Sequel® II and IIe Systems: Data Files (v10.1)

Introduction

This document is for Customer IT or SMRT® Link Administrators, and describes the data files generated by Sequel II Systems and Sequel IIe Systems, and how to work with those files.

PacBio Read Files Format Description

PacBio uses unaligned BAM files as the native format to store read information.

hifi_reads.bam

`hifi_reads.bam` files contains PacBio HiFi Reads (≥QV 20) and can be used **directly** as input for PacBio and third-party analysis tools designed to work with HiFi Reads. Additional filtering for higher read quality can be applied using the `rq` tag.

The typical size for Sequel II and Sequel IIe Systems `hifi_reads.bam` files is <50 GB. If kinetic information is optionally included, files can be 5 times larger. More information about the PacBio BAM format can be found here.

hifi_reads.fastq

`hifi_reads.fastq` files includes the same reads as `hifi_reads.bam` files, but contain less information about individual reads. `hifi_reads.bam` files can be used **directly** as input for PacBio and third-party analysis tools designed to work with HiFi Reads.

The typical size for Sequel II and Sequel IIe Systems `hifi_reads.fastq` file (gzipped) is <50 GB. More information about the PacBio FASTQ quality encoding can be found here.

reads.bam

`reads.bam` files contains one read per productive ZMW and consist of **both** HiFi Reads (≥QV 20) **and** non-HiFi reads (<QV 20). It is the native output file of the Sequel IIe System when running on-instrument CCS analysis. A `reads.bam` file is also generated when running CCS analysis in SMRT Link v10.0 or v10.1.

The typical size for Sequel II and Sequel IIe Systems `reads.bam` file is in the range of 50 GB. If kinetic information is optionally included, files can be 5 times larger.

subreads.bam

`subreads.bam` files are the native output data file of the Sequel System and the Sequel II Systems. `subreads.bam` files are also produced by the Sequel IIe System if users choose to skip on-instrument CCS analysis.

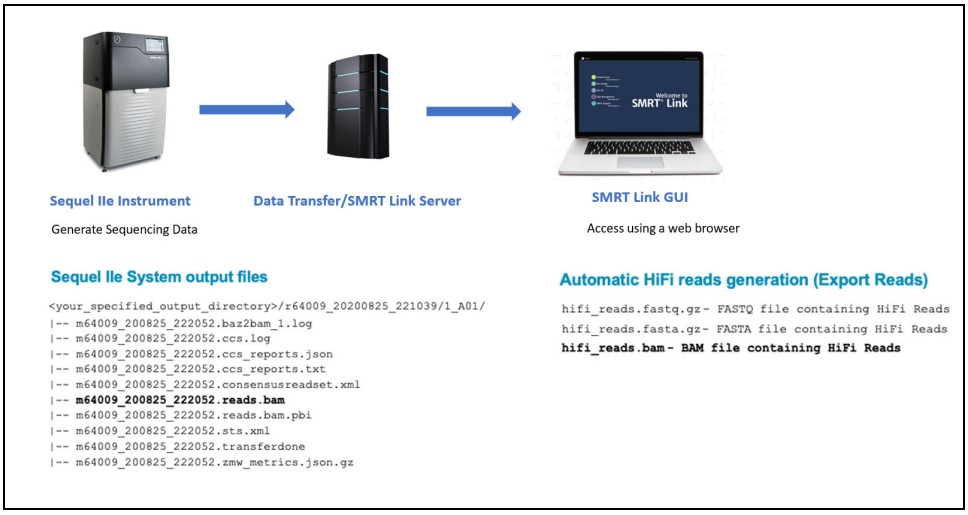
`subreads.bam` files contain the individual sequencing passes (subreads) from every productive ZMW. Subreads and HiFi Reads have different error

models, and subreads should **not** be used in HiFi Read applications or vice versa.

Notes:

- **All** PacBio read files are accompanied by a `*.pbi` index file, and a `*.xml` Data Set file.
- SMRT® Cell Data Sets transferred from Sequel II Systems also include additional files.
- The `subreads.bam` file size can range from 0.5 TB to 1.5 TB.

Data Flow From the Sequel IIe System to SMRT Link



Sequel IIe System Output Files

The run directory output by the Sequel IIe System includes a subdirectory for each collection (SMRT Cell) associated with a sample well. In the above example figure, `m64009_200825_222052` is the movie ID, including the instrument number (`64009`), date, and time. The collection subdirectory includes the following output files:

- `baz2bam_1.log`: Log file for post-primary analysis processing.
- `ccs.log`: Log file from CCS Analysis. This file is used internally for debugging and performance tracking by PacBio.
- `ccs_reports.json`, `ccs_reports.txt`: Contains processing metrics summarizing how many ZMWs generated HiFi Reads, and how many ZMWs failed CCS Reads generation. These files contain the same information and are used internally by PacBio Technical Support.
- `consensusreadset.xml`: This file is needed to import data into SMRT Link.
- `sts.xml`: Contains summary statistics about the collection and its post-processing.
- `transferdone`: Contains a list of files successfully transferred.
- `zmw_metrics.json.gz`: Contains processing information used to generate RunQC plots.

- `reads.bam.pbi`: Provides backwards-compatibility with the APIs enabled for accessing the `cmp.h5` file.
- `reads.bam`: The Sequel IIe System outputs one `reads.bam` file per collection, containing one read per productive ZMW. This is the central read data file. The file includes:
  - HiFi Reads (QV 20 or higher)
  - Lower-quality but still polished consensus reads (QV 1 - QV 20)
  - Unpolished consensus reads (RQ=-1)
  - 0- or 1-pass subreads unaltered (RQ=-1)

Note: The `reads.bam` contains HiFi Reads **and** non-HiFi reads.

The BAM format is a binary, compressed, record-oriented container format for raw or aligned sequence reads. BAM files produced by all Sequel Systems are fully compatible with the BAM specification. PacBio BAM files are unaligned reads. More information about the PacBio BAM format can be found here.

Note: If CCS Analysis is run on the Sequel IIe System, the `subreads.bam`, `scraps.bam` and `scraps.bam.pbi` files are **no longer** generated or available. If CCS Analysis is run in SMRT Link, Sequel IIe System instrument output includes the `subreads.bam` file, and optionally, the `scraps.bam` and `scraps.bam.pbi` files.

Reads.bam Versus HiFi Reads

Once the `reads.bam` file is transferred from the Sequel IIe System, SMRT Link **automatically** generates files containing only HiFi Reads, using the **Export Reads** application in SMRT® Analysis. The following HiFi data files are **always** generated by default:

- `hifi_reads.fastq.gz` - FASTQ file containing HiFi Reads.
- `hifi_reads.fasta.gz` - FASTA file containing HiFi Reads.
- `hifi_reads.bam` - BAM file containing HiFi Reads.

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

Input Data File Requirements for SMRT Link Analysis Applications and Third-Party Tools

All SMRT Link GUI applications (as well as `pbcromwell` if working on the command line) accept as input the `consensusreadsset.xml` file that points to the `reads.bam` file. The analysis pipeline then filters the `reads.bam` file to use **only** HiFi Reads, with the exception of Iso-Seq analysis, which takes all reads at or above Q10. (See “Input for the Iso-Seq Analysis Application” on page 7 for more information about Iso-seq analysis with HiFi Reads input.) As a result, **no** manual Data Set filtering needs to be applied in SMRT Link.

This is conceptually different in SMRT Link v10.0 and later versions from previous versions. In previous SMRT Link versions, a filter for minimum QV and minimum number of passes was applied at the CCS-generation

step. These two parameters are **not** required as input in SMRT Link v10.1 and they are available as advanced parameters in SMRT Analysis.

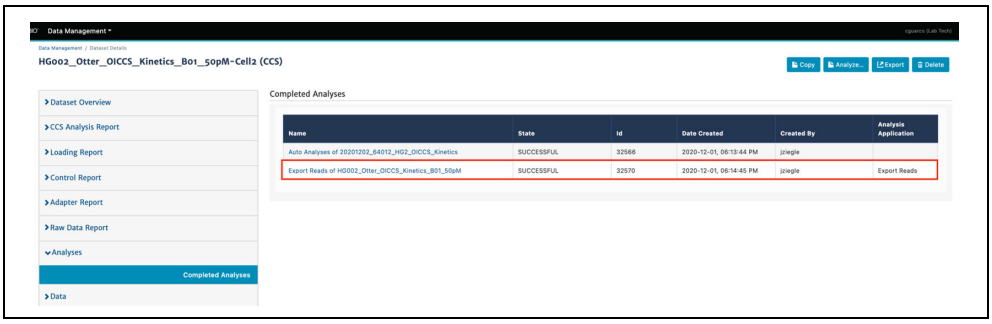
For analysis of Data Sets with tools outside of SMRT Link, PacBio **strongly** recommends that you use the `hifi_reads.bam` or `hifi_reads.fastq` files - **not** the `reads.bam` file.

Finding the Sequel IIe System HiFi Files

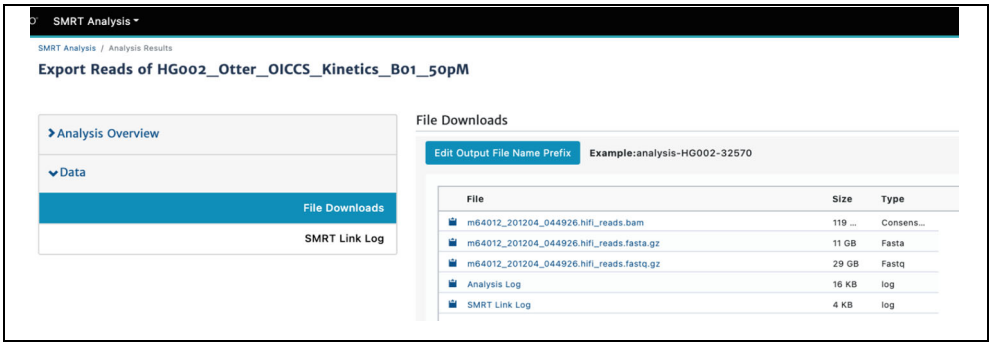
Using the SMRT Link GUI

To access the HiFi Read files generated from the `reads.bam` file:

1. Select **Data Management** and click on the desired Data Set Name.
2. On the **Dataset Details** page, click **Analyses > Completed Analyses**.
3. Click the name of a completed analysis to access the results of the **Export Reads** Analysis Application:



4. Click **Data > File Downloads** and locate and download the `hifi_reads.bam`, `hifi_reads.fasta.gz`, and `hifi_reads.fastq.gz` files.



Using the File System

1. Select SMRT Analysis, then click an analysis.
2. Select **Analysis Overview > Status**.
3. The **Status** section displays the file path, which points to the location of the output data files (including HiFi Reads) for the **Export Reads** analysis within the file system directory. The files `hifi_reads.bam`, `hifi_reads.fasta.gz`, and `hifi_reads.fastq.gz` can be found within this output directory.

Manually Generating HiFi Reads Files from the Sequel IIe System reads.bam File

If the **Export Reads** analysis application did **not** run automatically, you can run this application manually in SMRT Link. First go to **Data Management > Dataset Details** and click **Analyses > Completed Analyses** to determine if any Export Reads analysis application job has already been completed for your Data Set.

If **no** completed **Export Reads** analysis results are listed, follow the steps below to run the **Export Reads** analysis application for the Data Set of interest:

1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
3. Click **+ Create New Analysis**.
4. Enter a name for the analysis.
5. Select the type of data to use for the analysis: **HiFi Reads**.
6. In the Data Sets table select the Data Set to export to HiFi.
7. Click **Next**.
8. Select the **Export Reads** analysis application from the dropdown list.
9. Fill in the required parameters: **Output FASTA File** (ON or OFF), **Output BAM file** (ON on OFF), **Min CCS Predicted Accuracy** (Default QV 20.)
10. Click **Start**.

Extracting HiFi Reads from the Sequel IIe System reads.bam File Using the Command Line

The **Export Reads** analysis application in the SMRT Link GUI has its command line-version counterpart in our developmental repository at `pbbioconda/GitHub: extracthifi`.

The `extracthifi` tool extracts HiFi Reads (≥ Q20) from the full CCS `reads.bam` output. For more information, see https://github.com/PacificBiosciences/extracthifi/.

To use `extracthifi`, follow the installation instructions in the `pbbioconda` GitHub home page: https://github.com/PacificBiosciences/pbbioconda.

Usage:

```
extracthifi [options] <input.bam> <output.bam>
input.bam   STR  Input CCS BAM.
output.bam  STR  Ouput HiFi BAM.
```

Options:

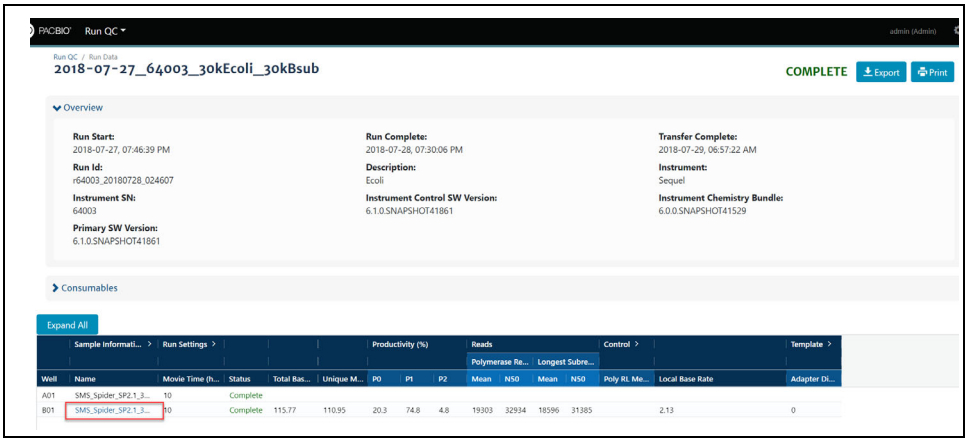
```
-h,--help         Show this help and exit.
--version         Show application version and exit.
```

Accessing Sequel IIe Raw Output (reads.bam) Files

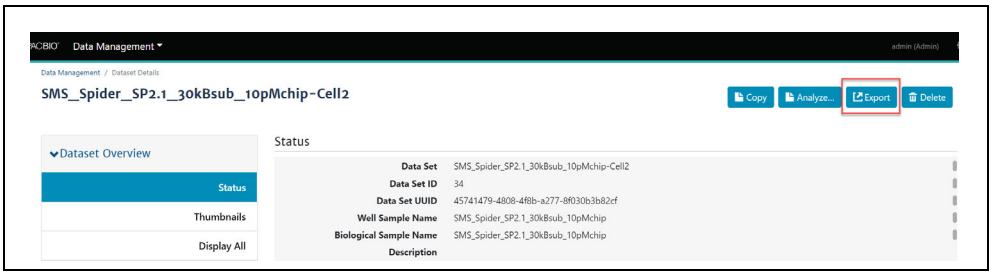
The SMRT Link GUI creates a report by default anytime it performs any action over data. A report is created for data transfer from the Sequel IIe instrument to SMRT Link, and another report is created once the **Export Reads** analysis application is run to generate HiFi-only data files as previously described.

To download the `reads.bam` file from SMRT Link, use the **Export Dataset** function:

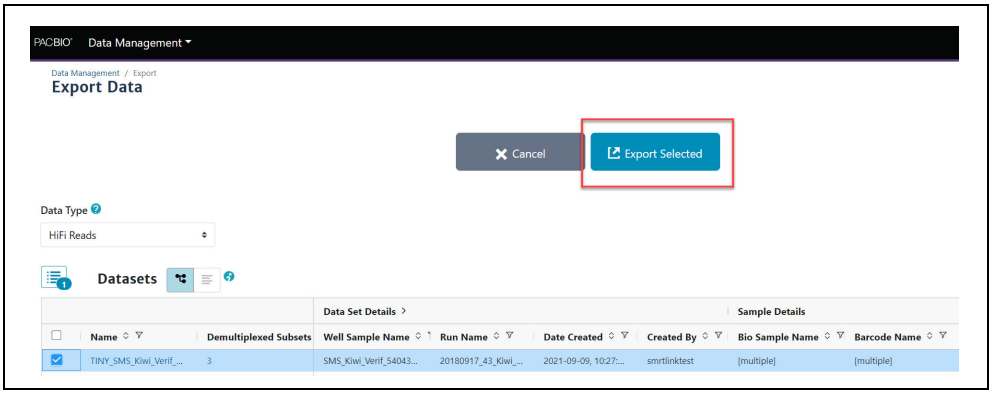
1. Access SMRT Link using the Chrome web browser.
2. Select **Run QC**.
3. Click the name of the desired run in the table; the **Run Data** page displays.



4. Click a sample name; the **Dataset Details** page displays.

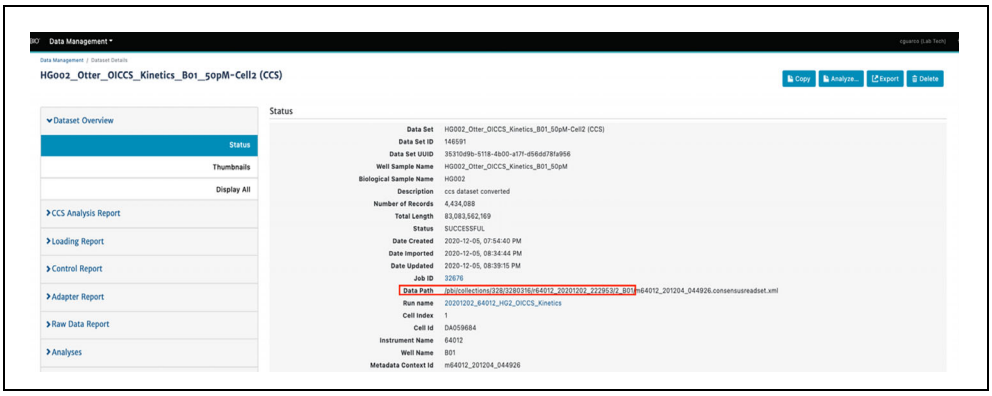


5. Click **Export** to export the data.
6. Another way to export data is by using **Data Management > Export Data > Export Selected**.



Downloading the reads.bam File using the Command Line

1. From the **Data Management > Dataset Details** page, go to the **Overview > Status** page and look for the **Data Path** line:



Using the **Data Path** line, find the path to the `reads.bam` file in the set of transferred files from the instrument. For the example data set shown above, the relevant directory path is:

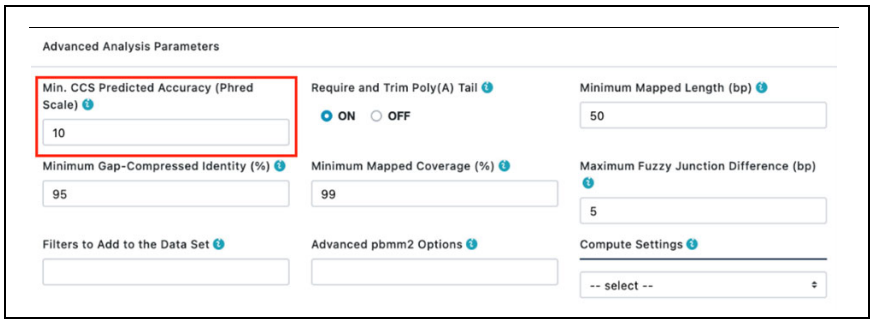
```
$ ls /pbi/collections/328/3280316/r64012_20201202_222953/2_B01/

m64012_201204_044926.baz2bam_1.log
m64012_201204_044926.ccs.log
m64012_201204_044926.ccs_reports.json
m64012_201204_044926.ccs_reports.txt
m64012_201204_044926.consensusreadset.xml
m64012_201204_044926.reads.bam
m64012_201204_044926.reads.bam.pbi
m64012_201204_044926.sts.xml
m64012_201204_044926.transferdone
m64012_201204_044926.zmw_metrics.json.gz
tmp-file-4d910bcf-760a-4be3-86b0-128714ee409c.txt
```

Input for the Iso-Seq Analysis Application

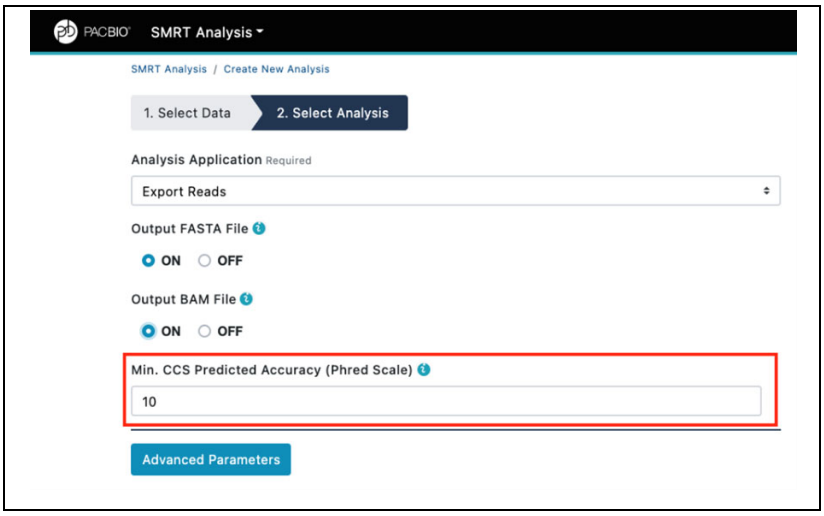
The Iso-Seq Analysis application does use some non-HiFi Reads as input, and will take reads at or above QV 10 as this can increase the sensitivity of the analysis. This QV 10 threshold is used as the default setting for Iso-Seq analyses and can be adjusted in the **Advanced Parameters** dialog of the Iso-Seq Analysis application. To access the dialog:

1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
3. Click **+ Create New Analysis**.
4. Enter a **name** for the analysis.
5. Select the type of data to use for the analysis: **HiFi Reads**.
6. In the **Data Sets** table, select one or more sets of data to be analyzed.
7. Click **Next**.
8. Select the **Iso-Seq Analysis** application from the dropdown list.

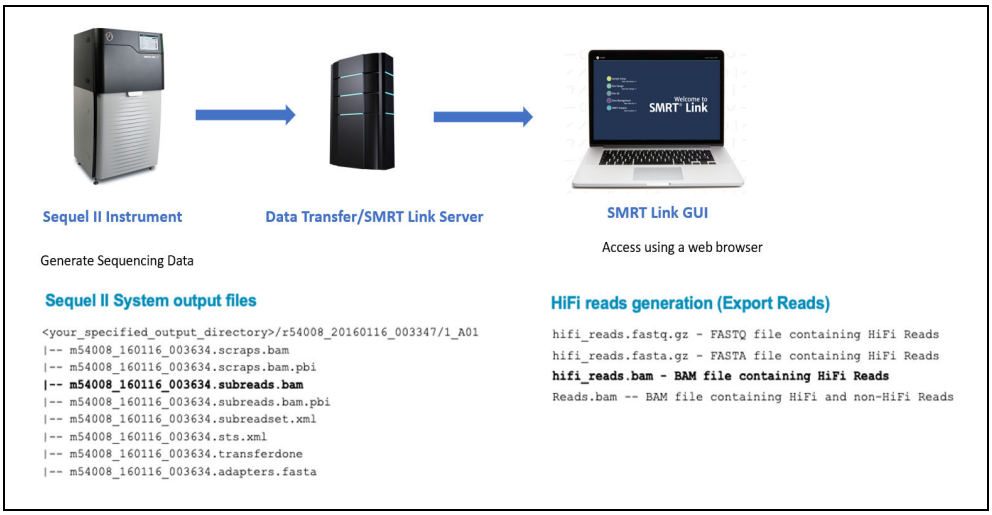
9. Fill in the relevant parameters, then click **Advanced Parameters**.



To obtain/share a BAM file that contains the **same** content as the input for the Iso-Seq Analysis application, we recommend using the **Export Reads** application and manually adjusting the **Min. CCS Predicted Accuracy (Phred Scale)** parameter to 10 as shown below:



Data Flow From the Sequel II Systems Without On-Instrument CCS to SMRT Link

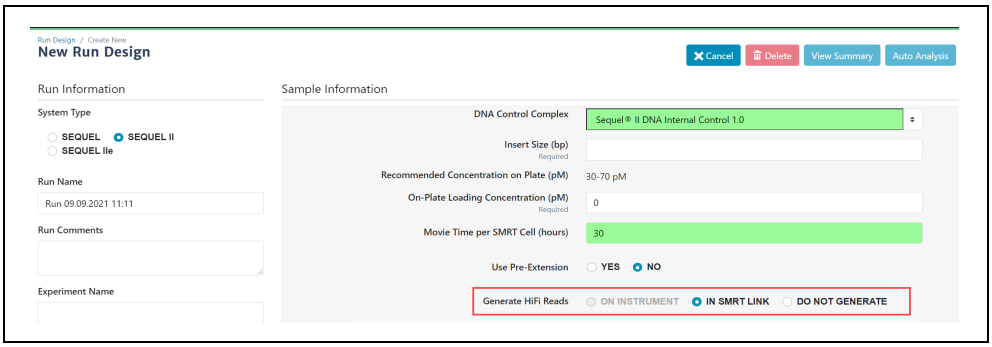


Generating HiFi Reads in SMRT Link for the Sequel II System

Starting a CCS Analysis in the SMRT Link GUI to Generate HiFi Reads

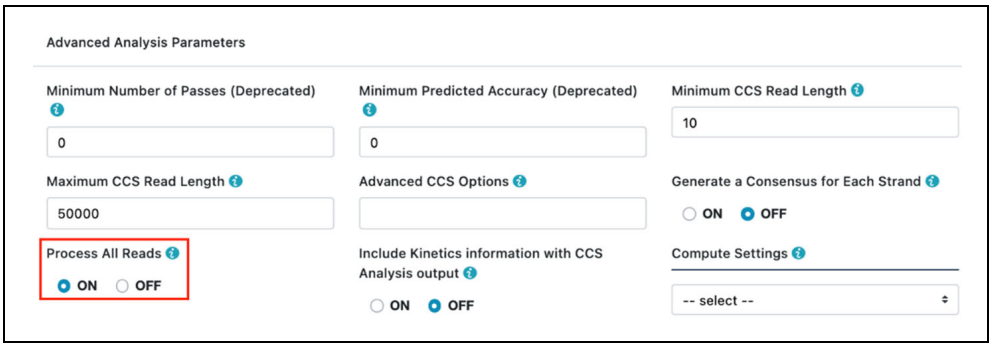
When running the CCS application in SMRT Link v10.1, whether automatically set up from Run Design or manually set up from the SMRT Analysis module, the reports and output files are the **same** as those produced by Sequel IIe instruments and on-instrument CCS analysis.

1. Generate HiFi Reads **automatically** by creating a CCS Pre Analysis Job in Run Design. (To learn more, see the SMRT Link User Guide Run Design section.) To create a CCS Pre Analysis job for a new Run Design, select **Run Design > Create a New Design**. The Run Design UI displays two options related to generation of HiFi Reads for Sequel II System users:
   - **In SMRT Link**: HiFi Reads are automatically generated after transfer to the compute cluster where SMRT Link is installed.
   - **Do Not Generate**: HiFi Reads are **not** generated automatically for this run. Only subread data are transferred to the local compute cluster where SMRT Link is installed. HiFi Reads may be manually generated later at the user's election.



Generating HiFi Reads Manually from the SMRT Analysis Module in SMRT Link

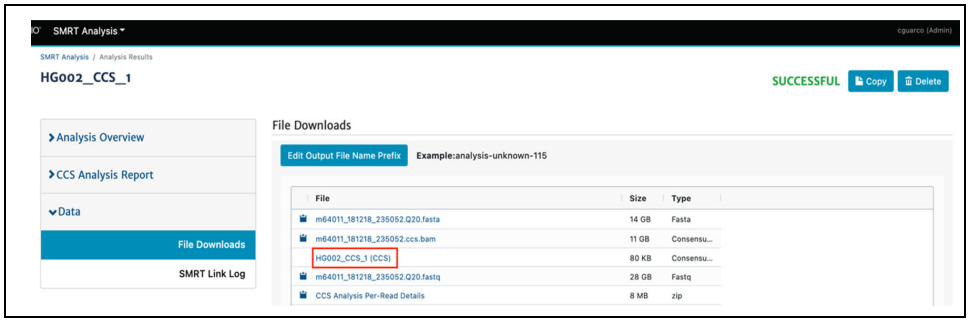
1. Access SMRT Link using the Chrome web browser.
2. Select **SMRT Analysis**.
3. Click **+ Create New Analysis**.
4. Enter a name for the analysis.
5. Select the type of data to use: **Continuous Long Reads**. The Data Sets table displays the corresponding Data Sets available for analysis.
6. In the Data Sets table, select one or more sets of data to be analyzed. (For multiple selection see the SMRT Link User Guide for instructions.)
7. Click **Next**.
8. Select the **Circular Consensus Sequencing (CCS)** application from the dropdown list.
9. Click **Advanced Parameters** and verify that **Process All Reads** is set to ON. This option creates a `reads.bam` file containing HiFi Reads (≥QV 20) and non-HiFi reads (<QV 20) as produced by on-instrument CCS Analysis on the Sequel IIe System.



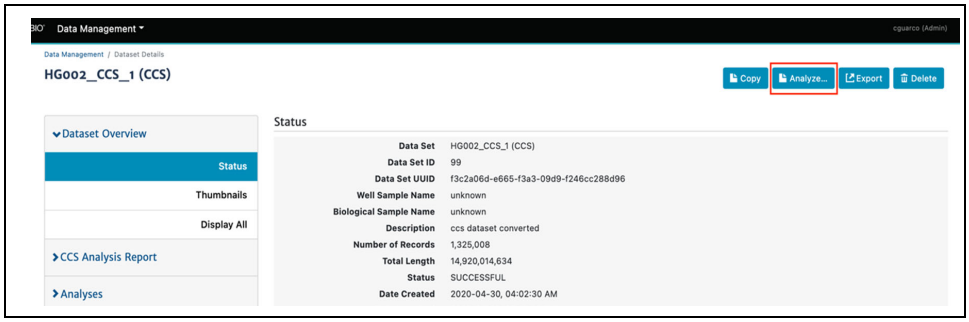
10. Click **Start** to submit the analysis.
11. After completion, the Circular Consensus Sequencing (CCS) application creates a CCS Analysis Report in the Analysis Results page that contains information and statistics about the HiFi Read generation process. (For a full interpretation of the report see the SMRT Link User Guide.) Click **Data > File Downloads** (on the left side menu of the report) to download the following files:
    - `Movie-name.Q20.fasta`: Contains HiFi Reads in FASTA format.
    - `Movie-name.ccs.bam`: Contains HiFi Reads in bam format.
    - `Run_name (CCS)`: Link to HiFi Reads report (See "Running an Analysis Using HiFi Reads from Sequel II Systems" on page 10.)
    - `CCS Analysis per-Read Details`: Summary of CCS Analysis performance and yield.
    - `Analysis Log`: CCS analysis log.
    - `SMRT Link Log`: `pbcromwell` log.

Running an Analysis Using HiFi Reads from Sequel II Systems

1. In **SMRT Analysis > Data > File Downloads**, click the `Run_name (CCS)` link that links to the **Data Management > Dataset Details** page containing the HiFi Reads report for your Data Set.



2. From the **Data Management > Dataset Details** page containing the HiFi Reads report for your Data Set, click **Analyze** to go the **SMRT Analysis > Create New Analysis** page.



3. From the **SMRT Analysis > Create New Analysis** page, proceed as shown in the SMRT Link User Guide to create a new analysis.

P/N 102-144-100 Version 01 (September 2021)