SMRT® Analysis Barcoding Overview (v5.1.0)

Introduction

This document applies to PacBio RS II and Sequel® Systems using SMRT Link v5.1.0.

Note: For SMRT Link v5.0.0 or v5.0.1, see the document **SMRT Analysis Barcoding Overview**, available here.

SMRT Link v5.1.0 includes a new streamlined end-to-end barcoding workflow for automated analysis of multiplexed samples. It enables users to define barcodes and their associated sample names for an experiment at Run Design.

This workflow uses a new and improved demultiplexing algorithm and provides extensive QC metrics for evaluation of barcoding performance.

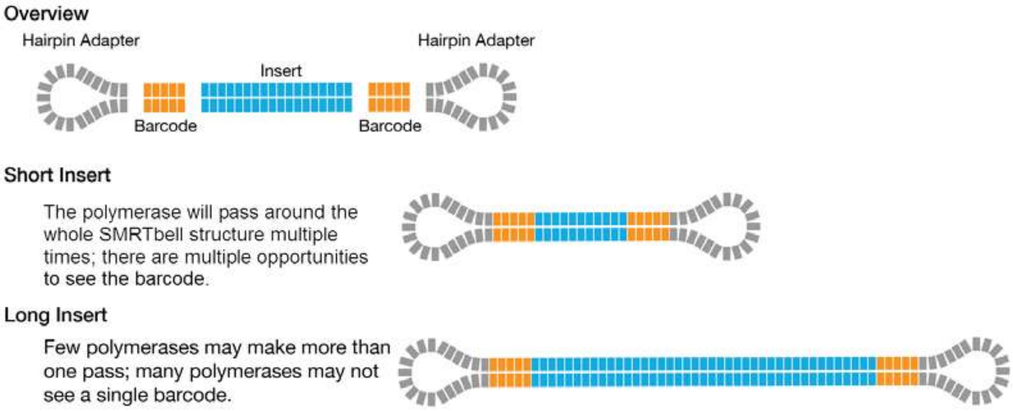
- The SMRT Link graphical user interface supports up to 384 barcodes per sample. Support for more than 384 samples is available using the command-line.

This document covers:

- Barcoding designs, strategies and modes for preparing barcoded samples and data analysis for multiplexed sequencing for the following applications: Amplicons (Resequencing, Minor Variants, LAA and CCS) and microbial assembly (HGAP 4).
- Identifying barcodes using SMRT Link.
- Performing downstream data analysis using SMRT Link.

This document does **not** include information about barcoding designs and analysis for the Iso-Seq® application. For details about barcoding samples for Iso-Seq analysis, click here.

Barcode Background

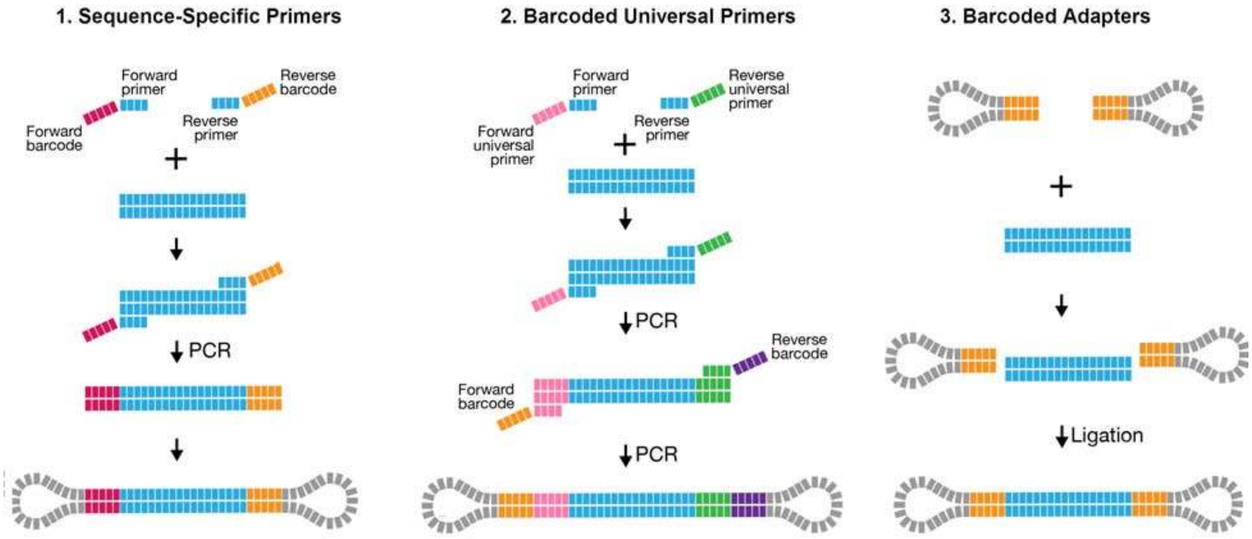

Overview
Hairpin Adapter
Insert
Hairpin Adapter
Barcode
Barcode
Short Insert
The polymerase will pass around the whole SMRTbell structure multiple times; there are multiple opportunities to see the barcode.
Long Insert
Few polymerases may make more than one pass; many polymerases may not see a single barcode.

Barcoding Strategies

There are three main strategies for barcoding samples using PacBio technology:

1. **Barcoded Primers**
   - Recommended for large projects where the target-specific primer and barcode are validated at the start of the project and used for many samples.
   - Used with non-degenerate primers.
2. **Barcoded Universal Primers**
   - Recommended for amplicons with easily modified primer designs.
   - Used with degenerate complex-primer designs.
3. **Barcoded Adapters**
   - Recommended for validated PCR systems and smaller projects.
   - Recommended for initial evaluation.
   - For use with off-the-shelf assays.
   - Additionally, the barcoded adapters strategy can be used for sheared libraries when multiplexing for microbial assembly projects.

Three Ways to Barcode Samples Using PacBio Technology



1. Sequence-Specific Primers

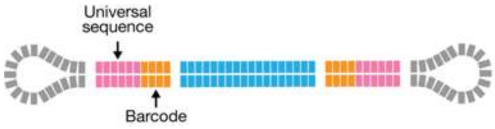
2. Barcoded Universal Primers

3. Barcoded Adapters

For each of the strategies, there is a different barcoded library design protocol.

For targeted captures **only**, use Linear Barcoded Adapters:

Linear Barcoded Adapters with Universal Sequences

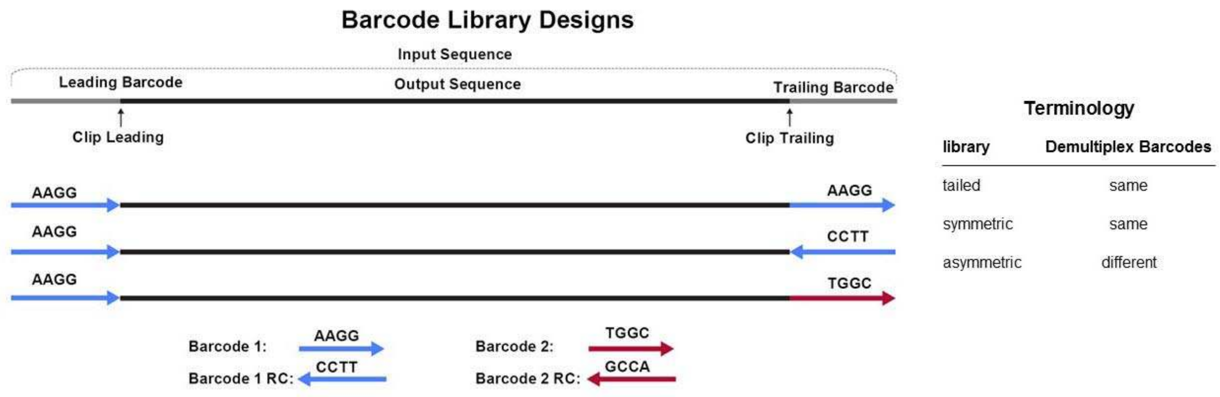


Long-read, probe-based capture solutions use linear barcoded adapters with universal primer sequences for amplification across all captured fragments. The barcoded adapters are ligated by AT ligation prior to the pooling and capture steps. This workflow supports up to 24 barcodes. Linear Barcoded Adapters can be ordered from any oligo synthesis provider using the order sheet on the PacBio Multiplexing page. Demultiplexing of such experiments is supported through SMRT Link command-line options.

Barcode Scoring Modes

The new demultiplexing algorithm included in SMRT Link v5.1.0 is significantly enhanced for efficiency and robustness. The only required information is if the barcoding sequences are the **same** or **different** on either side of the read. Barcode orientation does **not** need to be specified. The demultiplexing algorithm scores **all** barcode sequences in both orientations - forward and reverse complement.

This section describes barcode scoring modes and provides information about the different experiment options depending on barcoding strategy.



library	Demultiplex Barcodes
tailed	same
symmetric	same
asymmetric	different

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

Demultiplexing of CCS reads is possible on the command line.

Symmetric Mode

For **symmetric** and **tailed** library designs, the **same** barcode is attached to both sides of the insert sequence of interest. The only difference is the orientation of the trailing barcode. For barcode identification, one read with a single barcode region is sufficient. This is most commonly the case when using barcoded SMRTbell adapters and for target enrichment (non-hairpin) adapters. This is also the default scoring mode in SMRT Link v5.1.0 and later.

Asymmetric Mode

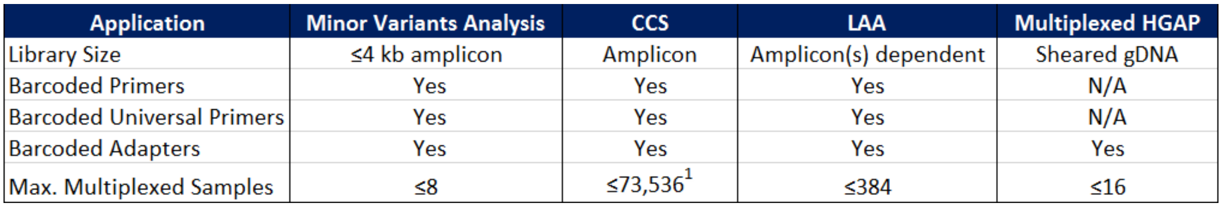
Barcode sequences are **different** on the forward and reverse ends of the insert. Asymmetric mode is most commonly used when appending barcodes during a single round of PCR with barcoded primers. Pacific Biosciences recommends using this mode **only** for cases when both ends of the insert are expected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

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

Mixed Mode

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

Which Barcoding Strategies to Use?

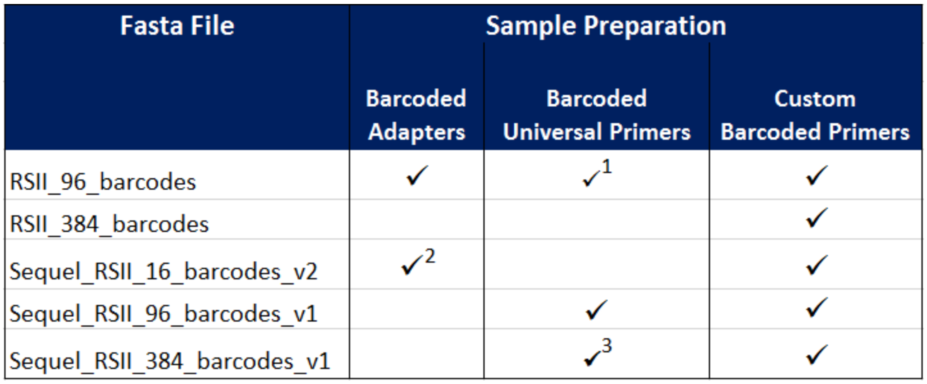


Application	Minor Variants Analysis	CCS	LAA	Multiplexed HGAP
Library Size	≤4 kb amplicon	Amplicon	Amplicon(s) dependent	Sheared gDNA
Barcoded Primers	Yes	Yes	Yes	N/A
Barcoded Universal Primers	Yes	Yes	Yes	N/A
Barcoded Adapters	Yes	Yes	Yes	Yes
Max. Multiplexed Samples	≤8	≤73,536[1]	≤384	≤16

[1] 384 unique barcodes used in an asymmetric design: (384 ^2 - 384)/2 = 73,536.

Note: For multiplexing microbial genomes experiments, click here to see details about the workflows, analysis and barcoding kits that we provide.

The following table includes the names of FASTA files used for analysis:



Fasta File	Sample Preparation		
	Barcoded Adapters	Barcoded Universal Primers	Custom Barcoded Primers
RSII_96_barcodes	✓	✓[1]	✓
RSII_384_barcodes			✓
Sequel_RSII_16_barcodes_v2	✓[2]		✓
Sequel_RSII_96_barcodes_v1		✓	✓
Sequel_RSII_384_barcodes_v1		✓[3]	✓

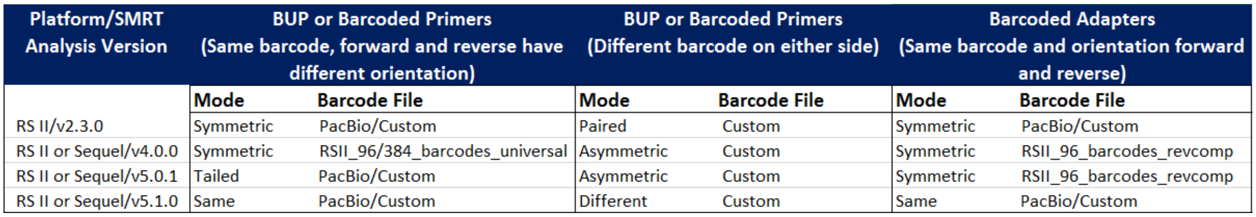
[1] Using Barcoded Universal F/R Primers Plate – 96 (P/N 100-466-100) and Barcoded Adapter Plate – 96 (P/N 100-466-000)

[2] Using Barcoded Adapter Kit 8A (P/N 101-081-300) and Barcoded Adapter Kit 8B. (P/N 101-081-400)

[3] Using oligos from PacBio recommended lists.

The following table describes the recommended combinations of:

- Platform and SMRT Analysis versions,
- Primers/Adapters, and
- Mode and Barcode files.

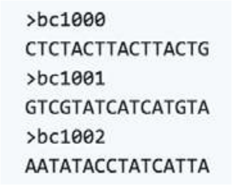


Platform/SMRT Analysis Version	BUP or Barcoded Primers (Same barcode, forward and reverse have different orientation)		BUP or Barcoded Primers (Different barcode on either side)		Barcoded Adapters (Same barcode and orientation forward and reverse)	
	Mode	**Barcode File**	**Mode**	**Barcode File**	**Mode**	**Barcode File**
RS II/v2.3.0	Symmetric	PacBio/Custom	Paired	Custom	Symmetric	PacBio/Custom
RS II or Sequel/v4.0.0	Symmetric	RSII_96/384_barcodes_universal	Asymmetric	Custom	Symmetric	RSII_96_barcodes_revcomp
RS II or Sequel/v5.0.1	Tailed	PacBio/Custom	Asymmetric	Custom	Symmetric	RSII_96_barcodes_revcomp
RS II or Sequel/v5.1.0	Same	PacBio/Custom	Different	Custom	Same	PacBio/Custom

Custom Symmetric Barcode FASTA

Independent of the barcoding strategy and mode, if you are using custom barcodes, you must generate a FASTA file of barcode sequences. This file should:

- Include **one** entry per barcode sequence.
- Include **no** duplicate sequences.
- Include only **upper-case** bases.
- Be orientation-agnostic (forward or reverse-complement, but not reversed.)



```
>bc1000
CTCTACTTACTTACTG
>bc1001
GTCGTATCATCATGTA
>bc1002
AATATACCTATCATTA
```

Please name your barcodes with an alphabetic character prefix to avoid later confusion of barcode name and index. Duplicate names are **not** permitted.

Working with Barcoded Data in SMRT Link v5.1.0

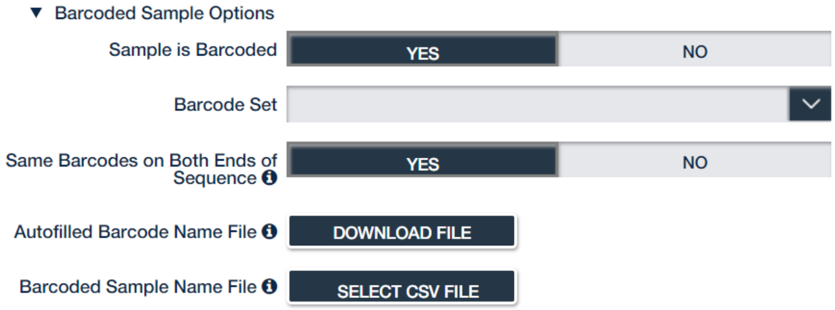
This section describes how to use SMRT Link to work with barcoded data. **Note**: There are major changes to barcoding functionality in this release.

The canned data provided with SMRT Link v5.1.0 includes 5 barcode sets:

- `RSII_384_barcodes`
- `RSII_96_barcodes`
- `Sequel_RSII_16_barcodes_v1`
- `Sequel_RSII_96_barcodes_v1`
- `Sequel_RSII_384_barcodes_v1`

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

1. In SMRT Link, create a new run design as described in “Creating a New Run Design” on page 10 of the **SMRT Link User Guide**.
   Before you finish the new Run Design, perform the following steps.



▼ Barcoded Sample Options

Sample is Barcoded: YES | NO

Barcode Set

Same Barcodes on Both Ends of Sequence: YES | NO

Autofilled Barcode Name File: DOWNLOAD FILE

Barcoded Sample Name File: SELECT CSV FILE

2. Click **Barcoded Sample Options** and then click **Yes** for **Sample is Barcoded**. Additional fields related to barcoding will be displayed.
3. Specify a **Barcode Set** using the dropdown list.
4. Specify if the **same** barcodes are used on both ends of the sequences.
   - Selecting **Yes** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
   - Selecting **No** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such experiments retains any barcode pair combination identified in the Data Set.
5. SMRT Link **automatically** creates a CSV-format **Autofilled Barcode Name File**. The barcode name is populated based on your choice of barcode set, and if the barcodes are the same at both ends of the sequence. The file includes a blank column for the biological sample names.
   - (**Optional**) If you want to specify the biological sample names corresponding to each barcode, click **Download File**, enter the biological sample names associated with the barcodes (Maximum: 40 characters) in the second column, and save the file. If you did **not** use all barcodes in the Autofilled Barcode Name file in the sequencing run, either leave the biological sample name column blank for those barcodes, or delete those rows.
   - If you **don’t** specify the biological sample name, it will automatically be set to the same value as the barcode name in SMRT Link.
   - **Note**: Open the CSV file in a text editor and check that the columns are separated by **commas**, not semicolons.
6. (**Optional**) Select the **Barcoded Sample Name File** you edited in **Step 5**. If you do **not** upload a Barcoded Sample Name File, the biological sample names for those barcodes will automatically be set to the barcode names.
7. Click **Create**.

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

1. Create a CSV file containing 2 columns.
2. The contents of the first row must be in the form of “Barcode Name,Bio Sample Name”. (Valid characters: Alphanumeric; space; dot; underscore; hyphen.)
3. Each row **must** contain a pair of barcode names that exist in the selected barcode set, separated by 2 hyphens. The Bio Sample name is entered after a comma.
   Example: `bc1001--bc1001,biological sample name 1`

Step 2: Perform the Sequencing Run

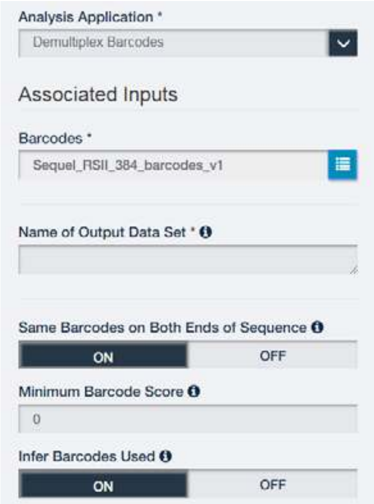
Load the samples and perform the sequencing run, using the Run Design you created in Step 1. The demultiplexing analysis is performed automatically on the SMRT Link Server once the data is transferred from the Sequel System. This creates an analysis of type `Demultiplex Barcodes (Auto)` in the SMRT Analysis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

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

Step 3: (Optional) Run the Demultiplex Barcodes Application

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

1. Click **+ Create New Analysis**.
2. Enter a **name** for the analysis.
3. Select **Demultiplex Barcodes** from the Applications list.



Analysis Application *
Demultiplex Barcodes

Associated Inputs

Barcodes *
Sequel_RSII_384_barcodes_v1

Name of Output Data Set *

Same Barcodes on Both Ends of Sequence
[✓ ON] OFF

Minimum Barcode Score
0

Infer Barcodes Used
[✓ ON] OFF

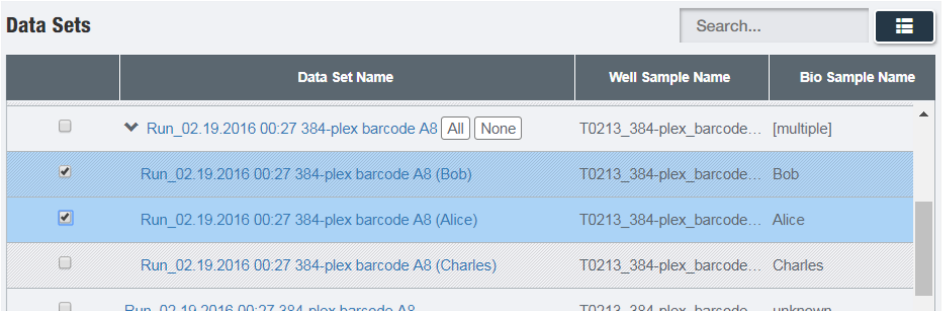
4. Specify a barcode sequence file.
5. Specify the name for the new demultiplexed Data Set that will display in SMRT Link.
6. Specify if the **same** barcodes are used on both ends of the sequences.
   - Selecting **On** specifies symmetric and tailed designs where **all** the reads have the same barcodes on both ends of the insert sequence. Barcode analysis of such experiments retains **only** data with the same barcode identified on both ends.
   - Selecting **Off** specifies asymmetric designs where the barcodes are **different** on each end of the insert. Barcode analysis of such data retains any barcode pair combination identified in the Data Set.
7. Specify the **Minimum Barcode Score**: Reads with barcode scores below the value are **not** included in downstream analysis. We recommend that you set this value to `26` for **all** applications.
8. Specify if you want to infer which barcodes were used:
   - **On** infers which subset of barcodes from the selected barcode set were used, and outputs **one** data set for **each** of those inferred barcodes.
   - **Off** outputs **one** data set with all barcodes in the selected barcode set.
9. Click **Start**. After the analysis is finished, a new demultiplexed Data Set is available.

Step 4: Run Applications Using the Demultiplexed Data as Input

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

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

1. Select the secondary analysis application to use.
2. Select the demultiplexed Data Set to use as input:



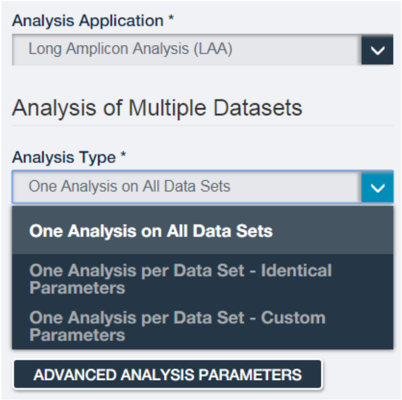
Data Sets

Search...

	Data Set Name	Well Sample Name	Bio Sample Name
□	▾ Run_02.19.2016 00:27 384-plex barcode A8 All None	T0213_384-plex_barcode...	[multiple]
☑	Run_02.19.2016 00:27 384-plex barcode A8 (Bob)	T0213_384-plex_barcode...	Bob
☑	Run_02.19.2016 00:27 384-plex barcode A8 (Alice)	T0213_384-plex_barcode...	Alice
□	Run_02.19.2016 00:27 384-plex barcode A8 (Charles)	T0213_384-plex_barcode...	Charles
□	Run_02.19.2016 00:27 384-plex barcode A8	T0213_384-plex_barcode	unknown

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

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



Analysis Application *

Long Amplicon Analysis (LAA)

Analysis of Multiple Datasets

Analysis Type *

One Analysis on All Data Sets

One Analysis on All Data Sets
One Analysis per Data Set - Identical Parameters
One Analysis per Data Set - Custom Parameters
ADVANCED ANALYSIS PARAMETERS

- **One Analysis on All Data Sets:** Runs **one** analysis using all the selected barcode Data Sets for a maximum of 30 Data Sets. Click **Start**.
- **One Analysis per Data Set - Identical Parameters:** Runs a separate analysis for **each** of the selected barcode Data Sets, using the **same** parameters, for a maximum of 384 Data Sets. Optionally click **Advanced Analysis Parameters** and modify parameters. Click **Start**.
- **One Analysis per Data Set - Custom Parameters:** Runs a separate analysis for **each** of the selected barcode Data Sets, using **different** parameters for each Data Set, for a maximum of 16 Data Sets. Click **Advanced Analysis Parameters** and modify parameters. Then click **Start and Create Next**. You can then specify parameters for each of the included barcode Data Sets.

Demultiplex Barcodes Input and Output

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

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

Barcode Set (Required):

- Specify a barcode sequence file to separate the reads.

Name of Output Data Set (Required)

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

Same Barcodes on Both Ends of Sequences

- Specify **On** to retain all the reads with the **same** barcodes on both ends of the insert sequence, such as symmetric and tailed designs.
- Specify **Off** to specify asymmetric designs where the barcodes are **different** on each end of the insert.
- See "Barcode Scoring Modes" on page 3 for information on barcode designs.

Minimum Barcode Score

- A **barcode score** measures the alignment between a barcode attached to a read and an ideal barcode sequence, and is an indicator how well the chosen barcode pair matches. It ranges between 0 (no match) and 100 (a perfect match). Specify that reads with barcode scores below this minimum value are **not** included in downstream analysis.

Infer Barcodes Used

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

Demultiplex Barcodes Reports and Data Files

The Demultiplex Barcodes application generates the following reports:

Barcodes > Summary Metrics

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

Barcodes > Barcode Data

- **Bio Sample Name:** The name of the biological sample associated with the barcode.
- **Barcode Index:** The index number associated with the barcode.
- **Barcode Name**: A string containing the pair of barcode indices for which the following metrics apply.
- **Polymerase Reads:** The number of polymerase reads associated with the barcode.

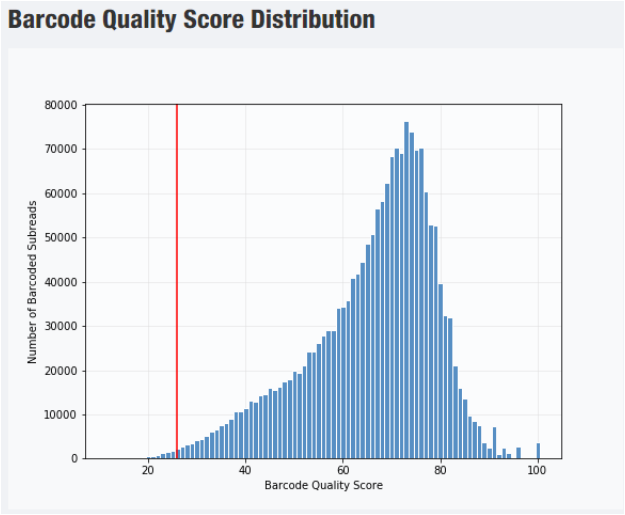
- **Subreads**: The number of subreads associated with the barcode.
- **Bases**: The number of bases associated with the barcode.
- **Mean Read Length:** The mean read length of reads associated with the barcode.
- **Longest Subread Length:** The longest subread length associated with the barcode.
- **Mean Barcode Quality:** The mean barcode quality associated with the barcode.
- **Rank Order (Num. Reads)**: The rank order of this barcode in terms of the number of reads.

Barcodes > Barcoded Read Statistics

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

Barcodes > Barcode Quality Scores

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



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

Barcodes > Barcoded Read Binned Histograms

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

Data > File Downloads

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

- **Analysis Log**: Log information for the analysis workflow.
- **Master Log**: Server-level analysis log information. (This file is displayed when you choose **Data > Analysis Log**.)
- **Barcode Files**: Barcoded subread Data Sets; one XML file per barcode.
- **Barcode Report Details**: Data displayed in the reports, in CSV format.

Note: You can get the demultiplexed BAM files using the Data Management module’s “Export Data Sets” feature. In the demultiplexed BAM output file a tag `bc` is added for each read, indicating the assigned barcode. The `bc` tag is the zero-based index of the barcodes in the FASTA file. For example, when using the barcodes `RSII_96_barcodes`, a subread with barcode `lbc1` identified on both sides will have the tag `bc:B:S,0,0` in the BAM output file.

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

For details on running the Demultiplex Barcodes application using the command-line, see “SMRT Tools Reference Guide”, available on the PacBio Downloads page.

Additional Information:

- Demultiplex Barcodes Algorithm: https://github.com/PacificBiosciences/barcoding
- PacBio BAM Format Specifications: http://pacbiofileformats.readthedocs.io/en/5.0/BAM.html

P/N 101-307-200 Version 03 (June 2018)