analysis output files</b> .....	131
<b>Using the PacBio self-signed SSL certificate</b> .....	134
<b>User management and configuration</b> .....	135
Adding and deleting SMRT Link users.....	136
Assigning SMRT Link user roles .....	136
<b>Hardware/software requirements</b> .....	137
Client hardware requirements.....	137

Client software requirements.....	137
<b>Appendix A - PacBio terminology .....</b>	<b>138</b>
General terminology.....	138
Read terminology .....	138
Secondary analysis terminology .....	139
Accuracy terminology .....	139
<b>Appendix B - Data search.....</b>	<b>140</b>
To search the entire table.....	140
To search for a value within a column.....	140
Numeric field operators.....	140
Text field operators.....	141
Date field operators .....	141
<b>Appendix C – Connecting Revio or Vega systems .....</b>	<b>142</b>
Adding a system to SMRT .....	142
Modifying an existing Revio or Vega system .....	143
<b>Appendix D – Specifying a file transfer location.....</b>	<b>144</b>
<b>Appendix E - BED file format Target Enrichment and HiFi Mapping applications</b> .....	<b>148</b>
<b>Appendix F - Additional information included in the CCS Data Set Export report</b> .....	<b>149</b>

## Introduction

This document describes how to use SMRT Link software from PacBio®. SMRT Link is the web-based end-to-end workflow manager for Revio and Vega systems. SMRT Link includes the following modules:

- **Instruments:** View information about instruments connected to SMRT Link. (See [“Instruments” on page 8](#) for details.)
- **Sample Setup:** Calculate loading concentrations for sequencing libraries on **all** Revio and Vega systems. (See [“Sample Setup” on page 9](#) for details.)
- **Runs:** View information about sequencing runs, monitor run progress, status and quality metrics, design sequencing runs and create and/or import run designs. (See [“Runs” on page 19](#) for details.)
- **Data Management:** Create Projects and Data Sets; generate QC reports for Data Sets; view, import, or delete sequence, reference, barcode, and BED files. (See [“Data Management” on page 44](#) for details.)
- **SMRT Analysis:** Perform secondary analysis on the basecalled data (such as sequence alignment, variant detection, and RNA analysis) after a run has completed. (See [“SMRT® Analysis” on page 54](#) for details.) **Note:** The SMRT Analysis module is **not** included when you install **SMRT Link Lite**.

Supported instruments:

- Revio systems with v13.3 instrument software.
- Vega systems with v1.1 instrument software.
- Sequel, Sequel II, and Sequel IIe systems are not supported and require a separate local installation of SMRT Link v13.1.

This document also describes:

- The data files generated by secondary analysis. (See [“Secondary analysis output files” on page 158](#) for details.)
- Configuration and user management. (See [“Configuration and user management” on page 163](#) for details.)
- SMRT Link client hardware/software requirements. (See [“Hardware/software requirements” on page 165](#) for details.)

Installation of SMRT Link **server** software is discussed in the document SMRT Link software installation guide (v25.2).

New features, fixed issues and known issues are listed in the SMRT Link release notes (v25.2).

When you first start SMRT Link, in Settings > Instrument Settings > Supported Instruments you must specify which system you are using: **Revio and/or Vega**. This choice affects some of the initial values used in the Sample Setup and Runs modules. In those modules, you can switch between the 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 161](#).

After accessing SMRT Link, the **Instruments module** displays.

- Click the **PacBio logo** at the top left to navigate back to the Instruments page from within the application.
  - Select a module name from the **Module** menu (next to the PacBio logo) to access that module.
  - Click the **Settings menu** to configure for the sequencing system, view version information, or perform administrative functions (**Admins only**).
  - Click **Help** to directed to PacBio SMRT Link documentation.
  - Select **User Name > Sign Out** to log out of SMRT Link.
- 
- **Instruments**: Displays the Instruments module.
  - **Sample Setup**: Displays the Sample Setup module.
  - **Runs**: Displays the Runs module.
  - **Data Management**: Displays the Data Management module.
  - **SMRT Analysis**: Displays the SMRT Analysis module. (**Note**: SMRT Link Lite does **not** include the SMRT Analysis module.)

---

## Settings menu

- General
  - **Admin users only:** Use the **Time Zone** control to specify the time zone to use with **all** instruments connected to this instance of SMRT Link.
  - To specify how numbers are formatted, click **Number Formatting** and select **Period** or **Comma** as the decimal separator.
- Instrument settings
  - Select the instrument type(s) to support with this instance of SMRT Link
  - **Admin users only.** Set up a file transfer location and connect an instrument. See [“Appendix C - Adding and modifying Revio or Vega systems” on page 171](#) for more details.
- Notification settings
  - Specify the number of days when to automatically archive notifications.
  - Specify the types(s) of notifications to include in the red notification count on the main page: **Informational**, **Warning**, **Error**, and **Critical**. (**Note:** All notification types still display in the **Notification Center** dialog; only the checked types are reflected in the red notification count.)
- User management
  - **Admin users only:** Add/delete SMRT Link users and specify their roles. See [“Adding and deleting SMRT Link users” on page 164](#) for details.
- Updates (
  - To update **Human Genome hg38 with Gencode v39 annotations**, or **Mouse Genome mm39 with Gencode vM28 annotations**, click the appropriate button.
  - View version numbers for **all** software components of SMRT Link.
- About
  - SMRT Link Status displays status information about SMRT Link Server directories
  - Space Available on Server is provided for the Analysis Directory and Temporary Directory

- 
- Analysis Usage Information: Click **Send** to send analysis usage information. Data does not include sample names or sequence data
  - Troubleshooting Information: Click **Send** to send troubleshooting information to PacBio Technical Support for help in troubleshooting failed jobs.
  - Additional Information: Displays information about this instance of SMRT Link.
  - Usage Log: Click **Download** to obtain a log file listing user actions that modified the SMRT Link server, such as adding/editing/deleting records. (This log file does **not** include any view-only operations.)
- **Sign Out:** Sign out and display the initial login page.

#### Notifications menu

- Displays SMRT Link system-level notifications. Click a notification to see additional information.
- To clear **all** notifications, click **Mark All As Read**. You can also select individual notifications and mark them as **Read** or **Unread**.
- To **save** the notification log as a text file, click **View Notification Log**.

#### 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 169.](#))

#### 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: **Settings > About > Troubleshooting Information > Send**.
- From a SMRT Link “Failed” analysis Results page: Click **Send Log Files**.



---

## Instruments

Use the SMRT Link **Instruments** module to **view** information about sequencing systems connected to SMRT Link, as well as any ongoing runs.

### Viewing instrument information

1. When you first login to SMRT Link, you are directed to the Instruments module automatically.
2. (**Optional**) If three or more instruments are connected to SMRT Link, click **Expand All** to view information about all instruments. Click **Collapse All** for a more compact display.
3. For **each** instrument connected to the instance of SMRT Link, the Instruments page displays:
  - The instrument name, as defined by the user.
  - The instrument type (Revio or Vega).
  - The current instrument current status (**Starting, WarmUp, SelfTest, Ready, Running, ShuttingDown, Problem.**) A **red alarm bell** symbol displays next to the instrument status if any system-critical errors appear during a sequencing run. Move the mouse over the alarm bell to see a brief status message; check the Notifications center for more information.
  - The time until preload available; that is, the time when you can access the work deck (Revio) or drawer (Vega) and queue up the next run while the current run is sequencing.
  - The individual run name. (Click the run name link to see information about the specific run in the Runs module.)
  - SMRT<sup>®</sup> Cell status (**Pending, Loading, Sequencing, Complete.**) This also displays the number of SMRT Cells in each status category, as well as the total number of SMRT Cells used in the run.
    - Click a small black SMRT Cell icon to display a plot of how many ZMWs on that SMRT Cell are actively sequencing during a movie collection.
    - **Note:** A red box with an **exclamation point** indicates that the cell **failed**. A **crossed out** red box indicates that the run was **stopped** by the user.
  - **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.

---

## Sample Setup

To prepare your samples for sequencing, use the SMRT Link **Sample Setup** module to calculate the final loading dilution for prepared polymerase-bound SMRTbell libraries. Sample Setup for annealing, polymerase binding, and cleanup (ABC) can only be used for Revio non-SPRQ/Vega polymerase kits. You can print the instructions for use in the lab.

### Creating a new calculation

1. Select **Sample Setup** from the Module menu.
2. Click **+Add Calculation** and select one of two choices from the drop-down menu. (**After** you select a choice from the drop-down menu, it becomes the **default** value for adding new calculations. You can also export the calculated values to a CSV file for laboratory automation.)

Annealing, binding, cleanup (ABC) calculator

✓ Loading calculator

- **Annealing, binding, and cleanup (ABC) calculator:** This setting should be selected for ABC can only be used for Revio non-SPRQ/Vega polymerase kits. Process multiple samples with similar library properties (such as mean insert size and DNA concentration) in parallel. (This was called **Revio polymerase kit** mode in the previous releases). This setting calculates the amount of sample required to load a specified number of cells; it outputs protocols for primer annealing, polymerase binding and cleanup. The entire volume of the prepared sample is loaded.
- **Loading calculator:** This setting should be used with the **Revio polymerase kit 96**, **Revio SPRQ polymerase kit**, and **Vega polymerase kit**. The annealing, binding and cleanup steps should have already been completed. The Loading calculator provides instructions for the final loading dilution for previously-prepared polymerase-bound SMRTbell libraries. Input the number of sample wells being prepared, followed by the input concentration, average insert size, and loading concentration of each sample. The tool will return instructions for making the final dilution for each sample well.

Using the Annealing, binding, and cleanup (ABC) calculator for the Revio polymerase kit

1. Enter the sample **name**.
2. Select a sequencing **application** for the sample. The following fields are **auto-populated** and display in green:
  - Library type
  - Polymerase kit

3. 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.
4. Enter the **number of SMRT® Cells** per sample.
5. Enter the available **volume per sample**, in µL. When preparing multiple samples, this should be the **minimum** volume available for any sample.
6. 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).
7. 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

8. 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.)
9. Specify the concentration on plate, in pM.
10. 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 µL, though in some cases, for example if sample availability is very limited, it may be appropriate to set a value below 1 µL. Some protocol steps include fixed values of 1 µL that will **not** be affected by this setting.
11. **(Optional)** Enter a comment, such as a batch identifier for a LIMS, or information about the sample.

- 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 **Remove** to delete the current calculation.
- Click **Lock** to lock the calculation. This is **required** before samples can be imported into the Runs module, and also sends a finalized version of the instructions to the server for use in Data Set reports. **Note:** You can lock calculations, but you **cannot** import sample calculations into Revio Run Designs.

After locking, no further changes can be made to a calculation. (Click **View** to see the locked Sample Setup 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 **Download CSV** to generate a CSV file. This exports the calculated values to a CSV file for lab automation.

12. To **print** the calculation(s) and instructions, click the **Print** button.

Using the Loading Calculator for Revio polymerase kit 96, Revio SPRQ polymerase kit, and Vega polymerase kit.

Use this setting to calculate the final sample dilution for the Revio sequencing plate when you are starting with polymerase-bound SMRTbell libraries prepared using liquid handler automation or previously prepared manually. The setting produces instructions for making the final dilution for each sample well.

1. Specify the **number** (1-4) of sample wells to use per Revio sequencing plate. **Note:** If you are using only **one** Revio sequencing plate, specify 0 for Plate 2.
2. For **each** well ID, enter the sample name, the concentration (ng/ul), the average insert size (in base pairs), the loading concentration (in pico- molar), and any comments. Use the tab key to move between fields. **Note:** If using a partially-used Revio sequencing plate, change a Well ID by clicking on it and using the drop-down menu. For example, wells A and B are used and you want to start with C01.
3. Repeat Step 2 for any additional sample wells.  
**Note:** All sample wells must be filled in for the instructions to display.
4. To **print** the calculation(s) and instructions, click the **Print** button.

If using the **Revio polymerase kit 96** setting and pooling multiple biological samples together in one well, click **Pooling Calculator**:

Sample Setup / Loading calculator

[Pooling Calculator](#) [Print](#)

The Loading calculator provides instructions for the final loading dilution for previously-prepared polymerase-bound SMRTbell libraries. Input the number of sample wells being prepared, followed by the input concentration, average insert size, and loading concentration of each sample. The tool will return instructions for making the final dilution for each sample well.

Sequencing plates ⓘ

Polymerase kit

Plate 1 wells

Plate 2 wells

1. Enter the number of samples to be multiplexed together (between 2 and 384.)
2. Specify the pooled library target volume (in  $\mu\text{L}$ ).
3. Specify the unit to use for the output concentration.
4. Specify the pooled library concentration, using the unit specified.
5. For the each sample, specify the concentration and the sample volume. (Use the tab key to move between fields.)
6. Repeat Step 6 for the rest of the samples to be multiplexed.
7. To **print** the pooling calculation, click **Print**.
8. To **export** the pooling calculations as a CSV file, click **Export**.

---

## Custom ABC calculations

1. To accommodate new or unique sample types, choose **Application > Other > 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 > Other > Custom**, those settings are retrieved.

## Editing or printing calculations

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

## Deleting calculations

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

## Importing/exporting calculations

Sample Setup supports importing and exporting calculations in CSV format.

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

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

## Note:

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

## 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`.

---

## Runs

Use the SMRT Link **Runs** module to:

- View information about sequencing runs performed on sequencing systems connected to the instance of SMRT Link. This includes completed runs, as well as runs still in progress.
- 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.

### Viewing information about runs

1. Select **Runs** from the Module menu.
2. 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.
3. To view some **basic** information about a run, click the **magnifying glass** next to the run name. This displays instrument and run details including run state and cell status.
4. Run information displayed in the table includes:
  - The name of the run.
  - The status of the run: **Ready, Started, Running, Stopped, Terminated** or **Complete**.
  - The name of the instrument and what type of instrument.
  - Who created the run, when it was created, when it was started, and when it was completed.
  - Any run comments.
  - The number of samples in the run.
  - The number of SMRT Cells used in the run.
  - The number of files successfully transferred to the network; one per SMRT Cell.
5. To **export** run information to a CSV file: Click the checkbox next to the run to export, then click **Export Selected**.
6. Click a run link.
  - If the run is **not completed**, this displays summary information about the run design used for the run.
  - If the run is **completed**, this displays information used to monitor performance trends and perform run QC remotely:

#### Table fields

- **Run Created:** The date and time when the run was created.
- **Run Start:** The date and time when the run was started.
- **Run Complete:** The date and time the run was completed.

- **Created By:** The name of the user who started the run.
  - **Instrument Name:** The name of the instrument.
  - **Completed Cells:** The number of successfully completed SMRT Cells.
  - **Failed Cells:** The number of SMRT Cells that did not successfully generate data.
  - **Time remaining for PostProcessing:** The time needed, after movie acquisition ends, to convert sequencing data to HiFi reads.
  - **Transfer Status:** Whether or not the data was successfully transferred from the instrument to the network. Possible values are: **Transferring, Failed, Complete**, or blank if the cell is still acquiring or has not started acquiring yet.
  - **Run ID:** An internally-generated ID number identifying the run. (This is different from a UUID, which identifies individual Data Sets.)
  - **Instrument SN:** The serial number of the instrument.
  - **Instrument software:** The versions of Instrument Control Software (ICS) installed on the instrument.
  - **Transfer Directory:** The path to run sequencing data determined by the configured transfer scheme, and the transfer subdirectory specified during run design.
7. Click the > arrow at the top of the **Consumables** table to see, for each sample well, the consumable type, lot number, expiration date, and other information.

Notes:

- Click **Expand All** to expand **all** of the table columns. Click **Collapse All** to collapse the table columns.
- To see additional table fields for a **specific column**, click the > symbol next to a column title.

Run settings and metrics

- **Well**
  - **Plate well:** The plate number and well ID of an individual well used for this sample.
  - **Well name:** The name of the individual well used for this sample. Allowed characters: alphanumeric, hyphen, underscore.
  - **Well comment:** User-specified comment for the individual well.
  - **Movie:** m<instrument-number\_date\_time>
- **Run**
  - **Status:** The current collection status for the SMRT Cell, which can be one of the following: **Complete, Collecting, Paused, Queued, Stopped, Failed, Running, or Pending**.
  - **Movie time:** The length of the movie, in hours, associated with this SMRT Cell.
  - **Loading concentration:** The on-plate loading concentration, in picomolarity.

- 
- **Workflow:** The instrument robotics workflow used for the run.
  - **Loading time:** The time the system took for loading to progress before proceeding to sequencing.
  - Productivity
    - **Total bases:** Calculated by multiplying the number of productive (P1) ZMWs by the mean polymerase read length; displayed in Gigabases.
    - **P0:** Empty ZMW; no signal detected. (Revio only)
    - **P1:** ZMW with a high quality read detected. (Revio only)
    - **P2:** Other, signal detected but no high quality read. (Revio only)
    - **Loading level:** Estimated percent loading on a Vega SMRT Cell (Vega only)
  - **HiFi reads:** CCS reads whose quality value is equal to or greater than 20.
    - **Reads:** The total number of CCS reads whose quality value is equal to or greater than 20.
    - **Yield:** The total yield (in base pairs) of the CCS reads whose quality value is equal to or greater than 20.
    - **Length (mean):** The mean read length of the CCS reads whose quality value is equal to or greater than 20.
    - **Read quality (median):** The median QV of the CCS reads whose quality value is equal to or greater than 20.
    - **Q30+ bases:** The percentage of bases in CCS reads whose quality value is equal to or greater than 30.
  - **Polymerase 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.
    - **Pol. 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.
    - **Pol. 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.
    - **Base rate:** The average base incorporation rate, excluding polymerase pausing events.
  - Control reads
    - **Reads:** The number of control reads obtained.
    - **Read length (mean):** The mean read length of the control reads.
    - **Concordance (mean):** The average concordance (agreement) between the control raw reads and the control reference sequence.
    - **Concordance (mode):** The median concordance (agreement) between the control raw reads and the control reference sequence.



- 
- File transfer
    - **Status:** Whether or not the data was successfully transferred from the instrument to the network. Possible values are: **Transferring, Failed, Complete**, or blank if the cell is still acquiring or has not started acquiring yet.
    - **Action:** If the Status is **Failed**, the **Retry File Transfer** button becomes available. Click the button to retry the file transfer. (If the file transfer doesn't work after several tries, contact PacBio Technical Support for help.)

## Run preview

Run previews are **estimates** of run performance at two different time points: 4 hours after sequencing acquisition begins and 1 hour before the end of acquisition. These estimates are based on a subsample of ZMWs. This information is **approximate** and intended to guide future runs by providing early information on loading, library fragment size, and representation of barcodes in the pool.

## Preview metrics

Click the > arrow at the top of the **Preview metrics** table to see estimated metrics for **all** SMRT Cells within a given run. **Note:** These preview metrics can also be seen by clicking on a SMRT Cell on the Instruments page.

## Basic preview (estimates)

Note: Estimate is displayed at 4 hours after acquisition begin

- **P1:** An estimate of the percentage of ZMWs rated as P1, meaning that a high-quality read was detected. **Note:** The P1% in the run preview table will be **greater** than the percent active ZMW's in the Sequencing ZMW's plot in the same table. This is because the Sequencing ZMW's plot reports percent active ZMWs at a given time point, while the run preview estimates the percent of ZMW's that will produce a high-quality read over the **entire** acquisition.
- **HiFi read length, mean:** An estimate of the mean length of the HiFi reads per SMRT Cell for the sequencing run. **Note:** This value is typically an underestimate as some of the longer molecules require more time to be fully sequenced.

## Full preview (estimates)

Note: Estimate is displayed 1 hour before end of acquisition

- **HiFi yield:** An estimate of HiFi yield generated per SMRT Cell for the sequencing run. **Note:** This value is based on a subsample of ZMWs and may be an overestimate or an underestimate.
- **HiFi read length, mean:** An estimate of the mean length of the HiFi reads per SMRT Cell for the sequencing run.
- **HiFi read quality, median:** An estimate of the median HiFi read quality per SMRT Cell for the sequencing run.

---

## Barcode counts

Click the > arrow at the top of the **Barcode Counts** table to see **estimated** metrics for **all** barcoded and unbarcoded reads included in the run. Select the desired SMRT Cell and time point under the **Well name** and **Time point** drop-down menus, respectively.

**Note:** The values displayed may **overestimate** the number of unbarcoded reads. In addition, all estimates may be **less accurate** for barcodes at low frequency (<10%) due to sample size. Any barcodes below a 1% frequency are **not** displayed, and are grouped into the “Other” category.

- **Barcode:** An individual barcode detected in the sample, as well as unbarcoded reads.
- **HiFi reads:** An estimate of the percent of reads with each barcode, as well as the percent of unbarcoded reads.
- **HiFi read length, mean:** An estimate of the average HiFi read length for each barcode or for unbarcoded reads.

## Plots

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  $\geq 20$ ), other CCS reads (three or more passes, but QV  $< 20$ ), and other reads, by read length. **Note: Other reads** means single-pass subreads and anything else that the software could not determine a consensus sequence for.
- **Read Quality Distribution:** Displays a histogram distribution of HiFi reads (QV  $\geq 20$ ) 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:** This plot displays a reverse cumulative distribution of all detected CpG motifs according to their predicted probability of methylation.

## Run Design

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.

After a run design is created, it becomes available from for selection on the instrument.

Run designs created in SMRT Link are accessible from **all** sequencing systems linked to the same SMRT Link server.

SMRT Link includes **two** different ways to create a run design:

- Use the SMRT Link **New Run Design** screen to create a new run design.
- Create a CSV file, then import it using the SMRT Link **Runs** module. See [“Creating run designs by importing a CSV file” on page 31](#) for details.

### Character requirements:

Several fields in Run Design have character requirements. This ensures a consistent user experience and compatibility with downstream SMRT Link and 3<sup>rd</sup> party tools and workflows. Please reference the Run Design CSV file structure table for each field’s requirements. Requirements apply when creating or editing via GUI or CSV import.

## Creating run designs for the Revio system



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

1. Select **Runs** from the Module menu and click **+ Create New Run**.
2. In Run Information, ensure that **Revio** is selected as the instrument type.
3. Enter a **Run Name**. Supported characters: alphanumeric, space, hyphen, underscore, colon, period, and apostrophe. Comma and newline are not permitted. By default, the software creates a new run name based on the current date and time; edit the name as needed.

- 
- Select a sequencing plate to specify **Plate 1** as the sequencing plate for position 1 on the Revio work deck associated with this new run design. You can locate the 6-digit lot number, 5-digit serial number (labeled SN XXXXX and located under the QR code) and 8-digit expiration date (YYYYMMDD) on the sequencing plate label.

**Note:** You can also scan the QR code on the sequencing plate label by using a laptop or webcam camera, then clicking the **Scan** button. This fills in the **Lot**, **Serial** and **Expiry** fields.

Plate 1 Required 

Revio SPRQ sequencing plate 		
<i>Lot</i>	<i>Serial</i>	<i>Expiry</i>

- 
5. **(Optional)** Specify **Plate 2** as the sequencing plate for position 2 on the Revio work deck associated with this run design.
  6. **(Optional)** Enter **Run Comments** as needed. Supported characters: alphanumeric, space, hyphen, underscore, colon, period, and apostrophe. Comma and newline are not permitted.
  7. **(Optional)** Enter a **Transfer Subdirectory**, which specifies a subdirectory within the transfer location. Run files are transferred to <TransferRoot>/<SubDirectory>/<RunDirectory> instead of <TransferRoot>/<RunDirectory>.
  8. Specify whether to use Adaptive Loading. Default is ON which is recommended. Adaptive Loading uses active monitoring of the ZMW loading process to predict a favorable loading end point. Note: Adaptive Loading applies to the entire run, not just individual SMRT Cells
  9. In **Sample Information**, select a sequencing **Application** from the list.
  10. Specify the plate well **position** to use for samples. Use wells on each plate sequentially and do **not** leave unused wells between samples.
  11. Specify the **Well Name** and any Well comments. Well Name allowed characters: alphanumeric, hyphen, underscore. Well Comment allowed characters: alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters.
  12. Specify the **Standard** Library type. **Note:** This can be set to **Kinnex** or **Adeno-associated Virus** based on the application you select in Step **Library Type** identifies the structure of the molecules to be sequenced, which determines how the instrument performs adapter calling and consensus read generation.
    - **Standard** specifies a single sequence for adapter calling. Standard libraries consist of a single DNA insert with the same SMRTbell adapter loop on each end of the molecule.
    - **Kinnex** specifies two sequences for adapter calling. Kinnex™ libraries consist of concatenated smaller inserts with different SMRTbell adapter loops on each end of the molecule. (For a video on using all Kinnex kits with the Runs module, click [here](#).)
    - **Adeno-associated Virus** disables adapter correction (that is, it disables splitting molecules with an adapter on only one end) and enables generating a consensus sequence per strand, considering only passes from that strand (by-strand mode). Adeno-associated Virus libraries include a variety of structures, which may have: 1) A SMRTbell adapter loop on only end or on both ends, and 2) either complementary or unique forward and reverse strand sequences.

- 
13. 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.
  14. Specify Library Concentration in picomoles.
  15. Movie Acquisition Time defaults to the recommendation for the Application selected.
  16. Under **Samples**, specify whether the sample is **indexed**.
    - If **No**, enter a **Bio Sample Name**. This is the name of the biological sample contained in the sequencing library, such as HG002. Go to Step 21. **Allowed characters**: alphanumeric, hyphen, underscore. **Character limit**: ≤40 characters
    - If **Yes**, go to the next step.
  17. Specify the **Indexes file**. (**SMRTbell Adapter Indexes** is the default).

**Note**: Demultiplexing on instrument is only supported for symmetric indexes. If Twist Universal Adapters with UDI is selected, a Demultiplexing Barcodes job will be scheduled to run in SMRT Analysis.
  18. Select **Biosample Names**, either interactively or by downloading a file:

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. **Allowed characters**: alphanumeric, hyphen, underscore. **Character limit**: ≤40 characters
- (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**.
  - Enter the Bio Sample Name associated with the barcodes in the second column, then save the file. **Allowed characters**: alphanumeric, hyphen, underscore. **Character limit**: ≤40 characters
- If you did **not** use all barcodes in the Autofilled Barcode Sample 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.

19. Under **Data Options**, specify, for this sample only, whether to include kinetics information (used for epigenetics analysis) in the HiFi reads BAM. Note: 5mC-CpG calls and 6mA calls for FiberSeq analysis are always provided in the HiFi reads BAM. Adding kinetics information can increase the amount of storage used by the output BAM files by up to 5 times.
20. Under **Data Options**, specify the **consensus mode**.
  - Strand separately generates a consensus sequence from the forward and reverse strands, and is useful for identifying heteroduplexes formed during

---

sample preparation.

– Molecule generates a single consensus sequence from both stands.

21. Under **Data Options**, specify if Data Sets generated by SMRT Cell(s) using this run design will be associated with a specific Project. (By default, all Data Sets are assigned to General Project, which is accessible to all users.)
22. Under **Analysis Options**, specify whether or not to add an analysis job that will run in SMRT Analysis using the generated sample data as input. If Yes, enter the analysis name and the select the analysis workflow to use with the data generated by this sample. Analysis name limited to alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters only. Comma and newline are not permitted. (For details on the a SMRT Analysis applications provided by PacBio and their parameters, see “SMRT Analysis applications” on page 63.).
23. Click Save at the top of the New Run Design page.


## Creating run designs for the Vega™ system

Select **Runs** from the Module menu.

1. Click **+ Create New Run**.
2. In Run Information, ensure that **Vega** is selected as the instrument type.
3. Enter a **Run Name**. By default, the software creates a new run name based on the current date and time; edit the name as needed.  
Supported characters: alphanumeric, space, hyphen, underscore, colon, period, and apostrophe. Comma and newline are not permitted.  
Select a Vega sequencing plate to specify **Plate 1** as the sequencing plate. You can locate the 6-digit lot number, 5-digit serial number (labeled SN XXXXX and located under the QR code) and 8-digit expiration date (YYYYMMDD) on the sequencing plate label.

**Note:** You can also scan the QR code on the sequencing plate label by using a laptop or webcam camera, then clicking the **Scan** button. This fills in the **Lot**, **Serial** and **Expiry** fields.

Plate 1 Required ⓘ

Vega sequencing plate ▾ 

Lot Serial Expiry

4. (Optional) Enter **Run Comments** as needed. Supported characters: alphanumeric, space, hyphen, underscore, colon, period, and apostrophe. Comma and newline are not permitted.
5. (Optional) Enter a **Transfer Subdirectory**, which specifies a subdirectory within the transfer location. Run files are transferred to <TransferRoot>/<SubDirectory>/<RunDirectory> instead of <TransferRoot>/<RunDirectory>
6. In **Sample Information**, select a sequencing **Application** from the list.
7. Specify the **Well Name** and any Well comments. Well Name allowed characters: alphanumeric, hyphen, underscore.
8. Specify the **Standard** Library type. **Note:** This can be set to **Kinnex** or **Adeno-associated Virus** based on the application you select in Step **Library Type** identifies the structure of the molecules to be sequenced, which determines how the instrument performs adapter calling and consensus read generation.
  - **Standard** specifies a single sequence for adapter calling. Standard libraries consist of a single DNA insert with the same SMRTbell adapter loop on each end of the molecule.
  - **Kinnex** specifies two sequences for adapter calling. Kinnex™ libraries consist of concatenated smaller inserts with different SMRTbell adapter loops on each end of the molecule. (For a video on using all Kinnex kits with the Runs module, click [here](#).)
  - **Adeno-associated Virus** disables adapter correction (that is, it disables splitting molecules with an adapter on only one end) and enables generating a consensus sequence per strand, considering only passes from that strand (by-strand mode). Adeno-associated Virus libraries include a variety of structures, which may have: 1) A SMRTbell adapter loop on only end or on both ends, and 2) either complementary or unique forward and reverse strand sequences.



- 
9. 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.
  10. Specify Library Concentration in picomoles.
  11. Movie Acquisition Time defaults to the recommendation for the Application selected.
  12. Under **Samples**, specify whether the sample is **indexed**.
    - If **No**, enter a **Bio Sample Name**. This is the name of the biological sample contained in the sequencing library, such as HG002. Go to Step 17. **Allowed characters**: alphanumeric, hyphen, underscore. **Character limit**: ≤40 characters.
    - If **Yes**, go to the next step.
  13. Specify the **Indexes file**. (**SMRTbell Adapter Indexes** is the default).

**Note**: Demultiplexing on instrument is only supported for symmetric indexes. If Twist Universal Adapters with UDI is selected, a Demultiplexing Barcodes job will be scheduled to run in SMRT Analysis.
  14. Select **Bio sample Names**, either interactively or by downloading a file:

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. **Allowed characters**: alphanumeric, hyphen, underscore. **Character limit**: ≤40 characters.
    - **(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**.
    - Enter the Bio Sample Name associated with the barcodes in the second column, then save the file. **Allowed characters**: alphanumeric, hyphen, underscore. **Character limit**: ≤40 characters.

If you did **not** use all barcodes in the Autofilled Barcode Sample 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.
  15. Under **Data Options**, specify, for this sample only, whether to include kinetics information (used for epigenetics analysis) in the HiFi reads BAM.

**Note**: 5mC-CpG calls and 6mA calls for Fiber-seq analysis are always provided in the HiFi reads BAM. Adding kinetics information can increase the amount of storage used by the output BAM files by up to 5 times.

- 
16. Under **Data Options**, specify the **consensus mode**.
    - Strand separately generates a consensus sequence from the forward and reverse strands, and is useful for identifying heteroduplexes formed during sample preparation.
    - Molecule generates a single consensus sequence from both stands.
  17. Under **Data Options**, specify if Data Sets generated by SMRT Cell(s) using this run design will be associated with a specific Project. (By default, all Data Sets are assigned to General Project, which is accessible to all users.)
  18. Under **Analysis Options**, specify whether or not to add an analysis job that will run in SMRT Analysis using the generated sample data as input. If Yes, enter the analysis name and the select the analysis workflow to use with the data generated by this sample. Analysis name limited to alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters only. Comma and newline are not permitted. (For details on the a SMRT Analysis applications provided by PacBio and their parameters, see “SMRT Analysis applications” on page 63.).
  19. Click Save at the top of the New Run Design.

### Editing or deleting run designs

1. Select **Runs** from the Module menu.
2. Click the name of the run design to edit or delete.
  - If the run status is **Ready**: Click **Edit**.
  - If the run status is **Completed**: Click **View Run Design**, then click **Edit**.
3. (**Optional**) Edit any of the fields.
4. (**Optional**) Click **Delete** to delete the current run design.
5. Click **Save**.
6. Select **Runs** from the Module menu.
7. Click the name of the run design to duplicate.
  - If the run status is **Ready**: Click **Duplicate**.
  - If the run status is **Completed**: Click **View Run Design**, then click **Duplicate**.
8. Edit the Run Name. (The default name is Copy of...).
9. Click **Save**.
10. On a remote workstation, open the sample CSV file included with the installation.

---

Duplicating run designs or creating run designs by importing a CSV file

To obtain the sample CSV files

1. Select **Runs** from the Module menu.
2. Click **Import Run**.
3. Click **Download Template**. The ZIP file containing run design templates and will download to your local computer.

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 fields in the tables below.)
2. Save the edited CSV file.
3. Import the file into SMRT Link.
4. Select **Runs** from the Module menu.
5. Click **Import Run**.
6. Select the saved CSV file designed for the run and click **Open**, then click **Done**. The file is now imported and available for selection on the instrument.
7. If Full Resolution Base Qual or Subread To HiFi Pileup is TRUE, the imported run design will display these options under Advanced in Data Options.

▼ Data Options

Include Base Kinetics ⓘ  YES  NO

Consensus Mode  MOLECULE  STRAND

Assign Data To Project ⓘ

Advanced ⓘ  Full Resolution Base Qual  Subread To HiFi Pileup

## Run Design CSV

### file structure

The Revio and Vega Run Design CSV file format is divided into three main sections:

1. [Run Settings] - Settings that apply to the entire run.
2. [SMRT Cell Settings] - Settings that apply to a specific collection, plate well, or SMRT Cell.
3. [Samples] - Settings that apply to a specific barcoded sample.

Each section begins with a line that starts with the name of the section surrounded by square brackets.

### Run Settings section

Each line in this section (after the [Run Settings] line) represents a **single** setting,

Setting name	Required	Description
Instrument Type	Yes	Must be Revio or Vega.
Run Name	Yes	A human-readable name used to refer to the run in SMRT Link interfaces. Must contain only alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters. Comma and newline are not permitted. <b>Example:</b> 20170530_A6_VVnC
Run Comments	No	A human-readable comment attached to run. Must contain only alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters. Comma and newline are not permitted. <b>Example:</b> My first Revio run1
Plate 1	Yes	A part number or plate identifier that specifies the sequencing plate in workdeck position 1 (See <a href="#">"Plate identifier requirements" on page 43.</a> )
Plate 2	No	A part number or plate identifier that specifies the sequencing plate in workdeck position 2. (See <a href="#">"Plate identifier requirements" on page 43.</a> ). Not applicable to Vega.
Transfer Subdirectory	No	Specifies a subdirectory within the transfer location. Run files are transferred to <TransferRoot>/<TransferSubdirectory>/<RunDirectory> instead of <TransferRoot>/<RunDirectory>. Enter alphanumeric characters, hyphens, underscores, or forward slash <b>only</b> .
CSV Version	Yes	Must be 1 for v25.2.
Compute Settings	No	Specifies compute cluster options for auto-analysis jobs. Enter values automatically generated by the SMRT Link installer in the form installer.user.config.computecfg_00, installer.user.config.computecfg_01 and so on. The numbering corresponds to the order in which they display in the <b>Advanced Analysis Parameters</b> dialog's <b>Compute Settings</b> drop-down menu when specifying a new job.

---

## SMRT Cell Settings section

This section is represented as a table. Each row in represents a **single** setting and includes the setting name. In the CSV, the setting name is followed by a comma separator, and then the setting value for the respective well sample as indicated in the section start line.

### Section start line - Well sample identifier

A well sample identifier is not required when using a Vega system because a run can only contain one sample. As a Revio run may contain between 1 and 8 SMRT Cells, the identifier of the well sample associated with each SMRT Cells used in the run **must** be listed on the first line of the section following [SMRT Cell Settings]. This forms the section table header.

Well sample identifiers are written in the format <plate number>\_<plate well name> where:

- <plate number> is either 1 or 2, representing plate 1 and plate 2, respectively.
- <plate well name> is one of the following: A01, B01, C01, D01.

For example, 1\_A01 is the identifier for the well sample that is loaded on plate 1 well A01.

Example section start line for a Revio run containing the maximum of 8 SMRT Cells:

```
[SMRT Cell Settings],1_A01,1_B01,1_C01,1_D01,2_A01,2_B01,2_C01,2_D01
```

Setting name	Required	Description
<b>Well Name</b>	Yes	Enter alphanumeric characters, hyphens, or underscores only. <b>Example:</b> A6_3230046_A01_SB_ChemKitv2_8rxnKit
<b>Well Comment</b>	No	Enter alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters only. Comma and newline are not permitted.
<b>Library Type</b>	Yes	Must be Standard, Kinnex, or Adeno-associated virus. <b>Library Type</b> identifies the structure of the molecules to be sequenced, which determines how the instrument performs adapter calling and consensus read generation.
<b>Insert Size (bp)</b>	Yes	Enter an integer greater than 500.
<b>Movie Acquisition Time (hours)</b>	Yes	Enter 12, 24, or 30 (Revio only). Time is in hours.
<b>Application</b>	No	<ul style="list-style-type: none"> <li>• Human WGS</li> <li>• Microbial assembly</li> <li>• Other WGS</li> <li>• Iso-Seq method</li> <li>• MAS-Seq single cell</li> <li>• Kinnex single-cell RNA</li> <li>• Kinnex full-length RNA</li> <li>• Adeno-associated virus</li> <li>• Kinnex 16S rRNA</li> <li>• Full-Length 16S rRNA sequencing</li> <li>• Shotgun metagenomic profiling or assembly</li> <li>• Hybrid Capture</li> <li>• &lt;3kb amplicons</li> <li>• &gt;=3kb amplicons</li> <li>• PureTarget repeat expansion</li> <li>• PureTarget custom</li> <li>• Other</li> </ul> <p>If blank or contains invalid values, default is Other.</p>
<b>Sample is indexed</b>	No	Enter TRUE or FALSE. Default = TRUE.
<b>Bio Sample Name</b>	No	Required <b>only</b> if Sample is indexed is FALSE for a given collection. Enter Bio Sample Names in the same row as their associated Barcode Names. Use alphanumeric characters, hyphens, or underscores only. Character limit: ≤40 characters. <b>Example:</b> sample1 <b>Note:</b> This field is used for collections for non-multiplexed data.

Setting name	Required	Description
Same Barcodes on Both Ends of Sequence	No	Enter TRUE if barcode indexes used for demultiplexing are symmetric, FALSE if asymmetric. Default = TRUE. Demultiplexing of asymmetric barcodes is not supported on-instrument.
Indexes	Yes	Must be a UUID for an Index Set present in the database. <b>Example:</b> dad4949d-f637-0979-b5d1-9777eff62008
Library Concentration (pM)	Yes	Enter a positive integer. Units are in parts per million.
Use Adaptive Loading	Yes	Revio only. Enter TRUE or FALSE. <b>Note:</b> All collections <b>must</b> use the same value.
Include Base Kinetics	Yes	Enter TRUE or FALSE. Enter TRUE to specify that CCS analysis output includes kinetics information (used for epigenetics analysis.) <b>Note:</b> Adding kinetics information can increase the amount of storage used by the output BAM files by up to <b>5 times</b> .
Consensus mode	Yes	Enter molecule to generate a consensus sequence per ZMW, considering passes from <b>both</b> strands. Enter strand to generate a consensus sequence per strand, considering only passes from that strand (--by-strand option).
Assign Data to Project	Yes	Enter an integer 1 or greater. Default is 1 for General Project. The number increases by 1 for each Project displayed in the <b>Data Management &gt; Projects</b> list.
Full Resolution Base Qual	No	Enter TRUE or FALSE. Default = FALSE. Enter TRUE to output base quality values (BAM QUAL column) without binning. If TRUE, Full Resolution Base Qual will be shown under Advanced Data Options in the Edit Run Design menu.
Subread To HiFi Pileup	No	Enter TRUE or FALSE. Default = FALSE. Enter TRUE to output three tags – sa, sm, and sx – that summarize the subread-to-HiFi consensus read alignment. If TRUE, Subread To HiFi Pileup will be shown under Advanced Data Options in the Edit Run Design menu. <ul style="list-style-type: none"> <li>sa: Number of subread alignments that span each position in the consensus read. Represented as a series of run-length encoded spans sa:B:I:&lt;LENGTH1&gt;,&lt;COVERAGE1&gt;,...,&lt;LENGTHN&gt;,&lt;COVERAGE N&gt;</li> <li>sm: Number of aligned subread bases that match the consensus base at each position in the consensus read. Output as one value per position: sm:B:C,&lt;MATCH1&gt;,&lt;MATCH2&gt;,...,&lt;MATCHN&gt;</li> <li>sx: Number of aligned subread bases that differ from the consensus base at each position in the consensus read. Output as one value per position: sm:B:C,&lt;MISMATCH1&gt;,&lt;MISMATCH2&gt;,...,&lt;MISMATCHN&gt;</li> </ul>

Setting name	Required	Description
Pipeline Id	No	<p><b>Note:</b> This is <b>only</b> required if specifying analysis settings for a well sample.</p> <ul style="list-style-type: none"> <li>• <b>Demultiplex Barcodes:</b> cromwell.workflows.pb_demux_ccs</li> <li>• <b>Export Reads:</b> cromwell.workflows.pb_export_ccs</li> <li>• <b>HiFi Mapping:</b> cromwell.workflows.pb_align_ccs</li> <li>• <b>Target Enrichment:</b> cromwell.workflows.pb_target_enrichment</li> <li>• <b>Iso-Seq Analysis:</b> cromwell.workflows.pb_isoseq</li> <li>• <b>Mark PCR Duplicates:</b> cromwell.workflows.pb_mark_duplicates</li> <li>• <b>Microbial Genome Analysis:</b> cromwell.workflows.pb_microbial_analysis</li> <li>• <b>PureTarget repeat expansion:</b> cromwell.workflows.pb_puretarget_re_panel_v2</li> <li>• <b>Read Segmentation:</b> cromwell.workflows.pb_segment_reads</li> <li>• <b>Read Segmentation and Iso-Seq Analysis:</b> cromwell.workflows.pb_segment_reads_and_isoseq</li> <li>• <b>Read Segmentation and Single-Cell Iso-Seq Analysis:</b> cromwell.workflows.pb_segment_reads_and_sc_isoseq</li> <li>• <b>Single-Cell Iso-Seq Analysis:</b> cromwell.workflows.pb_sc_isoseq</li> <li>• <b>Trim Ultra-Low Adapters:</b> cromwell.workflows.pb_trim_adapters</li> <li>• <b>Variant Calling:</b> cromwell.workflows.pb_variant_calling</li> </ul>
Analysis Name	No	<p><b>Note:</b> This is <b>only</b> required if specifying analysis settings for a well sample. Enter alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters only. Comma and newline are not permitted. <b>Example:</b> sample 1 analysis</p>
Entry Points	No	<p><b>Note:</b> This is <b>only</b> required if specifying analysis settings for a well sample when the selected pipeline requires Entry Points. Entry Points only apply to Barcode Sets and Reference Sets. Enter an ASCII string in the format file_type;entry_id;uuid, with parameters separated by   characters.</p> <ul style="list-style-type: none"> <li>• To find the UUID: Click <b>Data Management &gt; View Data &gt; HiFi Reads</b>. Click the Data Set of interest, then view the UUID.</li> <li>• See the <b>SMRT® Tools reference guide</b> section <b>Appendix A - Application entry points and output files</b> to see the entry point names for each application.</li> </ul> <p><b>Example:</b> PacBio.DataSet.BarcodeSet;eid_barcode;afe89e3f-17ca-e9b8-eae9-b701dbb1f02d PacBio.DataSet.ReferenceSet;eid_ref_dataset;6b8db144-a601-4577-ab04-ba64cad0548</p>
Task Options	No	<p><b>Note:</b> This is <b>only</b> required if specifying analysis settings for a well sample. 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. <b>Example:</b> pbmm2_align.task_options.minalnlength;integer;50</p> <p><b>Note:</b> This field can be left blank - any task options not specified will use pipeline defaults.</p>



## Sample Settings section

This section is represented as a table. The first line after the [Samples] line is the table header and includes these columns:

Column name	Required	Description
<b>Bio Sample Name</b>	<b>Yes</b>	Enter Bio Sample Names in the same row as their associated Barcode Names. Use alphanumeric characters, hyphens, or underscores only. Character limit: ≤40 characters. <b>Example:</b> sample1 <b>Note:</b> This field is used for collections for barcoded samples in multiplexed data.
<b>Plate well</b>	<b>Yes</b>	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. The well sample identifier that a given barcoded sample belongs to. <b>Example:</b> 1_A01
<b>Adapter</b>	<b>Yes</b>	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. This is the name of the <b>left</b> adapter used for a barcoded sample.
<b>Adapter2</b>	<b>Yes</b>	Enter alphanumeric characters, spaces, hyphens, underscores, colons, or periods only. This is the name of the <b>right</b> adapter used for a barcoded sample.
<b>Pipeline ID</b>	No	<b>Note:</b> This is <b>only</b> required if specifying analysis settings for an individual barcoded sample. <ul style="list-style-type: none"> <li>• <b>Demultiplex Barcodes:</b> cromwell.workflows.pb_demux_ccs</li> <li>• <b>Export Reads:</b> cromwell.workflows.pb_export_ccs</li> <li>• <b>HiFi Mapping:</b> cromwell.workflows.pb_align_ccs</li> <li>• <b>Target Enrichment:</b> cromwell.workflows.pb_target_enrichment</li> <li>• <b>Iso-Seq Analysis:</b> cromwell.workflows.pb_isoseq</li> <li>• <b>Mark PCR Duplicates:</b> cromwell.workflows.pb_mark_duplicates</li> <li>• <b>Microbial Genome Analysis:</b> cromwell.workflows.pb_microbial_analysis</li> <li>• <b>PureTarget repeat expansion:</b> cromwell.workflows.pb_puretarget_re_panel_v2</li> <li>• <b>Read Segmentation:</b> cromwell.workflows.pb_segment_reads</li> <li>• <b>Read Segmentation and Iso-Seq Analysis:</b> cromwell.workflows.pb_segment_reads_and_isoseq</li> <li>• <b>Read Segmentation and Single-Cell Iso-Seq Analysis:</b> cromwell.workflows.pb_segment_reads_and_sc_isoseq</li> <li>• <b>Single-Cell Iso-Seq Analysis:</b> cromwell.workflows.pb_sc_isoseq</li> <li>• <b>Trim Ultra-Low Adapters:</b> cromwell.workflows.pb_trim_adapters</li> <li>• <b>Variant Calling:</b> cromwell.workflows.pb_variant_calling</li> </ul>
<b>Analysis Name</b>	No	<b>Note:</b> This is <b>only</b> required if specifying analysis settings for an individual barcoded sample. Enter alphanumeric, space, hyphen, underscore, colon, period, and apostrophe characters only. Comma and newline are not permitted. <b>Example:</b> sample 1 analysis

Column name	Required	Description
Entry Points	No	<p><b>Note:</b> This is <b>only</b> required if specifying analysis settings for an individual barcoded sample and the selected pipeline requires Entry Points.</p> <p>Entry Points only apply to Barcode Sets and Reference Sets. Enter an ASCII string in the format file_type;entry_id;uuid, with parameters separated by   characters.</p> <ul style="list-style-type: none"> <li>To find the UUID: Click <b>Data Management &gt; View Data &gt; HiFi Reads</b>. Click the Data Set of interest, then view the UUID.</li> <li>See the <b>SMRT Tools reference guide</b> section <b>Appendix A - Application entry points and output files</b> to see the entry point names for each application.</li> </ul> <p><b>Example:</b> PacBio.DataSet.BarcodeSet;eid_barcode;afe89e3f-17ca-e9b8-eae9-b701dbb1f02d PacBio.DataSet.ReferenceSet;eid_ref_dataset;6b8db144-a601-4577-ab04-ba64cad0548</p>
Task Options	No	<p><b>Note:</b> This is only required if specifying analysis settings for an individual barcoded sample.</p> <p>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.</p> <p><b>Example:</b> pbmm2_align.task_options.minInlength;integer;50</p> <p><b>Note:</b> This field can be left blank - any task options not specified will use pipeline defaults.</p>

Each row following the table header line represents a **single barcoded sample** in the run. The comma-separated values in each row correspond to the column names described in the table above.

Example Revio run design CSV file

[Run Settings]

Instrument Type,Revio

Run Name,Example Revio Run

Plate 1,102118800

Plate 2,102118800

Transfer Subdirectory,Example\_Transfer\_Subdirectory

CSV Version,1

[SMRT Cell Settings],1\_A01,1\_B01,2\_A01

Well Name,Sample1,Sample2,Sample3

Application,Other,Other,Other

Library Type,Standard,Standard,Standard

Movie Acquisition Time (hours),24,24,24

Insert Size (bp),15000,15000,15000

Assign Data To Project,1,1,1

Library Concentration (pM),200,200,200

Include Base Kinetics,FALSE,FALSE,FALSE

---

Indexes,43f950a9-8bde-3855-6b25-c13368069745,,  
Sample is indexed,TRUE,FALSE,FALSE  
Bio Sample Name,,BioSampleB,BioSampleC  
Use Adaptive Loading,TRUE,TRUE,TRUE  
Consensus Mode,molecule,molecule,molecule

[Samples]

Bio Sample Name,Plate Well,Adapter,Adapter2  
BioSample1,1\_A01,bc2001,bc2001  
BioSample2,1\_A01,bc2002,bc2002

Example Vega run design CSV file

[Run Settings]

Instrument Type,Vega  
Run Name,Example Vega Run  
Plate 1,102118800  
Transfer Subdirectory,Example\_Transfer\_Subdirectory  
CSV Version,1

[SMRT Cell Settings]

Well Name,A01  
Application,Human WGS  
Library Type,Standard  
Movie Acquisition Time (hours),24  
Insert Size (bp),15000  
Assign Data To Project,1  
Library Concentration (pM),200  
Include Base Kinetics,FALSE,  
Indexes,43f950a9-8bde-3855-6b25-c13368069745,,  
Sample is indexed,TRUE  
Consensus Mode,molecule

[Samples]

Bio Sample Name,Plate Well,Adapter,Adapter2  
BioSample1,1\_A01,bc2001,bc2001  
BioSample2,1\_A01,bc2002,bc2002

---

### CSV file general requirements

- Each line in the CSV file represents **one setting or** 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.

### Plate identifier requirements

Sequencing plates are designed in run designs through a “plate identifier” string. There are two acceptable formats: part number only (“anonymous run”), and full plate identifier (“linked run”).

#### *Part number*

1. 9-digit part number without dashes (Example: 103496700)

#### *Plate identifier*

1. 9-digit part number without dashes (Example: 103496700)
2. 6-digit lot number (Example: 036175)
3. 5-digit serial number (Example: 00347)
4. 8-digit expiration date as YYYYMMDD (Example: 20250321)

The full plate identifier is entered as a concatenation of its components, for example “1034967000361750034720250321”)

---

## 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, barcode and BED files.

### Understanding Data Sets

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 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 47](#) 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.

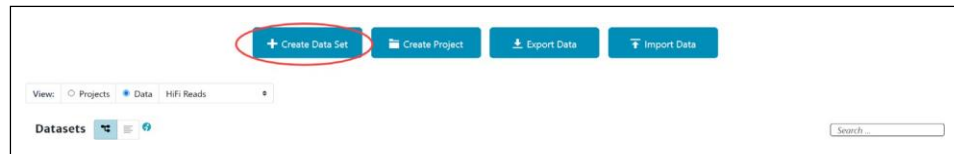
Some Data Sets can contain **basecalled** data, while others can contain **analyzed** data:

- **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 consensus reads, in aligned or unaligned format.

## Creating a Data Set

1. Select **Data Management** from the Module menu.
2. 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 169](#) for details.



3. Click **+ Create Data Set**.
4. In the **Data Sets** table, select one or more sets of sequence data.
5. Enter a name for the new Data Set.

A screenshot of a form for creating a new Data Set. It has three input fields: 'Data Set Name Required' with the placeholder 'Data Set Example', 'Well Sample Name Required' with the placeholder 'A01', and 'Bio Sample Name Required' with the placeholder 'A01'. Below these fields are two filter options: 'Filter Reads by QV >= [input] qv' and 'Filter Reads by Length [input] bp to [input] bp'. At the bottom, there is a 'Datasets' section with a search bar and the text 'Displaying rows 1 to 10 out of 101 (scroll to load more) Search...'.

6. **(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.
7. **(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**.

8. (Optional) Use the Search function to search for specific Data Sets. See [“Appendix B - Data search” on page 169](#) for details.
9. (Optional) If you selected **one** Data Set **only**, click the **Filter Reads by QV<sub>≥</sub>** box above the Data Set list. Enter the minimum quality value to retain in the new Data Set.
10. (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 read length to retain in the new Data Set.
11. Click **Save Data Set**. The new Data Set becomes available for starting analyses, viewing, or generating reports.
12. 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 47](#) for details.



#### Viewing Data Set information

1. Select **Data Management** from the Module menu.
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.
 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 169](#) 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.

**Note:** Some of the plots displayed include the **Movie Name**, in <instrument-number\_date\_time> format.

**Example:** m64263e\_211008\_2133059.

#### Copying a Data Set

1. Select **Data Management** from the Module menu.
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.
 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 169](#) 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 Data Set deletion 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. Select **Data Management** from the Module menu.
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.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 169](#) 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 and click **Next**.
2. Follow the instructions starting at Step 13 of [“Creating and starting a job” on page 54](#).

#### Data Set reports

The Data Set reports are generated when run metadata data is imported into SMRT Link, 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 report elements 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**.

### Run Preview - 4hr

- See “Basic preview” from Runs section for a description of report fields.

### Run Preview - 23hr

- See “Full preview” from Runs section for a description of report fields.

### CCS Analysis Report Summary Metrics

- 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 Length (median, bp): The median read length of the CCS reads whose quality value is equal to or greater than 20.
- HiFi Read Length N50 (bp): 50% of all CCS reads whose quality value is equal to or greater than 20 are longer than this value.
- HiFi Read Quality (median): The median number of CCS reads whose quality value is equal to or greater than 20.
- Base Quality  $\geq$ Q30 (%): The percentage of CCS reads whose quality value is equal to or greater than 30.
- 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.

### HiFi Read Length Summary

- Read Length (Kb): The HiFi read length, ranging from  $\geq 0$  to  $\geq 40,000$  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 (Gb): 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.

### HiFi Read Quality Summary

- Read Length (Kb): The HiFi read length, ranging from  $\geq 0$  to  $\geq$

---

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 (Gb): 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 > 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.

#### Methylation

HiFi methylation callers estimate the likelihood of a modification at a specific motif. All models have some false positives and negatives, so consider the known modifications in the sample to avoid mistaking false positives for real modifications.

#### 5mC/6mA modifications:

- Reports the modification context and the cumulative of percentage of modified sites in the sample mapped against the predicted probability of methylation.
- Plots the percentage of CpG sites in the sample versus the predicted probability of methylation.
- For 6mA, the model only reports 6mAs where probability of modification is  $\geq 0.98$  in order to reduce false positives.

Note: On-instrument 6mA calling is intended to support Fiber-Seq applications and is not intended for use as a general 6mA caller for microbial genome analysis application.

#### Barcodes

- See “Reports and data files” from Demultiplex Barcodes section for a description of report fields.

#### 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.

### **Analyses**

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

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

## Understanding Projects

- 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

The screenshot shows the 'Create Project' form in the Data Management interface. The form is titled 'Create Project' and has a breadcrumb 'Data Management / Projects'. In the top right corner, there is a purple button labeled 'Projects: All My Pr' and two buttons: 'Cancel' and 'Save'. The form is divided into several sections:

- Project Name Required:** A text input field containing 'QC\_Group'.
- Description:** A larger text input field.
- Associated Data Sets:** A section with a search icon and the text 'Select Data Sets'.
- Members:** A section for user access. It includes a dropdown menu for 'Access for All SMRT® Link users' set to 'None'. Below it, 'Access for Individual SMRT® Link Users' is shown with a list of users: 'Administrator (Administrator@p...' and 'EPMAdmin2', each with a 'View' button and a plus icon. At the bottom of this section, there is a search bar for 'QA' with a 'Search By Name' button.
- User Selection Table:** A table with columns for 'User Name' and 'Email'. It contains one entry: 'sappi-admin-qa' with email 'sappi-admin-qa@pacificbiosciences.com'. Below the table is an 'Add Selected User' button.

1. Select **Data Management** from the Module menu.
2. Click **Create Project**.
3. Enter a name for the new project.

- 
4. (Optional) Enter a description for the project.
  5. Click **Select Data Sets** and select one or more sets of sequence data to associate with the project.
    - (Optional) Use the Search function to search for Data Sets. See [“Appendix B - Data search” on page 169](#) for details.
  6. (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**.
  7. Click **Save**. The new project becomes available for SMRT Link users who now have access.

#### Editing a Project

1. Select **Data Management** from the Module menu.
2. Click **View > Projects**.
3. Projects can be sorted and searched for:
  - To sort Projects: Click a **column title**.
  - To search for a Project, use the Search function. See [“Appendix B - Data search” on page 169](#) 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.

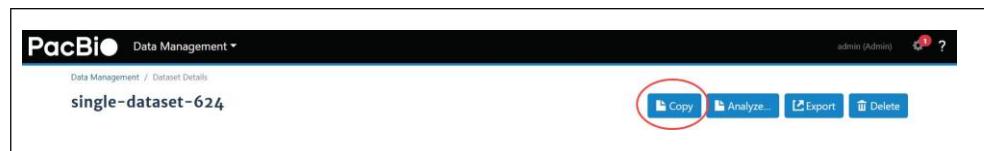
#### Deleting a Project

1. Select **Data Management** from the Module menu.
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, references, barcodes and target region files

1. Select **Data Management** from the Module menu.
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.
  - **Barcodes**: Barcodes from barcoded samples.
  - **References**: Reference sequence FASTA files used when creating certain analyses.
  - **Target regions**: BED files that specify target genes or regions for analysis; used with the **PureTarget™ repeat expansion panel** application and optionally with the Target Enrichment workflow.
3. (Optional) Use the Search function to search for specific Data Sets, barcode files, reference sequence files or target region files. See [“Appendix B - Data search” on page 169](#) for details.
4. Click the name of the file of interest. Details for that file display.
5. (Optional) To delete the sequence data, reference sequence, barcode file, or target region file, click **Delete**.

**Note:** The **Copy** button is available for HiFi reads, but **not** for Reference, Barcode, and Target data.



## Importing and exporting data

### Importing sequence, reference, barcode and target regions data

**Note:** If your sequencing 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 and target regions files, can be imported for use in SMRT Link.

1. Select **Data Management** from the Module menu.
2. Click **Import Data**.
3. Specify whether to import data from the **SMRT Link Server**, or from a **Local File System**. **Note:** Only references, barcodes and target region files are available if you select **Local File System**.

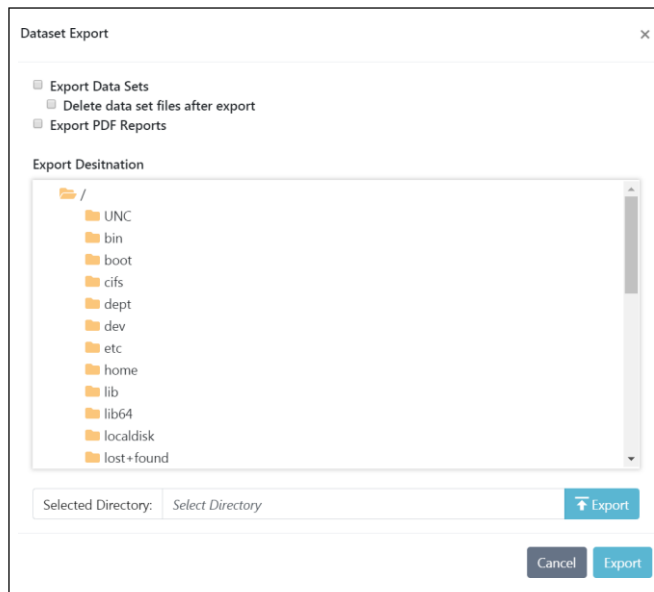
- 
4. Select the data type to import:
    - **HiFi reads:** XML file (.consensusreadset.xml) or ZIP file containing information about HiFi reads (reads generated with CCS analysis whose quality value is equal to or greater than 20.)  
  
Use **only** ZIP files created by SMRT Link.
    - **Barcodes:** FASTA (.fa or .fasta), XML (.barcodeset.xml), or ZIP files containing barcodes. **Note:** Barcode names must be  $\leq 40$  characters, only contain alphanumeric characters or underscores, and unique within the set.
    - **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.)
    - **Target Regions (BED):** BED files that specify target genes or regions for analysis; used with the **PureTarget repeat expansion panel** application.
    - **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

HiFi reads can be exported, as well as barcode and reference files.

1. Select **Data Management** from the Module menu.
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.

- **Barcodes:** Files containing barcodes.
  - **References:** Files containing a reference sequence for use in starting analyses.
4. **(Optional)** Use the Search function to search for Data Sets, barcode files, or reference files. See [“Appendix B - Data search” on page 169](#) 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. See [“Appendix E - Additional information included in the CCS Data Set Export report” on page 178](#) for details. Note: Exporting PDF Reports requires that pop ups are allowed.
10. Click **Export**.



---

## SMRT® Analysis

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

**Note:** The SMRT Analysis module is **not** included when you install SMRT Link Lite.

### Creating and starting a job

1. Select **SMRT Analysis** from the Module menu.
2. 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 169](#) 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**.



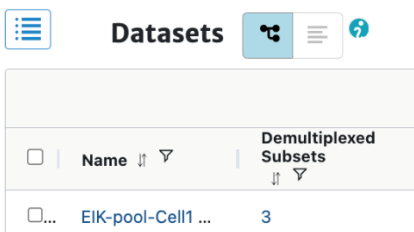
- To filter the list of jobs based on the **Project(s)** that they are associated with: Click the **Projects** menu (located at the top-right of the main SMRT Analysis page) and select a Project. See [“Understanding Projects” on page 49](#) for details.
3. Click **+ Create New Job**.
  4. **(Optional)** 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).
  5. Enter a **name** for the job.

6. **(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.

7. 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 169](#) 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):

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

Analysis of Multiple Data Sets

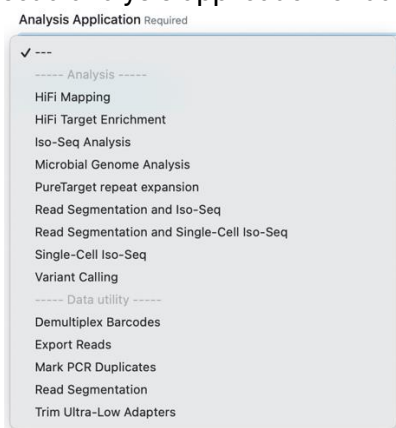
- One Analysis for All Data Sets
- One Analysis per Data Set - Identical Parameters
- One Analysis per Data Set - Custom Parameters

- **One Analysis for All Data Sets:** Runs one 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 (v25.2)**.

9. Click **Next**.

10. Select an analysis application or data utility from the drop-down menu.



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

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

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

12. 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 ["Import-](#)

ing sequence, reference, barcode and target regions data” on page 51 for details.)

SMRT Analysis / Create New Analysis

1. Select Data 2. Select Analysis

Analysis Application Required  
Iso-Seq Analysis

Analysis Name  
Job 99

Import Analysis Settings Export

Associated Inputs

Primer Set Required  
Iso-Seq v2 Barcoded cDNA Primers

Reference Set

Cluster of Barcoded Samples  
Pool reads and cluster together

Advanced Parameters

ID #	Name #
422	SLT-4_test-2023-1

13. (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.)

14. To see information about parameters for **all** secondary analysis applications and data utilities provided by PacBio, see “[Analysis applications](#)” on page 63 or “[Data utilities](#)” on page 134 for details.

Advanced Analysis Parameters

Min. CCS Predicted Accuracy (Phred Scale) 20

Filters to Add to the Data Set

Require and Trim Poly(A) Tail  ON  OFF

Minimum Mapped Length (bp) 50

Minimum Gap-Compressed Identity (%) 95

Minimum Mapped Coverage (%) 99

Advanced pbmm2 Options

Maximum Fuzzy Junction Difference (bp) 5

Advanced pigeon filter options

Add task memory (MB) 0

Compute Settings -- select --

OK Cancel

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

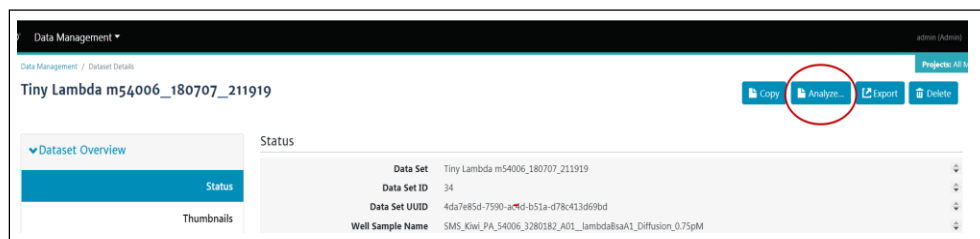
16. (Optional) Click **Back** if you need to change any of the analysis attributes you selected earlier.
17. Click **Start** to submit the job. (If you selected multiple Data Sets as input, click **Start Multiple Jobs** or **Start and Create Next**.)
18. Select **SMRT Analysis** from the Module menu to navigate to the main SMRT Analysis screen. There, the status of the job you started displays. When the job has **completed**, click on its name - reports are available for the completed job.
19. (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. Select **Data Management** from the Module menu.
2. Click **View > Data** and select in the Name column, click the name of the sequence dataset of interest. Details for the selected sequence data will be displayed.



3. To **start** a job using this sequence data, click **Analyze...**, name the job, click **Next**, then follow the instructions starting at Step 13 of **“Creating and starting a job”** on page 54.

## Canceling a running job

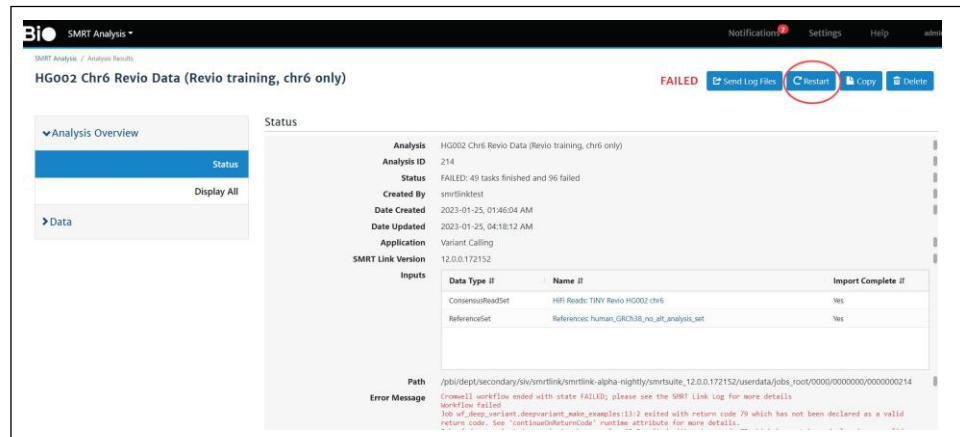
1. Select **SMRT Analysis** from the Module menu.
2. Click the funnel in the **State** column header, then click **Running**. This displays **only** currently-running jobs.
3. Select a currently-running job to cancel.
4. Click **Cancel**.
5. Click **Yes** in the confirmation dialog. The cancelled job 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. Select **SMRT Analysis** from the Module menu.
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. Select **SMRT Analysis** from the Module menu. 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 169](#) for details.
4. Click the job link of interest.

5. Click **Analysis Overview > Status** to see job information status, including which application 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 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.
8. For information on the reports and data files produced by analysis applications/data utilities, see [“Analysis applications” on page 63](#) or [“Data utilities” on page 134](#) for details.
9. 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. You can also click the small icon to the left of the file name to copy the file’s path to the Clipboard.
10. **(Optional)** In **Data > File Downloads**, specify prefix(es) 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**.
11. To view job log details: Click **Data > SMRT Link Log**.

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

1. Select **SMRT Analysis** from the Module menu. 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 169](#) 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 [“Analysis applications” on page 63](#) or [“Data utilities” on page 134](#) for further details.
9. Click **Start**.

Copying and running an existing job

### 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. Select **SMRT Analysis** from the Module menu.
  2. Click **Export Jobs**.
  3. (Optional) Use the Search function to search for specific analyses. See [“Appendix B - Data search” on page 169](#) 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. Select **SMRT Analysis** from the Module menu.
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.

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

## Summarizing Microbial Genome Analysis jobs

- 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. Select **SMRT Analysis** from the Module menu.
2. Click **Summarize Jobs**. This displays all the successfully-completed Microbial Genome Analysis jobs.
3. (Optional) Use the Search function to search for specific jobs. See [“Appendix B - Data search” on page 169](#) 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 Analysis jobs you selected in Step 5.



---

## Importing and exporting job settings

For all workflows, job settings included advanced parameters can be exported and imported as follows.

1. 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.
2. 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.

## SMRT Analysis applications

Following are the secondary analysis applications provided with SMRT Analysis v25.2. 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.

### HiFi Mapping

- Align (or map) data to a user-provided reference sequence.
- See [“HiFi Mapping” on page 68](#) for details.

### Target Enrichment

- Analyze multiplexed samples prepared with a target enrichment workflow.
- See [“Target Enrichment” on page 72](#) for details.

### Iso-Seq® Analysis

- Characterize full-length transcript isoforms.
- See [“Iso-Seq® Analysis” on page 81](#) for details.

### Microbial Genome Analysis

- 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 89](#) for details.

### PureTarget repeat expansion

- Analyze multiplexed samples prepared with the targeted sequencing PureTarget repeat expansion panel.
- See [“PureTarget repeat expansion” on page 95](#) for details.

---

## Read Segmentation and Iso-Seq® Analysis

- Characterize full-length transcript isoforms, using HiFi reads sequenced on PacBio instruments. Includes the Read Segmentation step to split arrayed HiFi reads at adapter positions, generating segmented reads (S-reads) comprised of multiple fragments.
- See [“Read Segmentation and Iso-Seq® Analysis” on page 99](#) 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), that were sequenced on PacBio instruments.
- The application is for use when using concatenation-based library preparations such as the Kinnex™ libraries.
- See [“Read Segmentation and Single-Cell Iso-Seq® Analysis” on page 108](#) 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 [“Single-Cell Iso-Seq® Analysis” on page 117](#) details.

## Variant Calling

- Identify single-nucleotide variants, short insertions and deletions, and structural variants for a single sample against a specific reference genome.
- See [“Variant Calling” on page 128](#) for details.

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

Reference Set (Required)

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

Parameters

Advanced parameters	Default value	Description
<b>Filters to Add to the Data Set</b>	NONE	A semicolon-separated ( <b>not</b> comma-separated) list of other filters to add to the Data Set.
<b>Minimum Mapped Length (bp)</b>	50	The minimum required mapped read length, in base pairs.
<b>Bio Sample Name of Aligned Dataset</b>	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. <b>Note:</b> Use alphanumeric characters, hyphens, or underscores only. Character limit: ≤40 characters.
<b>Minimum Gap-Compressed Identity (%)</b>	70	The minimum required gap-compressed alignment identity, in percent. Gap-compressed identity counts consecutive insertion or deletion gaps as one difference.
<b>Min. CCS Predicted Accuracy (Phred Scale)</b>	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for <b>all</b> applications is 20 (QV 20), or 99% predicted accuracy.
<b>Advanced pbmm2 Options</b>	NONE	Space-separated list of custom pbmm2 options. Not all supported command-line options can be used, and HPC settings <b>cannot</b> be modified. See <b>SMRT Tools reference guide</b> for details.
<b>Target Regions (BED file)</b>	NONE	<b>(Optional)</b> Specifies a BED file that defines regions for a <b>Target Regions</b> report showing coverage over those regions. See <a href="#">“Appendix D - BED file format for Target Enrichment and HiFi Mapping applications” on page 177</a> for details. Note: BED file is uploaded from your local file system.
<b>Add task memory (MB)</b>	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
<b>Compute Settings</b>	Select	<b>(Optional)</b> 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 to a reference sequence.

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

## Mapping Report > Mapping Statistics Summary

Displays mapping statistics per movie.

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

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

#### Mapping Report > Mapped Read Length

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

#### Mapping Report > Mapped 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 read 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](#).

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

---

## Target Enrichment

Use this application to process and analyze multiplexed target enrichment samples including amplicons, hybrid capture, and PureTarget libraries. Files output for each sample include mapped BAM files, target enrichment statistics, and variant call sets.

- Demultiplexing must be performed before the Target Enrichment workflow.

The application includes the following main steps:

1. (Optional) Mark PCR duplicate HiFi reads by sample using [pbmarkdup](#).
2. Align HiFi reads to the reference using [pbmm2](#).
3. (Optional) Call small variants using [DeepVariant](#).
4. (Optional) Phase small variants and haplotag BAMs using [whatshap](#).
5. (Optional) Call structural variants using [pbsv](#).
6. Produce statistics using [picard](#).

Preparing input data for the Target Enrichment analysis application

- Run the Demultiplex Barcodes data utility; see [“Demultiplex Barcodes” on page 138](#) for details.
- The recommended barcoded set is **Twist Universal Adapter with UDI**. Select **No** for **For Same Barcodes on Both Ends**. You may specify a different barcode set for demultiplexing.

Reference Set (Required)

- Specify a reference genome against which to align the reads and call variants. The default set is **Human Genome hg38, with Gencode v39 annotations**.

Target BED file (Required)

- Specify a BED-format file containing the target genes or regions of interest. See [“Appendix D - BED file format for Target Enrichment and HiFi Mapping applications” on page 177](#) for details.

## Parameters

Advanced parameters	Default value	Description
Include Fail Reads	OFF	For PureTarget panels that include repeat expansions, this option should be set to YES.
Perform variant calling	OFF	DeepVariant and pbsv are ON to make small variant and SV calls, respectively. <b>Note:</b> your SMRT Link server needs to be configured with Singularity and a job management system to perform SNV and INDEL variant calling
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 <b>Well Sample Name</b> field in the Runs module. By convention, different LB tags correspond to different library preparations. Use this option when the LB tag does <b>not</b> follow this convention to treat <b>all</b> reads as from the same sequencing library.
Target order in boxplots	BED Order	Specify how targets are ordered in the box plots generated: <ul style="list-style-type: none"> <li>• <b>Bed order:</b> Targets display in the order in which they appear in the Target BED file.</li> <li>• <b>Alphabetical:</b> Targets display in alphabetical order.</li> <li>• <b>Genomic coordinate:</b> Targets display in genomic coordinate order.</li> <li>• <b>Mean coverage:</b> Targets display in mean coverage order.</li> </ul>
Padding around regions (bp)	3000	Padding (in base pairs) to add to each side of the regions specified in the Target BED file. This setting impacts the percentage of on-target bases in the Sample Summary table.
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for <b>all</b> applications is 20 (QV 20), or 99% predicted accuracy.
Mark PCR duplicates	OFF	Mark any duplicate reads in the output mapped BAM file. On is recommended when PCR is used during sample preparation.
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.
Minimum Mapped Length (bp)	50	The minimum required mapped read length, in base pairs.
Minimum Length of Structural Variant (bp)	20	The minimum length of structural variants, in base pairs.
Advanced pbsv Options	NONE	Additional pbsv command-line arguments.
Minimum % of Reads that Support Variant (any one sample)	10	Ignore calls supported by <N% of reads in every sample.
Minimum Reads that Support Variant (any one sample)	3	Ignore calls supported by <N reads in every sample.
Minimum Reads that Support Variant (total over all samples)	3	Ignore calls supported by <N reads total across samples.
Use GPU if available	ON	Send GPU-accelerated tasks to an HPC queue that includes GPU resources. <b>Note:</b> This option requires additional setup by the SMRT Link Administrator, and is <b>not</b> supported on AWS or local compute backends.
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Maximum number of parallel tasks	5120	Limits the number of simultaneous DeepVariant make_examples tasks, which helps ensure stable performance in SMRT Analysis. Raising this value (or setting it to zero) allows greater parallelism.
Compute Settings	Select	<b>(Optional)</b> Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.



---

## Reports and data files

The Target Enrichment application generates the following reports:

### Target Enrichment > Summary Metrics

- **HiFi Bases:** The number of bases of HiFi data analyzed.
- **HiFi Reads:** The number of HiFi reads analyzed.
- **HiFi Read length (median, bp):** The median length of the HiFi reads analyzed, in base pairs.
- **HiFi Read Quality (median):** The median Phred-scale quality value of the HiFi reads analyzed.
- **Sample Count:** The number of samples analyzed.
- **Total Target Length (bp):** The total number of bases contained in the target regions specified in the BED file, including padding.
- **Target regions:** The number of target regions specified in the input Target BED file.

### Target Enrichment > Sample Summary

- **Sample:** The name of the sample for which the following statistics apply.
- **HiFi Reads:** The number of HiFi reads in the sample.
- **HiFi read length (median, bp):** The median length of the HiFi reads in the sample, in base pairs.
- **HiFi Read Quality (median):** The median Phred-scale quality value of the HiFi reads in the sample.
- **Mean Target Coverage:** The mean coverage across the target regions for the sample. Equivalent to MEAN\_TARGET\_COVERAGE in hs\_metrics of Picard Tools.
- **Percent Target  $\geq$ 10-fold Coverage:** The percent of the bases in the target regions with 10-fold or more coverage for the sample. Equivalent to PCT\_TARGET\_BASES\_10X in hs\_metrics of Picard Tools.
- **Percent Target  $\geq$ 20-fold Coverage:** The percent of the bases in the target regions with 20-fold or more coverage for the sample. Equivalent to PCT\_TARGET\_BASES\_20X in hs\_metrics of Picard Tools.
- **Percent Target  $\geq$ 30-fold Coverage:** The percent of the bases in the target regions with 30-fold or more coverage for the sample. Equivalent to PCT\_TARGET\_BASES\_30X in hs\_metrics of Picard Tools.
- **Percent Target Zero Coverage:** The percent of the bases in the target region with no coverage for the sample. Equivalent to ZERO\_CVG\_TARGETS\_PCT in hs\_metrics of Picard Tools.
- **Target Coverage Uniformity:** The fold over-coverage necessary to raise 80% of bases in "non-zero-cvg" targets to the mean coverage level in those targets. Equivalent to FOLD\_80\_BASE\_PENALTY in hs\_metrics of Picard Tools.
- **Percent On-Target Bases:** The percent of bases in the sample that map to the target region (including the padding). Equivalent to PCT\_EXC\_OFF\_TARGET in hs\_metrics of Picard Tools.
- **Percent Duplicate Reads:** The percent of the duplicate reads in the sample. Equivalent to PCT\_EXC\_DUPE in hs\_metrics of Picard Tools.

### Target Enrichment > Read Categories

- Stacked histogram mapping the length of 3 categories of reads in the sample, on-target, off-target, and unmapped.

### Target Enrichment > Target Coverage

- Box plot for each target regions of mean coverage across all samples analyzed. Use this plot to quickly compare coverage across all target regions and identify any targets with low or high coverage.

---

## Target Enrichment > Sample Coverage

- **Box plot** for each sample of mean coverage across all target regions. Use this plot to quickly compare performance across samples.

## 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**.)
- **Target Coverage by Sample CSV:** Comma-delimited text file of the matrix of all mean coverage values for each target (rows) and sample (columns). The data in this file is used to generate the Target Coverage and Sample Coverage box plot.
- **Sample Summary CSV:** CSV version of the data displayed in the Target Enrichment Sample Summary table.
- **Sample VCFs for SVs:** Zipped archive of PBSV structural variant calls in VCF format for each sample. (See [here](#) for details.)
- **Sample Mapped BAMs:** Zipped archive of haplotagged, mapped reads in BAM format.
- **Sample VCFs for small variants:** Zipped archive of [DeepVariant](#) small variant calls in VCF format for each sample.

## Data > IGV Visualization Files

The following files are used for visualization using [IGV](#).

- **Small Variants VCF:** For each sample, [DeepVariant](#) small variant calls in VCF Format (VCF).
- **Structural Variants VCF:** For each sample, PBSV structural variant calls in VCF format. (See [here](#) for details.)
- **Mapped+Haplotagged BAM:** For each sample, mapped, haplotagged HiFi reads in BAM format.
- **Mapped+Haplotagged BAM Index:** The BAI index file for the corresponding Mapped BAM file.

---

## 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 is for analyzing regular, unconcatenated Iso-Seq libraries. Alternatively, HiFi reads from Kinnex full-length RNA libraries can first be deconcatenated using the **Read Segmentation** data utility, and the segmented reads (S-reads) can run through the **Iso-Seq Analysis** workflow. (For a video on using the Kinnex full-length RNA kit, click [here](#).)

Note on barcoded libraries:

- The Read Segmentation and Iso-Seq workflow will **only** process barcoded libraries at the cDNA level (such as using **Iso-Seq v2 Barcoded cDNA Primers** as part of the MAS-Seq for bulk Iso-Seq kit). Demultiplexing of barcoded adapters (also part of the **Kinnex full-length RNA kit**) should first run through the Demultiplexing workflow in SMRT Link.

---

## Workflow

1. **Iso-Seq Analysis (all use cases)** - Full-length, non-concatemer (FLNC) reads are then identified by the presence of cDNA primers and polyA tails. Reads are then clustered *de novo* at the isoform level to generate high-quality, full-length consensus isoform sequences. If **no** reference genome is selected, the workflow stops here.
2. **Iso-Seq Analysis, with reference genome only** – If a reference genome is provided, high-quality, full-length isoform sequences are mapped to the genome and collapsed into unique isoforms (GFF, FASTA output). If the reference bundle contains **only** a reference genome, the workflow stops here.
3. **Iso-Seq Analysis, reference genome + annotation (human/mouse only)**  
- If using the human or mouse reference and annotation package that is preloaded in SMRT Link, the collapsed isoforms are classified and filtered against reference annotation using pigeon, which is a transcript classification and filtering tool based on the [SQANTI3](#) software.

Segmentation Adapter Set (Default = MAS-Seq Adapter v3 (MAS8))

- 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 = Iso-Seq v2 Barcoded 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:** Iso-Seq v2 Barcoded cDNA Primers.

```
>IsoSeqX_bc01_5p
CTACACGACGCTCTCCGATCTACTACACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc02_5p
CTACACGACGCTCTCCGATCTACTAGTAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc03_5p
CTACACGACGCTCTCCGATCTAGTGTACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc04_5p
CTACACGACGCTCTCCGATCTATCACTAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc05_5p
CTACACGACGCTCTCCGATCTCAGCTGTGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc06_5p
CTACACGACGCTCTCCGATCTCAGTCACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc07_5p
CTACACGACGCTCTCCGATCTCATGTATGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc08_5p
CTACACGACGCTCTCCGATCTCGTATGTGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc09_5p
CTACACGACGCTCTCCGATCTGACATGTGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc10_5p
CTACACGACGCTCTCCGATCTGAGTCTAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc11_5p
CTACACGACGCTCTCCGATCTGTAGATAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc12_5p
CTACACGACGCTCTCCGATCTGTATGACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_3p
AAGCAGTGGTATCAACGCAGAGTAC
```

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 Genome hg38, with Gencode v39 annotations** and **Mouse Genome mm39, with Gencode vM28 annotations**.
- Alternatively, choose other reference genomes (but not with annotations) that were custom-uploaded to SMRT Link.
- The Reference Set can be left blank. If blank, the workflow will **stop** after the isoform clustering step (isoseq cluster).

Cluster of Barcoded Samples

- This option specifies barcoded samples that were barcoded at the cDNA level, where the (barcoded) cDNA primers are specified in the Primer Set option. This option does **not** address libraries that were barcoded using barcoded adapters.
- Specify whether all FLNC reads will be pooled for clustering, then demultiplexed based on pooled result. **Note:** This setting does **not** apply to non-barcoded samples.

- Specify **Pool reads and cluster together** 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 **Cluster reads separately** 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	<b>ON</b> means that polyA tails are required for a sequence to be considered full length. <b>OFF</b> 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. <b>Note:</b> This is applicable <b>only</b> 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. <b>Note:</b> This is applicable <b>only</b> 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. <b>Note:</b> This is applicable <b>only</b> 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 <b>all</b> less than the provided value, they will <b>all</b> be collapsed. Setting to 0 requires all junctions to be exact to be collapsed into a single isoform. Applicable <b>only</b> 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 ( <b>not</b> 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 <b>cannot</b> be modified.
Advanced pigeon filter options	NONE	Space-separated list of custom pigeon filter options. Example: --min-cov N to reduce minimum coverage for low abundance isoforms (default value 3).
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> 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:

### 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 (FLNC Reads):** 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 FLNC 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.

### 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.

### 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.

### Read Classification > Primer Quality Scores

- Histogram of primer scores.

---

## 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.

## Transcript Clustering > Length of Consensus Isoforms

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

## Transcript Mapping and Classification > Summary Metrics

- **Sample Name:** Sample name for which the following metrics apply.
- **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 isoforms:** The total number of unique isoforms across all cells.
- **Total unique isoforms, filtered:** The total number of unique isoforms across all cells, after filtering out reads based on the SQANTI transcript filtering criteria.

## Transcript Mapping and Classification > Transcript Classifications

- **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 <b>same</b> 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 <b>fewer</b> 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 <b>not</b> 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 <b>not</b> have a FSM or ISM match, and has at least <b>one</b> donor or acceptor site that is <b>not</b> annotated.
Antisense	The query isoform does <b>not</b> 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 Mapping and Classification > 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 Mapping and Classification > 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 Mapping and Classification > 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 Mapping and Classification > Gene Saturation

- **Gene Saturation, all genes, filtered**
  - Saturation plot showing the level of gene saturation for all genes, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Gene Saturation, known genes only, filtered**
  - Saturation plot showing the level of gene saturation, for unique known genes only (genes annotated in the reference annotation) per cell, after filtering out reads based on the SQANTI transcript filtering criteria.

#### 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  $\geq 10$  counts and mean scores  $\geq 45$ .
- **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.
- Pigeon summary, Pigeon filtered summary, and Pigeon filtering report files: txt and json files summarizing input transcripts, unique genes, unique transcripts, and isoform and read based classification summaries

#### Data > IGV Visualization Files

The following files are used for visualization using [IGV](#).

- **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 [isoseq.how](#) 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 previous releases.

The Microbial Genome Analysis application:

- 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 4mC and 6mA and associated DNA sequence motifs. **Note:** This requires kinetic information.
- Facilitates assembly of larger genomes (yeast) as well.

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
Genome Length	10M	The approximate number of base pairs expected in the genome. This is used only for downsampling; if the value is $\leq 0$ , downsampling is disabled. 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.
Advanced parameters	Default value	Description
Downsampled coverage	0	The input Data Set can be downsampled to a desired coverage for the assembly steps, provided that the <b>Downsampled coverage</b> value is $>0$ . This parameter selects reads randomly, using a fixed random seed for reproducibility. Note: For very high coverage datasets, recommend to manually downsample prior to selecting as input for Microbial Genome Assembly, to prevent job failures in polishing step.
Advanced Assembly Options for chromosomal stage	NONE	A semicolon-separated list of KEY=VALUE pairs. New line characters are not accepted. See <b>Appendix C</b> in <b>SMRT Tools reference guide</b> for details.
Advanced Assembly Options for plasmid stage	NONE	A semicolon-separated list of KEY=VALUE pairs. New line characters are not accepted. See <b>Appendix C</b> in <b>SMRT Tools reference guide</b> for details.
Maximum plasmid length, bp	300,000	This value should be set <b>higher</b> 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 <b>all</b> applications is 20 (QV 20), or 99% predicted accuracy.
Filters to Add to the Data Set	NONE	A semicolon-separated ( <b>not</b> 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.
Use chemistry model	NONE	Override the default behavior of ipdSummary.py and specify a chemistry model by string name.
Minimum Qmod Score	35	Specify the minimum Qmod score to use in motif-finding.
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> 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 to a reference sequence.

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

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

#### Mapping Report > Mapping Statistics Summary

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

#### Mapping Report > Mapped Read Length

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

#### Mapping Report > Mapped Reads Concordance

- Histogram distribution of the number of HiFi reads by the percent concordance with the reference sequence. Concordance for HiFi 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 HiFi 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 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](#).

- **Mapped BAM:** The BAM file of read 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.



---

## PureTarget repeat expansion analysis

Use this application to analyze multiplexed samples prepared with the PureTarget repeat expansion panel. The analysis produces target enrichment summary statistics and uses the tandem repeat genotyping tool (**TRGT**) for variant-calling (trgt genotype) and for visualization (trgt plot).

- **Revio and Vega systems:** The application accepts demultiplexed **HiFi reads** (BAM format) plus **Fail reads** (BAM format) with 5mC calls for one or more datasets.
  - **HiFi reads** are reads generated with CCS analysis whose quality value is equal to or greater than 20.
  - **Fail reads** include the median-length subread from ZMWs with at least one full-length subread.

Reference genome (Required)

- Specify a reference genome against which to align the reads. The default set is **Human Genome hg38, with Gencode v39 annotations**.

Target and repeat definitions (Required)

- Specify a target and repeat definition BED dataset. The default set is PureTarget repeat expansion panel.
- Only reads that map within the target regions in the BED file are included in the analysis. To create a "virtual panel" excluding targets in the repeat expansion panel, create a new BED dataset **without** that target.
- To analyze data from a custom panel, create a new BED dataset with those additional targets. Note that this analysis workflow will **only** work for tandem repeat regions. See [here](#) for more information on the required BED file format for tandem repeats.

## Parameters

Advanced parameters	Default value	Description
Sample karyotypes	NONE	Ploidy is considered when genotyping X-chromosome repeats. User may specify a csv file with header 'biosample,karyotype' followed by one biosample name and karyotype (comma-separated) per line. Default karyotype is XX. Analogous to -k or --karyotype for trgt genotype command.
Target order in boxplots	BED Order	Specify how targets are ordered in the box plots generated: <ul style="list-style-type: none"> <li>• <b>BED order:</b> Targets display in the order in which they appear in the Target BED file.</li> <li>• <b>Alphabetical:</b> Targets display in alphabetical order.</li> <li>• <b>Genomic coordinate:</b> Targets display in order of genomic coordinates.</li> <li>• <b>Mean coverage:</b> Targets display in order of increasing coverage.</li> </ul>
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

### Reports and data files

The PureTarget repeat expansion application generates the following reports:

Note: On- and off-target rates are based on unique mapped reads only.

Target Enrichment > Summary Metrics

- **Total Bases:** The total number of bases analyzed, including both HiFi reads and Fail reads from Revio and Vega systems.
- **Total Reads:** The total number of reads analyzed, including both HiFi reads and Fail reads from Revio and Vega systems.
- **Median Read Length (bp):** The median length of the reads analyzed, in base pairs.
- **Median Read Quality:** The median Phred-scale quality value of the reads analyzed.
- **Sample Count:** The number of samples analyzed.
- **Target Regions:** The number of target regions specified in the input Target BED dataset.

Target Enrichment > Sample Summary

- **Sample:** The sample name for which the following metrics apply.
- **Number of reads:** The number of reads for the sample.
- **Median Mapped Read Length:** The median length of the mapped reads for the sample, in base pairs.
- **Median Mapped Read Quality:** The median RQ tag value as PHRED quality score.
- **Mean Target Coverage:** The mean coverage across all target regions for the sample.
- **Percent 10X/20/30X Coverage:** The percent of targets with  $\geq 10X/20X/30X$  coverage.
- **Percent Target Low Coverage (<5X):** The percent of targets in the sample with low coverage, less than 5X.
- **Percent On Target Reads:** The percent of mapped reads that have a non-empty alignment overlap with the target defined by the provided BED.
- **Percent Duplicate Reads:** The percent of mapped reads that have the 0x400 SAM flag.

---

## Target Enrichment > Read Categories

- Histogram mapping the length of 3 categories of reads (On target, Off target, Unmapped) in the sample.

**Note:** A read is defined as being **on-target** if its alignment region in the reference genome has a non-empty overlap with any defined target in the input BED file. Histogram bars for the different categories are stacked.

## Target Enrichment > Target Coverage

- Box plot for each target regions of mean coverage across all samples analyzed. Use this plot to quickly compare coverage across all target regions and identify any targets with low or high coverage.

## Target Enrichment > Sample Coverage

- Box plot for each sample of mean coverage across all target regions. Use this plot to quickly compare performance across samples.

## 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.**)
- **VCF files:** Zipped archive <Prefix>-trgt\_vcfs.zip for all targets for each sample, generated by the trgt.
- **Repeat expansion genotypes:** Comma-delimited text file <Prefix>-puretarget\_repeat\_expansion\_genotypes.csv where samples are in rows and the following data are in columns for each panel target allele:
  - **Repeat unit:** Sequence of repeat for the target.
  - **[Target] read count allele 0:** Number of reads assigned to allele 0.
  - **[Target] consensus size allele 0:** Length in base pairs of the tandem repeat array for the consensus sequence of allele 0.
  - **[Target] min size allele 0:** Minimum length in base pairs of the tandem repeat array in the reads assigned to allele 0.
  - **[Target] max size allele 0:** Maximum length in base pairs of the tandem repeat array in the reads assigned to allele 0.
  - **[Target] motif counts allele 0:** Number of repeat units in the consensus sequence of allele 0.
  - **[Target] motif spans allele 0:** This column describes how the consensus sequence of the allele is split into motif units.

For example, if the column says: 0(0-64)\_1(64-104)\_2(104-146) **and** the repeat unit column says CAGG:CAGA:CA, **then:**

- motif 0 is CAGG
- motif 1 is CAGA
- motif 2 is CA

The column can be read as:

- Bases 0 (inclusive) to 64 (exclusive) are spans of CAGGs
- Bases 64 (inclusive) to 104 (exclusive) are spans of CAGAs
- Bases 104 (inclusive) to 146 (exclusive) are spans of CAs

### Notes:

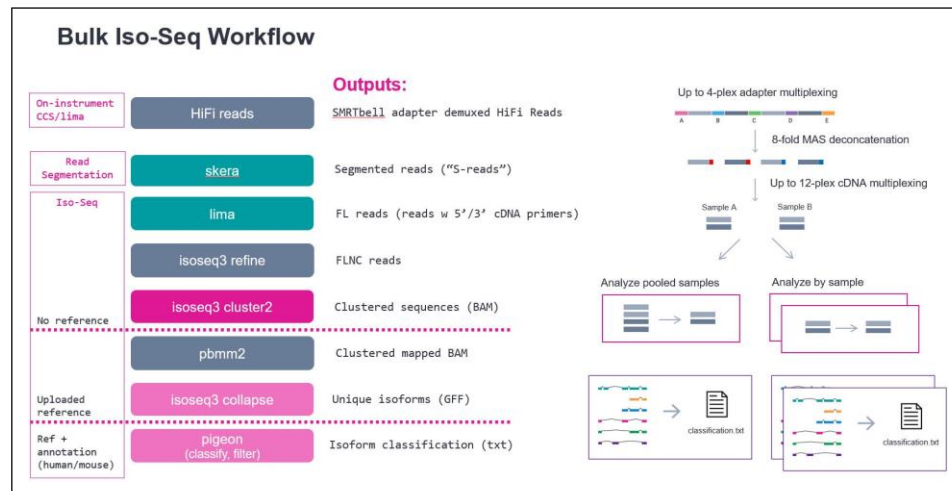
- Motif allele spans don't need to be multiples of motif length due to interruptions and other deviations from the sequence motif.
- Column 2 through 7 are the same information for allele 1.

- 
- **Sample summary:** CSV <prefix>-sample\_summary.csv contains the per sample metrics in the Target Enrichment Summary report.
  - **TRGT plots meth-allele:** Zipped archive  
<Prefix>-trgt\_meth\_allele.zip of plots for all targets for each sample, generated by the trgt plot tool (formerly TRVZ). This plot depicts consensus repeat alleles and reads aligning to them. Bases in repeats are colored by methylation levels.
  - **TRGT plots meth-waterfall:** Zipped archive  
<Prefix>-trgt\_meth\_waterfall.zip of plots for all targets for each sample, generated by the trgt plot tool (formerly TRVZ). This plot depicts portions of reads spanning the repeat **without** aligning them, which is convenient for showing mosaicism. Bases in repeats are colored by methylation levels.
  - **TRGT plots motifs-allele:** Zipped archive  
<Prefix>-trgt\_motifs\_allele.zip of plots for all targets for each sample, generated by the trgt plot tool (formerly TRVZ). This plot depicts consensus repeat alleles and reads aligning to them. Bases in repeats are colored by repeat motif.
  - **TRGT plots motifs-waterfall:** Zipped archive  
<Prefix>-trgt\_motifs\_waterfall.zip of plots for all targets for each sample, generated by the trgt plot tool (formerly TRVZ). This plot depicts portions of reads spanning the repeat **without** aligning them, which is convenient for showing mosaicism. Bases in repeats are colored by repeat motif.
  - **Mapped BAM:** Zipped archive  
<Prefix>-mapped\_sorted\_BAM.zip of BAM-format files and their index files for each sample. This file contains reads with any non-empty overlap with any locus in the input target BED file, plus/minus 3000 base pairs flanking each side. These are meant to be visualized using IGV or other visualization tool. There is **no** annotation on which locus it overlaps, and **all** BAM tags from PacBio tools (such as rq, barcodes, methylation on MM/ML tags, num passes, and so on) are carried over to this file.
  - **Input - Target BED file used in run:** User-supplied Target and repeat definitions BED file, <Prefix>PureTarget\_repeat\_expansion\_panel.bed.
  - **TRGT spanning BAM:** Zipped archive  
<Prefix>-trgt\_spanning\_bams.zip of BAM-format file generated by the trgt tool. This file includes **only** reads that were used to genotype alleles, and contains the tandem repeat sequence, plus/minus 50 base pairs flanking each side. Each read contains tags that annotate the read in reference to its genotype. For example, the TR tag shows the locus it was assigned to (such as CNBP) and the AL tag shows whether the read was assigned to a shorter allele (0) or longer (1). This file can be used as input when using **trgt plot** command-line tools for customizing visualizations.

## Read Segmentation and Iso-Seq® Analysis

The Read Segmentation and Iso-Seq Analysis application enables analysis and functional characterization of full-length transcript isoforms that were sequenced on PacBio instruments.

The application is for use when using concatenation-based library preparations such as the Kinnex full-length RNA libraries. (For a video on using the Kinnex full-length RNA kit, click [here](#).) If analyzing regular, unconcatenated Iso-Seq libraries, use the **Iso-Seq Analysis** workflow instead. Alternatively, HiFi reads from Kinnex full-length RNA libraries can first be deconcatenated using the **Read Segmentation** data utility, and the segmented reads (S-reads) can run through the Iso-Seq Analysis workflow.



The Read Segmentation and Iso-Seq Analysis application:

- HiFi reads should be generated using the **Kinnex full-length RNA library protocol**. If the library is regular Iso-Seq **without** MAS-Seq concatenation, use the Iso-Seq Analysis workflow instead.

Note on barcoded libraries:

- The Read Segmentation and Iso-Seq workflow will **only** process barcoded libraries at the cDNA level (such as using **Iso-Seq v2 Barcoded cDNA Primers** as part of the MAS-Seq for bulk Iso-Seq kit). Demultiplexing of barcoded adapters (also part of the **Kinnex full-length RNA kit**) should first run through the Demultiplexing workflow in SMRT Link.

Workflow

4. **Read Segmentation** data utility - Split arrayed HiFi reads at adapter positions, generating **segmented reads** (S-reads) which are the comprising fragments. For each input HiFi read, this step creates multiple

---

BAM records, one for each fragment. An arrayed HiFi read can contain many fragments.

5. **Iso-Seq Analysis (all use cases)** - Full-length, non-concatemer (FLNC) reads are then identified by the presence of cDNA primers and polyA tails. Reads are then clustered *de novo* at the isoform level to generate high-quality, full-length consensus isoform sequences. If **no** reference genome is selected, the workflow stops here.
6. **Iso-Seq Analysis, with reference genome only** – If a reference genome is provided, high-quality, full-length isoform sequences are mapped to the genome and collapsed into unique isoforms (GFF, FASTA output). If the reference bundle contains **only** a reference genome, the workflow stops here.
7. **Iso-Seq Analysis, reference genome + annotation (human/mouse only)**  
- If using the human or mouse reference and annotation package that is preloaded in SMRT Link, the collapsed isoforms are classified and filtered against reference annotation using pigeon, which is a transcript classification and filtering tool based on the [SQANTI3](#) software.

Segmentation Adapter Set (Default = MAS-Seq Adapter v3 (MAS8))

- 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 = Iso-Seq v2 Barcoded 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:** Iso-Seq v2 Barcoded cDNA Primers.

```
>IsoSeqX_bc01_5p
CTACACGACGCTCTTCCGATCTACTACACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc02_5p
CTACACGACGCTCTTCCGATCTACTAGTAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc03_5p
CTACACGACGCTCTTCCGATCTAGTGTACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc04_5p
CTACACGACGCTCTTCCGATCTATCACTAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc05_5p
CTACACGACGCTCTTCCGATCTCAGCTGTGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc06_5p
CTACACGACGCTCTTCCGATCTCAGTCACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc07_5p
CTACACGACGCTCTTCCGATCTCATGTATGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc08_5p
CTACACGACGCTCTTCCGATCTCGTATGTGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc09_5p
CTACACGACGCTCTTCCGATCTGACATGTGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc10_5p
CTACACGACGCTCTTCCGATCTGAGTCTAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc11_5p
CTACACGACGCTCTTCCGATCTGTAGATAGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_bc12_5p
CTACACGACGCTCTTCCGATCTGTATGACGCAATGAAGTCGCAGGGTTGGG

>IsoSeqX_3p
AAGCAGTGGTATCAACGCAGAGTAC
```

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 Genome hg38, with Gencode v39 annotations** and **Mouse Genome mm39, with Gencode vM28 annotations**.
- Alternatively, choose other reference genomes (but not with annotations) that were custom-uploaded to SMRT Link.
- The Reference Set can be left blank. If blank, the workflow will **stop** after the isoform clustering step (isoseq cluster).

Cluster of Barcoded Samples

- This option specifies barcoded samples that were barcoded at the cDNA level, where the (barcoded) cDNA primers are specified in the Primer Set option. This option does **not** address libraries that were barcoded using barcoded adapters.
- Specify whether all FLNC reads will be pooled for clustering, then demultiplexed based on pooled result. **Note:** This setting does **not** apply to non-barcoded samples.

- Specify **Pool reads and cluster together** 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 **Cluster reads separately** 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
Adapters FASTA	NONE	Specify a custom FASTA file containing segmentation adapters. If <b>not</b> specified, the adapters specified in the XML metadata are used. Adapters <b>must</b> 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. <b>Example:</b> >A AGCTTACTTGTGAAGA >B ACTTGTAAGCTGTCTA >C ACTCTGTCAGGTCCGA >D ACCTCCTCCTCCAGAA >E AACCGGACACACTTAG
Min. CCS Predicted Accuracy (Phred Scale)	20	Phred-scale integer QV cutoff for filtering HiFi reads. The default for <b>all</b> applications is 20 (QV 20), or 99% predicted accuracy.
Filters to Add to the Data Set	NONE	A semicolon-separated ( <b>not</b> comma-separated) list of other filters to add to the Data Set.
Require and Trim Poly(A) Tail	ON	<b>ON</b> means that polyA tails are required for a sequence to be considered full length. <b>OFF</b> 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. <b>Note:</b> This is applicable <b>only</b> 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. <b>Note:</b> This is applicable <b>only</b> 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. <b>Note:</b> This is applicable <b>only</b> if a reference genome is provided.
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 <b>cannot</b> be modified.
Advanced pigeon filter options	NONE	Space-separated list of custom pigeon filter options. Example: --min-cov N to reduce minimum coverage for low abundance isoforms (default value 3).



Advanced parameters	Default value	Description
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 <b>all</b> less than the provided value, they will <b>all</b> be collapsed. Setting to 0 requires all junctions to be exact to be collapsed into a single isoform. Applicable <b>only</b> if a reference genome is provided.
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

## Reports and data files

The Read Segmentation and 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 (concatenation factor):** 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 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 (FLNC Reads):** 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 FLNC 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.

#### 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.

#### 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.

#### Read Classification > Primer Quality Scores

- Histogram of primer scores.

#### Read Classification > Length of Full Non-Concatemer Reads

- Per-sample histograms 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.

#### Transcript Clustering > Length of Consensus Isoforms

- Per-sample histograms of the consensus isoform lengths and the distribution of isoforms exceeding a read length cutoff.

#### Transcript Mapping and Classification > Summary Metrics

- **Sample Name:** Sample name for which the following metrics apply.
- **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 isoforms:** The total number of unique isoforms across all cells.

- **Total unique isoforms, filtered:** The total number of unique isoforms across all cells, after filtering out reads based on the SQANTI transcript filtering criteria.

### Transcript Mapping and Classification > Transcript Classifications

- **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 <b>same</b> 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 <b>fewer</b> 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 <b>not</b> 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 <b>not</b> have a FSM or ISM match, and has at least <b>one</b> donor or acceptor site that is <b>not</b> annotated.
Antisense	The query isoform does <b>not</b> 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 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

- **Gene Saturation, all genes, filtered**
  - Saturation plot showing the level of gene saturation for all genes, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Gene Saturation, known genes only, filtered**
  - Saturation plot showing the level of gene saturation, for unique known genes only (genes annotated in the reference annotation) per cell, after filtering out reads based on the SQANTI transcript filtering criteria.

#### 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**.)
- **Collapsed Filtered Isoform Counts By Sample**: Report of read count information for each collapsed filtered isoform.
- **Non-passing reads, unaligned**: BAM file containing HiFi reads that did **not** generate S-reads.
- **Segmented Reads, passing, unaligned**: BAM file containing the generated S-reads that passed filtering.
- **Collapsed Filtered Isoforms (All Samples)**: 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, GFF (All Samples)**: Mapped, unique isoforms, in GFF 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.
- **Isoform Counts by Barcode (All Samples)**: For each isoform, report supportive FLNC reads for each barcode.
- **High-Quality Isoforms (All Samples)**: 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.

- 
- **Full-length Non-Concatemer Read Assignments (All Samples):** Report of full-length read association with collapsed filtered isoforms, in text format.
  - **Full-length Non-Concatemer Reads <Sample>:** Per-sample BAM file containing full-length non-concatemer reads.
  - **Mapped High-Quality Isoforms (All Samples):** BAM file containing Isoforms with high consensus accuracy that mapped to the reference sequence.
  - **Mapped High-Quality Isoforms (BAM Index) (All Samples):** BAM index file associated with the Mapped High-Quality Isoforms BAM file.
  - **Collapsed Filtered Isoform Counts (All Samples):** Report of read count information for each collapsed filtered isoform.
  - **Collapsed Filtered Isoforms Groups (All Samples):** Report of isoforms mapped into collapsed filtered isoforms.
  - **Isoform Counts by Barcode (All Samples):** For each isoform, report supportive FLNC reads for each barcode.
  - **Collapsed Filtered Isoforms GFF (All Samples):** Mapped, unique isoforms, in GFF format. This is the Mapping step output that is the recommended output file to work with.
  - **Cluster Report (All Samples):** Report of each full-length read into isoform clusters.
  - Pigeon summary, Pigeon filtered summary, and Pigeon filtering report files: txt and json files summarizing input transcripts, unique genes, unique transcripts, and isoform and read based classification summaries

#### Data > IGV Visualization Files

The following files are used for visualization using [IGV](#).

- **Non-passing reads, unaligned:** BAM file containing HiFi reads that did **not** generate S-reads.
- **Segmented Reads, passing, unaligned:** BAM file containing the generated S-reads that passed filtering.
- **Mapped High-Quality Isoforms (All Samples):** BAM file containing Isoforms with high consensus accuracy that mapped to the reference sequence.
- **Mapped High-Quality Isoforms (BAM Index) (All Samples):** BAM index file associated with the Mapped High-Quality Isoforms BAM file.

---

## 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 Kinnex™ libraries.

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

### 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.

### Segmentation Adapter Set (Default = MAS-Seq Adapter v1 (MAS16))

- 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 1:** The 10x Chromium single cell 3' cDNA primer set.

```
>5p
AAGCAGTGGTATCAACGCAGAGTACATGGG

>3p
AGATCGGAAGAGCGTCGTGTAG
```

**Example 2:** The 10x Chromium single cell 5' cDNA primer set.

```
>5p
CTACACGACGCTCTTCCGATCT

>3p
GTACTCTGCGTTGATACCACTGCTT
```

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 Genome hg38, with Gencode v39 annotations** and **Mouse Genome mm39, with Gencode vM28 annotations**.

Kit Type

- Specify the **10x 3' Kit**, or **10x 5' Kit**. This determines which set of 10x primers and barcode sequences to use, and also affects the UMI and single-cell barcode design settings.

## Parameters

Advanced parameters	Default value	Description
<b>10x Barcodes for 3' (text, gzipped)</b>	3M-february-2018.REVERSE.txt.gz	A gzipped text file containing known 10x single-cell barcodes, one per line, for the 10x 3' Kit. This include list specifies the barcodeset to which raw cell barcodes are remapped by minimum edit distance. The workflow chooses the appropriate barcodes for the kit type. <b>Note:</b> The barcodes are shown in reverse-complement compared to how it was provided in 10x's software pipeline. Additional barcode whitelists can be found <a href="#">here</a> .
<b>10x Barcodes for 5' (text, gzipped)</b>	737K_august_2016.txt.gz	A gzipped text file containing known 10x single-cell barcodes, one per line, for the 10x 5' Kit. This include list specifies the barcodeset to which raw cell barcodes are remapped by minimum edit distance. The workflow chooses the appropriate barcodes for the kit type. <b>Note:</b> The barcodes are shown in reverse-complement compared to how it was provided in 10x's software pipeline. Additional barcode whitelists can be found <a href="#">here</a> .

Advanced parameters	Default value	Description
<b>Adapters FASTA</b>	NONE	Specify a custom FASTA file containing segmentation adapters. If <b>not</b> specified, the adapters specified in the XML metadata are used. Adapters <b>must</b> 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. <b>Example:</b> >A AGCTTACTTGTGAAGA >B ACTTGTAAGCTGTCTA >C ACTCTGTCAGGTCCGA >D ACCTCCTCCTCCAGAA >E AACCGGACACACTTAG
<b>Single-cell barcode and UMI design for 3'</b>	T-12U-16B	Single-cell barcode and UMI design for 3' kits. The workflow chooses the appropriate barcodes for the kit type. <ul style="list-style-type: none"> <li>T indicates the transcript position and is <b>mandatory</b>. It is the anchor that determines whether tags are located on the 3' or 5' side.</li> <li>U, as in UMI, <b>must</b> be preceded by the length of the UMI.</li> <li>B, as in cell barcode, <b>must</b> be preceded by the length of the cell barcode.</li> </ul> <b>Example:</b> T-12U-16B indicates a 12bp UMI and 16bp cell barcode after the transcript (with polyA tail).
<b>Single-cell barcode and UMI design for 5'</b>	16B-10U-13X-T	Single-cell barcode and UMI design for 5' kits. The workflow chooses the appropriate barcodes for the kit type. <ul style="list-style-type: none"> <li>T indicates the transcript position and is <b>mandatory</b>. It is the anchor that determines whether tags are located on the 3' or 5' side.</li> <li>U, as in UMI, <b>must</b> be preceded by the length of the UMI.</li> <li>B, as in cell barcode, <b>must</b> be preceded by the length of the cell barcode.</li> <li>X must be preceded by the length of an extra sequence to clip, such as a TSO sequence.</li> </ul> <b>Example:</b> T-12U-16B indicates a 12bp UMI and 16bp cell barcode after the transcript (with polyA tail).
<b>Output prefix</b>	scisoseq	The output file name prefix for all non-matrix files. If <b>not</b> specified, the primary input classification file is used to determine the prefix. <b>Example:</b> my_isoforms_classification.txt inputs yields an output prefix of my_isoforms.
<b>Cell Barcode Finding Method</b>	knee	Select the knee or percentile method to determine the number of real cells based on the cell barcode ranking plot.
<b>Cell Barcode Percentile Cutoff</b>	99	An integer between 0-100 for the percentile cutoff if using the percentile cell barcode-finding method. (This value is <b>ignored</b> if using the knee cell barcode-finding method.)
<b>Advanced pigeon filter options</b>	NONE	Space-separated list of custom pigeon filter options. Example: --min-cov N to reduce minimum coverage for low abundance isoforms (default value 3).
<b>Advanced pigeon make-seurat options</b>	NONE	Space-separated list of customer pigeon make-seurat options. Example: --keep-ribo-mito-genes to turn off exclusion of ribosomal and mitochondrial gene outputs.
<b>Add task memory (MB)</b>	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.



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 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 (concatenation factor):** 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 <b>same</b> 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 <b>fewer</b> 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 <b>not</b> 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 <b>not</b> have a FSM or ISM match, and has at least <b>one</b> donor or acceptor site that is <b>not</b> annotated.
Antisense	The query isoform does <b>not</b> 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 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

- **Gene Saturation, all genes, filtered**
  - Saturation plot showing the level of gene saturation for all genes, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Gene Saturation, known genes only, filtered**
  - Saturation plot showing the level of gene saturation, for unique known genes only (genes annotated in the reference annotation) per cell, after filtering out reads based on the SQANTI transcript filtering criteria.

#### 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, tar-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.

#### Data > IGV Visualization Files

The following files are used for visualization using [IGV](#).

- **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.
  - **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.
  - **Collapsed transcript groups:** Text file containing the isoseq collapse output associating redundant transcripts into unique isoforms.
  - **Deduplicated transcripts:** FASTA file containing unmapped reads after cell barcode correction and UMI deduplication.

---

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:

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

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

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 Genome hg38, with Gencode v39 annotations** and **Mouse Genome mm39, with Gencode vM28 annotations**.

Kit Type

- Specify the **10X 3' Kit**, or **10X 5' Kit**. This determines which set of 10x primers and barcode sequences to use, and also affects the design settings.

## Parameters

Advanced parameters	Default value	Description
10x Barcodes for 3' (text, gzipped)	3M-february-2018.REVERSE.txt.gz	A gzipped text file containing known 10x single-cell barcodes, one per line, for the 10x 3' Kit. This include list specifies the barcodeset to which raw cell barcodes are remapped by minimum edit distance. The workflow chooses the appropriate barcodes for the kit type. <b>Note:</b> The barcodes are shown in reverse-complement compared to how it was provided in 10x's software pipeline. Additional barcode whitelists can be found <a href="#">here</a> .
10x Barcodes for 5' (text, gzipped)	737K_august_2016.txt.gz	A gzipped text file containing known 10x single-cell barcodes, one per line, for the 10x 5' Kit. This include list specifies the barcodeset to which raw cell barcodes are remapped by minimum edit distance. The workflow chooses the appropriate barcodes for the kit type. <b>Note:</b> The barcodes are shown in reverse-complement compared to how it was provided in 10x's software pipeline. Additional barcode whitelists can be found <a href="#">here</a> .
Output prefix	scisoseq	The output file name prefix for all non-matrix files. If <b>not</b> specified, the primary input classification file is used to determine the prefix. <b>Example:</b> my_isoforms_classification.txt inputs yields an output prefix of my_isoforms.
Single-cell barcode and UMI design for 3'	T-12U-16B	Single-cell barcode and UMI design for 3' kits. The workflow chooses the appropriate barcodes for the kit type. <ul style="list-style-type: none"> <li>T indicates the transcript position and is <b>mandatory</b>. It is the anchor that determines whether tags are located on the 3' or 5' side.</li> <li>U, as in UMI, <b>must</b> be preceded by the length of the UMI.</li> <li>B, as in cell barcode, <b>must</b> be preceded by the length of the cell barcode.</li> </ul> <b>Example:</b> T-12U-16B indicates a 12bp UMI and 16bp cell barcode after the transcript (with polyA tail).
Single-cell barcode and UMI design for 5'	16B-10U-13X-T	Single-cell barcode and UMI design for 5' kits. The workflow chooses the appropriate barcodes for the kit type. <ul style="list-style-type: none"> <li>T indicates the transcript position and is <b>mandatory</b>. It is the anchor that determines whether tags are located on the 3' or 5' side.</li> <li>U, as in UMI, <b>must</b> be preceded by the length of the UMI.</li> <li>B, as in cell barcode, <b>must</b> be preceded by the length of the cell barcode.</li> <li>X must be preceded by the length of an extra sequence to clip, such as a TSO sequence.</li> </ul> <b>Example:</b> 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 <b>ignored</b> if using the knee cell barcode-finding method.)
Advanced pigeon filter options	NONE	Space-separated list of custom pigeon filter options. Example: --min-cov N to reduce minimum coverage for low abundance isoforms (default value 3).
Advanced pigeon make-seurat options	NONE	Space-separated list of customer pigeon make-seurat options. Example: --keep-ribo-mito-genes to turn off exclusion of ribosomal and mitochondrial gene outputs.



<b>Add task memory (MB)</b>	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
<b>Compute Settings</b>	Select	<b>(Optional)</b> 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 <b>same</b> 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 <b>fewer</b> 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 <b>not</b> 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 <b>not</b> have a FSM or ISM match, and has at least <b>one</b> donor or acceptor site that is <b>not</b> annotated.
Antisense	The query isoform does <b>not</b> 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

- **Gene Saturation, all genes, filtered**
  - Saturation plot showing the level of gene saturation for all genes, after filtering out reads based on the SQANTI transcript filtering criteria.
- **Gene Saturation, known genes only, filtered**
  - Saturation plot showing the level of gene saturation, for unique known genes only (genes annotated in the reference annotation) per cell, after filtering out reads based on the SQANTI transcript filtering criteria.

## 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.

#### Data > IGV Visualization Files

The following files are used for visualization using [IGV](#).

- **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.

## Variant Calling

Use this application to identify single-nucleotide variants, short insertions and deletions, and structural variants for a single sample against a specific reference genome. Variants are automatically phased and haplotagged in the aligned BAM output file.

The application performs read mapping, structural variant-calling using pbsv, small variant calling using [DeepVariant](#), and phasing using [whatshap](#).

- In order to run this workflow with SNVs and indel calling your SMRT Link server must be configured with Singularity. Please see the [SMRT Link installation guide](#) for additional details: <https://www.pacb.com/support/software-downloads/>.

### Reference Set (Required)

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

### Call SNVs and INDELS

Specify if DeepVariant for SNV and INDEL calling should be run, or if only structural variant calling should be run. **Default: ON**

**Note:** Specifying ON requires that your SMRT Link server is configured with Singularity and a job management systems. Please see SMRT Link Install Guide for additional details.

### 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 <b>all</b> applications is 20 (QV 20), or 99% predicted accuracy.
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. Use alphanumeric characters, hyphens, or underscores only. Character limit: ≤40 characters.
Advanced pbmm2 Options	NONE	Space-separated list of custom pbmm2 options. Not all supported command-line options can be used, and HPC settings <b>cannot</b> be modified.
Minimum Length of Structural Variant (bp)	20	The minimum length of structural variants, in base pairs.
Advanced pbsv Options	NONE	Additional pbsv command-line arguments.
Minimum % of Reads that Support Variant (any one sample)	10	Ignore calls supported by <N% of reads in every sample.
Minimum Reads that Support Variant (any one sample)	3	Ignore calls supported by <N reads in every sample.

Advanced parameters	Default value	Description
Minimum Reads that Support Variant (total over all samples)	3	Ignore calls supported by <N reads total across samples.
Use GPU if available	OFF	Send GPU-accelerated tasks to an HPC queue that includes GPU resources. <b>Note:</b> This option requires additional setup by the SMRT Link Administrator, and is <b>not</b> supported on AWS or local compute backends.
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

### To launch a multi-sample analysis

1. Select **SMRT Analysis** from the Module menu.
2. Click + **Create New Job**.
3. Enter a **name** for the analysis.
4. Ensure that **Analysis** is selected as the workflow type.
5. Select all the Data Sets for all the input samples.
6. In the **Analysis of Multiple Data Sets** list, select **One Analysis for All Data Sets**.
7. Click **Next**.
8. Select **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 Variant Calling application generates the following reports:

#### Mapping Report > Summary Metrics

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

- **Mean Concordance (mapped):** The mean concordance of reads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).
- **Number of Alignments:** The number of alignments that mapped to the reference sequence.
- **Number of reads (total):** The total number of CCS reads in the sequence.
- **Number of reads (mapped):** The number of CCS reads that mapped to the reference sequence.



- **Number of reads (unmapped):** The number of CCS reads not mapped to the reference sequence.
- **Percentage of reads (mapped):** The percentage of CCS reads that mapped to the reference sequence.
- **Percentage of reads (unmapped):** The percentage of CCS reads not mapped to the reference sequence.
- **Number of Bases (mapped):** The number of CCS bases that mapped to the reference sequence.
- **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.
- **Read N50 (mapped):** The read length at which 50% of the mapped bases are in CCS reads longer than, or equal to, this value.
- **Read Length 95% (mapped):** The 95th percentile of read length of CCS reads that mapped to the reference sequence.
- **Read Length Max (mapped):** The maximum length of CCS reads that mapped to the reference sequence.

#### Mapping Report > Mapping Statistics Summary

- **Sample:** The sample name for which the following metrics apply.
- **Number of Reads (mapped):** The number of CCS reads that mapped to the reference sequence. This includes adapters.
- **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.
- **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 Bases (mapped):** The number of CCS bases that mapped to the reference sequence.
- **Mean Concordance (mapped):** The mean concordance of reads that mapped to the reference sequence. Concordance for alignment is defined as the number of matching bases over the number of alignment columns (match columns + mismatch columns + insertion columns + deletion columns).

#### Mapping Report > Mapped Read Length

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

#### Mapping Report > Mapped 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.

## Variant Report > Count by type

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

- **SNVs:** The count, total length (in base pairs), and heterozygous-to-homozygous ratio of **all** single-nucleotide variants.
- **Indels:** The count, total length (in base pairs), and heterozygous-to-homozygous ratio of **all** insertions and deletions.
- **SVs:** The count, total length (in base pairs), and heterozygous-to-homozygous ratio of **all** structural variants.
- **Total:** The count, total length (in base pairs), and heterozygous-to-homozygous ratio of **all** called variants.

## Variant Report > SV Type Count by Sample

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.

## Variant Report > SV 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.

#### Variant Report > SV 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  $\geq 1$  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**.)
- **Structural Variants VCF**: PBSV structural variant calls in Variant Call Format (VCF) format. (See [here](#) for details.)
- **Structural Variants VCF Index**: Index file associated with the Structural Variants VCF file.
- **DeepVariant Phased VCF**: DeepVariant variant calls in VCF format that were phased with WhatsHap.
- **DeepVariant Phased VCF Index**: Index file associated with the DeepVariant Phased VCF file.
- **Small Variant Statistics**: Statistics describing DeepVariant calls.
- **Phasing Statistics**: Statistics describing DeepVariant WhatsHap phasing.
- **Haplotagged Mapped BAM**: HiFi reads, aligned to the reference, tagged with HP (1 or 2, local haplotype) and PS (local haplotype block).
- **Haplotagged Mapped BAM Index**: Index file associated with the Haplotagged Mapped BAM file.
- **DeepVariant gVCF**: DeepVariant variant calls in Genomic Variant Call Format ([gVCF](#)) gVCF files are required as input for analyses that create a set of variants in a cohort of individuals, such as cohort merging or joint genotyping. Note that the calls in this file are **not** phased.

#### Data > IGV Visualization Files

The following files are used for visualization using [IGV](#).

- **DeepVariant Phased VCF**: DeepVariant variant calls in VCF format that were phased with WhatsHap.
- **DeepVariant Phased VCF Index**: Index file associated with the DeepVariant Phased VCF file.
- **Haplotagged Mapped BAM**: HiFi reads, aligned to the reference, tagged with HP (1 or 2, local haplotype) and PS (local haplotype block).
- **Haplotagged Mapped BAM Index**: Index file associated with the Haplotagged Mapped BAM file.
- **DeepVariant gVCF**: DeepVariant variant calls in Genomic Variant Call Format (gVCF). gVCF files are required as input for analyses that create a set of variants in a cohort of individuals, such as cohort merging or joint genotyping. Note that the calls in this file are **not** phased.
- **Structural Variants VCF**: PBSV structural variant calls in Variant Call Format (VCF) format. (See [here](#) for details.)

---

## Data utilities

Following are data processing utilities provided with SMRT Analysis v25.2. 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.

### Demultiplex Barcodes

- Separate reads by barcode.
- See [“Demultiplex Barcodes” on page 138](#) 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** utility to create BAM files **before** using this utility.
- See [“Export Reads” on page 144](#) for details.

### Mark PCR Duplicates

- Remove duplicate reads from a Data Set created using an ultra-low DNA sequencing protocol.
- See [“Mark PCR Duplicates” on page 146](#) for details.

### Read Segmentation

- Splits arrayed HiFi reads at adapter positions, generating segmented reads (S-reads) comprised of multiple fragments.
- See [“Read Segmentation” on page 148](#) for details.

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

- Trim PCR Adapters from a Data Set created using an ultra-low DNA sequencing library.
- See [“Trim Ultra-Low Adapters” on page 150](#) for details.

### Undo Demultiplexing

- Recreate the original BAM file **before** any demultiplexing processing was performed on the instrument or using the **Demultiplex Barcodes** utility.
- See [“Undo Demultiplexing” on page 152](#) for details.

---

## Demultiplex Barcodes

Use this utility to separate sequence reads by barcode.

- Barcoded or indexed 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.
- Barcode names must be  $\leq 40$  characters and only contain alphanumeric characters or underscores.
- 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.

Multiplexed method	Run Demultiplex Barcodes utility?
Not multiplexed	No
Barcoded adapters	Yes
Barcoded cDNA primer	No

#### Barcode Set (Required)

- Specify a barcode sequence file to separate the reads.

#### 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.

#### 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.
- Specify **No** to specify asymmetric designs where the barcodes are **different** on each end of the insert sequence.

#### Parameters

Advanced parameters	Default value	Description
Min. CCS Predicted Accuracy (Phred Scale)	-1	Phred-scale integer QV cutoff for filtering HiFi reads. The default for <b>all</b> applications is -1, which does <b>not</b> apply any QV filtering.
Minimum Barcode Score	-1	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 <b>not</b> included in the analysis. This affects the output BAM file and the output demultiplexed Data Set XML file. The default for <b>all</b> applications is -1 which does <b>not</b> apply any additional filtering based on barcode score.
Advanced lima Options	NONE	Space-separated list of custom lima options. Not all supported command-line options can be used, and HPC settings <b>cannot</b> be modified.

Advanced parameters	Default value	Description
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

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 (Gb):** The number of bases in HiFi sequence data reads that contain barcodes.
- **Unbarcoded HiFi Yield (Gb):** 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, Kb):** The mean read length of HiFi reads per barcode combination, in base pairs.
- **Unbarcoded HiFi Read Length (mean, Kb):** 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:** The barcode name.
- **Reads %:** The percent of reads out of the first 35,000 that are inferred to be assigned to the barcode combination.
- **Barcode score, mean:** The mean barcode score associated with the reads inferred to be associated with the barcode combination.

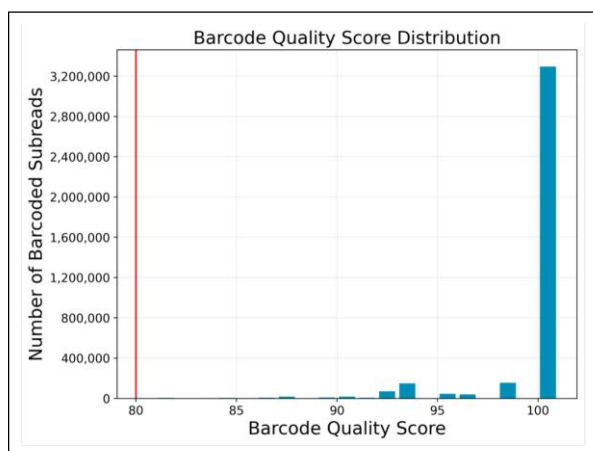
#### Barcodes > Barcoded Read Statistics

- **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  $\geq 95$ .





- **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 read 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  $\geq 10$  counts and mean scores  $\geq 45$ .
- **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.

- 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.

Output FASTQ File (Default = ON)

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

Output FASTA File (Default = ON)

- Outputs a single FASTA 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
<b>Filters to Add to the Data Set</b>	NONE	A semicolon-separated ( <b>not</b> comma-separated) list of other filters to add to the Data Set.
<b>Output file prefix</b>	NONE	The file name prefix for all output files. <b>Note:</b> The default prefix is automatically determined.
<b>Add task memory (MB)</b>	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
<b>Compute Settings</b>	Select	<b>(Optional)</b> 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 Data Set created using an Ampli-Fi DNA sequencing procedure.

**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. If analyzing Ampli-Fi libraries, prior to running this utility, Demultiplex Barcodes must be run to demultiplex and remove adapters first.

### 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 LBtag, which is set using the Well Sample Name field in the Runs module. By convention, different LBtags 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.
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
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.

### 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 <b>not</b> specified, the adapters specified in the XML metadata are used. Adapters <b>must</b> 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 <b>not</b> allowed. <b>Example:</b> >A AGCTTACTTGTGAAGA >B ACTTGTAAGCTGTCTA >C ACTCTGTCAGGTCCGA >D ACCTCCTCCTCCAGAA >E AACCGGACACACTTAG
Add task memory (MB)	0	Increasing this value allocates extra memory per task when submitting the job to the compute backend.
Compute Settings	Select	<b>(Optional)</b> 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 (concatenation factor):** 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 Data Set created using an ultra-low DNA sequencing library.

**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.

### 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 <b>all</b> applications is 20 (QV 20), or 99% predicted accuracy.
Compute Settings	Select	<b>(Optional)</b> 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.

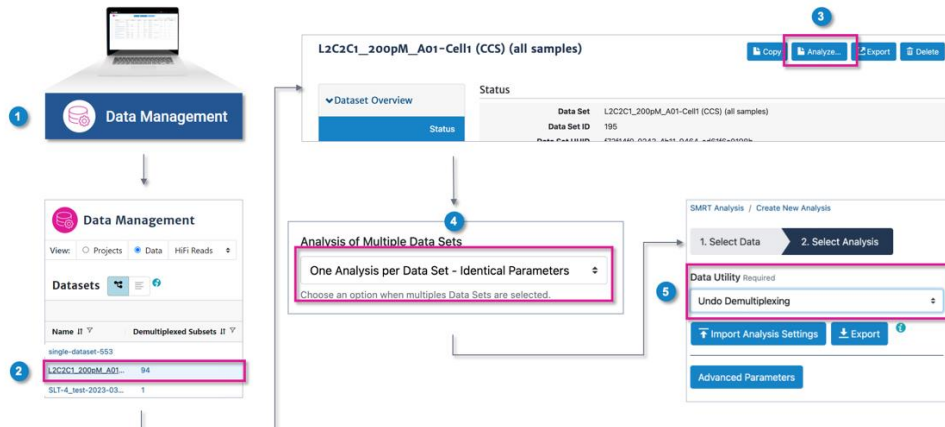


## Undo Demultiplexing

Use this utility to recreate the original BAM file **before** any demultiplexing processing was performed on the instrument or using the **Demultiplex Barcodes** data utility.

The utility accepts a set of demultiplexed BAM files (one per barcode) in which barcodes were removed from reads, and outputs a single BAM file with all reads and the barcode sequence added back to the ends of the reads.

To use Undo Demultiplexing feature, first select the desired (demultiplexed) data set in Data Management



### Barcode Set for re-analysis (Optional)

- Optionally specify barcode sequences to use **after** the original BAM file is recreated to generate an **Inferred Barcodes** report. (The report displays information about reads inferred to be assigned to the barcode combination.)

### Parameters

Advanced parameters	Default value	Description
Output File Prefix	all_samples_with_barcode	The output file name prefix for the Data Set containing the restored BAM file.
Compute Settings	Select	<b>(Optional)</b> Specify the distributed computing cluster settings configuration, if made available by the site SMRT Link administrator.

### Reports and data files

The Undo Demultiplexing utility generates a CCS Analysis data set report (see Data set report page 47) and the following reports.

#### Inferred Barcodes > Inferred Barcodes

- Barcode:** The barcode name.
- Reads, %:** The percent of reads out of the first 35,000 that are inferred to be assigned to the barcode combination.
- Barcode score, mean:** The mean barcode score associated with the reads inferred to be associated with the barcode combination.

---

## 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.**)
- **<Data Set> (CCS):** Data Set containing the original HiFi reads, with all reads and the barcode sequence added back to the ends of the reads.

---

## Secondary analysis output files

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

- SMRT analysis jobs are organized within directories named for the job ID in the `$$SMRT_ROOT/userdata` directory. Directory contents are organized by type of output or task within the workflow. Files may be symlinked so check the path when working with files.
- Every job result has a similar file structure.

### Job directories:

- `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.)

---

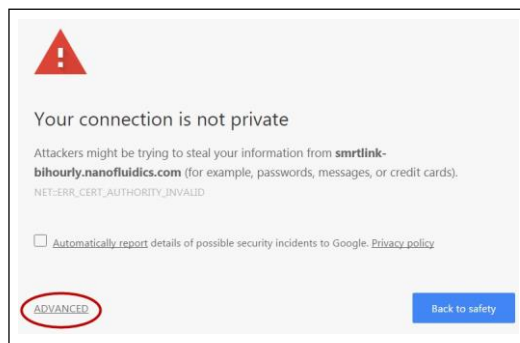
To download data files created by SMRT Link:

1. Select **SMRT Analysis** from the Module menu. 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 left of the file name to copy the file's path to the Clipboard.

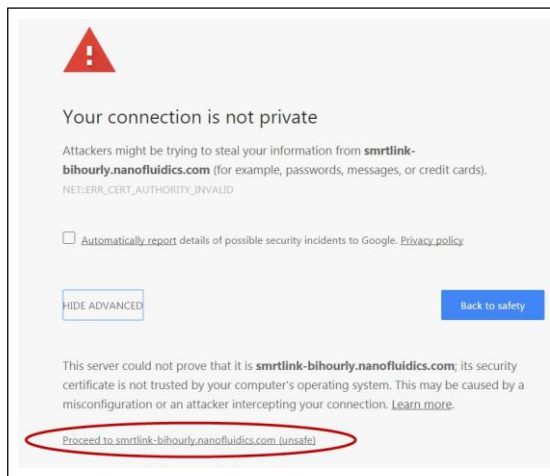
## Using the PacBio self-signed SSL certificate

SMRT Link v25.2 ships with a PacBio self-signed SSL certificate. If this is used at your site, security messages display when you try to login to SMRT Link for the **first time** using the Chrome browser. These messages may also display **other times** when accessing SMRT Link.

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



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**.

---

## User management and configuration

**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 Keycloak **without** LDAP integration.

- For details on integrating LDAP and SMRT Link, see the document [SMRT Link software installation guide \(v25.2\)](#).

**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 (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 \(v25.2\)](#).

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

## Adding and deleting SMRT Link users

1. Choose **Settings > 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**.

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:

## Assigning SMRT Link 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
Add/update instruments	Y	N	N
Access <b>Instruments</b> module	Y	Y	Y
Access <b>Sample Setup</b> module	Y	Y	N
Access <b>Runs</b> module	Y	Y	N
Access <b>Data Management</b> module	Y	Y	Y
Access <b>SMRT Analysis</b> module	Y	Y	Y

1. Choose **Settings > 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** an up-to-date version of the Google® Chrome web browser.

**Note:** SMRT Link **server** hardware and software requirement are listed in the document **SMRT Link software installation guide (v25.2)** at <https://www.pacb.com/support/software-downloads/>.

---

## Appendix A - PacBio terminology

### General terminology

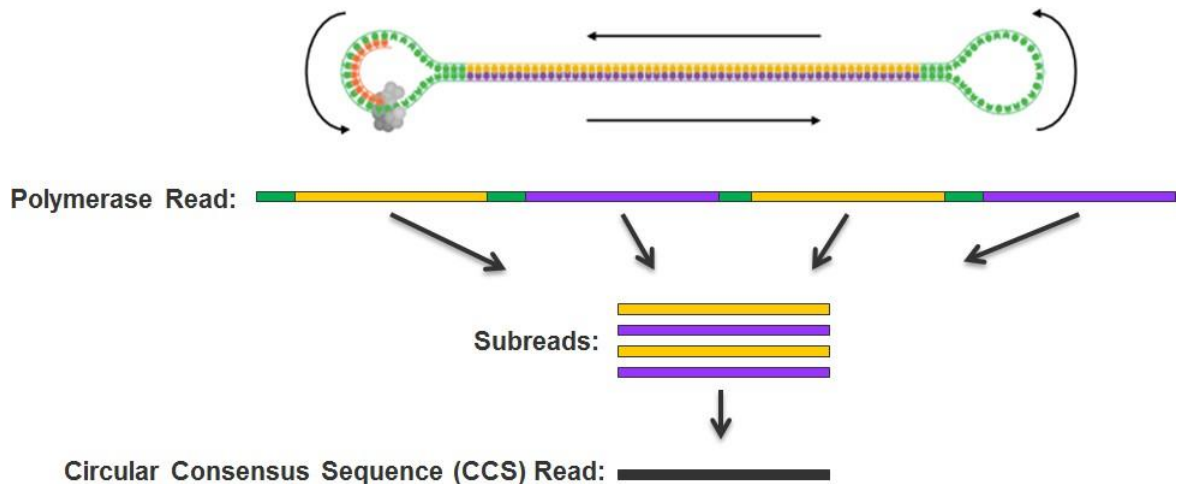
- **SMRT® Cell:** Consumable substrates comprising arrays of zero-mode waveguide nanostructures. SMRT Cells are used in conjunction with the DNA sequencing kit for on-instrument DNA sequencing.
- **SMRTbell® template:** A double-stranded DNA template capped by hairpin adapters (i.e., SMRTbell adapters) 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.

### 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 a full set of quality values and kinetic measurements.
- **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.
- **HiFi reads:** Reads generated with CCS analysis whose quality value is

---

equal to or greater than 20.



### 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.

### Accuracy terminology

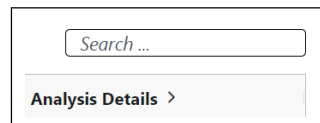
- **circular consensus accuracy:** Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.
- **consensus accuracy:** Accuracy based on aligning multiple sequencing reads together.
- **polymerase read quality:** A trained prediction of a read's 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, reference files, or target regions files.

To search the entire table

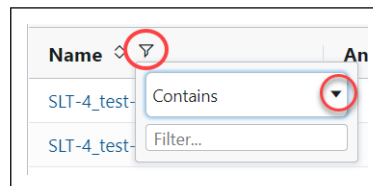
1. Enter a text query into the Search box. This searches **every field** in the table, and displays **all** table rows containing the search characters.



A search interface with a text input field containing the placeholder text "Search ...". Below the input field is a button labeled "Analysis Details >".

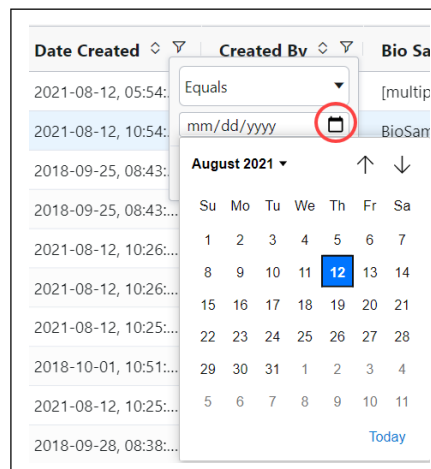
To search for a value within a column

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 drop-down menu and select another search operator. Different search operators are available, based on the column's data type.)



A table with columns "Name" and "An". The "Name" column has a filter icon (a funnel) circled in red. A dropdown menu is open, showing "Contains" and "Filter..." options. The "Contains" option is also circled in red.

- 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.



A table with columns "Date Created", "Created By", and "Bio Sa". The "Date Created" column has a filter icon (a funnel) circled in red. A dropdown menu is open, showing "Equals" and "mm/dd/yyyy" options. A calendar is displayed, showing the month of August 2021. The date "12" is highlighted in blue.

Numeric field operators

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

---

Text field operators

- Less than, Less than or equals
- In range

- Contains, Not contains
- Equals, Not equal
- Starts with, Ends with

Date field operators

- Equals, Not equal
- Greater than, Less than
- In range

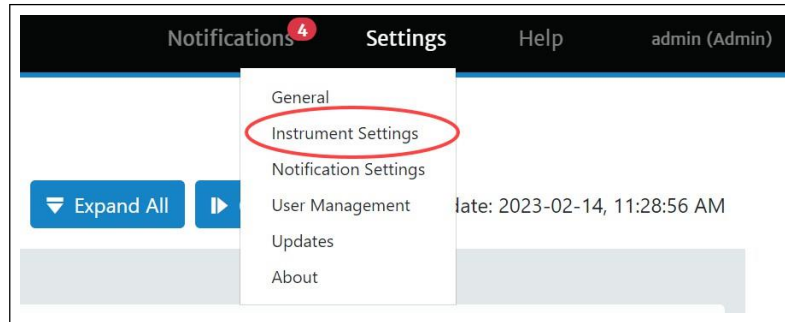
## Appendix C – Connecting Revio or Vega systems

Adding a system to SMRT

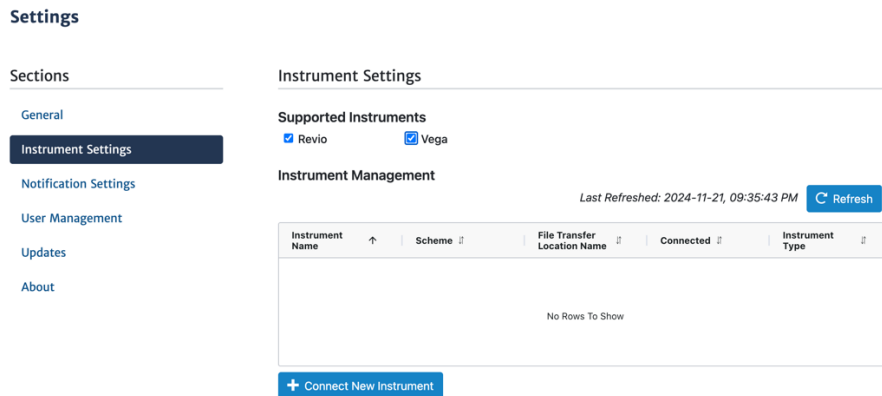
**Note:** The following procedure is available **only** for SMRT Link users whose role is **Admin**.

Link

1. Choose **Settings > Instrument Settings**.



2. Click **+ Connect New Instrument**.



3. Enter the **IP address** of the new Revio or Vega system. This is available on the instrument touchscreen; please write it down for later reference.

Connect new instrument ■ ■ ■ ■

Enter IP and secret key

Instrument IP Address Required

Instrument Secret Key Required

Cancel Continue

4. Enter the **Instrument Secret Key**. This is available on the instrument touchscreen, please write it down for later reference. Then, click **Continue**.
5. Enter the **name** of the new instrument, then click **Continue**.  
**Note:** The name must contain **only** alphanumeric characters, spaces, hyphens (-), underscores (\_), or apostrophes ('). In addition, the instrument name must be **unique** for a given SMRT Link installation.

6. Select a **File Transfer Location** - this is a network location where the sequencing data generated by the instrument will be transferred to. (These locations are defined using the + **New File Transfer Location** button. See [“Specifying a new file transfer location”](#) on page 173 for details.)

	Name ↑	Scheme ↓	destPath ↓
<input checked="" type="checkbox"/>	Revio s5	s5cmd	/collections

7. Click **Continue**.
8. Verify the information, click **Confirm**, and then click **Close**. The new instrument displays in the **Instrument Management** table.

**Note:** The following procedure is available **only** for SMRT Link users whose role is **Admin**.

Modifying an existing Revio or Vega system

1. Choose **Settings > Instrument Settings**.
2. In the **Instrument Management** table, click an existing instrument. The **Edit Instrument** dialog displays.

Edit Instrument  
 Name  
 Test\_Revio\_23  
 51 characters left  
 Select the File Transfer Location

Name	Scheme	destPath
<input checked="" type="checkbox"/> Revio s5	s5cmd	/collections

Apply Cancel

3. (Optional) Edit the instrument name. **Note:** The name must contain **only** alphanumeric characters, spaces, hyphens (-), underscores (\_), or apostrophes (').
4. (Optional) Select a different **File Transfer Location**. **Note:** If a run is currently in progress, the file transfer location will be updated **after** the run is completed.
5. Click **Apply**.

## Appendix D – Specifying a file transfer location

A **file transfer location** is your cloud or network location where sequencing data generated by the instrument will reside. (The instrument delivers data to the designated location **after** completion of sequencing and post- processing.) **After** you have defined a new file transfer location, it displays in the **File Transfer Location** table, and becomes available when adding new instruments.

There are five different transfer schemes that you can choose from to transfer sequencing data from the instrument to your network:

- **ssh (srs):** This method provides transfers to your network storage over an encrypted connection provided by SSH. Your FSE or Tech Support can provide the public key from the instrument, which must then be installed on the storage server by your network/IT administrator to allow transfer.
- **Amazon S3:** Provides secure file transfers between your Revio system and your Amazon S3 cloud storage bucket
- **Google Cloud Storage:** Provides secure file transfers between your Revio system and your Google Cloud Storage.
- **Microsoft Azure Blob storage:** Provides secure file transfers between your Revio system and your Microsoft Azure Blob storage.
- **S3-compatible storage:** Supports transfer for cloud-based systems that use an S3-compatible API, including Oracle Cloud Object Storage and Cloudflare R2.

Note, while setting up transfer schemes using the rsync daemon will continue to be available for current Revio customers currently using this method, this transfer scheme is deprecated and will not be supported in future software, as it does not provide in-transit data encryption. It is not available for Vega users.



---

## Setting up a file transfer location

1. Go to Settings > Instrument Settings
2. Click + New File Transfer Location.
3. Select the desired Scheme
4. Fill in the required fields.
  - Name: User-specified text string that displays in the Data directory dialog to identify the transfer scheme.
  - Description: User-specified text string that describes the transfer scheme.
  - Scheme-specific fields (see below).
5. Click Save.
6. Once the transfer scheme is associated with a connected instrument, click Test settings to ensure that the scheme works.

### ssh (srs) scheme specific fields

- Host: DNS name or IP address of your storage server. This may be the SMRT Link server or another storage location on the network. The name of this system can be obtained from your system administrator. Example: mp-srs.
- Destination path: File system location that contains all data transferred via srs.
- (Optional) Relative path: Path used to place run data in a specific sub- directory underneath the location specified in Destination path, on a per-instrument basis. A common value for this field is the instrument serial number or name. This field can contain only alphanumeric, "-", "\_", and "/". This field allows separation of run data from different instruments, which allows for easier location of particular run data when browsing the file system.
- Username: Name of the service account used for transferring datasets to the remote file server.
- SSH Key: Full path to the SSH private key. The SSH key must be manually installed on your instrument server by your FSE or Tech Support.)

### Amazon S3 scheme specific fields and requirements

- Bucket: Bucket name without the "s3://" prefix. Example: mp-s3
- Region: Region where the bucket is hosted. To find your region run

```
curl --silent --head https://s3.amazonaws.com/<your-bucket> |
grep 'x-amz-bucket-region'
```
- (Optional) Path: Path used to place run data in a specific sub- directory within the specified bucket, This field can contain only alphanumeric, "-", "\_", and "/". This field allows separation of run data from different instruments, which allows for easier location of particular run data when browsing the file system.
- Access key ID.
- Secret access key

### S3 Bucket configuration

1. Create a private S3 bucket, e.g. `mylabname-pacbio-instrument-data` in your preferred geographical region.
2. Create an IAM policy `pacbio-instrument-file-transfer` granting the necessary permissions on the bucket, including the ability to read, write, and delete objects.

- 
3. Create an IAM user with the new policy attached.
  4. Create static access credentials for the new IAM user and save them for step 5.
  5. Log in to SMRT Link as an admin user, navigate to Instrument Settings, and create a new Amazon S3 transfer scheme with the bucket name and region from step 1, and access credentials from step 4.

Example IAM policy, where `$BUCKET` is replaced by the actual destination bucket name (e.g. `mylabname-pacbio-instrument-data`).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor0",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject",
        "s3:GetObject",
        "s3:GetBucketWebsite",
        "s3:GetObjectVersionTagging",
        "s3:ListBucketVersions",
        "s3:GetObjectAttributes",
        "s3:GetObjectTagging",
        "s3:ListBucket",
        "s3:GetBucketVersioning",
        "s3:GetObjectVersionAttributes",
        "s3:GetObjectVersion",
        "s3>DeleteObject"
      ],
      "Resource": [
        "arn:aws:s3:::$BUCKET",
        "arn:aws:s3:::$BUCKET/*"
      ]
    }
  ]
}
```

---

### Google Cloud Storage specific fields

- Bucket: Bucket name with no “gs://” prefix
- Path: /path/to/subfolder in bucket to which to post
- Access key: Create HMAC key in GCP console (see <https://cloud.google.com/storage/docs/authentication/managing-hmackeys#console>)
- Secret key: Create HMAC key in GCP console (see <https://cloud.google.com/storage/docs/authentication/managing-hmackeys#console>)

### Microsoft Azure Blob Storage specific fields

- Account name: Account name, for example <account-name> from <https://<account-name>.blob.core.windows.net>
- Container: The container name from <https://<account-name>.blob.core.windows.net/<container>>
- Path: /path/to/subfolder in container
- Account key: Key associated with account

### S3-compatible storage specific fields

Note: Follow S3 Bucket configuration requirements when setting up S3 compatible storage.

- Endpoint: URL for object storage endpoint, for example <https://storage.googleapis.com>
- Bucket: Bucket name with no prefix, for example /bucketname .
- Region (Optional): Region name for cloud-based services.
- Path (Optional): Path used to place run data in a specific sub- directory underneath the location specified by your Bucket, on a per-instrument basis. A common value for this field is the instrument serial number or name. This field can contain only alphanumeric, “-”, “\_”, and “/”. This field allows separation of run data from different instruments, which allows for easier location of particular run data when browsing the file system.
- Access key: Access key for cloud-based services.
- Secret key: Secret key for cloud-based servicesfor

---

## Appendix E - BED file format Target Enrichment and HiFi Mapping applications

The **Target Enrichment and PureTarget repeat expansion** application requires an input BED file to specify the target genes or regions for analysis.

The **HiFi Mapping application** includes an optional **Target Regions** report that displays the number (and percentage) of reads that hit specified target regions.

The BED file needed in both cases 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. Region Name: The name of the region. Must be in the format NAME=VALUE, for example ID=GENE1. This column is required for the Target Enrichment and PureTarget repeat expansion workflows. If no region name is provided NA will be used.

**Example:** lambda\_NE3011 15000 25000 ID=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 F - 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 52](#) for details on exporting Data Sets.
- The other fields and plots in this report are described in the appropriate Reports sections of [“Analysis applications” on page 63](#).
- **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.