BD Single Cell Genomics Bioinformatics Handbook

Doc ID: 54169 Rev. 6.0
10/2018

Becton, Dickinson and Company
BD Biosciences
2350 Qume Drive
San Jose, CA 95131 USA
Tel 1.877.232.8995, prompt 2, 2
researchapplications@bd.com

Copyrights/trademarks

Trademarks are the property of their respective owners.

The information in this guide is subject to change without notice. BD Biosciences reserves the right to change its products and services at any time to incorporate the latest technological developments. Although this guide has been prepared with every precaution to ensure accuracy, BD Biosciences assumes no liability for any errors or omissions, nor for any damages resulting from the application or use of this information. BD Biosciences welcomes customer input on corrections and suggestions for improvement.

Regulatory information

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

Doc ID: 54169 Rev. 6.0

History

Revision	Date	Changes made
Doc ID: 54169 Rev. 1.0	09/2017	Initial release.
Doc ID: 54169 Rev. 2.0	11/2017	—Added content on sample multiplexing. See Step 7. Determine the sample of origin (sample multiplexing only) (page 32) and Reviewing sequencing analysis output files (page 38). —Added content on specifying gene targets. See Specifying the gene targets (page 12).
Doc ID: 54169 Rev. 3.0	01/2018	—Updated content of BD™ Data View to v1.1, which includes these new features: New color options for plots, highlight selected annotated groups in plots, filter data table by cells based on gene expression, new calculation on fold changes and mean gene expression, new option to modify data table names and annotation list names. See BD™ Data View (page 87). —Added two examples for use with BD Data View. See Analyzing multiple samples with BD Data View (page 161) and Designing a targeted panel from whole transcriptome amplification (WTA) RNA-seq data (page 177). —Expanded information on selecting a transcript, selecting primers, and output files. See Primer design (page 9).

Revision	Date	Changes made
Doc ID: 54169 Rev. 4.0	04/2018	—Added another example for use with BD Data View. See Analyzing a multiplexed sample with BD Data View (page 191).
Doc ID: 54169 Rev. 5.0	07/2018	—Added metrics outputs for BD™ AbSeq. See BD Rhapsody™ Targeted sequencing analysis (page 9) —Updated to BD Data View v1.2. See BD™ Data View (page 87). Some new features include: –Automatic detection of AbSeq markers in data tables –New Gene A v. Gene B feature to compare two gene markers –Combine one or more data tables –Differential expression of >1,500 genes
Doc ID: 54169 Rev. 6.0	10/2018	—Updated cross references from system user guides to instrument user guides. —Changed content to say that a BAM file is sorted according to the alignment coordinates of R2 reads on each chromosome. See BAM (page 50). —Added recommendation to analyze datasets that are ≤1 TB in size. See Understanding the BD Rhapsody Analysis pipeline step-by-step (page 11). —Updated output file name in example to Combined_<sample_multiplex_name>_DBEC_MolsPerCell.csv. See Analyzing a multiplexed sample with BD Data View (page 191).

Contents

1

Introduction

About this handbook

Introduction

This handbook is a comprehensive reference to help you prepare and analyze single cell libraries with the BD Rhapsody™ Single-Cell Analysis system or the BD Rhapsody™ Express Single-Cell Analysis system. Major aspects of the BD single cell genomics bioinformatics workflow are covered. This reference explains the BD single cell genomics sequencing and clustering algorithms to deepen your understanding of how single cell mRNA and protein (AbSeq) expression profiles are generated and clustered. In addition, the handbook defines every analysis metric. Finally, the handbook contains step-by-step instructions and worked-out examples on how to use the visualization tool, BD™ Data View. BD Data View will speed your bioinformatics analysis, help you create custom primer panels, and empower you to make new insights in biology through genomics.

The BD single cell genomics team

2

BD Rhapsody™ Targeted sequencing analysis

How to use this chapter

This chapter provides in-depth information on the process, output metrics, and interpretation of output from BD Rhapsody Targeted sequencing analysis:

Section	Information
Understanding the BD Rhapsody Analysis pipeline step-by-step (page 11)	Detailed description of each step in the BD Rhapsody Analysis pipeline
Reviewing sequencing analysis output files (page 38)	Definitions of the sequencing analysis output metrics
Interpreting output metrics (page 65)	Recommended solutions to possible problems during sequencing analysis

For definitions of the clustering analysis metrics, see BD Rhapsody™ Targeted clustering analysis (page 73).

Understanding the BD Rhapsody Analysis pipeline step-by-step

Introduction

This section provides an in-depth description of each step in the BD Rhapsody Analysis pipeline.

For instructions on running the pipeline, see the *BD Single Cell Genomics Analysis Setup User Guide* (Doc ID: 47383).

Genomics technical publications are available for download from the BD Genomics Resource Library at bd.com/genomics-resources.

BD Biosciences recommends analyzing datasets that are ≤1 TB in size. For datasets (compressed FASTQ FILES from all libraries) >1 TB, contact BD Biosciences technical support at researchapplications@bd.com.

Overview

The BD Rhapsody™ targeted assays are used to create sequencing libraries from single cell transcriptomes. After sequencing, the analysis pipeline takes the FASTQ file, an mRNA reference file, and an AbSeq reference file (if the latter is required) for gene alignment. See Figure 1.

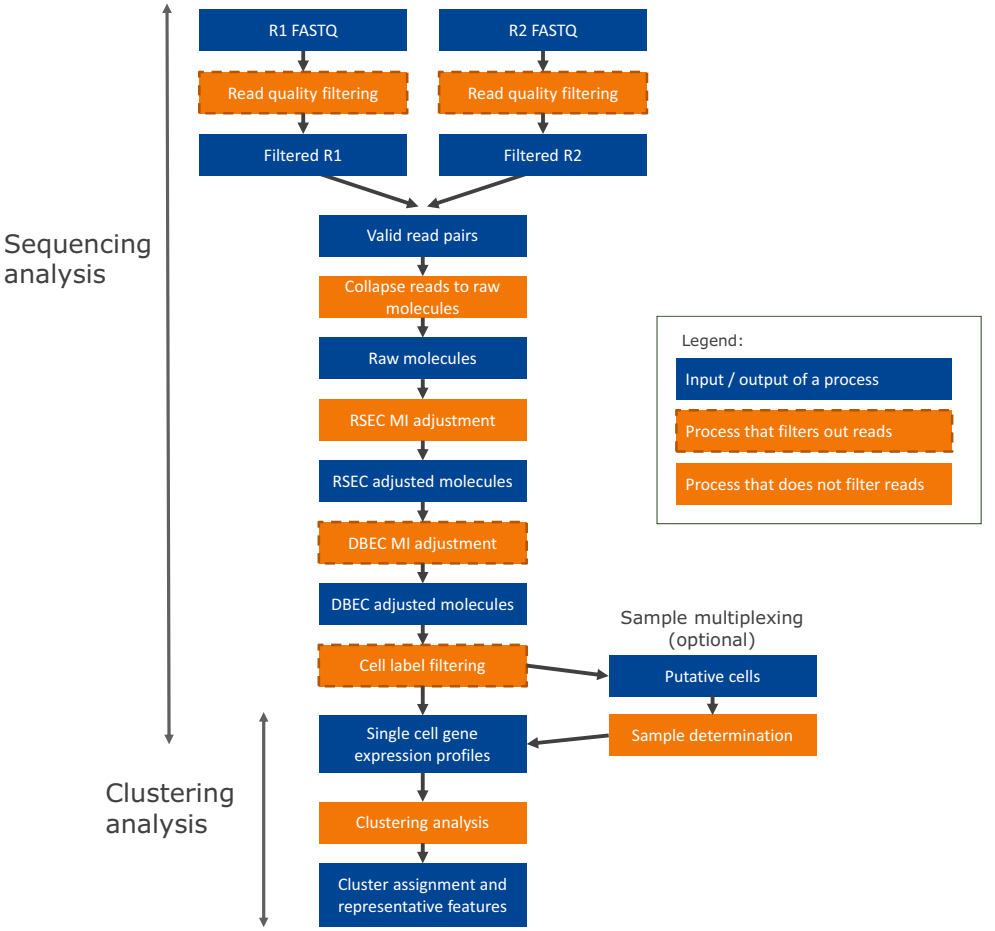


Figure 1. Overview of the steps in the analysis pipeline. For definitions of terms, see Glossary (page 201).

The analysis pipeline works with paired-end FASTQ R1 and R2 files generated from Illumina sequencers. The minimum read length required is 75 bp on each end of the library molecule. R1 reads contain information on the cell label and molecular identifier, and R2 reads contain information on the gene. See Figure 2.

Figure 2. Structure of a read pair that is generated by sequencing the libraries prepared with the BD Rhapsody Targeted assays.

The next sections describe the analysis pipeline step-by-step.

Step 1. Filter by read quality

Filtering criteria

Read pairs with low sequencing quality are first removed. This step reduces the influence of poor sequencing quality from the metrics that are specific to the BD Rhapsody Targeted assays.

The following filtering criteria are applied to each read pair:

- Read length: If the length of R1 read is <66 or R2 read is <64, the R1/R2 read pair is dropped.
- Mean base quality score of the read: If the mean base quality score of either R1 read or R2 read is <20, the read pair is dropped.
- Highest Single Nucleotide Frequency (SNF) observed across the bases of the read: If SNF is ≥0.55 for the R1 read or SNF ≥0.80 for the R2 read, the read pair is dropped. This criterion removes reads with low complexity such as strings of identical bases and tandem repeats.

  The thresholds for each filter are determined empirically.

Step 2. Annotate R1 reads

R1 structure

The quality-filtered R1 reads are analyzed to identify the cell label section sequence (CLS), common sequences (L), Unique Molecular Identifier (UMI) sequence, and poly(T) tail. See Figure 3.

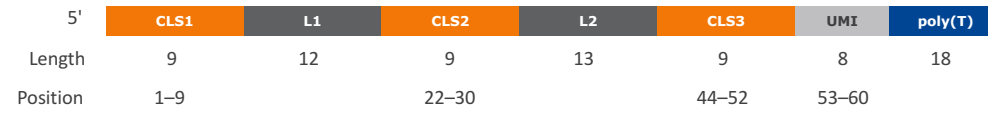


Figure 3. Structure of R1 read.

Cell label

Information of the cell label is captured by bases in three sections (CLS1, CLS2, CLS3) along each R1 read. Two common sequences (L1, L2) separate the three CLSs, and the presence of L1 and L2 relates to the way the capture oligonucleotide probes on the beads are constructed. By design, each CLS has one of 96 predefined sequences, which has a Hamming distance of at least four bases and an edit distance of at least two bases apart. A cell label is defined by the unique combination of predefined sequences in the three CLSs. Thus, the maximum possible number of cell labels is 96^3 (884,736). A cell label is represented by an index between 1–96^3.

Reads are first checked for perfect matches in all three pre-designed CLS sequences at the expected locations, CLS1: position 1–9, CLS2: position 22–30, and CLS3: position 44–52. Reads with perfect matches are kept.

The remaining reads are subjected to another round of filtering to recover reads with base substitutions, insertions, deletions caused by sequencing errors, PCR errors, or errors in oligonucleotide synthesis.

UMI

By design, the UMI is a string of eight randomers immediately downstream of CLS3. If the CLSs have perfect matches or base substitutions, the UMI sequence is at position 53–60. For reads with insertions or deletions within the CLSs, the UMI sequence is eight bases immediately following the end of the identified CLS3.

Poly(T) tail

Following the UMI, a poly(T) tail, the polyadenylation [poly(A)] complement of an mRNA molecule, is expected. Each read with a valid cell label is kept for further consideration only if ≥6 out of 8 bases after UMI are found to be Ts.

Step 3. Annotate R2 reads

Criteria for a valid R2 read

The pipeline uses Bowtie2 version 2.2.9 to map the filtered R2 reads to the reference panel sequences. Option `--norc` is enabled to map all of the reads only to the forward strand of the provided reference. The default setting of the local alignment mode is used for all other parameters.

An R2 read is a valid gene alignment if all of these criteria are met:

- The read aligns uniquely to a transcript sequence in the reference.
- The R2 alignment begins within the first five nucleotides. This criterion ensures that the R2 read originates from an actual PCR priming event.
- The length of the alignment that can be a match or mismatch in the CIGAR (Compact Idiosyncratic Gapped Alignment Report) string is >60, where CIGAR is a sequence of base lengths to indicate base alignments, insertions, and deletions with respect to the reference sequence. See samtools.github.io/hts-specs/SAMv1.pdf.
- The read does not align to phiX174.

Step 4. Combine information from R1 and R2 annotations

Retain R1 and R2 reads

Read pairs with a valid R1 read and a valid R2 read are retained for further analyses. A valid R1 read requires identified CLSs, a UMI sequence with non-N bases, and a poly(T) tail. A valid R2 read requires reads uniquely mapped to a gene in a panel with the correct PCR2 primer sequence at the start and an alignment of >60 bases in length.

Step 5. Annotate molecules

Collapse reads into raw molecules

Reads with the same cell label, same UMI sequence, and same gene are collapsed into a single raw molecule. The number of reads associated with each raw molecule is reported as the *raw adjusted sequencing depth*.

Remove artifact molecules using RSEC and DBEC UMI adjustment algorithms

PCR and sequencing often generate errors. If the error occurs within the UMI sequence, the R1/R2 read pair is called a unique molecule but is, in fact, an artifact. Artifact molecules contribute to an over-estimated molecule count of a gene in a cell. As sequencing depth increases, the number of raw molecules rises and never plateaus due to these artificial molecules.

To remove the effect of UMI errors on molecule counting, BD Biosciences has developed a set of UMI adjustment algorithms. UMI errors that are single base substitution errors are identified and adjusted to the parent UMI barcode using recursive substitution error correction (RSEC). Other UMI errors derived from library preparation steps or sequencing base deletions are later adjusted using distribution-based error correction (DBEC).

Figure 4 shows the workflow of the two algorithms used on data generated from BD Rhapsody Targeted assays. Figure 5 shows how the two algorithms are applied to example results to correct the apparent counts of molecules.

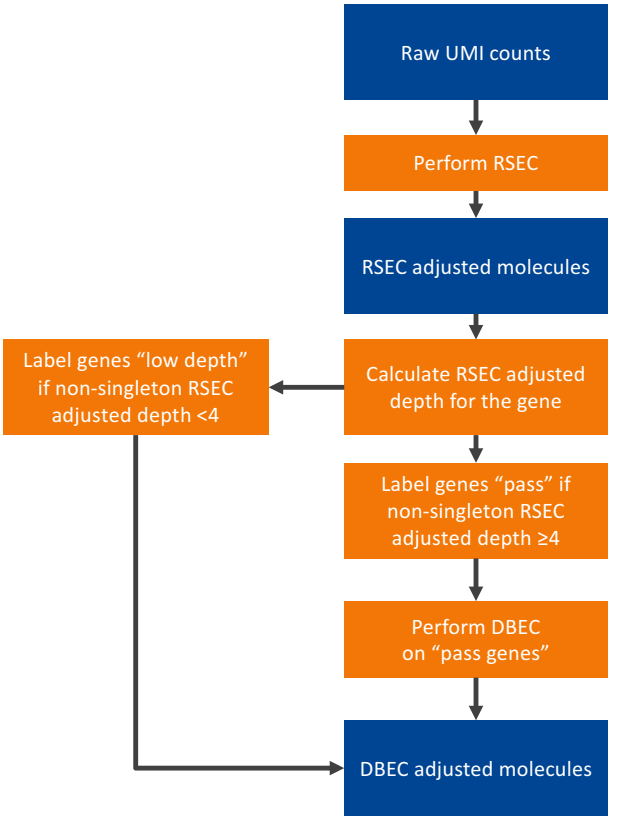


Figure 4. Workflow of UMI count adjustment.

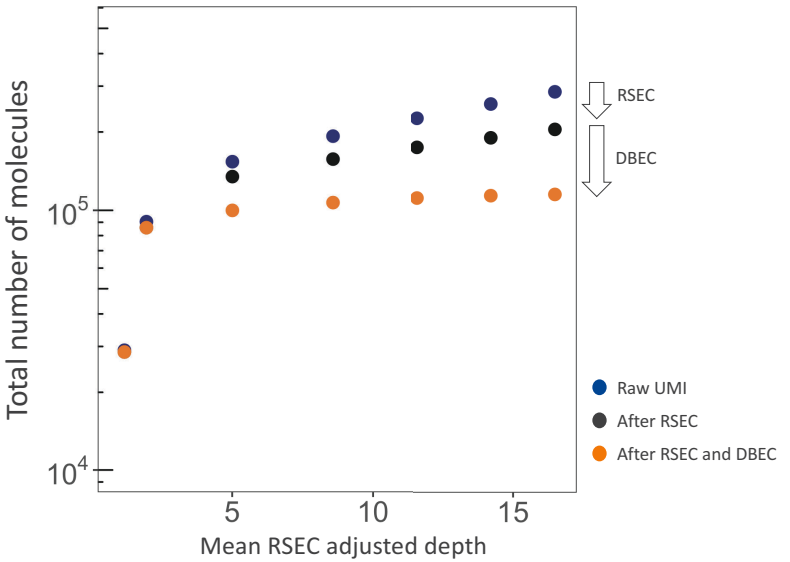


Figure 5. Example results after applying RSEC and DBEC algorithms. If we consider only raw UMIs, the apparent total number of molecules continues to rise with sequencing depth, because the presence of sequencing and PCR errors contribute to unique UMIs. RSEC removes artifact molecules from single base substitutions in the UMI sequence. Further adjustment by DBEC removes artifact molecules originated from PCR errors. As a result, the number of molecules stabilizes with additional sequencing, indicating the library is sequenced to saturation.

Collapse molecules that differ by one base in the UMI sequence using RSEC

RSEC considers two factors in error correction: 1) similarity in UMI sequence and 2) raw UMI coverage or depth. See Figure 6.

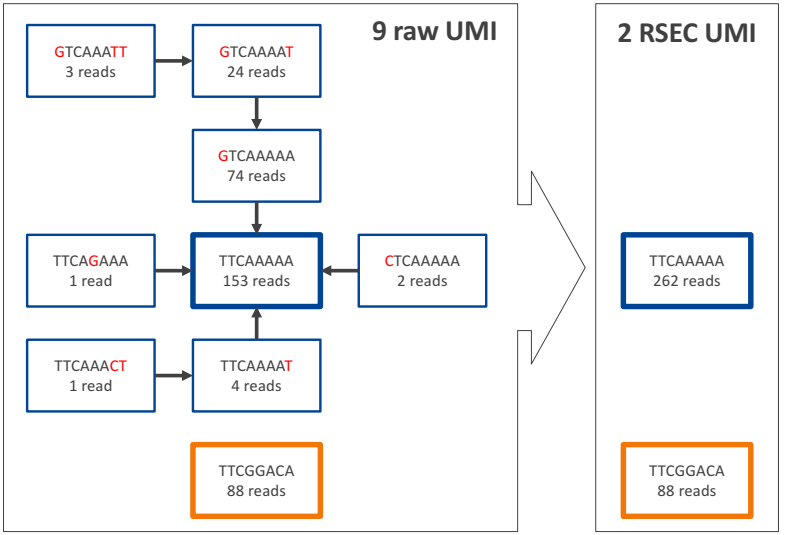


Figure 6. Example of the RSEC algorithm. Nine raw UMIs are collapsed into two UMIs.

For each target gene, UMIs are connected when their UMI sequences are matched to within one base (Hamming distance = 1). For each connection between UMI x and y, if Coverage(y)> 2 * Coverage(x) – 1, then y is Parent UMI and x is Child UMI. Based on this assignment, child UMIs are collapsed to their parent UMI. This process is recursive until there are no more identifiable parent-child UMIs for the gene. See Figure 6.

The number of reads for each child UMI is added to the parent, so no reads are lost. The sum of the reads is the *RSEC-adjusted depth* of the *RSEC-adjusted molecule*.

Adjust molecule counts by DBEC

The RSEC-adjusted molecule counts are further corrected by DBEC.

DBEC is applied on a per-gene basis. The algorithm is based on the assumption that the pre-amplified set of molecules of the same gene, regardless of the cell of origin, is subject to the same amplification efficiency and, therefore, should have similar read depth. Artifact molecules created later in the PCR cycles, such as those derived from PCR chimera formation, will likely have less read depth.

DBEC considers the distribution of RSEC-adjusted depth distribution, not UMI sequence. The sequencing depth of RSEC-adjusted molecules for each gene is a bimodal distribution. See Figure 7. The lower mode of the distribution likely represents artifact molecules, and the upper mode likely represents true molecules. The algorithm fits two negative binomial distributions to statistically distinguish between the two modes. Molecules in the upper mode are retained (*DBEC-adjusted molecules)*, while the molecules in the lower mode are discarded. The average depth of the molecules in the upper mode is known as the *DBEC-adjusted depth*, and the depth of molecules in the lower mode is the metric *error depth*. The cutoff between the two modes is the *DBEC minimum depth*.

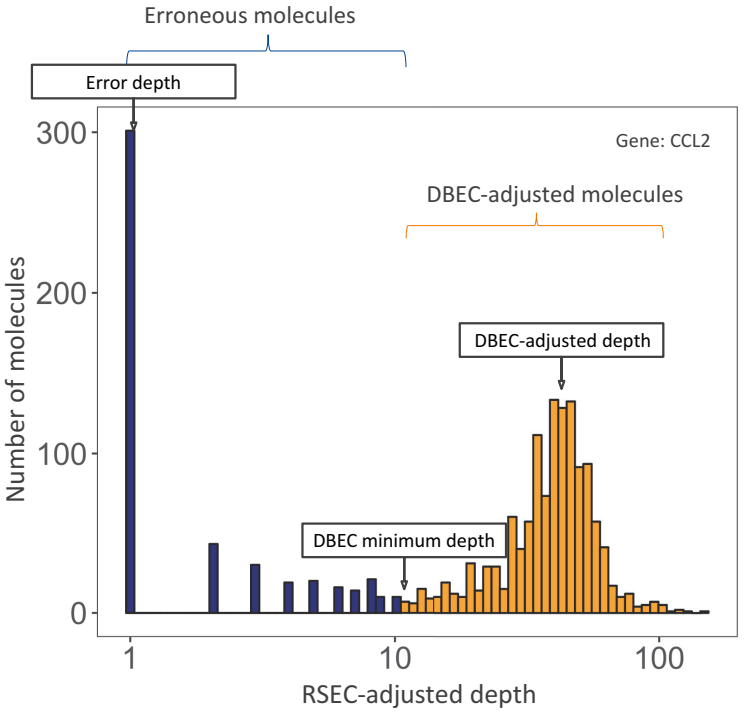


Figure 7. Example of the DBEC algorithm for gene CCL2. Counts under the orange bars are kept and labelled as DBEC-adjusted molecules. Counts under the blue bars are labelled as erroneous molecules and are discarded. The error depth and DBEC-adjusted depth arrows point to the respective average depths.

DBEC is applied to genes with an average non-singleton RSEC sequencing depth ≥4. This means that the depth is calculated after removing RSEC UMIs with only one representative read of ≥4. According to the Poisson distribution, if the average UMI depth is <4, more signal UMIs are removed than error UMIs. As a result, a gene is marked *pass* if its average RSEC depth ≥4 and is subject to DBEC; otherwise, it is marked *low depth* and bypasses DBEC. If no count is associated with the gene, it is labelled as *not detected*.

DBEC removes molecules and the reads associated with the removed molecules from consideration in downstream analyses. The percentage of reads retained by DBEC is reported together with the other pipeline metrics.

The RSEC and DBEC metrics associated with each gene are reported in the file, <sample_name>_UMI_Adjusted_Stats.csv.

Step 6. Determine putative cells

Excessive cell labels

In theory, the number of unique cell labels detected by the bioinformatics pipeline should be similar to the number of cells captured and amplified by the BD Rhapsody™ workflow. However, various processes throughout the workflow can introduce noise that contribute to excessive cell labels generated during sequencing analysis, including:

- Hybridizing polyadenylated [poly(A)] oligonucleotides to beads residing in neighboring wells when the cell lysis step is too long
- Underloading beads in BD Rhapsody™ Cartridges resulting in cells without beads and the RNA from the cells diffusing to adjacent wells
- Experiencing low-level contamination during oligonucleotide and bead synthesis
- Generating errors during the PCR amplification steps of the workflow

To distinguish cell labels associated with putative cells from those associated with noise, a multi-step algorithm was designed for filtering cell labels. See Figure 8.

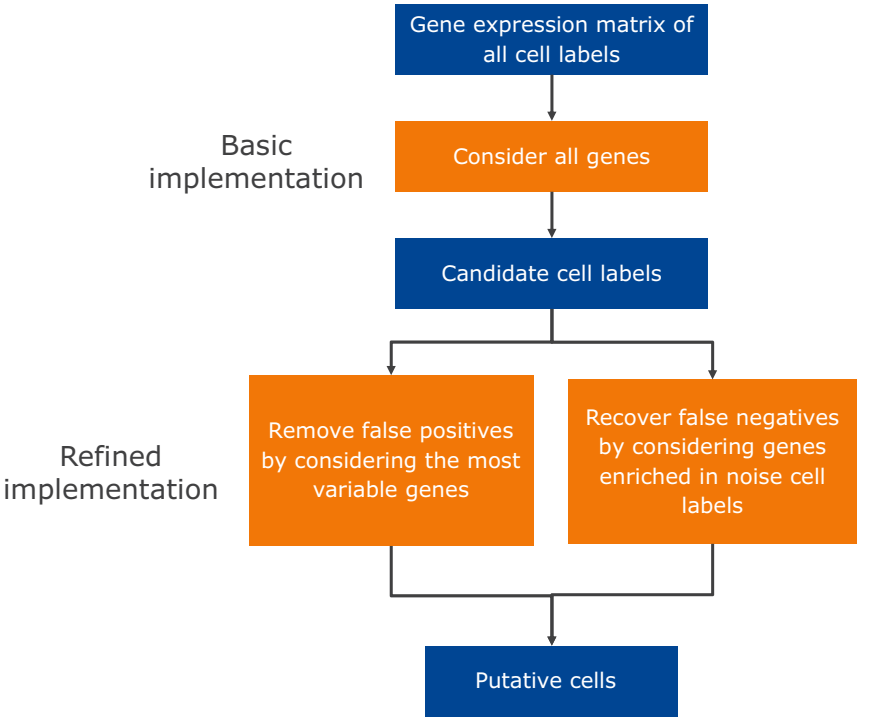


Figure 8. Workflow for determining putative cells.

Putative cell identification using second derivative analysis (basic implementation)

The principle of the cell label filtering algorithm is that cell labels from actual cell capture events should have many more reads associated with them than noise cell labels. All reads associated with DBEC-adjusted molecules from all genes are taken into account. The number of reads (post-DBEC) of each cell is plotted on a $\log_{10}$-transformed *cumulative* curve, with cells sorted by the number of reads in descending order. See Figure 9, left. In a typical experiment, a distinct inflection point is observed, indicated by the red vertical line. The algorithm finds the minimum second derivative along the cumulative reads curve as the inflection point. See Figure 9, right. Cell labels to the left of the red vertical line (Figure 9, left) are most likely derived from a cell capture event and are considered as signal (labeled as *cell labels set A* or *candidate cell labels*). The remaining cell labels to the right of the red line are noise. Up to this point, the analysis is the *basic* implementation of the second derivative analysis.

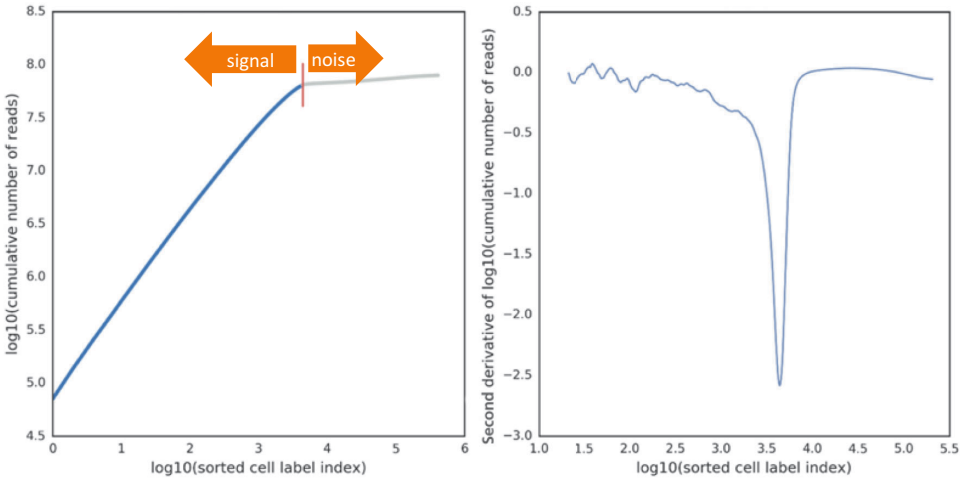
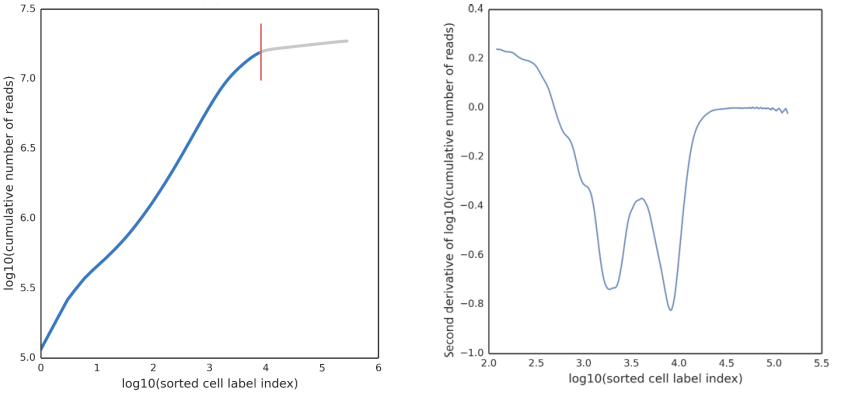


Figure 9. Results of the basic implementation of the second derivative analysis applied to a typical BD Rhapsody™ library.

If every cell in the sample is well represented by panel genes, there is only one inflection point. The number of reads of the putative cells is a single distribution well separated from the noise distribution.

There are situations, however, when a sample contains cells with a very wide range of number of molecules of genes in the panel. If subpopulations of cells with high and low mRNA content are considerably large, multiple inflection points can be observed. Example scenarios include biological samples such as peripheral blood mononuclear cells (PBMCs) with plasma cells being much larger and active carrying thousands of molecules in the panel and lymphocytes being smaller and less active carrying tens of molecules in the panel (see Figure 10A) or artificial mixtures of cell lines cells and primary cells (see Figure 10B). The basic implementation of the second derivative analysis chooses the inflection point that includes all distributions beyond the usual noise distribution. Specifically, inflection points are considered valid if the second derivative minimum corresponding to the inflection point is ≤ -0.3. The smoothing window of the second derivative curve increases until there are two valid inflection points. The inflection point corresponding to the larger cell number is deemed the better one.

A. PBMCs containing myeloid cells with high mRNA content and lymphocytes with low mRNA content



B. Jurkat and Ramos cell lines (high mRNA content) mixed with PBMCs (low mRNA content)

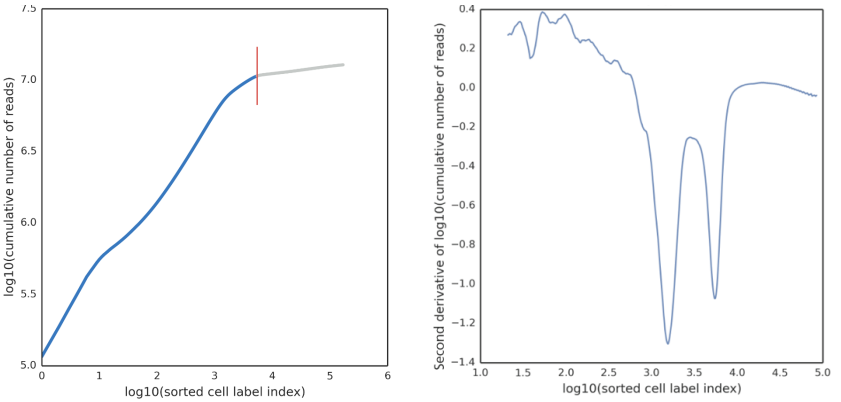


Figure 10. Results of basic implementation of the second derivative analysis on libraries with very different levels of mRNA content. A. PBMCs with myeloid (high mRNA content) and lymphoid (low mRNA content) cells. B. Mixture of Jurkat and Ramos cells (cell lines, high mRNA content) and PBMCs (low mRNA content). Both libraries were analyzed with the BD Rhapsody™ Immune Response Panel Hs (human).

Removing false positives and false negatives (refined implementation)

In some cases, the basic implementation of the second derivative analysis might include small numbers of false positive and false negative cell labels. Additional refinement steps are implemented to identify these false positive cell labels in order to generate a final set of cell labels for further analysis.

Removing false positives

Consider the case where the chosen inflection point includes the populations of cell labels with wide ranges of number of reads per cell label. Then, the signal population with lower reads per cell label might also include noise cell labels derived from residual mRNA molecules from the cells with very high mRNA content. The number of reads associated with these noise cell labels derived from high-expressing cells can be indistinguishable from low-expressing cells, which have similar reads per cell.

Since these false positive cells can be hard to identify with reads alone, the relative gene expression profile of cell labels can be used to identify them. For example, a false positive cell label that is derived from a high mRNA-expressing, true positive cell label would likely have a similar gene expression profile but with a lower read signal. Therefore, a second derivative analysis is done on the most variable genes to identify these false positive cell labels.

The most variable genes are defined by a process similar to that described by Macosko, EZ, et al. [see References (page 70)]:

a. Log-transform read counts of each gene within each cell to get the gene expression: $\log_{10}(\text{count} + 1)$.

b. Calculate the mean expression and dispersion (defined as variance/mean) for each gene.

c. Place genes into 20 bins based on their average expression.

d. Within each bin, calculate the mean and standard deviation of the dispersion measure of all genes, and then calculate the normalized dispersion measure of each gene using the following equation:

   Normalized dispersion =
   (dispersion – mean)/(standard deviation)

e. Apply a cutoff value for the normalized dispersion to identify genes for which expression values are highly variable even when compared to genes with similar average expression.

A second derivative analysis is applied on variable gene sets defined by a different cutoff value for the normalized dispersion to derive the *cell label filtered set B*. For each dispersion cutoff, the noise cell labels are determined as A – B. For instance, for three cutoff values, noise cell labels are N1 = A – B1, N2 = A – B2, and N3 = A – B3, where the minus sign represents the set difference. The common noise cell labels detected among N1, N2, and N3 are subtracted from cell labels set A. The resultant set is denoted as *cell label filtered set C* = A – intersection(N1, N2, N3).

Recovering false negatives

Cells with low numbers of molecules might be missed by the basic implementation of the second derivative analysis algorithm, because a cell subset might express very few of the genes in the panel. The cell labels carry a very low number of reads, and the size of the cell population is small enough that their cell labels do not form a distinct second inflection point. These cell labels might be mistaken as noise.

If there are genes specific to the false negative cell label subset (for example, marker genes), they can be identified by comparing the number of reads for each gene from all detected cell labels to those from cell labels deemed as signal. The assumption is that the relative abundance of reads for each gene from all of the noise cell labels should be no different than that from all of the cell labels considered as signal. If a specific cell subset is missed initially, there is a set of genes that appears as enriched in the noise cell labels in the basic implementation.

This enriched set of genes is detected by the following steps:

a. For each gene, calculate the total read counts from all detected cell labels and from cell labels in set C.

b. Identify the genes that have the biggest discrepancy in representation by cell labels in set C versus all cell labels. This is done by plotting and finding the line of best fit to detect the genes with the largest residuals at least one standard deviation away from the median of residuals of all genes. See Figure 11.

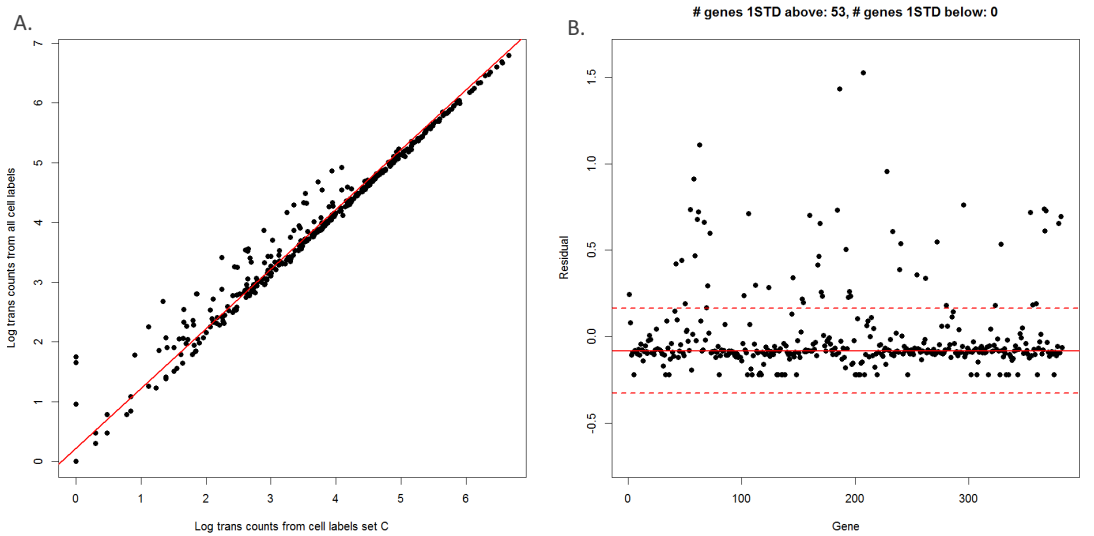


Figure 11. A. and B. Detecting genes enriched in noise as determined by the basic implementation of the second derivative analysis. Each dot represents a gene. B. The two red dashed lines correspond to one standard deviation above and below the median (red solid line). In this example, 53 genes are enriched in the noise population.

The second derivative analysis algorithm is run again with this enriched set of genes. The recovered cell labels (*cell label filtered set D*) are combined with cell labels in *set C* to form *set E*. As a final cleanup step, cell labels carrying less than the minimum threshold number of molecules are removed. The number of cell labels in the final set is *the number of putative cells*.

Reporting putative cells

The category of each cell label is listed in the file <sample_name>_Putative_Cells_Origin.csv. The cell label is marked *basic* if it is considered a putative cell in the basic implementation when the second derivative analysis is run using data from all genes in the panel. A cell label is marked as *refined* if it is considered a putative cell in the refined implementation and is a recovered false negative. In most cases, most putative cell labels originate from the basic implementation. See Putative cells origin (page 57).

Step 7. Determine the sample of origin (sample multiplexing only)

Sample multiplexing option

Up to 12 samples of cell suspension can be loaded into a BD Rhapsody Cartridge using a BD™ Single-Cell Multiplexing Kit. Each sample is labelled with a separate Sample Tag from the kit.

When you start the BD Rhapsody Analysis pipeline, you can select the sample multiplex option. You can associate a name with a Sample Tag before the pipeline starts, and the specified sample names will be used in the output files.

To account for every Sample Tag, each Sample Tag sequence in the kit is considered during pipeline analysis, whether the Sample Tags are used in the experiment or specified with a sample name.

The pipeline automatically adds the Sample Tag sequences to the FASTA reference file. Reads that align to a Sample Tag sequence and associate with a putative cell are used to identify the sample for that cell.

Sample determination algorithm

The algorithm first identifies high quality singlets. A high quality singlet is a putative cell where more that 75% of Sample Tag reads are from a single tag. When a singlet is identified, the counts for all the other tags are considered Sample Tag noise. See Figure 12. Sources of low-level noise can be PCR and sequencing errors and residual Sample Tag labelling during cell preparation.

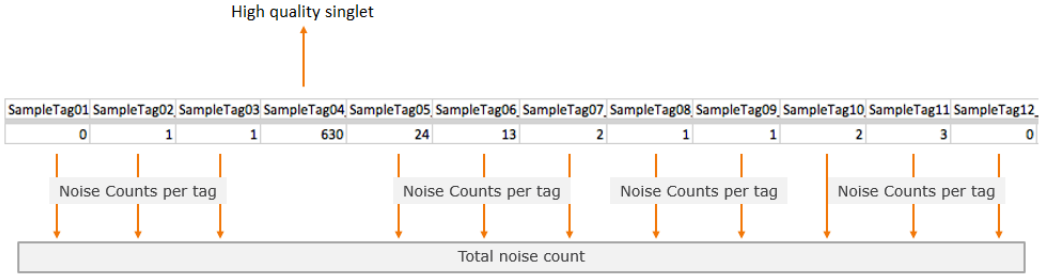


Figure 12. Example of Sample Tag read counts for a putative cell that is considered a high quality singlet, labelled SampleTag04. All of the other Sample Tag counts are recorded as separate noise counts and are summed to find the noise read count for that putative cell.

The minimum Sample Tag read count for a putative cell to be positively identified with a Sample Tag is defined as the lowest read count of a high quality singlet for that Sample Tag. See Figure 13.

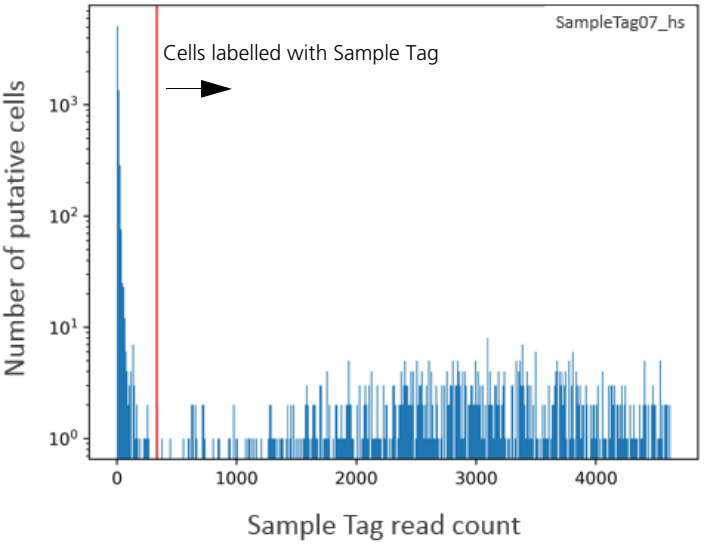


Figure 13. Histogram of number of Sample Tags per putative cell for one of the 12 Sample Tags. The red vertical line indicates the threshold of minimum Sample Tag read count. Putative cells with Sample Tag read counts greater than the threshold (to the right of the red line) are considered labelled with this Sample Tag. In addition to singlets, these putative cells can include multiplets, which are cell labels associated with more than one Sample Tag.

The percentage noise contribution of each Sample Tag of all cells is calculated by dividing the total per tag noise by the total overall noise. In addition, the total amount of noise versus the total Sample Tag count per putative cell is recorded so that a trend line can be established to estimate the total per-cell noise given an observed number of total Sample Tag count for a cell. See Figure 14. The level of antigen expression across cells can vary, contributing to variation in Sample Tag count per cell. Generally, cells with higher total Sample Tag counts have higher noise Sample Tag counts.

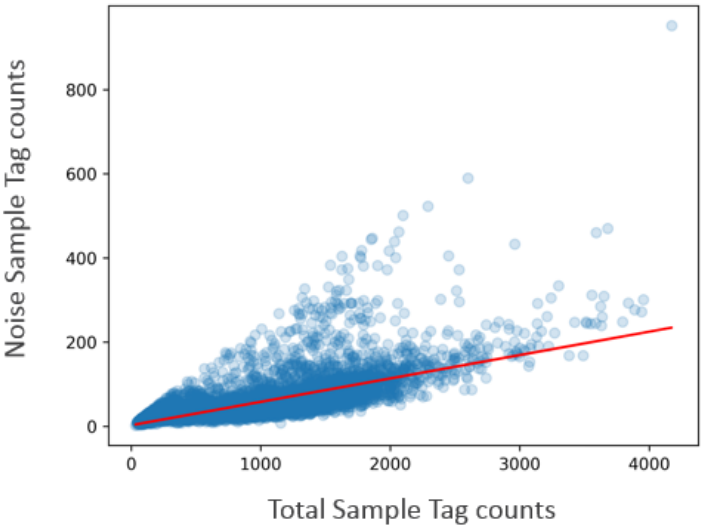


Figure 14. Overall noise profile where each dot is a cell. A trend line (in red) is fitted and used to establish the expected amount of noise given a total Sample Tag count. Cells that are off the trend line are likely multiplets.

To improve sample determination and recover singlets that are not initially considered high quality, the algorithm subtracts the expected number of per-cell noise counts from each Sample Tag. The total expected per-cell noise, derived from the trend line, is multiplied by the percentage noise contribution of each Sample Tag to determine the expected noise per Sample Tag.

After subtracting the expected per tag noise, any Sample Tag that has a count higher than its minimum read count is called for that cell, and the putative cell is considered a *called* cell.

When the counts of two or more Sample Tags exceed their minimum thresholds, then that putative cell is called as a cross-sample *Multiplet*, indicating more than one actual cell in the microwell, and the cells are of different samples of origin. Some putative cells might not have enough Sample Tag counts to definitively call their sample of origin, and those are labeled as *Undetermined*.

Reporting sample origin

If you chose the sample multiplexing option, the main top-level RSEC and DBEC data tables contain counts for putative cells from all samples combined. The sample of origin for each putative cell is listed in the file <sample_name>_Sample_Tag_Calls.csv. This file can be used to annotate the combined data tables. The file, <sample_name>_Sample_Tag_Metrics.csv reports the metrics from the sample determination algorithm. Per sample data tables and cluster analysis are output in folders contained in <sample_name>_Sample_Tag<number>.zip.

Step 8. Generate expression matrices

Reporting RSEC and DBEC metrics

RSEC-adjusted molecule counts and associated reads of each gene for each putative cell and DBEC-adjusted molecule counts and associated reads are presented in either .csv or .st format. See Expression data (page 53) and Data tables (page 51).

Step 9. Annotate SAM

Annotating SAM

The SAM file output by Bowtie2 is further annotated to summarize the results of the BD Rhapsody Analysis pipeline. The table lists the tags appended to the annotation of each read. The SAM file is then converted to BAM format and output. For BAM tags, see BAM (page 50), samtools.github.io/hts-specs/SAMv1.pdf, and bowtie-bio.sourceforge.net/bowtie2/manual.shtml#sam-output.

Step 10. Generate metrics summary

Summary

A summary .csv file documenting the metrics of each of the analysis steps is generated. See Metrics summary (page 41).

Step 11. Clustering analysis

Clustering algorithm

The measured single cell gene expression profiles go through a clustering analysis pipeline. See BD Rhapsody™ Targeted clustering analysis (page 73).

Reviewing sequencing analysis output files

Before you begin

Obtain the output files after running the appropriate pipeline on the Seven Bridges Genomics platform or on a local installation. See the *BD Single Cell Genomics Analysis Setup User Guide* (Doc ID: 47383).

Sequencing analysis outputs

Most outputs contain a header summarizing the pipeline run. Headers contain all of the information needed to re-run the pipeline with the same settings.

Output	File	Content
Metrics summary (page 41)	<sample_name>_Metrics_Summary.csv	Report containing sequencing, molecules, and cell metrics
BAM (page 50)	<sample_name>.final.BAM	Alignment file of R2 and associated R1 annotations
Data tables (page 51)	<sample_name>_RSEC_MolsPerCell.csv <sample_name>_RSEC_ReadsPerCell.csv <sample_name>_DBEC_MolsPerCell.csv <sample_name>_DBEC_ReadsPerCell.csv	Reads per gene per cell and molecules per gene per cell, based on RSEC or DBEC
	<sample_name>_RSEC_MolsPerCell_Unfiltered.csv.gz <sample_name>_RSEC_ReadsPerCell_Unfiltered.csv.gz <sample_name>_DBEC_MolsPerCell_Unfiltered.csv.gz <sample_name>_DBEC_ReadsPerCell_Unfiltered.csv.gz	Unfiltered tables containing all cell labels of ≥5 reads

Output (continued)	File	Content
Expression data (page 53)	<sample_name>_Expression_Data.st	The expression sparse matrix, a table of counts in sparse format
	<sample_name>_Expression_Data _Unfiltered.st.gz	Compressed file containing all cell labels of ≥5 reads
Cell label filtering (page 55)	<sample_name>_Cell_Label_Filter.png	Visualization of cell label filtering results
Second derivative curve (page 56)	<sample_name>_Cell_Label _Second_Derivative_Curve.png	
Putative cells origin (page 57)	<sample_name>_Putative_Cells _Origin.csv	Algorithm that found the putative cell: basic or refined
UMI metrics (page 58)	<sample_name>_UMI_Adjusted_Stats.csv	Metrics from RSEC and DBEC molecular identifier adjustment algorithms on a per-gene basis
Sample Tag metrics (sample multiplexing option selected) (page 60)	<sample_name>_Sample_Tag_Metrics.csv	Metrics from the sample determination algorithm

Output (continued)	File	Content
Sample Tag calls (sample multiplexing option selected) (page 62)	<sample_name>_Sample_Tag_Calls.csv	Assigned Sample Tag for each putative cell
Per sample folder (sample multiplexing option selected) (page 63)	<sample_name> _Sample_Tag<number>.zip <sample_name>_Multiplet_and _Undetermined.zip	Data tables, expression matrix, and clustering analysis files for a particular sample. **Note:** For putative cells that could not be assigned a specific Sample Tag, a Multiplet_and _Undetermined.zip file is also output.
Clustering analysis	ClusteringAnalysis.zip	See Clustering analysis outputs (page 79)

Metrics summary

File: <sample_name>_Metrics_Summary.csv

The Metrics summary provides statistics on sequencing, molecules, cells, and targets. Here is an example of a portion of the output:

Note: Sample Tag and AbSeq metrics display only when they are used in an experiment.

#Sequencing Quality#					
Total_Reads_in_FASTQ	Pct_Reads_Too_Short	Pct_Reads_Low_Base_Quality	Pct_Reads_High_SNF	Pct_Reads_Filtered_Out	Total_Reads_After_Quality_Filtering
1571225	0.04	5.44	2.71	7.69	1450457
66394695	1.9	1.05	0.18	2.96	64429542
67965920	1.85	1.15	0.24	3.07	65879999
#Library Quality#					
Total_Filtered_Reads	Pct_Contaminating_PhiX_Reads_in_Filtered_R2	Pct_Q30_Bases_in_Filtered_R2	Pct_Assigned_to_Cell_Labels	Pct_Cellular_Reads_Aligned_Uniquely_to_Amplicons	Library
1450457	0	67.83	93.38	85.74	J80FC1G
64429542	0	85.31	96.28	92.49	J80FC1O
65879999	0	84.93	96.21	92.34	Combined_stats
#Reads and Molecules#					
Aligned_Reads_By_Type	Total_Raw_Molecules	Total_RSEC_Molecules	Total_DBEC_Molecules	Mean_Raw_Sequencing_Depth	Mean_RSEC_Sequencing_Depth
1216821	104733	104733	51686	11.62	11.62
59616320	24717877	22724878	22724878	2.41	2.62
60833141	24822610	22829611	22776564	2.45	2.66
#Cells RSEC#					
Putative_Cell_Count	Pct_Reads_from_Putative_Cells	Mean_Reads_per_Cell	Mean_Molecules_per_Cell	Median_Molecules_per_Cell	Mean_Targets_per_Cell
898	89.39	1211.27	65.95	61	65.95
898	74.62	49539.76	18792.04	17385.5	38.53
898	74.92	50751.03	18857.99	17434	104.47
#Cells DBEC#					
Putative_Cell_Count	Pct_Reads_from_Putative_Cells	Mean_Reads_per_Cell	Mean_Molecules_per_Cell	Median_Molecules_per_Cell	Mean_Targets_per_Cell
898	95.85	1153.4	50.34	45	50.34
898	74.62	49539.76	18792.04	17385.5	38.53
898	73.37	49590.1	18857.99	17434	104.47
#Targets#					
Number_of_Pass_Targets	Number_of_Undersequenced_Targets	Number_of_Targets_in_Panel	Target_Type		
296	52	399	mRNA		
0	40	40	AbSeq		
#Sample_Tags#					
Sample_Tag_Filtered_Reads	ST_Pct_Reads_from_Putative_Cells				
639495	73.65				

Section/metric	Definition	Major contributing factors
Sequencing Quality		
Total_Reads_in_FASTQ	Number of read pairs in the input FASTQ files	Sequencing amount
Pct_Reads_Too_Short	Percentage of read pairs filtered out due to length of either R1<66 bp or R2<64 bp	Sequencing quality
Pct_Reads_Low_Base_Quality	Percentage of reads filtered out due to average base quality score of R1 reads <20 or R2 reads <20	Sequencing quality
Pct_Reads_High_SNF	Percentage of read pairs filtered out due to single nucleotide frequency ≥55% for R1 or ≥80% for R2	Sequencing quality
Pct_Reads_Filtered_Out	Percentage of reads removed by the combination of length, quality, and SNF filters	Sequencing quality
Total_Reads_After_Quality_Filtering	Number of read pairs after length, quality, and SNF filtering	• Sequencing amount • Sequencing run quality • Library quality
Library	Name of library	Name of library
Library Quality **Note:** Sequencing contains an extra row if BAM input was used. The first row contains the metrics of the FASTQ inputs only. The last row contains the combined metrics for FASTQ and BAM inputs.		
Total_Filtered_Reads	Number of read pairs after length, quality, and SNF filtering	• Sequencing amount • Sequencing run quality • Library quality

Section/metric (continued)	Definition	Major contributing factors
Pct_Contaminating _PhiX_Reads_in _Filtered_R2	Percentage of read pairs after quality filtering that are aligned to the PhiX control	• Sequencing run quality • Amount of PhiX spiked in
Pct_Q30_Bases_in _Filtered_R2	Percentage of R2 bases with quality score >30, averaged across all read pairs retained after quality filtering	Sequencing quality
Pct_Assigned_to_Cell _Labels	Percentage of read pairs containing a valid cell label	• Sequencing quality • Library quality
Pct_Cellular_Reads _Aligned_Uniquely_to _Amplicons	Percentage of read pairs containing a valid cell label and UMI that aligned uniquely to an amplicon presented in the panel reference	• Sequencing quality • Library quality
Library	Name of library	Name of library

Section/metric (continued)	Definition	Major contributing factors
Reads and Molecules		
Aligned_Reads_By _Type	Number of filtered read pairs aligned to target type	• Sequencing quality • Library quality • Panel compatibility with sample composition
Total_Raw_Molecules	Total number of molecules as defined by the unique combination of cell label, gene identity, and UMI	• Sequencing depth • Panel compatibility with sample composition
Total_RSEC_Molecules[a]	Total number of molecules detected after the RSEC molecular identifier adjustment algorithm	• Sequencing depth • Panel compatibility with sample composition
Total_DBEC_Molecules[a]	Total number of molecules detected after RSEC and DBEC molecular identifier adjustment algorithms	• Sequencing depth • Panel compatibility with sample composition
Mean_Raw_Sequencing _Depth	Average number of read pairs per molecule before molecular identifier adjustment algorithms	Sequencing depth
Mean_RSEC _Sequencing_Depth	Average number of read pairs per molecule after the RSEC molecular identifier adjustment algorithm	Sequencing depth
Mean_DBEC _Sequencing_Depth	Average number of read pairs per molecule after RSEC and DBEC molecular identifier adjustment algorithms	Sequencing depth
Sequencing_Saturation	Percentage of read pairs representing RSEC-adjusted molecules that are sequenced more than once	Sequencing depth

Section/metric (continued)	Definition	Major contributing factors
Pct_Cellular_Reads _with_Amplicons _Retained_by_DBEC	Percentage of read pairs with valid cell labels and gene alignment retained after the DBEC molecular adjustment algorithm	Sequencing depth
Target_Type	Type of target in library (mRNA, AbSeq, or mRNA + AbSeq)	Panel composition
Cells RSEC **Note:** Cells RSEC contains the metrics from cell label filtering based on molecule data generated from the RSEC molecular index adjustment algorithm.		
Putative_Cell_Count[b]	Number of cell labels detected by the cell label filtering algorithm	• Number of cells input and captured by cartridge workflow • Bead handling • Panel compatibility with sample composition
Pct_Reads_from _Putative_Cells	Percentage of reads that are assigned to putative cells	• Cell viability • Cartridge workflow performance • Sequencing depth (for DBEC-derived metric only) • Panel compatibility with sample composition
Mean_Reads_per_Cell	Average number of reads representing the molecules detected in each cell	• Sequencing depth • Panel compatibility with sample composition
Mean_Molecules_per _Cell	Average number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition

Section/metric (continued)	Definition	Major contributing factors
Median_Molecules_per_Cell	Median number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Mean_Targets_per_Cell	Average number of targets detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Median_Targets_per_Cell	Median number of targets detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Total_Targets_Detected	Number of targets detected from all cells	• Sequencing depth • Panel compatibility with sample composition
Target_Type	Type of target in library (mRNA, AbSeq, or mRNA + AbSeq)	Panel composition
Cells DBEC **Note:** Cells contains the metrics from cell label filtering based on molecule data generated from the RSEC and DBEC molecular index adjustment algorithm.		
Putative_Cell_Count[b]	Number of cell labels detected by the cell label filtering algorithm	• Number of cells input and captured by cartridge workflow • Bead handling • Panel compatibility with sample composition
Pct_Reads_from_Putative_Cells	Percentage of reads that are assigned to putative cells	• Cell viability • Cartridge workflow performance • Sequencing depth (for DBEC-derived metric only) • Panel compatibility with sample composition

Section/metric (continued)	Definition	Major contributing factors
Mean_Reads_per_Cell	Average number of reads representing the molecules detected in each cell	• Sequencing depth • Panel compatibility with sample composition
Mean_Molecules_per _Cell	Average number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Median_Molecules_per _Cell	Median number of molecules detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Mean_Targets_per_Cell	Average number of targets detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Median_Targets_per _Cell	Median number of targets detected per cell label	• Sequencing depth • Panel compatibility with sample composition
Total_Targets_Detected	Number of targets detected from all cells	• Sequencing depth • Panel compatibility with sample composition
Target_Type	Type of target in library (mRNA, AbSeq, or mRNA + AbSeq)	Panel composition
Targets		
Number_of_Pass _Targets	Number of targets with pass status: the targets have sufficient sequencing depth to be considered for adjustment by the DBEC molecular identifier algorithm	• Sequencing depth • Panel compatibility with sample composition

Section/metric (continued)	Definition	Major contributing factors
Number_of _Undersequenced _Targets	Number of targets not having sufficient sequencing depth to be considered for adjustment by the DBEC molecular identifier algorithm	• Sequencing depth • Panel compatibility with sample composition
Number_of_Targets _in_Panel	The number of targets featured in the panel	Panel choice
Target_Type	Type of target in library (mRNA, AbSeq, or mRNA + AbSeq)	Panel composition

Section/metric (continued)	Definition	Major contributing factors
Sample Tags (If used in the experiment)		
Sample_Tag_Filtered _Reads	Number of filtered read pairs aligned to Sample Tags	• Sequencing depth • Panel compatibility with sample composition
ST_Pct_Reads_from _Putative_Cells	Percentage of Sample Tag reads that are assigned to putative cells	• Cell viability • Sample Tag labelling and wash protocols • Cartridge workflow performance • Sequencing depth (for DBEC-derived metric only) • Panel compatibility with sample composition

a. For more information on RSEC and DBEC molecular identifier adjustment algorithms, see Step 5. Annotate molecules (page 17).
b. For further information on how putative cells are defined in terms of the number of reads associated with true and noise cell labels, see Cell label filtering (page 55).

BAM

File: <sample_name>.final.BAM

BAM is an alignment file in binary format that is generated by the aligner Bowtie2. Bowtie2 aligns R2 reads to the reference file and outputs tags related to alignment quality. This BAM file is sorted according to the alignment coordinates of R2 reads on each chromosome.

The BD Rhapsody Analysis pipeline adds the following tags:

Tag	Definition
CB	A number between 1 and 96^3 (884,736) representing a unique cell label sequence (CB = 0 when no cell label sequence is detected)
MR	Raw molecular identifier sequence
MA	RSEC-adjusted molecular identifier sequence. If not a true cell, the raw UMI is repeated in this tag.
PT	T if a poly(T) tail was found in the expected position on R1, or F if poly(T) was not found
CN	Indicates if a sequence is derived from a putative cell, as determined by the cell label filtering algorithm (*T*: putative cell; *x*: invalid cell label or noise cell) **Note:** You can distinguish between an invalid cell label and a noise cell with the CB tag (invalid cell labels are 0).
ST	The value is 1–12, indicating the Sample Tag of the called putative cell, or *M* for multiplet, or *x* for undetermined.

Note: A BAM file can be converted to a tab-delimited text file (SAM format) by using SAMtools (see samtools.sourceforge.net).

Data tables

Files containing filtered data:

<sample_name>_RSEC_MolsPerCell.csv

<sample_name>_RSEC_ReadsPerCell.csv

<sample_name>_DBEC_MolsPerCell.csv

<sample_name>_DBEC_ReadsPerCell.csv

Compressed files containing unfiltered data:

<sample_name>_RSEC_MolsPerCell_Unfiltered.csv.gz

<sample_name>_RSEC_ReadsPerCell_Unfiltered.csv.gz

<sample_name>_DBEC_MolsPerCell_Unfiltered.csv.gz

<sample_name>_DBEC_ReadsPerCell_Unfiltered.csv.gz

Eight Data Table .csv files, four filtered and four unfiltered, are output. They contain reads per gene per cell and molecules per gene per cell.

For example:

Cell_Index	ADA	ADGRE1	ADGRG3	ADM	AIM2	ALAS2	ANXA5	AOC3
525435	5	0	0	0	0	0	0	0
268870	3	0	0	0	0	0	0	0
38817	22	0	0	0	0	0	0	0
24642	19	0	0	0	0	0	1	0
444017	5	0	0	0	0	0	0	0
771197	2	0	0	0	0	0	0	0
480465	8	0	0	0	0	0	1	0
161815	0	0	0	0	0	0	0	0
379509	2	0	0	0	0	0	0	0
757154	3	0	0	0	0	0	0	0
25539	4	0	0	0	0	0	0	0
548867	2	0	0	0	0	0	0	0
297014	0	0	0	0	0	0	0	0
714491	1	0	0	0	0	0	0	0
604203	0	0	0	0	0	0	0	0

- Each row represents the number of reads or molecules in a cell for each gene in the panel. A cell is identified with a unique cell index number under Cell_Index.
- The cell index is sorted in descending order based on the total number of reads. The cell order in the four files is the same.
- Genes are sorted alphabetically in the panel.
- For PerCell.csv files: Reads and molecules are counted only if they have passed all pipeline filters and have been determined to be from putative cells.
- For PerCell_Unfiltered.csv.gz: The files contain unfiltered tables with cell labels of ≥5 reads.

Note: It is generally recommended to use <sample_name>_DBEC_MolsPerCell.csv for clustering analysis. Read counts for DBEC, read counts for RSEC, and molecule counts for RSEC are provided for reference. The RSEC files can be used when sequencing depth is so low that most genes do not pass the threshold for the DBEC molecular identifier adjustment algorithm to be applied; that is, low_depth in <sample_name>_UMI_Adjusted_Stats.csv.

Expression data

File: <sample_name>_Expression_Data.st

Unfiltered file: <sample_name>_Expression_Data_Unfiltered.st.gz

Information is presented in sparse notation.

- Data.st: Reads and molecules are counted only if they have passed all pipeline filters and have been determined to be from putative cells.
- Unfiltered.st.gz: Compressed file containing all cell labels of ≥5 reads.

Open the .st file in a text editor.

Each row records counts for cell-gene combinations that have non-zero RSEC molecule counts.

For example:

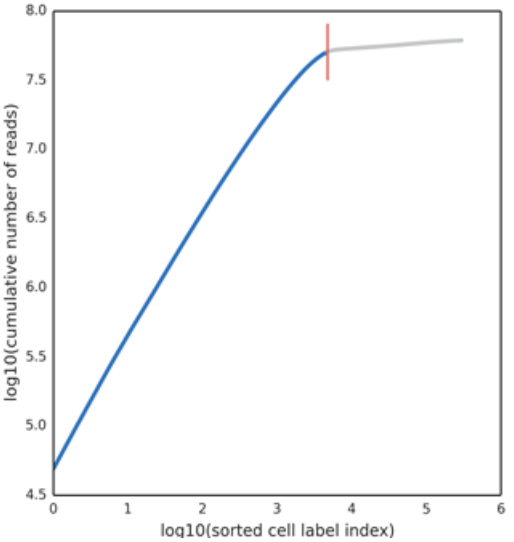
```
####################
## BD Targeted Rhapsody Analysis Pipeline Version 1.0
## Analysis Date: 2017-08-03 23:57:20
## Sample: mySample
## Reference: Immune_Response_Panel_Hs
####################
Cell_Index    Gene  RSEC_Reads      Raw_Molecules     RSEC_Adjusted_Molecules  DBEC_Reads  DBEC_
525435        ADA   5       5       5       5
525435        ATF6B         1       1       1       1       1
525435        AURKB         1       1       1       1       1
525435        BACH2         3       2       2       3       2
525435        BCL6  1       1       1       1
525435        BLNK_ALT      1       1       1       1       1
525435        BTG1_ALT1     1       1       1       1       1
525435        BTLA  1       1       1       1
525435        CD1C_ALT      1       1       1       1       1
525435        CD22  4       4       3       4       3
525435        CD27  2       2       2       2       2
525435        CD3D  4       3       3       4       3
```

Metric	Definition
Cell_Index	Unique cell index sorted by total number of reads per cell in descending order
Gene	Genes in panel listed in alphabetical order
RSEC_Reads	Number of reads after the RSEC molecular identifier adjustment algorithm
Raw_Molecules	Number of UMIs before molecular identifier adjustment algorithms
RSEC_Adjusted _Molecules	Number of UMIs after RSEC molecular identifier adjustment algorithm
DBEC_Reads	Number of reads remaining after the DBEC molecular identifier adjustment algorithm
DBEC_Adjusted _Molecules	Number of UMIs after RSEC and DBEC molecular identifier adjustment algorithms

Cell label filtering

File: <sample_name>_Cell_Label_Filter.png

This is an example output plot from a high quality BD Rhapsody™ experiment:

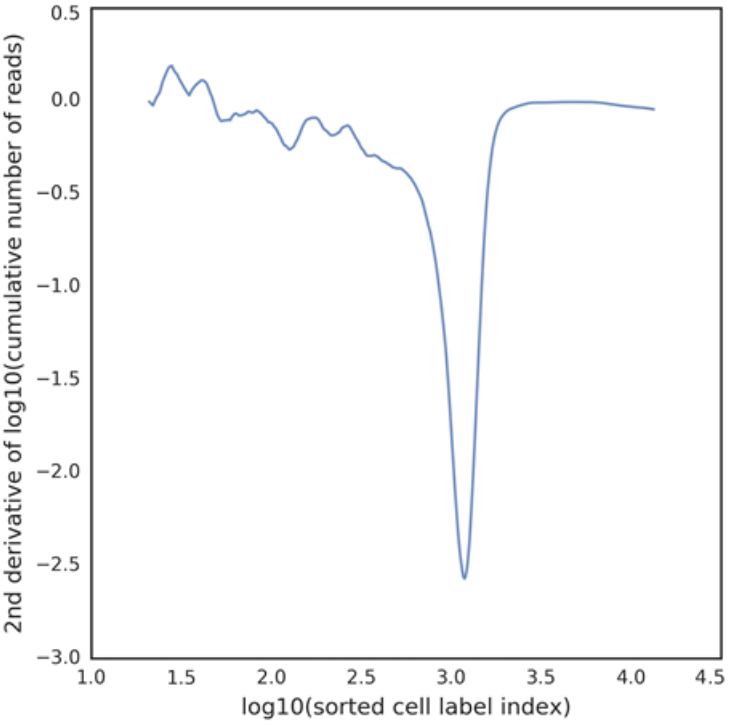


The cell label filter plot and the second derivative curve (see Second derivative curve (page 56)) are outputs from the basic implementation of the second derivative analysis algorithm for determining putative cells. For details on determining putative cells, see Step 6. Determine putative cells (page 23).

Second derivative curve

File: <sample_name>_Cell_Label_Second_Derivative_Curve.png

This plot is the second derivative of the cell label filter output plot:



Putative cells origin File: <sample_name>_Putative_Cells_Origin.csv

The output lists the step in the cell label filtering algorithm that determined a particular cell is a putative cell. If the cell label is categorized as putative in the basic implementation of the second derivative analysis, it is labeled *Basic*. If the cell label is a recovered false negative in the refined implementation, it is labeled *Refined*. See Step 6. Determine putative cells (page 23). For example:

##########	
## BD Targeted Rhapsody Analysis Pipeline Version 1.0	
## Analysis Date: 2017-08-03 23:57:20	
## Sample: mySample	
## Reference: Immune_Response_Panel_Hs	
##########	
Cell_Index	Algorithm
525435	Basic
268870	Basic
38817	Basic
24642	Basic
444017	Basic
771197	Basic
480465	Basic
161815	Basic
379509	Basic
757154	Basic
25539	Basic
548867	Basic

UMI metrics

File: <sample_name>_UMI_Adjusted_Stats.csv

The molecular identifier adjustment algorithms RSEC and DBEC are applied to each gene. The molecular identifier metrics file lists the metrics from RSEC and DBEC on a per-gene basis. For more information on RSEC and DBEC molecular identifier adjustment algorithms, see Step 5. Annotate molecules (page 17). For example:

```
####################
## BD Targeted Rhapsody Analysis Pipeline Version 1.0
## Analysis Date: 2017-08-14 23:27:15
## Sample: mySample
## Reference: immune_response_panel_hs_with_phix
####################
```

Gene	Status	Raw_Reads	Raw_Molecu	Raw_Seq_De	RSEC_Adjust	RSEC_Adjust	RSEC_Adjust	DBEC_Minim	DBEC_Adjust	DBEC_Adjust	DBEC_Adjust	Pct_Error_Re	Error_Depth
ADA|NM_0(	low_depth	20706	19762	1.05	19669	1.05	2.06	1	20706	19669	1.05	0	0
ADGRE1|NM	low_depth	5	5	1	5	1	0	1	5	5	1	0	0
ADGRG3|NM	low_depth	3	3	1	3	1	0	1	3	3	1	0	0
AIM2|NM_0	not_detecte	0	0	0	0	0	0	0	0	0	0	0	0
ALAS2|NM_(	not_detecte	0	0	0	0	0	0	0	0	0	0	0	0

Metric	Definition
Gene	Gene in panel listed in alphabetical order
Status	Gene status across all reads and molecules: • Not detected: Gene is in the panel but was not detected, because it has zero reads • Low depth: Minimum sequencing depth not achieved • Pass: Minimum sequencing depth has been achieved
Raw_Reads	Number of reads before molecular identifier adjustment algorithms
Raw_Molecules	Number of UMIs before molecular identifier adjustment algorithms
Raw_Seq_Depth	Number of raw reads ÷ the number of raw molecules
RSEC_Adjusted _Molecules	Number of molecules detected after RSEC molecular identifier adjustment algorithm

Metric (continued)	Definition
RSEC_Adjusted_Seq _Depth	Number of raw reads ÷ the number of RSEC-adjusted molecules
RSEC_Adjusted_Seq _Depth_without _Singletons	Number of raw reads ÷ the number of RSEC-adjusted molecules without considering molecules represented by only one read
DBEC_Minimum _Depth	Threshold of RSEC depth for a molecule to be considered a putative molecule by DBEC
DBEC_Adjusted_Reads	Number of reads retained after DBEC molecular identifier adjustment algorithm
DBEC_Adjusted _Molecules	Number of molecules retained after RSEC and DBEC
DBEC_Adjusted_Seq _Depth	Number of DBEC-adjusted reads ÷ the number of molecules detected after RSEC and DBEC
Pct_Error_Reads	Percentage of reads removed by DBEC molecular identifier adjustment algorithm
Error_Depth	RSEC depth of molecules that are removed by DBEC correction

Sample Tag metrics (sample multiplexing option selected)

File: <sample_name>_Sample_Tag_Metrics.csv

The Sample Tag metrics file contains statistics on the reads aligned to each Sample Tag and cells called for each sample. For example:

##							
## BD Targeted Multiplex Rhapsody Analysis Pipeline Version 1.01							
## Analysis Date: 2017-10-27 08:07:06							
## Sample: T26FC1NB							
## Reference: onco_bc_panel_hs_with_phix							
## Sample Tags Version: Hs							
##							
Sample_Tag	Sample_Nam	Raw_Reads	Pct_of_Raw_Reads	Cells_Called	Pct_of_Putative_Ce	Raw_Reads_in_Called	Mean_Reads_per
All_Tags		16163862	100	1787	100	0	0
SampleTag01_hs	Jurkat_1	2938864	18.18	262	14.66	1616700	6170.61
SampleTag02_hs	Jurkat_2	3928186	24.3	273	15.28	2175688	7969.55
SampleTag03_hs	Ramos_1	4052350	25.07	265	14.83	1997990	7539.58
SampleTag04_hs	Ramos_2	4171232	25.81	278	15.56	2126098	7647.83
SampleTag05_hs	T47D_1	484744	3	356	19.92	315126	885.19
SampleTag06_hs	T47D_2	588480	3.64	291	16.28	377908	1298.65
Multiplet		0	0	59	3.3	0	0
Undetermined		0	0	3	0.17	0	0

File	Description	Major contributing factors
Sample_Tag	List of the Sample Tags in the pipeline run	—
Sample_Name	User-provided sample name	—
Raw_Reads	Number of reads aligned to each Sample Tag	Sample Tag sequencing amount
Pct_of_Raw_Reads	Percentage of Sample Tag reads aligned to each Sample Tag	Sample Tag sequencing amount
Cells_Called	Number of putative cells called for each Sample Tag	• Number of cells input and captured by cartridge workflow • Sample Tag sequencing amount

File (continued)	Description	Major contributing factors
Pct_of_Putative_Cells _Called	Percentage of putative cells called for each Sample Tag	• Number of cells input and captured by cartridge workflow • Sample Tag sequencing amount
Raw_Reads_in_Called _Cells	Number of Sample Tag reads that are assigned to called cells	Sample Tag sequencing amount
Mean_Reads_per _Called_Cell	Average number of Sample Tag reads representing each called cell	Sample Tag sequencing amount

Sample Tag calls (sample multiplexing option selected)

File: <sample_name>_Sample_Tag_Calls.csv

The Sample Tag calls file contains the determined sample call for every putative cell. Sample names that you provided are included in a separate column. The Sample Tag calls file can be used to annotate the main data tables, which contain results from all samples. For example:

#####################		
## BD Targeted Multiplex Rhapsody Analysis Pipeline Version 1.01		
## Analysis Date: 2017-10-27 08:07:06		
## Sample: T26FC1NB		
## Reference: onco_bc_panel_hs_with_phix		
## Sample Tags Version: Hs		
#####################		
Cell_Index	Sample_Tag	Sample_Name
205097	SampleTag05_hs	T47D_1
165394	SampleTag05_hs	T47D_1
855569	SampleTag01_hs	Jurkat_1
249537	SampleTag03_hs	Ramos_1
323327	SampleTag04_hs	Ramos_2
696623	Multiplet	Multiplet
635228	SampleTag05_hs	T47D_1
314225	SampleTag02_hs	Jurkat_2
4570	SampleTag01_hs	Jurkat_1
570473	Undetermined	Undetermined
199238	SampleTag02_hs	Jurkat_2
293711	SampleTag03_hs	Ramos_1

File	Description
Cell_Index	Unique cell identifier
Sample_Tag	List of the Sample Tags in the pipeline run
Sample_Name	User-provided sample name

Per sample folder (sample multiplexing option selected)	File: <sample_name> _Sample_Tag<number>.zip or <sample_name>_Multiplet_and_Undetermined.zip Either zipped file includes: • <sample_name> _Sample_Tag<number>_DBEC_MolsPerCell.csv • <sample_name> _Sample_Tag<number>_DBEC_ReadsPerCell.csv • <sample_name> _Sample_Tag<number>_RSEC_MolsPerCell.csv • <sample_name> _Sample_Tag<number>_RSEC_ReadsPerCell.csv • <sample_name> _Sample_Tag<number>_Expression_Data.st • ClusteringAnalysis/ Each sample with at least one called putative cell will generate a sample-specific folder containing data tables and a cluster analysis. The formats of the files are the same as described in Data tables (page 51) and Clustering analysis outputs (page 79). Data for putative cells that could not be assigned to a specific sample are found in the Multiplet and Undetermined folder.

Assessing BD Rhapsody library quality with skim sequencing

Introduction

Several output metrics from the BD Rhapsody Analysis pipeline can be evaluated while performing skim sequencing to assess library and sequencing run quality. Output metrics are stable at low sequencing depth (~2 million sequencing reads or higher).

Metrics for evaluation with skim sequencing

Read quality
• Pct_Reads_Too_Short • Pct_Reads_Low_Base_Quality • Pct_Reads_High_SNF • Pct_Reads_Filtered_Out
Sequencing alignment
• Pct_Q30_Bases_in_Filtered_R2 • Pct_Assigned_to_Cell_Labels • Pct_Cellular_Reads_Aligned_Uniquely_to_Amplicons
Cells detected
• Putative_Cell_Count (RSEC)[a] • Pct_Reads_from_Putative_Cells (RSEC)[b] • Putative_Cell_Count (DBEC)[a]

a. By metric definition, Putative_Cell_Count (RSEC) has the same value as Putative_Cell_Count (DBEC). Putative_Cell_Count (RSEC) and Putative_Cell_Count (DBEC) might vary by up to ±5% from one sequencing run to the next due to differences in sequencing depth.

b. While Pct_Reads_From_Putative_Cells (RSEC) is stable at low sequencing depth, Pct_Reads_From_Putative_Cells (DBEC) is sequencing-depth dependent.

Interpreting output metrics

Introduction

This topic describes possible problems and recommended solutions for sequencing analysis issues. Issues with sequencing metrics might be related to issues that can be resolved in the experimental workflow.

Percentage reads assigned to cell label and percentage cellular reads aligned uniquely to amplicons are low

Possible causes	Recommended solutions
Low sequencing quality	• Ensure that the appropriate PhiX % is used for the type of sequencer used. • Ensure that the Illumina sequencing flow cell is not over-clustered. • Repeat the sequencing run if sequencing quality is suspected to be the reason.
Low library quality	• Ensure that the correct gene panel is used to amplify the sample and the correct amplification protocol and PCR product purification protocols are used. • Repeat amplification from leftover PCR1 products, if necessary.

High percentage assigned to cell labels but low percentage cellular reads aligned uniquely to amplicons

Possible causes	Recommended solutions
Incorrect FASTA file panel used for mapping	• If <50% alignment, then the wrong panel was likely used. • Verify that the correct panel reference file was used.
Incorrect number of sequencing cycles	Run at least 75 x 2 sequencing cycles.
Low sequencing quality	Rerun sequencing, and use at least the minimum recommended concentration of PhiX.

Low percentage reads mapped to putative cells

Possible causes	Recommended solutions
Some cells in the samples are not well represented by the panel. Their associated cell labels have very few detectable molecules, so they are classified as noise cell labels.	• Ensure that the panel matches the sample and species. • Ensure that the panel of genes provides good representation across the cells in the sample tested if all cells are to be detected.
Lysis time too long	Ensure that lysis time is exactly 2 minutes and lysis buffer is cold.
Automated pipette settings are incorrect	Ensure that the correct setting is used for the specific step in the cartridge workflow.
Wrong buffer used for bead retrieval from the cartridge	Use only lysis buffer, as indicated in the protocol for bead retrieval.
Mixed species in experiment	Ensure that the panel used contains genes that cover both species.
Excessive dead or dying cells	Proceed with the experiment if cell viability is ≥50%.
Very low bead loading density. The bead loading efficiency on the BD Rhapsody™ Scanner likely reported failed.	See bead loading density troubleshooting in the *BD Rhapsody™ Single-Cell Analysis System Instrument User Guide* (Doc ID: 214062) or the *BD Rhapsody™ Express Single-Cell Analysis System Instrument User Guide* (Doc ID: 214063).

Batch effects across multiple libraries

Possible causes	Recommended solutions
Variations in sequencing depth	Examine the status of each gene in <sample_name>_UMI_Adjusted _Stats.csv across samples. If there are highly abundant genes with a *pass* status in one library but a *low depth* status in another, consider using <sample_name>_RSEC _MolsPerCell.csv for analysis. Or, use <sample_name>_DBEC _MolsPerCell.csv for analysis after removal of genes that do not have *pass* status in any of the libraries under consideration.
Variations in cell sample handling protocol	Use a similar cell sample handling protocol for all samples to be analyzed together, noting that temperature, duration of handling, and handling method can affect gene expression.
Differences in thermal cycling	For samples to be analyzed together, it is recommended to perform the PCR amplification of the Cell Capture Beads of those samples in parallel.
Low sequencing depth	Use <sample_name>_RSEC _MolsPerCell.csv or use <sample_name>_DBEC _MolsPerCell.csv after removal of genes that do not have *pass* status.

Number of cells detected in sequencing is much lower than the expected cell number based on imaging results

Possible causes	Recommended solutions
Some cells in the samples are not well represented by the panel. Their associated cell labels have very few detectable molecules, so they are classified as noise cell labels.	• If all of the cells are to be detected, ensure that the panel of genes provides good representation across the cells in the sample tested. • Ensure that the panel matches the sample and species.
Cell Capture Beads settled to the bottom of the tube before the start of PCR1.	Ensure that Cell Capture Beads are well suspended just before starting PCR1, and the thermal cycler lid is pre-heated when the PCR tubes are placed on the thermal cycler.
Cell Capture Beads are lost during handling after cartridge use.	Ensure maximum recovery of Cell Capture Beads by using low retention tips and tubes. See product information in *the BD Rhapsody™ Single-Cell Analysis System Instrument User Guide* (Doc ID: 214062) or the *BD Rhapsody™ Express Single-Cell Analysis System Instrument User Guide* (Doc ID: 214063).

References

Bioinformatics analysis tools

- broadinstitute.github.io/picard/. The website contains a set of command line tools for working with high throughput sequencing data and formats, including SAM/BAM/CRAM, and VCF.
- Li H, et al. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. doi: 10.1093/bioinformatics/btp352.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9(4):357–60. doi: 10.1038/nmeth.1923.
- Fan J, Tsai J, Shum E. Technical Note: Molecular Index counting adjustment methods. BD Biosciences. This is an introduction to RSEC (recursive substitution error correction) and DBEC (distribution-based error correction). For more information, contact BD Biosciences technical support at researchapplications@bd.com.
- Li H. Toolkit for processing sequences in FASTA/Q formats. github.com/lh3/seqtk.

Expression profiling

Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161:1202–1214.

t-distributed stochastic neighbor embedding (t-SNE)

- van der Maaten, LJP. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*. 2014; 15(Oct):3221–3245 (PDF).
- van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*.2008; 9(Nov):2579–2605. jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

This page intentionally left blank

3

BD Rhapsody™ Targeted clustering analysis

Clustering Analysis Workflow

Workflow

The BD Rhapsody™ Clustering Analysis app on the Seven Bridges Genomics platform or on a local installation clusters gene expression profiles of cells and is part of the BD Rhapsody™ Analysis pipeline. See Figure 1. While sequencing analysis is required before clustering analysis, clustering analysis can be performed independently.

The clustering algorithm is based on hierarchical clustering and identifies statistically significant clusters. To aid visualization, the bh-tSNE algorithm is also performed to project the high-dimensional profiles to 2D space, using perplexity of 15 and dimension of 50. See van der Maaten, LJP, in References (page 86).

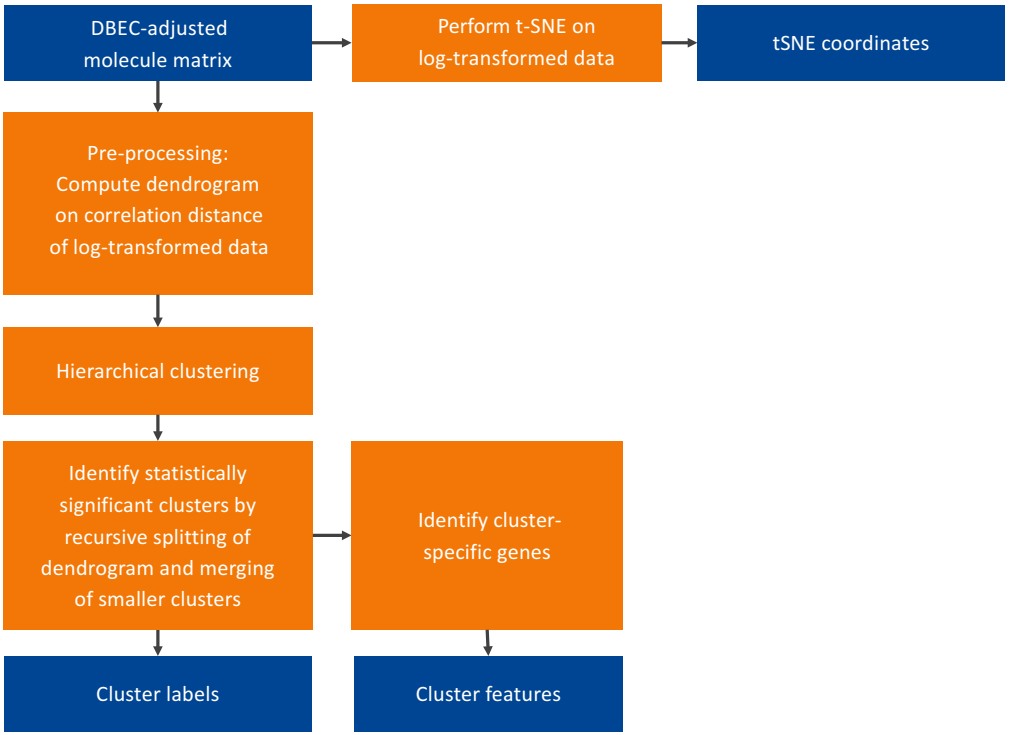


Figure 1. The clustering analysis pipeline.

Pre-processing of the gene expression matrix	A count matrix is log-transformed after a pseudo-count of 1 is applied to each entry. Correlation distance is used to describe the pairwise dissimilarity between each pair of cells.
Hierarchical clustering	Hierarchical clustering iteratively merges the two closest clusters. All clusters are initiated as individual points with pairwise distances determined as described in Pre-processing of the gene expression matrix. Computing the distance between clusters is done by using complete linkage, and a full dendrogram is obtained.
Splitting and testing	Starting from the top of the dendrogram, the tree is split into two candidate sub-trees under the constraint that the intra-cluster median correlation of the two sub-trees should be higher than the inter-cluster median correlation. The split is scored with the smallest p-value when performing Welch's t-tests for every gene. All possible splits are performed, and their scores are recorded. Various thresholds of $-\log_{10}$(p-value) cutoffs are attempted as the split criterion to generate multiple versions of the clustering results. A graph of number of clusters versus $-\log_{10}$(p-value) cutoff can be plotted to inspect the stable cut of the dendrogram (see Figure 2). A stable cut is defined as a plateau on the curve over a range of *5* on a $\log_{10}$-transformed p-value scale. Splitting results (sets of labels) corresponding to all stable cuts are kept and subjected to the next merging step.

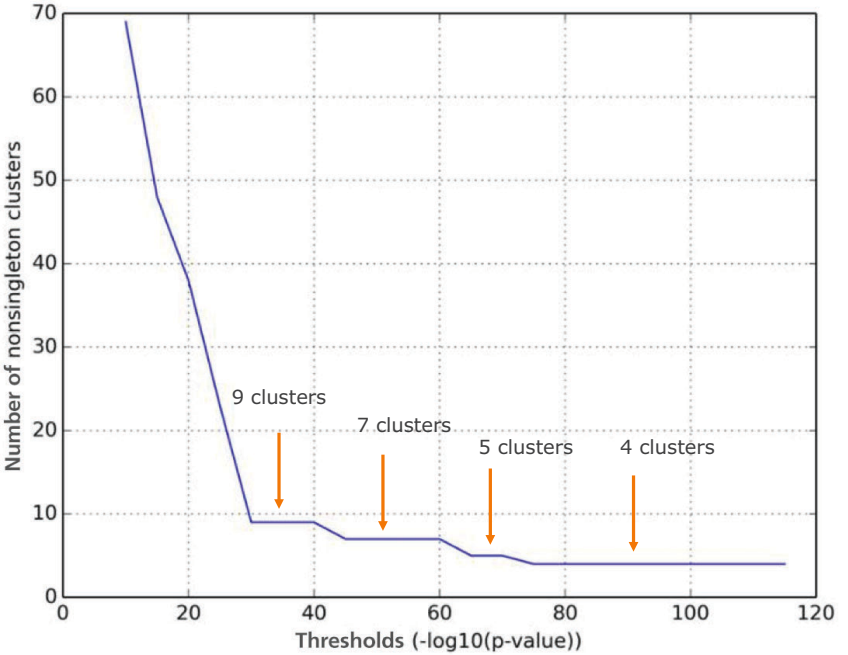


Figure 2. Example results of the dendrogram splitting step. The stable numbers of clusters found are 4, 5, 7, and 9.

Merging

Using the labels generated during splitting and testing, the merging phase determines if any of these clusters should be combined to form one cluster. The splitting phase can produce small clusters of a few data points each. This merging phase cleans up the smaller clusters by merging them with larger clusters. Specifically, all pairs of clusters are compared against each other, and then a p-value from Welch's t-test for each gene is generated. If the $-\log_{10}$(smallest p-value from all gene comparisons) is less than the threshold, which is defined as $-\log_{10}$(p-value threshold for the stable split)/2, the smaller cluster is merged with the larger one. The labels are updated, and all pairs are tested again until all pairs have the smallest p-value lower than the threshold.

Reporting the cluster assignment

Each cluster is denoted by an integer. Cells that cannot be merged with any other clusters (singletons) are given the label –1. The file, <sample_name>_<num_clusters>_Labels.csv, records the cluster assignment of each cell in the same order as in the loaded data table.

Reporting marker genes of each cluster

For each cluster, one-versus-rest tests are done using only the genes that have higher means in the cluster of interest. A table of important genes for each cluster is output as <sample_name>_<num_clusters>_Cluster_Features.csv along with additional information about each gene, including p-value, fold-change, and mean expression level within the cluster.

For each pair of clusters, Welch's t-test is performed to generate the gene list to differentiate two clusters the most. The list of results from all pairs is output as <sample_name>_<num_clusters>_Pairwise_Cluster_Features.csv.

To review clustering analysis metrics outputs, proceed to Reviewing clustering analysis output files (page 79).

Reviewing clustering analysis output files

Before you begin

Obtain the output files after running clustering analysis on the Seven Bridges Genomics platform or on a local installation. See the *BD Single Cell Genomics Analysis Setup User Guide* (Doc ID: 47383).

Clustering analysis outputs

The BD Rhapsody Clustering Analysis app outputs one or more sets of four files (cluster labels, t-SNE projection labelled by cluster, cluster features, and pairwise cluster features) that describe levels of clustering:

Output	File	Content
t-SNE coordinates (page 80)	<sample_name> _bh-tSNEcoordinates.csv	Coordinates of the t-SNE projection
Cluster labels (page 82)	<sample_name>_<num_clusters> _Labels.csv	Cluster membership per cell
t-SNE plot (page 83)	<sample_name>_<num_clusters> _tSNE.png	Visualization of the t-SNE projection with cells colored by cluster labels
Over-represented genes in each cluster to all clusters (page 84)	<sample_name>_<num_clusters> _Cluster_Features.csv	Top 50 statistically over-represented genes compared to all clusters

Output (continued)	File	Content
Over-represented genes in each cluster to every other cluster (page 85)	<sample_name>_<num_clusters> _Pairwise_Cluster_Features.csv	Top 50 statistically over-represented genes compared to every other cluster
(Optional) Concatenated data tables (page 86)	<sample names>_MolsPerCell.csv or <sample names>_Expression_Data.st	Combined data table; output only if multiple inputs specified
(Optional) Sample IDs (page 86)	SampleIDs.csv	Table of sample IDs; output only if multiple inputs specified

t-SNE coordinates

File: <sample_name>_bh-tSNEcoordinates.csv

The output is the projection of the data using the t-SNE algorithm. See der Maaten and Hinton in References (page 70).The output file contains coordinates that you can use to generate other visualizations. The order of cells/rows listed in the output file is in the same order of cells/rows listed in the input file.

For example:

#####################	
## BD Targeted Rhapsody Analysis Pipeline Version 1.0	
## Analysis Date: 2017-08-03 23:57:20	
## Sample: mySample	
## Reference: Immune_Response_Panel_Hs	
#####################	
Coordinate_1	Coordinate_2
11.89186	24.79593
5.98289	29.08972
27.13341	3.54942
25.44122	3.78159
12.18246	23.77134
12.78139	24.09827

Cluster labels

File: <sample_name>_<num_clusters>_Labels.csv

The output is the assignment of an integer representing the cluster label to each cell. The order of cells/rows listed in the output file is in the same order of cells/rows listed in the MolPerCell.csv input file. The value –1 means singletons, which are cells not assigned to any of the clusters. You can use this file and the coordinate file for additional clustering analysis.

For example:

#####################
BD Targeted Rhapsody Analysis Pipeline Version 1.0
Analysis Date: 2017-08-03 23:57:20
Sample: mySample
Reference: Immune_Response_Panel_Hs
#####################
Cluster_Label
2
2
2
2
2
2

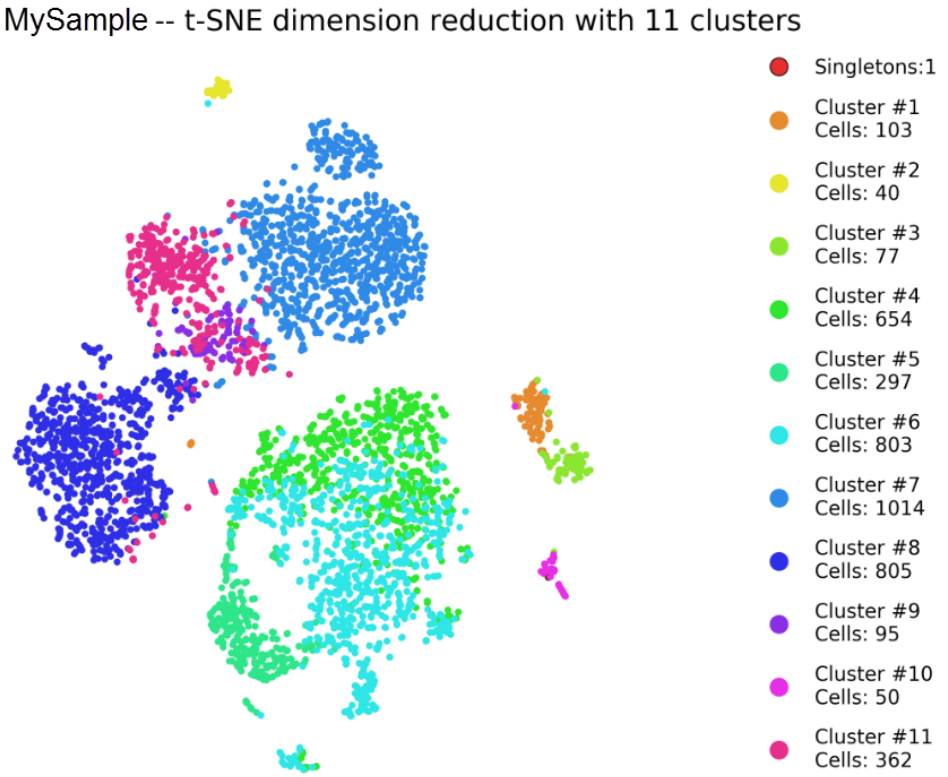
t-SNE plot

File: <sample_name>_<num_clusters>_tSNE.png

The output is a visualization of the t-SNE plot with cells colored according to cluster label. The visualization shows the number of clusters that have been identified from the analysis.

Singletons are not associated with any cluster due to the low pair-wise correlation between the singleton and other cells in the sample. Singletons are infrequent.

For example:



Over-represented genes in each cluster to all clusters

File: <sample_name>_<num_clusters>_Cluster_Features.csv

The output is a list of up to the top 50 statistically over-represented genes in each cluster as compared to all other clusters.

For example:

#####################				
## BD Targeted Rhapsody Analysis Pipeline Version 1.0				
## Analysis Date: 2017-08-09 13:18:06				
## Sample: mySample				
## Reference: Immune_Response_Panel_Hs				
#####################				
Cluster	Gene	p-Value	Mean_of_Expression	Fold_Change_of_Expression
1	GAPDH|NM_	167.152	161.336	3.089
1	LGALS1|NM_	165.651	147.816	3.182
1	LGALS3|NM_	164.283	34.697	3.616
1	ANXA5|NM_	141.459	31.73	2.424
1	GATA3|ENST	140.168	17.434	5.295
1	PYCR1|NM_(	137.654	17.243	3.356

Metric	Description
Cluster	Identified cluster
Gene	Over-expressed gene in this cluster compared to other clusters
p-Value	This is $-\log_{10}$ of the p-value. The larger the value, the more significant the differential expression of the gene within the cluster.
Mean_of _Expression	Mean number of molecules in all cells in that cluster
Fold_Change_of _Expression	Fold change in mean expression of the gene in that cluster and all of the remaining cells

Over-represented genes in each cluster to every other cluster

File: <sample_name>_<num_clusters>_Pairwise_Cluster_Features.csv

The output is a list of up to the top 50 statistically over-represented genes in each cluster as compared to every other cluster. The output shows the pairwise differential expression between all pairs of clusters:

#####################							
## BD Targeted Rhapsody Analysis Pipeline Version 1.0							
## Analysis Date: 2017-08-09 13:18:06							
## Sample: mySample							
## Reference: Immune_Response_Panel_Hs							
#####################							
Comparison	Gene	p-Value	Larger_Cluster	Fold_Change_of_Expression_for_Larger_Cluster			
Cluster1_vs_	AURKB|NM_	107.874	2	48.241			
Cluster1_vs_	HMGB2|NM	87.207	2	11.112			
Cluster1_vs_	TOP2A|NM_	68.18	2	11.625			
Cluster1_vs_	UBE2C|NM_	57.632	2	16.4			
Cluster1_vs_	HMMR|NM_	48.925	2	22.132			
Cluster1_vs_	TYMS|NM_0	46.727	2	3.988			
Cluster1_vs_	CCNB1|NM_	46.314	2	9.951			
Cluster1_vs_	MKI67|NM_	38.276	2	21.613			

Metric	Description
Comparison	The two clusters being compared
Gene	Over-expressed gene in this cluster compared to the paired cluster
p-Value	This is $-\log_{10}$ of the p-value. The larger the value, the more significant the differential level of the gene with the cluster.
Larger_Cluster	The cluster with the higher mean expression level of the gene
Fold_Change_of _Expression_for _Larger_Cluster	Fold change of expression of the gene in that cluster and another cluster

(Optional) Concatenated data tables

File: < sample names>_MolsPerCell.csv or <sample names>_Expression_Data.st

The output is a concatenated data table of all inputs (only if multiple data files are input).

(Optional) Sample IDs

File: SampleIDs.csv

The output is the Sample ID and the sample name associated with each molecule in the concatenated data file (only if multiple data files are input):

Sample ID	Sample name
1	MySample
1	MySample
1	MySample
1	MySample
1	MySample

References

Clustering algorithm

Zhang JM, Fan J, Fan HC, Rosenfeld D, and Tse DN. An interpretable framework for clustering single-cell RNA-Seq datasets. *BMC Bioinformatics*. 2018;19:93–105. doi: doi.org/10.1186/s12859-018-2092-7.

t-SNE

van der Maaten LJP. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*. 2014;15(Oct):3221–3245.

4

BD™ Data View

BD Data View applications

BD Data View is a software tool for visualization and exploratory analysis of output files generated following bioinformatics analysis. Some of applications of BD Data View include:

- View output results from the BD Rhapsody™ Analysis pipeline or from other single cell gene expression analysis platforms.
- Explore across multiple sets of single cell experiments.
- Analyze single cell 3' RNA seq data to derive a list of genes for creation of custom targeted panels.
- Create figures for scientific presentations or publications.

How to use this chapter

Proceed to a section to learn about BD Data View functions or to work through the examples:

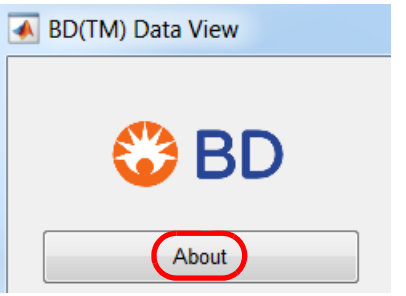
- Getting to know BD Data View v1.2 (page 89)
- Analyzing targeted sequencing output files from a single BD Rhapsody™ experiment with BD Data View (page 138)
- Analyzing multiple samples with BD Data View (page 161)
- Designing a targeted panel from whole transcriptome amplification (WTA) RNA-seq data (page 177)
- Analyzing a multiplexed sample with BD Data View (page 191)
- Managing sessions (page 197)
- Managing errors encountered with BD Data View (page 200)

For instructions on installing BD Data View, see the *BD Single Cell Genomics Analysis Setup User Guide* (Doc ID: 47383).

Getting to know BD Data View v1.2

Learn the workflow of BD Data View by reading the detailed descriptions, according to callout number, that follow the figure:

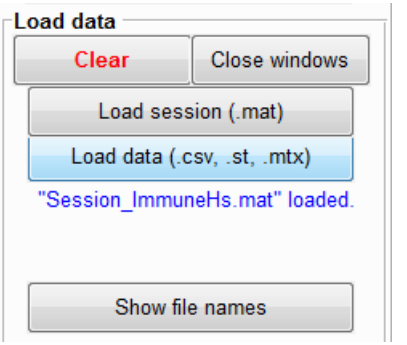
(1) About



About Displays the version number of the application.

(2) Load data

BD Biosciences recommends naming the data table files with the parameters used.



Clear Clears all data and variables. Unless the session is saved, the data and variables are not recoverable.

Close windows Closes all open windows.

Load session (.mat) Opens a saved .mat file from a previous session.

Load data (.csv, .st, .mtx)

Supported formats

Format	Properties
.csv	• BD Rhapsody Analysis pipeline output: Cell by gene, where row = cell and column = gene. Header lines between two lines of ##### are ignored. You must first select the column containing the cell identifier (for example, **Cell_Index**), and then select the column corresponding to the first gene to be loaded.
.st	• Sparse matrix format from BD Rhapsody Analysis pipeline output. Header lines between two lines of ##### are ignored.
.mtx	• Output from another single cell gene expression analysis platform. Matrix.mtx, barcodes.tsv, and genes.tsv are required. • If the barcodes.tsv and genes.tsv files are in the same folder as matrix.mtx, the application automatically loads the two files; otherwise, the application prompts you to browse and identify the folder with the .tsv files.

Supported formats for multiple file loading

Files multiply-loaded	Properties
.csv	Ensure that the data tables are in the same folder and are in the same format (cell by gene or gene by cell). Select multiple data table files by holding the control key and left-clicking files with the mouse. The number and order of genes in the data tables can vary. During loading, data tables are concatenated. If one data table does not contain a gene found in other loaded data tables, then all of the cell identifiers (IDs) in that data table are assigned zero counts for that gene.
.st	Select multiple .st files in the same format (column order) and folder by holding the control key and left-clicking files with the mouse.
.mtx	The .mtx files must be in the same folder. The application prompts you to load the barcodes.tsv and genes.tsv files for each .mtx file.

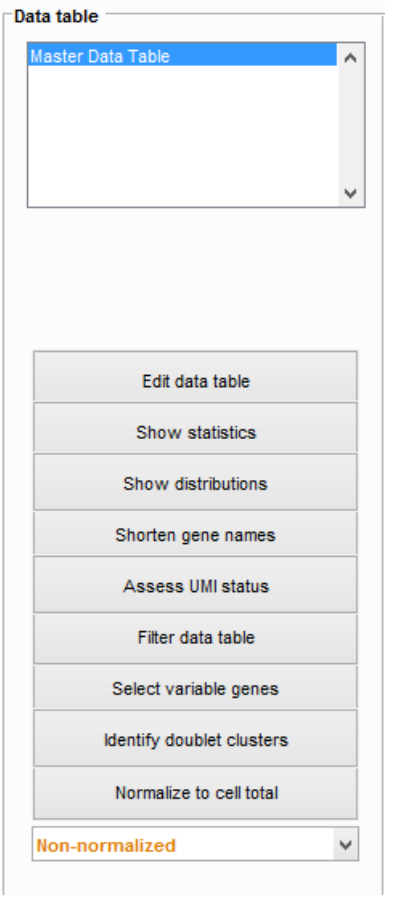
Properties of data table

- The concatenated data table is called the Master Data Table and is listed in the Data table list box (callout 3).
- The cell identifier in the data file is recorded under Cell ID in the Annotation list box (callout 5).
- The file of origin for each cell is recorded under File in the Annotation list box (callout 5).
- The gene names in the loaded data table are listed in the Single gene expression list box (callout 6).

Reloading new data clears data and variables.

Show file names Displays file names of loaded data tables. If multiple files are loaded, shortened file names are displayed.

(3) Data table



Edit data table

Removes, renames, or combines selected data tables. All data tables except the Master Data Table can be removed or renamed. You can select multiple files to rename, or you can select and combine two or more tables. The renamed or removed data table names are updated in BD Data View automatically.

Removal of the data table is permanent.

Show statistics

Displays statistics for each loaded data table, Master Data Table, or subset of the Master Data Table. For example: number of cells in the data table, number of genes in the file, and the mean and median molecules/cell. By selecting both the data table and annotation group, the statistics of the data table and the annotated group are displayed. The output for data table statistics is saved as DataTableStatistics.csv in the current directory, and the output for annotation statistics is saved as AnnotationStatistics.csv in the Annotation subfolder.

Show distributions

Displays histograms of the highlighted data table in the Data table list box. These plots can be used to guide cell and gene filtering. See Filter data table (page 96):

- Number of molecules/cell.
- Number of genes/cell.
- Number of molecules of mitochondrial genes/cell (human gene name starting with **MT-**; mouse gene name starting with **mt-**). This plot can serve as a guide to remove potentially dying cells with a high proportion of mitochondrial transcripts, particularly applicable to 3′ RNA seq data.
- Number of cells expressing each gene.
- Expression level/cell for each gene.
- Histograms saved as are in the *Distrib* subfolder.

Shorten gene names

Shortens gene names by selecting parts of names (gene symbols, accession numbers) that are separated by delimiters, including vertical bar, tab, underscore, space, and dot.

- Shorten gene names automatically updates the gene name for AbSeq markers and generates an AbSeq gene list in the gene set window of the Multi-gene expression panel. See Select gene set list box (page 125).

Assess UMI Status

Assesses sequencing depth of each gene.

To mitigate the effect of over-estimation of molecules from PCR and sequencing errors, the BD Rhapsody™ pipeline contains Unique Molecular Identifier (UMI) adjustment algorithms recursive substitution error correction (RSEC) and distribution-based error correction (DBEC). A gene is subjected to DBEC if it meets a certain threshold for sequencing depth. If a gene passes the threshold for DBEC, the status is *pass*. If a gene does not pass, the status is *low depth*. If a gene has zero counts across all cells, the status is *not detected*.

It is recommended to remove genes with insufficient depth, particularly when the genes contribute to a large number of molecules and comparison across multiple samples is intended.

If a gene has a pass status in one file for one library but low depth in another and is relatively abundant, the gene can contribute to batch effect.

To help you with assessing UMI status, keep in mind:

- For analysis of BD Rhapsody data, load <sample_name>_UMI_Adjusted_Stats.csv. Follow the windows to select the column representing the UMI adjustment status. See Evaluating sequencing depth (page 144).
- Genes with low depth status are labeled *genes with insufficient depth*.

- BD Data View displays three figures to assess sequence depth per gene. See Evaluating sequencing depth (page 144).
- See Filter data table (page 96) for filtering out genes with insufficient depth.

Filter data table

Performs filtering by cells or genes to reduce the data table. Select filtering by gene, cell, or both by selecting all.

- Use on any of the data tables listed in the Data table list box. Click to select the data table to filter from.
- The name of the reduced data table, given by the user, will appear in the Data table list box.

Cell filtering (Cells)

- You can specify these options:
  - **Based on gene expression**: Removes groups of cells based on expression conditions listed according to gene. For example: **GAPDH<10**. Select **AND** or **OR** with multiple conditions. For more information on writing conditional statements, see **Conditions on gene expression level (page 110)**.
  - **Based on current annotation**: Removes groups of cells in the highlighted annotation in the Annotation List Box.
  - **Random subsampling**: Subsamples (1) a specific number of cells from the entire highlighted data table or (2) a specific number of cells from each of the cell groups in the highlighted annotation. Select **Random subsampling** when comparing multiple data sets with widely varying numbers of cells in order to analyze data sets with the same number of cells.

Gene filtering (Genes)

- You can specify these options:
  - **Specify in window genes to keep or remove**: Enter in the pop-up window genes to keep or remove. Enter each gene on a new row.

Note: Any gene name not present in the loaded data table is ignored. This is true for uploading a list of genes to keep or remove and using pre-loaded gene sets.

- **Upload list of genes to keep or remove**: Gene names uploaded in a .csv file, one gene name per row. There are no header or trailing spaces.
- **Use pre-loaded gene sets to keep or remove**: Select a gene list in the Multi-gene expression list box (callout 7).
- **Based on expression level**: Enter the minimum number of cells expressing the gene or enter the minimum number of molecules from all cells. Both conditions need to be met for genes to be kept.

  Note: If normalized data has been computed [see Normalize to cell total (page 102)], this filtering criterion will apply to either non-normalized or normalized data, depending on which is chosen as displayed in the normalization drop-down menu at the bottom of the panel.

- **Based on UMI adjustment status**: Requires first running the function **Assess UMI status**. (Optional) Select genes with insufficient depth to remove.

Select variable genes

Reduces number of genes in large data sets (for example, whole transcriptome RNA seq data) to a subset of most variable genes (~100 to ~1000), depending on thresholds of parameters. Dispersion of $\log_{10}$-transformed gene expression data is calculated for each gene. Genes are divided into a number of bins based on mean gene expression per cell. In each bin, the z-score of the dispersion of each gene is calculated. Genes that exceed the dispersion threshold are retained. This is similar to the method presented in Macosko, EZ, et al. See References (page 70).

- Cells will also be filtered out if they do not have any molecule counts for the retained set of genes.
- This operation is conducted on the highlighted data table in the Data table list box, either normalized or non-normalized depending on the condition selected in the drop-down menu at the bottom of the panel.
- You can specify these parameters:

  - **Dispersion threshold** (z-score of dispersion): Genes greater than this threshold are kept.
  - **Number of bins**
  - **High expressor threshold**: Genes with expression levels greater than the threshold are removed.
  - **Low expressor threshold**: Genes with expression levels lower than the threshold are removed.

  Note: Changing the parameter changes the number of genes selected. For example, if you lower the threshold, more genes are selected.

- The name of the reduced data table that you enter will appear in the Data table list box.
- A scatter plot showing dispersion versus gene expression for each gene, together with the parameters stated in the legend, is saved in a new folder with the name of the reduced data table as the folder name.

Identify doublet clusters

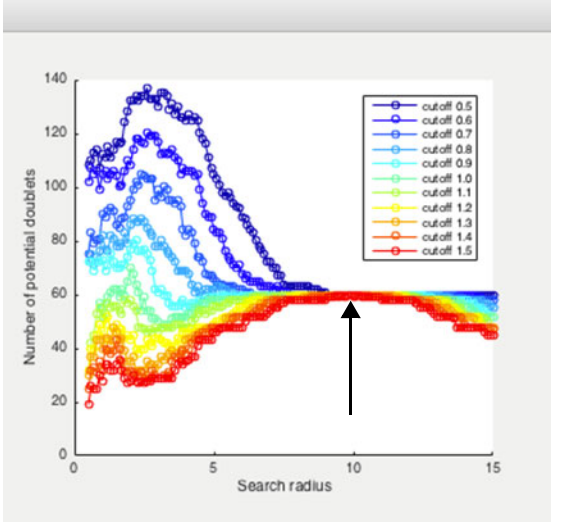
Identifies clusters that consist of doublets of cells from different cell types to distinguish biologically meaningful clusters from assay artifacts. The function was developed by BD Biosciences.

The function operates on the highlighted data table in the Data table list box and on non-normalized or normalized data (if previously computed) selected from the pull-down menu. For more information on normalization, see Normalize to cell total (page 102).

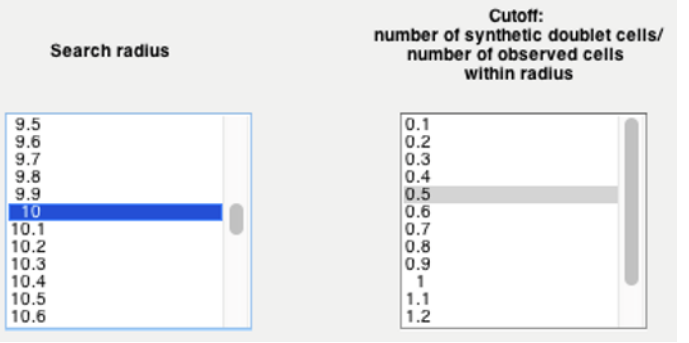
- For visualization, load or generate a set of coordinates before using this function. Highlight the set of coordinates in the Coordinates list box that corresponds to the highlighted data table to be computed on.
- Once started, Identify doublet clusters generates a synthetic doublet data table by random subsampling sets of two cells in the highlighted data table and sums their gene expression.

- You can specify these parameters:
  - **Expected percent of doublets**: Enter a doublet rate. The default is 5%, which is suitable for most analyses with <10,000 cells captured on the BD Rhapsody™ Single-Cell Analysis system. This parameter determines the number of random subsampled events in order to generate synthetic doublets.
  - **tSNE with synthetic doublets** and **tSNE with synthetic doublet perplexity**: The algorithm performs bh-tSNE on the synthetic doublet data table concatenated with the highlighted data table ($\log_{10}$-transformed with a pseudo-count of 1 added). You can specify bh-tSNE parameters, but the default values are suitable for most cases. See Calculate (page 103).
- For each cell in the highlighted data table, the algorithm computes the ratio of the observed data point counts to the synthetic doublet data point counts within a search radius on the t-SNE projection. The t-SNE projection with combined synthetic doublet and observed cells is not displayed.

- After the computation is completed, a pop-up window is displayed of the plotted ratio of synthetic doublets to actual observed cells within each search radius:



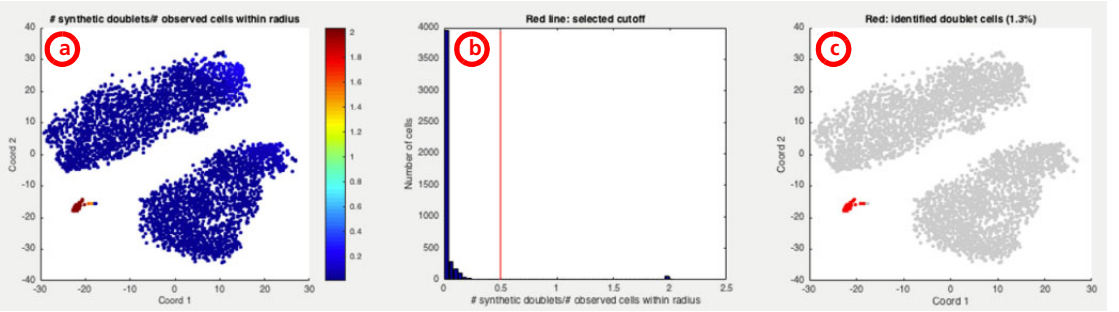
You can select values for the **search radius** and **cutoff** parameters to define the probable doublets:



The choice of the two parameters can be guided by the plot of synthetic doublets, where the number of potential doublets are computed for a range of values for both parameters.

For mixed cell type clusters from very distinct cell types, there is usually a range of search radius that forms a stable plateau (arrow). In this example, you can choose a radius between 9–11 and ratio cutoff between 0.5–1.5.

- An observed data point with more than the specified cutoff of (number of synthetic doublet cells)/(number of observed cells) within the defined search radius is labeled *doublet*. This is based on the expectation that an actual cell doublet will co-localize with the synthetic cell doublets when projected together. Synthetic doublets are not displayed on t-SNE coordinates.
- As you select different values for the two parameters **search radius** and **cutoff**, three plots are updated:



– Plot **a**: the t-SNE projection (highlighted in the Coordinates list box) with cells colored by the number of synthetic doublets found within the specified radius.

– Plot **b**: a histogram showing the distribution of the (number of synthetic doublet cells)/(number of observed cells) within a given search radius. The red line indicates the selected cutoff value.

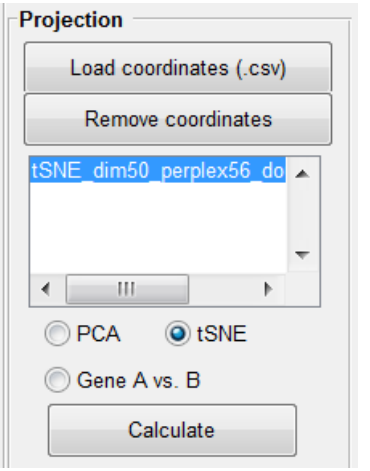
  - Plot **c**: the t-SNE projection (highlighted in the Coordinates list box) with cells passing the criteria of the two selected parameters (identified as doublets) in red.
- Click **Export Doublet Annotation** to export the doublet annotation to the Annotation list box in the main window.

Normalize to cell total

Normalizes each cell by taking the count of each gene and dividing it by the total molecule count of that cell and multiplies by the median of the total number of molecules per cell across all cells.

- This operation is done for all cells in the Master Data Table (regardless of which data table is highlighted in the Data table list box).
- The drop-down menu in the Data table panel will read **Normalized** instead of **Non-normalized**. Toggle between Normalized and Non-normalized to visualize and conduct analysis on normalized data or non-normalized data, respectively.

(4) Projection



Load coordinates Loads coordinates.

- .csv format
- Two columns with a column descriptor in each: column 1 = *x*-coordinate; column 2 = y-coordinate.
- Header lines between two lines of ##### are ignored.
- The coordinates can be pre-computed t-SNE coordinates from the BD Rhapsody cluster analysis pipeline or other analysis methods that generate 2D coordinates for single cell data.
- If multiple data files are loaded, you will not be able to load more than one coordinate file, because coordinates should be calculated for all cells in the concatenated Master Data Table using PCA or the t-SNE generation function in the application.
- For a single loaded data table, multiple pre-computed coordinates can be loaded, but a coordinates file must be loaded each time.

Remove coordinates Removes any of the loaded or calculated coordinates.

Calculate Calculates coordinates using the highlighted data table in the Data table list box, using either PCA, t-SNE, or a gene-gene comparison.

PCA

- This operation is limited to ≤1,500 genes due to computation time.
- The calculation is performed on $\log_{10}$-transformed data with a pseudo-count of 1 added.
- Select two principle components to display.
- A biplot and a table listing the variance is explained for each principle component displayed.

tSNE

- This operation is limited to ≤1,500 genes due to computation time.
- The calculation is performed on $\log_{10}$-transformed data with a pseudo-count of 1 added.

- Two implementations are available: **bh-tSNE** (Barnes-Hut), which is for large cell numbers (>100s), or **original**, which is for smaller cell numbers. bh-tSNE is much faster than original implementation for a large number of cells. For the Barnes-Hut implementation, see der Maaten and Hinton in References (page 70).
- For setting the parameter initial dimensions and perplexity, see van der Maaten, LJP in References (page 70). Default values are adopted from MATLAB t-SNE implementation. See lvdmaaten.github.io/tsne.
- The default perplexity value is 30. BD Biosciences recommends a starting point for perplexity that is the square root of the number of cells.
- If the number of genes is less than the default initial dimensions, the default initial dimensions are adjusted to the number of genes.
- For the original implementation, if the same seed (default at 0) is chosen, the same t-SNE projection is generated.

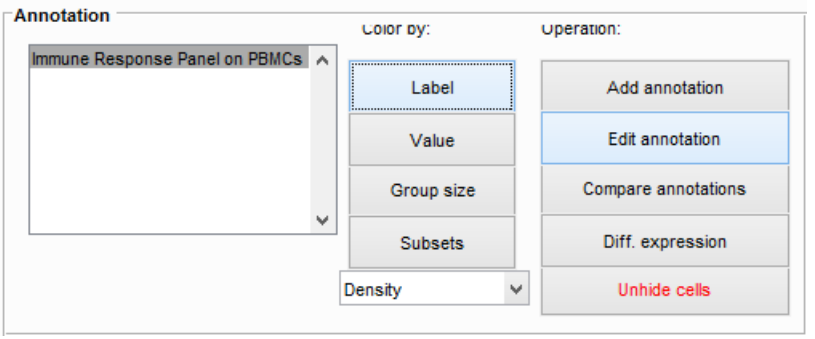
Gene A vs. B

- This function is similar to analysis of flow cytometry data, in which two genes are selected at a time for comparison and gating.
- The function opens a pop-up window where you select or enter two genes for comparison. The actual gene names are displayed along the plot axes.
- The jitter option helps to visualize low count data (<tens of molecules per cell) by adding random noise to each data point.
- Toggle between $\log_{10}$-transformed vs. untransformed count for plotting.
- Click **Export coordinates** to keep the desired plot for further visualization. The set of coordinates are listed in the Coordinates list box in the main window.

Projection list box Lists the loaded and computed coordinates. For each computed set of coordinates, the data table that is used to generate the coordinates and the parameters is the filename of the coordinate set.

- Highlight the desired set of coordinates, and the Annotation plot and Gene expression plot are updated. Note that if the highlighted coordinate is not generated from the highlighted data table, the plots will display the coordinates for only the cells that are in the highlighted data table.

(5) Annotation



Label **Note:** For categorical data, the unique categories are referred to as groups.

Colors cells by the cell groups defined in the highlighted annotation in the Annotation list box. Coloring by groups is suitable if annotation is in the form of discrete names/category labels (for example, distinct name of groups in annotation, such as cell type).

Value Colors cells by value. This is suitable if the highlighted annotation is a continuous variable (for example, fluorescence values from index sort data). Data points are colored based on the $\log_{10}$ or linear value, if **log10** is checked under the Gene expression plot.

Group size Colors cells based on the size of the group the cell belongs to ($\log_{10}$ or linear), if **log10** is checked under the Gene expression plot. Coloring cells by size is suitable for category labels.

Subsets Control-click up to 30 annotated groups in the annotation to display. Each annotated group is plotted with a unique color, and the remaining groups are plotted in gray.

Density In drop-down menu. Colors cells based on the density of data points on the projection. The more points there are within an area, the warmer the color.

Contour In drop-down menu. Contour based on the density of points on the projection. The more points there are within an area, the hotter the color.

Surface In drop-down menu. Heatmap based on the density of points on the projection. The more points there are within an area, the hotter the color.

Add annotation **Load from file**

The annotation files to be loaded must be in the same folder as the loaded data table.

- Format:
  - .csv file
  - Single to multiple columns
  - The first row should describe each column, containing the names of the annotations; for example, sample name, well name, cell type.
  - Row = cell
  - The file must have the same number of rows (or cells) as the loaded data table, and the cells must be sorted in the same order as the loaded data table.

  - Do not include commas in annotation names.
  - Header lines between two lines of ##### are ignored.
- Annotation can be categorical data (labels) or continuous values.
- Annotation files import:
  - Single data table loaded: Clicking **Load from file** can be performed multiple times to load multiple annotations. Import one annotation file at a time. For each loaded annotation, each selected column from the uploaded file displays in the Annotation list box in the main window.
  - Multiple data files are loaded: The same number of annotation files must be selected. Pair each annotation file with each of the loaded data tables.

    Note: During loading of the annotation files, select the columns to import from the file.

Hierarchical clustering

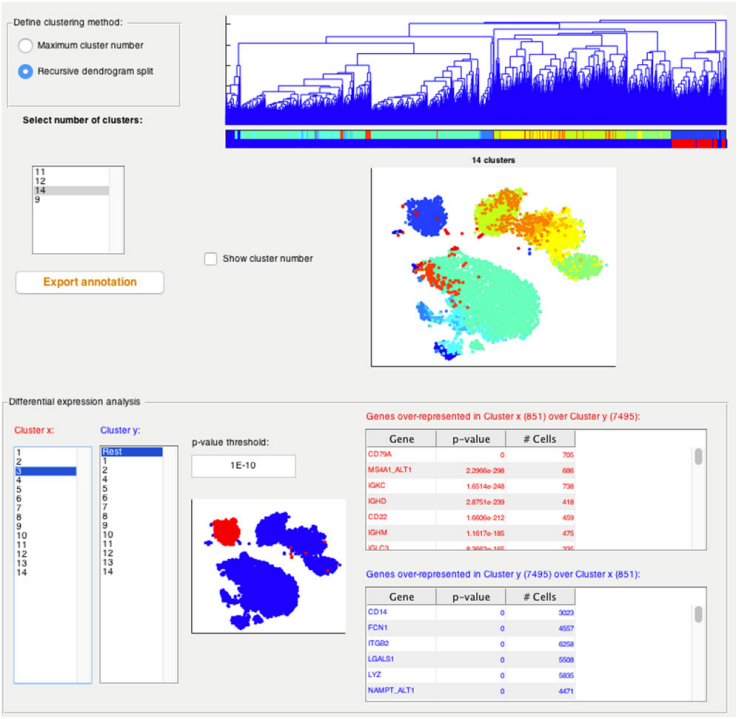
- A dendrogram is displayed in a pop-up window. The dendrogram is the calculation for the highlighted data table, without hidden cells and has been $\log_{10}$-transformed and normalized or non-normalized depending on the condition selected in the drop-down menu at the bottom of the Data table panel (callout 3). The dendrogram is computed using the MATLAB built-in dendrogram function, which uses correlation as the distance metric and complete linkage. See mathworks.com/help/stats/dendrogram.html and Hide/Unhide cells (page 115).
- This operation is limited to ≤1,500 genes.

- Two clustering methods are available:

Clustering method	Properties
Maximum cluster number	• Uses a MATLAB built-in cluster function with the **maxclust** option. See mathworks.com/help/stats/cluster.html. • Move the slider bar to define the maximum cluster number. • The algorithm finds the minimum vertical distance from the base of the dendrogram and draws a horizontal line that would result in the specified number of clusters or leaves defined below the line.
Recursive dendrogram split	• This is the same clustering algorithm as the one used in the BD Rhapsody Analysis pipeline. See BD Rhapsody™ Targeted clustering analysis (page 73). • The algorithm outputs a series of clustering results, each with a different number of clusters.

- When the slider bar for Maximum cluster number is moved or a cluster number for the Recursive dendrogram split is selected, the following features are updated:
  - If a projection is available, the highlighted projection is displayed with cells colored by the cluster assignment under the current selection.
  - The same color scheme is used to label cells ordered in the dendrogram as shown in the color bar under the dendrogram.

- A differential expression analysis tool is provided in the same window to guide selection of the number of clusters. The tool has the same features as the function **Diff. expression**. [See Differential expression (Diff. expression) (page 114).] A second color bar below the dendrogram is updated for each selection of clusters to reflect the two clusters in the dendrogram being compared. For example:



- After the number of clusters is selected, click **Export annotation** for further analysis. The annotation is listed in the Annotation list box in the main window.

Draw with mouse

- Enter the number of clusters to draw.
- A pop-up window with the current projection with data points colored by density is displayed. Draw polygons around points to define cell groups.

Add group to current annotation by drawing

- Same features as **Draw with mouse**, except the groups defined by drawing are added to the highlighted annotation.
- Use this function if an annotation exists, but more clusters are to be selected.

Combination of existing annotations

- Combinations of groups from two annotations are generated.

  Example: If the first annotation is a sample type with groups labeled sample 1 and sample 2, and the second annotation is cell type with groups labeled cell type 1 and cell type 2, then the resulting groups are: sample 1|cell type 1, sample 2|cell type 1, sample 1|cell type 2, sample 2|cell type 2.

Delete group from selected annotation based on group size

- Groups within the highlighted annotation are listed in descending order according to the number of cells in the group.
- Select the groups to remove. Cells with an annotation group removed are labeled as *unselected*.

Conditions on gene expression level

- A pop-up window is displayed for interactive analysis. Enter thresholds for a list of genes, and then cells matching the criteria are grouped and annotated.
- Conditions window:
  - Each line is a condition.
  - The format is a gene name or keyword (sumallgene, numgene, mito), followed by an operator (>, >=, <=, <, ==, ~=), followed by an integer or fraction.

  - Keywords:
    - **sumallgene**: total number of molecules from all genes.
    - **numgene**: total number of genes detected.
    - **mito**: total number of molecules from mitochondrial genes. Mitochondrial genes are those with names starting with *MT-* for human or *mt-* for mouse.
  - Conditions:
    - Integer: number of molecules.
    - Fraction: proportion of molecules in gene relative to cell total. Fraction only applies to the keyword *mito*.
  - Examples:
    - mito>0.2: the proportion of molecules from mitochondrial genes >20%.
    - GAPDH<=10: the number of GAPDH molecules is ≤10.
    - sumallgene<100: the total number of molecules from all genes <100.
    - numgene<50: the number of genes detected is <50.
- **Logic**: Each line entered is a condition. You are asked if all conditions are treated as an AND or OR condition. For AND, all conditions need to be met to be selected. For OR, cells meeting any one of the conditions is selected.
- **Display**: Implements the conditions and visualizes cells meeting the conditions. The following graphs are updated:
  - Histograms to guide selection of the threshold for each of the first eight genes listed: The distribution of molecules per cell is plotted for each gene, together with a red vertical bar indicating the threshold.
  - The current highlighted projection with cells passing the conditions labeled in red.

    Note: Repeatedly changing the thresholds and clicking **Display** changes the cell selection until the desired selection is displayed.

- **Add Group**:
  - Use this function to name the group of cells meeting the specified criteria. For the rest of the cells, they are labeled *unselected*.
  - Use this function repeatedly to add additional groups based on additional sets of criteria. Note that if a cell meets both current and previous conditions, it will receive a label based on the current condition.
- **Reset**: Clears all conditions and defined groups.
- **Export annotation**: The annotation is listed in the Annotation list box in the main window.

Conditions on gene expression level added to existing annotation

- Same as Conditions on gene expression level, except that the new groups defined are added to the highlighted annotation.

Edit annotation

Renames, reorders, or removes annotations or groups.

Rename Annotation

Select one or more annotations in the Select Annotation to Rename window to rename the annotation. The renamed annotation is updated in BD Data View automatically.

Rename Groups

- The groups of the highlighted annotation are listed under Cluster *x*. Select the group to rename. Enter a new name, and click **Update Cluster Name**.
- **Export annotation**: A new annotation is displayed in the Annotation list box in the main window.
- A differential expression tool is also displayed in the same window to guide the renaming of groups. This tool has similar features as the Differential expression function. See Differential expression (Diff. expression) (page 114).
- The function can also be used to collapse multiple clusters into one new or existing cluster by name. Rename all clusters to be collapsed with the new or existing name.

Reorder Groups

- The default order is alphabetical.

To order groups, click to select the group, and then click **Move Up** or **Move Down** to reorder the group. Click **Save Group Order**. Click **Label** to refresh the annotation plot and box plot.

Remove Annotation

Removes any of the loaded and generated annotations.

Compare annotations

Compares annotations, and requires two annotations.

- Choose to save the results as a .csv file in the *Annotation* subfolder.

  Example: Annotation 1 is sample name, and annotation 2 is cell type. The function calculates the number of cells in each cell type in each sample. Three tables are displayed: number of cells, percentage of cells relative to group total in annotation 1, and percentage of cells relative to group total in annotation 2.

Differential expression (Diff. expression)

Performs differential expression analysis on counts (not $\log_{10}$-transformed). It is based on the negative binomial test. See mathworks.com/help/bioinfo/ref/nbintest.html?searchHighlight=nbintest. Opens in a pop-up window for interactive analysis.

- You can specify these parameters:
  - **p-value threshold**: Enter the p-value threshold. The lower the p-value, the fewer the genes reported.
  - **Cluster *x* and cluster *y***: Select any two groups for comparison. If a projection is available, the highlighted projection is also displayed with cells colored in red (those in cluster *x*), blue (those in cluster *y*), or gray (not in *x* nor *y*).
  - **Correlation plot**: The scatter plot shows the mean ($\log_{10}$ molecules per cell with a pseudo-count of 1 added) for each gene of the two selected groups of cells.
- The analysis output can be saved:
  - .csv file output with columns:
    - Name of cell group
    - Gene name
    - Fold change
    - p-value
    - Total number of cells in cluster x expressing the gene
    - Total number of cells in cluster y expressing the gene
    - Mean molecules per cell in cluster x
    - Mean molecules per cell in cluster y
    - Total number of cells expressing the gene
  - The .csv file can be loaded into the Gene set list box with **Custom set**.
  - **Save results: mean(mol/cell) for selected cluster x and y** outputs the mean(mol/cell) and mean($\log_{10}$(mol/cell + 1)).
  - **Save results: selected cluster x vs. selected cluster y** outputs comparisons for only the highlighted cluster x and cluster y.

- **Save results: every cluster x vs. rest** outputs results for the analysis of every group in the list box under cluster X with the remaining cells.
- **Save correlation plot** saves the t-SNE plot and gene correlation plots.

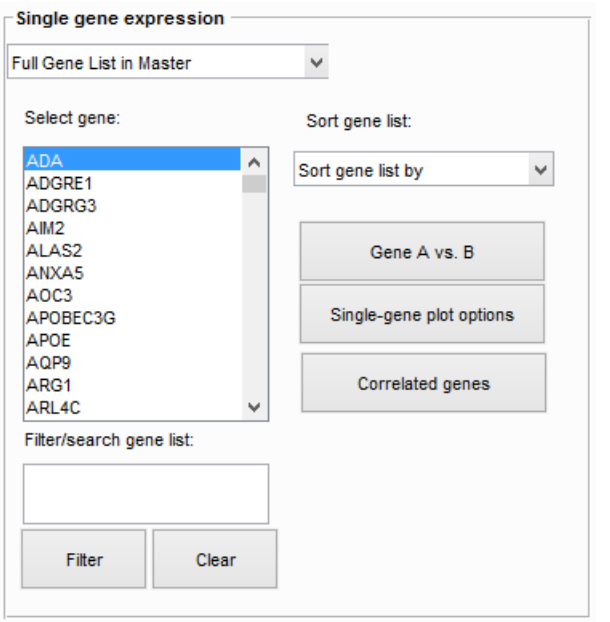
Hide/Unhide cells

Hides cells temporarily from view on scatter plots, box plots, histograms, heatmaps, and correlated genes calculations. The function is useful if there are many points on the plot.

This function is intended to temporarily hide cells for display purposes. It does not affect functions in the Data table, Projection, or Save data panels. Use Filter data table to remove cells from analyses and operations.

- Select the group of cells to hide based on groups in the highlighted annotation in the Annotation list box. The cells are hidden in the next refresh of the Gene expression plot and the Annotation plot.
- To display cells, highlight any annotation, and click **Unhide cells**.

(6) Single gene expression



Gene list box and Filter

Updates and displays data according to the data table selected in the Data table list box and the gene clicked in the Gene list box or by the gene entered and **Filter** clicked.

- The Gene expression plot is updated to display the number of molecules per cell by color. The top right hand corner of the Gene expression plot displays:
  - The number of cells and percentage of all cells with at least one molecule of that gene detected.
  - The number of molecules across all cells detected for the gene and the corresponding percentage across all detected molecules from all genes.
- Histogram, box plot, and lists of correlated and anti-correlated genes are displayed in the pop-up windows if their associated checkboxes are checked, and if necessary, a gene is reselected.

- If a gene list contains a non-unique gene name, BD Data View guides you to pick one for display.
- Filter/search gene list displays filtered gene list after entering a partial or full gene name in the Filter text box and clicking **Filter**. Click **Clear** to reset.

Sort gene list by drop-down menu

Alphabetical

Sorts gene names alphabetically. Select from drop-down menu.

Abundance

Sorts in descending order the total number of molecules from all cells for each gene. Select from drop-down menu.

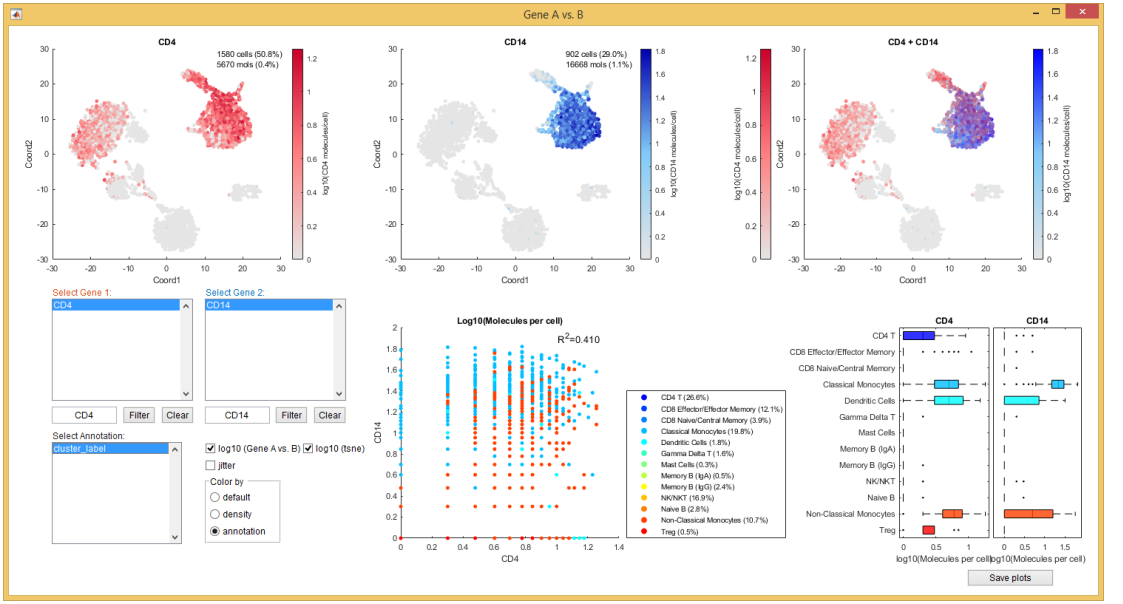
Variance

Sorts in descending order the variance in the number of molecules detected per cell for each gene. Select from drop-down menu.

Cell number

Sorts in descending order the total number of cells expressing the gene. Select from drop-down menu.

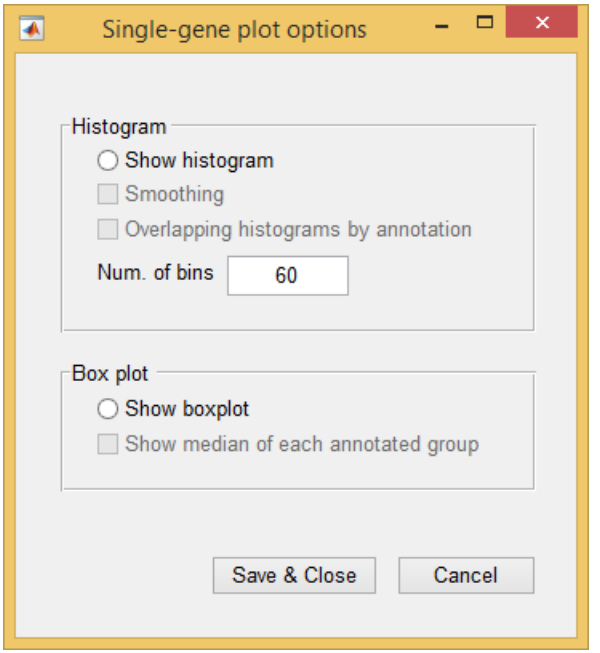
Gene A vs. B Compares mRNA and protein expression between two genes. For example:



- Select data table and projection coordinates, and then click **Gene A vs. B**.
- In the pop-up window, click to select the two genes for comparison. Three plots are displayed. The expression of each selected gene is shown in a plot. The third plot overlays expression of the two genes.
- The correlation of expression between the two genes is plotted (linear or $\log_{10}$) with or without annotation. Select $\log_{10}$ scales for the Gene A vs. B and/or t-SNE axis.
- If **Color by annotation** is selected, two box plots are displayed. Each box plot shows the expression of a selected gene and is organized according to the annotation.
- Click **Save plots** to save the results as PNG, PDF, or SVG. Default is PDF.

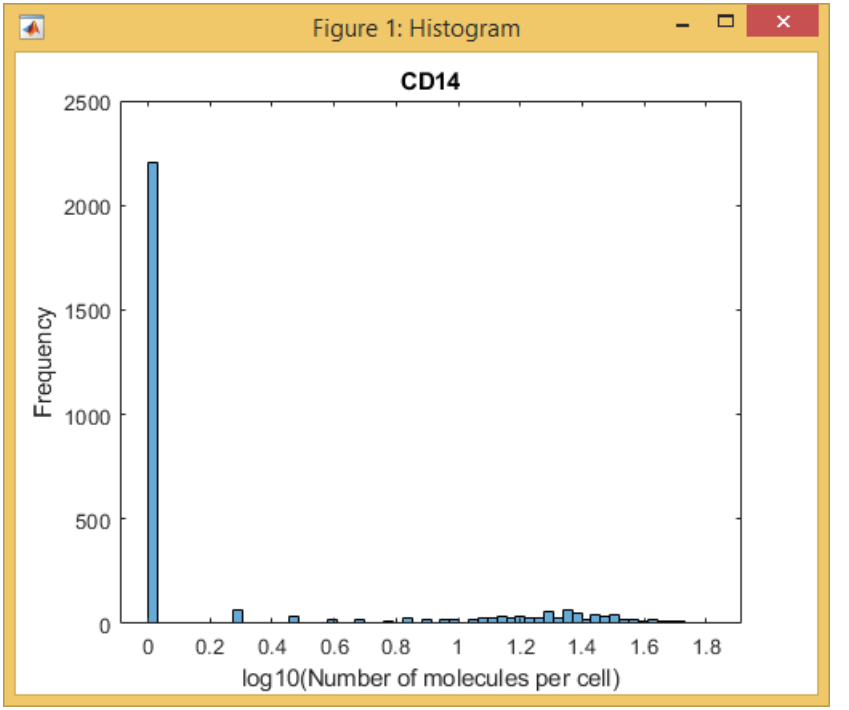
Single-gene plot options

Provides options for displaying gene data:

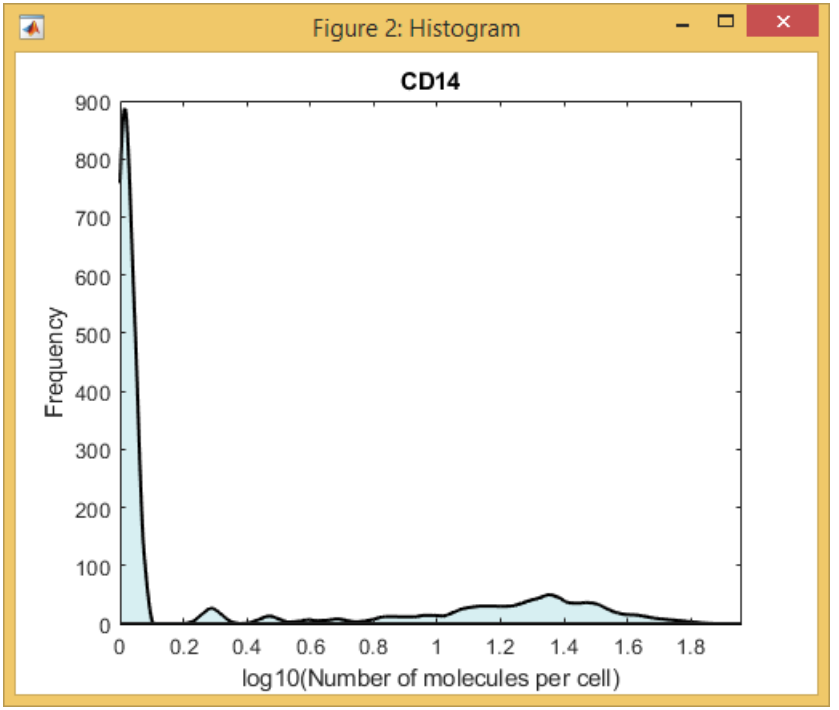


Histogram

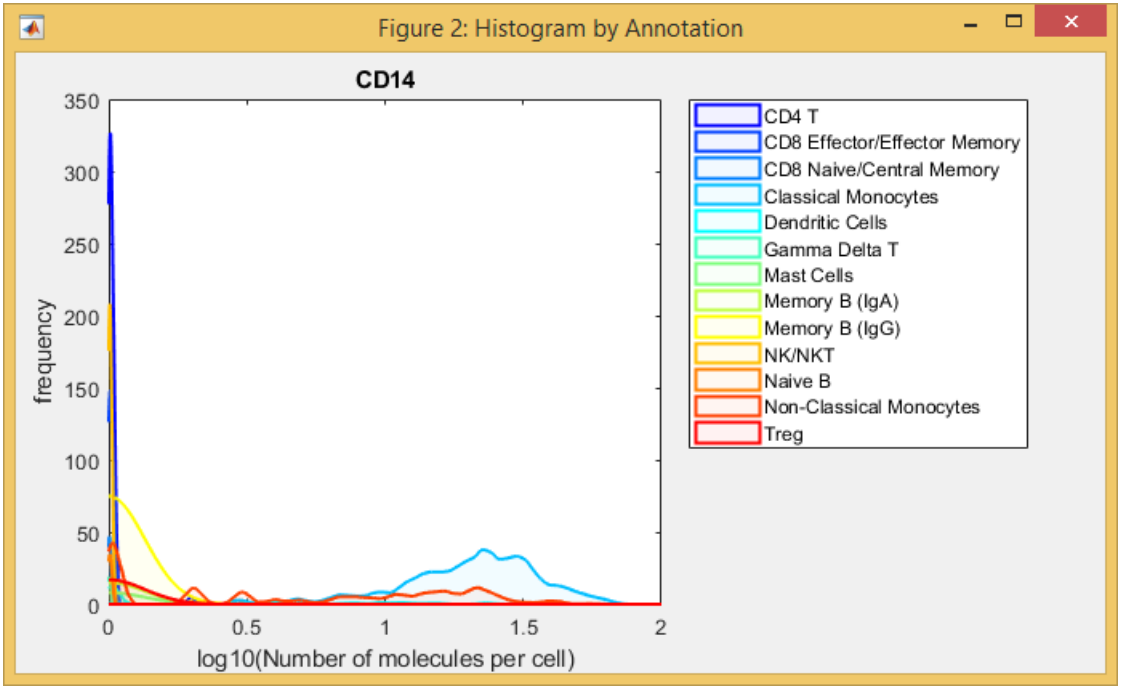
If **Show histogram** is checked, displays the expression level across cells each time a gene is highlighted and clicked in the Gene list box or when **Filter** is clicked after manually entering a gene name. The histogram is displayed in a pop-up window. For example:



- The histogram is automatically saved in the *Histograms* subfolder.
- Cells in the highlighted data table and unhidden cells are used in the calculation.
- Click **Show histogram** and **Smoothing** to display a smoothed histogram. For example:



- Click **Overlapping histograms by annotation** to display a set of histograms plotted according to the number of molecules per cell of a gene. For example:



Box plot

- If **Show boxplot** is checked, displays the expression level per group in the highlighted annotation each time a gene is clicked in the Gene list box or when **Filter** is clicked after manually entering a gene name. The box plot is displayed in a pop-up window.
- At least one annotation must be present.
- The box plot is automatically saved in the *BoxPlots* subfolder.

- Cells in the highlighted data table and unhidden cells are used in the calculation.
- If **Show median of each annotated group** is clicked, displays the median molecule count for each annotated group in box plot. A .csv file with mean, median, and standard deviations for each annotated group is saved automatically.

Correlated genes

Displays a list of genes correlated (or anti-correlated) with the selected gene if the number of genes in the highlighted data table is ≤1,500.

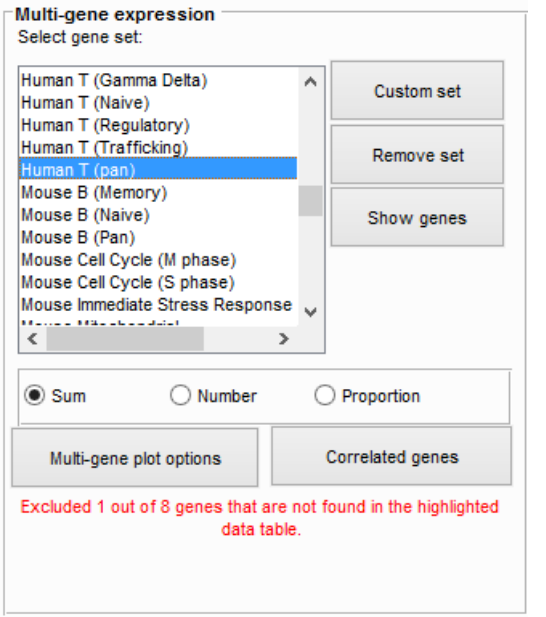
- The correlation coefficient is computed together with p-values for testing the hypothesis of no correlation. See the **corrcoef** function in MATLAB for more details at mathworks.com/help/matlab/ref/corrcoef.html.
- A .csv file with the same information as the displayed table is output and saved in the *Correlation* subfolder. The file can be loaded into the Select gene set list box using **Custom set**.

Gene list drop-down menu

Controls genes names displayed.

- Select from the drop-down menu:
  - **Full Gene List in Master:** Displays the full list of genes in the Master Data Table.
  - **Highlighted Gene Set:** Displays the list of genes in the gene set as selected in the Select gene set list box and present in the Master Data Table.

(7) Multi-gene expression



Select gene set list box

For all functions described in this section, check or uncheck log10 to transform the plots between log_{10} and linear scale.

Lists pre-loaded gene sets consisting of commonly examined human and mouse gene sets when BD Data View is launched.

When a gene set is selected:

- The Gene expression plot is updated. Cells are colored based on the parameter (sum, proportion, or number) of the highlighted gene set. The top right corner displays:
  - The number of cells and percentage of all cells with at least one molecule of the entire gene set.
  - The number of molecules across all cells detected for the gene set and the corresponding percentage across all detected molecules from all genes.
- The gene list box lists the genes in the highlighted gene set that are present in the current data table. If short gene names are used for AbSeq markers, then the gene list box includes an AbSeq gene list.
- If checked, a box plot and histogram are displayed for the selected parameters (sum, proportion, or number).

Sum

Displays the sum of the number of molecules from all genes in the highlighted gene set detected per cell. The sum can be displayed either in linear or log_{10} scale.

If you select another **Number** or **Proportion**, reselect the gene set to update the plot.

Number

Displays the number of genes in the highlighted gene set detected per cell. This can be displayed either in linear or log_{10} scale.

Proportion

Displays the (sum of the number of molecules from all genes in the highlighted gene set)/(sum of total number of molecules) per cell. The proportion is calculated only on the linear scale.

Custom set

Loads custom gene sets.

Load from file

- File must contain ≥2 columns. The first column has the gene set name, and the second column has the gene symbol with the correct case. The remaining columns are ignored.

  There should be no white space at the end of the gene name, and the gene names in the .csv file must exactly match in name and capitalization the gene names in the loaded data file. There should be one row to describe the columns. Header lines between two lines of ##### are ignored.

- Example files:
  - List of featured genes output by the BD Rhapsody Analysis pipeline (clustering analysis).
  - List of genes output and saved by this application using the function **Correlated genes** in the Single gene expression or Multi-gene expression panel and **Diff. expression** in the Annotation panel.

Paste/Enter in window

- Enter in the window one gene name per row without trailing white space.

Remove set

Removes custom gene sets.

Pre-loaded gene sets cannot be removed.

Show genes

Displays a table with a list of genes in the highlighted gene set.

Multi-gene plot options

Histogram

If **Show histogram** is checked, displays the distribution of the parameter (sum, proportion, or number) of the highlighted gene set across cells.

- Displays each time a gene set is selected in the select gene set list box.
- The histogram is automatically saved in the *Histograms* subfolder.

- Click **Smoothing** to display a smoothed histogram.
- Click **Overlapping histograms by annotation** to display the distribution of the highlighted gene set across the groups in the highlighted annotation.

Box plot

If **Show boxplot** is checked, displays the parameter (sum, number, or proportion) of the highlighted gene set across the groups in the highlighted annotation.

- Displays each time a gene set is selected in the select gene set list box.
- The box plot is automatically saved in the *BoxPlots* subfolder.
- Click **Show median** to display the median of the parameter (sum, number, or proportion).

Multi-gene box plots

If Multi-gene box plots is clicked, the function creates a series of box plots of the selected gene set (≤30 genes) for the cell groups in the selected annotation. The function can be used to provide a graphical overview of the distribution of the genes that are representative of each cell group.

- Molecule counts can be plotted in linear or $\log_{10}$ scale. Check **log10**.
- The plot is saved in the *MultiGenePerGroupPlots* subfolder.

If (**Show median**) is clicked, displays the median molecule count for each annotated group in the box plot. A .csv file with mean, median, and standard deviations for each annotated group is saved automatically.

Heatmaps

Heatmap (group)

Allows you to visualize gene expression levels per cell cluster in a heatmap format.

- This function is limited to ≤1,500 genes.
- Cells in the highlighted data table and unhidden cells are used for calculation.
- Non-normalized or normalized data can be used.
- **Log10**-transformed data can be used.
- Multiple gene sets can be selected and displayed.
- Genes can be ordered by: order in gene list, mean or median expression of all cells in a selected annotated group.
- Three heatmaps are displayed and saved as .png files in the *HeatMaps* subfolder:
  - Mean number of molecules per cell in each group of the highlighted annotation.
  - Median number of molecules per cell in each group of the highlighted annotation.
  - Relative expression: (mean number of molecules per cell for cells in each group)/(average number of molecules per cell from all cells in the highlighted data table).
- Creates .csv file with mean, median, relative expression, and standard deviation for each annotated group.

Heatmap (cell)

Allows you to visualize gene expression levels per cell in a heatmap format.

- This function is limited to ≤1,500 genes.
- Cells in the highlighted data table and unhidden cells are used for calculation.
- Non-normalized or normalized data can be used.
- **Log10**-transformed data can be used.

- Multiple gene sets can be selected and displayed.
- Single (cell or gene only) or bi-directional (cell and gene) hierarchical clustering can be calculated. Complete linkage and correlation distance are used for the clustering. See **clustergram** function in MATLAB for more details at mathworks.com/help/bioinfo/ref/clustergram.html.
- The order of genes and cells from the hierarchical clustering will define the order of how they are displayed in the heatmaps.
- Three heatmaps are generated and saved in the *HeatMaps* subfolder:
  - Heatmap showing gene by all cells (.png and .fig)
  - Heatmap showing cell-to-cell correlation (.png)
  - Heatmap showing gene-to-gene correlation (.png and .fig)
- Left-click to zoom in, right-click to zoom out, or select other zoom options.
- If the MATLAB application is installed, the .fig files can be opened to use the zoom function and inspect gene names. Use the zoom function when the number of genes plotted is large. If MATLAB is not installed, the .fig files cannot be opened.

Correlated genes

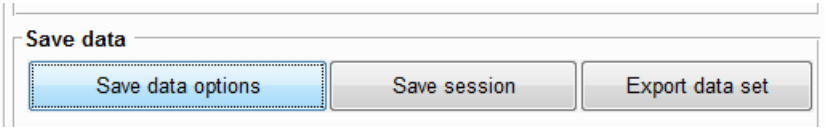
Computes groups of correlated genes based on a defined threshold correlation coefficient.

- This function is limited to ≤1,500 genes due to computation time.
- Cells in the highlighted data table and unhidden cells are used for the calculation.
- Non-normalized or normalized data can be used.
- **Log10**-transformed data can be used.
- An output .csv file is saved in the *Correlation* subfolder. This file can be loaded into the select gene set list box using **Custom set**. The file has the following columns:
  - Group number
  - Gene name

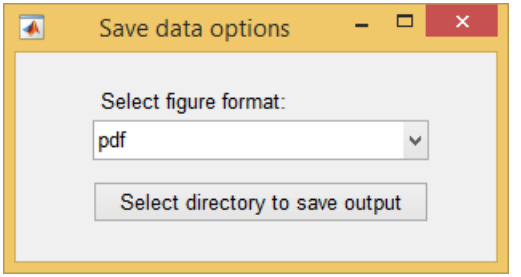
  - Total number of molecules expressed by all cells for each gene
  - Total number of cells expressing each gene
- The correlated gene sets are automatically added to the Select gene set list box. To see the genes, click **Show genes** or toggle to **Highlighted gene set** in the drop-down menu.

---

(8) Save data



Save data options Specifies format of figure output (PDF, SVG, or PNG). Default is PDF.



Select directory to save output

Specifies the directory for saving files.

- The default directory is the directory that contains the loaded data table.
- You can change the directory to save by using this function.
- You might need to change the working directory if you do not have write permission to the default directory containing the input data table files.

Save session

Saves all the variables of the session in a .mat file.

- If a session with the given name exists in the working directory, it is replaced by the current session.
- The .mat file can be reloaded into the application using **Load session** in the Data table panel.

Export data set

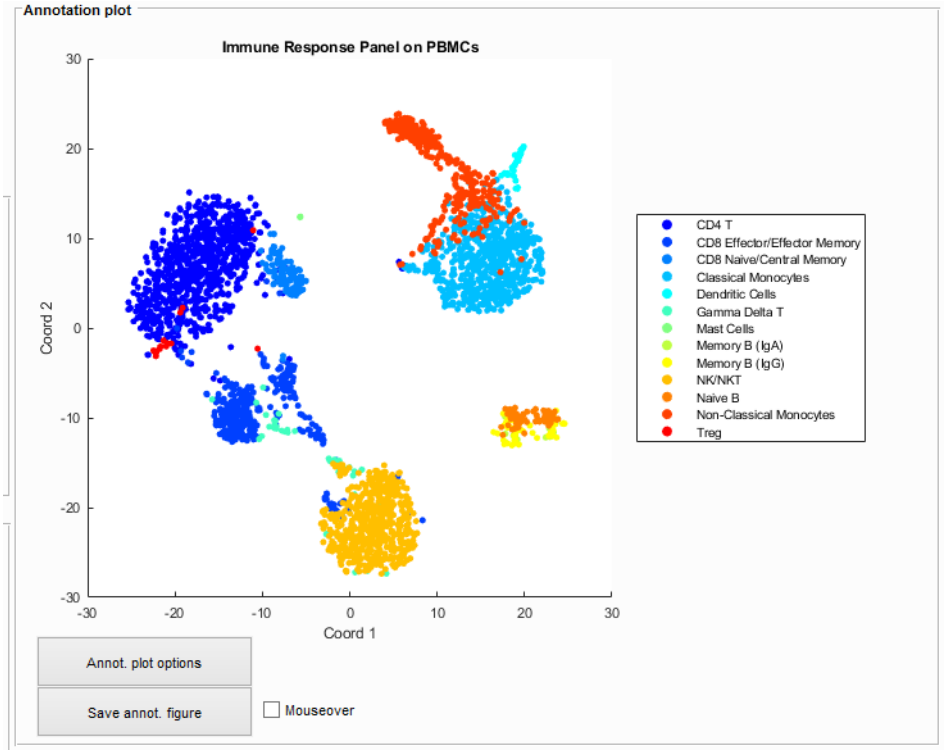
Exports data tables, coordinates, and annotations as .csv files.

The data table is saved in cell by gene .csv format, even if the original file loaded is gene by cell, or in .st or .mtx formats.

- Highlight the data table in the data table list box to be exported.
- If normalized data has been computed, you can choose to export a normalized data table.
- If the highlighted data table contains genes with zero counts from all cells, you can choose to export a data table with only non-zero genes. The list of non-zero genes is also saved as *_NonZeroGeneList.csv.
- If the highlighted data table is a result of **Select variable genes** or **Filter data table** function and contains a subset of genes of the Master Data Table, you can choose to export the accompanying data table with the full list of genes for the cells in the highlighted data table. The list of genes in the reduced data table is also exported as *_ReducedGeneList.csv.
- The first column of the exported data tables preserves the cell index or well label in the loaded data tables. If multiple files are loaded, the short file name given at data loading is added in front of the cell index or well label.

- If sets of coordinates have names starting with the name of the highlighted data tables (by default, coordinates generated from a data table start with the name of the data table), they are displayed to choose for exporting. All coordinate sets are displayed.
- If a chosen set of coordinates is derived from more cells than the selected data table, only the coordinate values of the cells present in the selected data table are exported.

(9) Annotation plot



Annotation plot options

Choose-color

Select from the drop-down menu:

- **Jet: default colors.**
- **Current random:** Uses a preset random color palette to color the annotation plot. Click **Label** to refresh plot and display a new color.
- **Generate new random color**: Generates a random color palette for t-SNE annotation plot. You can continue to use the current color palette. Click **Label** to refresh plot and display a new color.

Show marker edge

Displays a gray outline around each point on the Annotation plot. The change displays after refreshing the plot by clicking **Label**.

Show group number

Displays a number of each cell group in the highlighted annotation at the center of the cell group when checked or hides when unchecked. Refresh the plot by clicking **Label**.

Show population percentage

Displays a population percent of each cell group when checked or hides when unchecked. Refresh the plot by clicking **Label**.

log10

Transforms value or size of group by $\log_{10}$. The change displays after refreshing the plot by clicking **Label**.

Save annot. figure

Saves the current content of the Annotation plot. The figure is saved in the *Plots* subfolder.

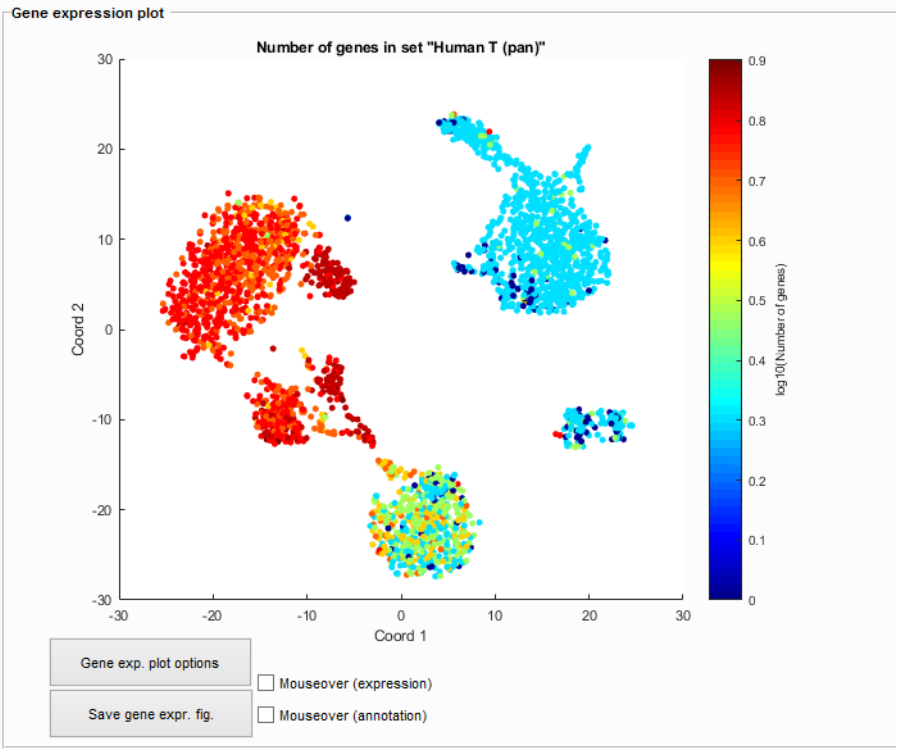
After the plot is saved, on the pop-up figure, you can left-click to zoom in, right-click to zoom out, or select other zoom options.

Mouseover

Displays the label or value in the selected annotation group or value when the cursor hovers over a point (cell). Select the annotation to show on mouseover.

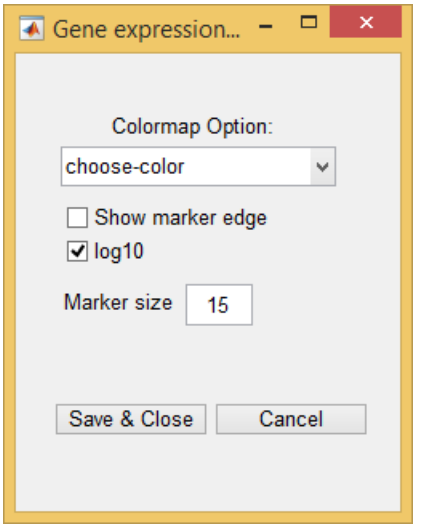
Note: Depending on computer configuration, this function can be slow when the plot contains many data points. It is intended for data tables with data points in the hundreds.

(10) Gene expression plot



Gene exp. plot options

Specifies color, marker properties, and $\log_{10}$ scale.



choose-color

Select from the drop-down menu to choose one of six color schemes. Jet is the default color scheme.

Show marker edge

Displays a gray outline around each point on the Gene expression plot. The change displays after refreshing the plot by selecting a single gene, entering a single gene, or selecting a multi-gene set.

log10

Transforms calculations or operations in Single gene expression and Multi-gene expression panels to $\log_{10}$. The change displays after refreshing the plot by selecting a single gene, entering a single gene, or selecting a multi-gene set. See (7) Multi-gene expression (page 124).

Marker size

Changes the size of the dots.

- The change is applied to both the Annotation plot and Gene expression plot.
- The change displays after refreshing the plot by selecting a single gene, entering a single gene, or selecting a multi-gene set.

Save gene expr. fig

Saves the current content of the Gene expression plot. The figure is saved in the *Plots* subfolder.

After the plot is saved, on the pop-up figure, you can left-click to zoom in, right-click to zoom out, or select other zoom options.

-

Mouseover (expression) and Mouseover (annotation)

Displays an annotation or the actual value of the gene expression value when the cursor hovers over a point (cell).

- Functions when one or the other box is checked.
- To display a gene expression value, check **Mouseover (expression),** and refresh the plot by selecting a single gene, entering a single gene, or selecting a multi-gene set. Untransformed (linear) count is displayed regardless of whether **log10** is selected.

Analyzing targeted sequencing output files from a single BD Rhapsody™ experiment with BD Data View

Introduction

This example illustrates how to view and analyze targeted sequencing output files from the BD Rhapsody Analysis pipeline for a single experiment with BD Data View.

The data set used in this example is derived from a mixture of Jurkat and Ramos cells. BD Data View is used to quickly identify cell clusters and generate differential gene expression heatmaps. For a link to download all the example data sets, see the readme.txt in the BD Data View zip file.

Before you begin

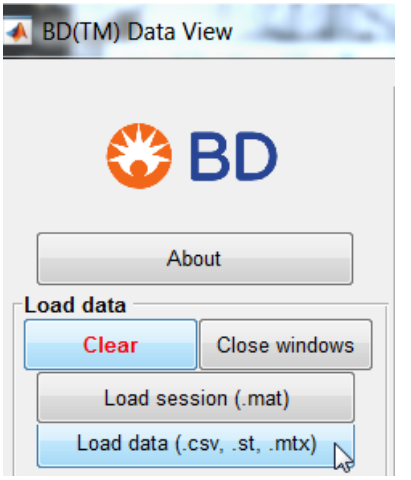
1. Install BD Data View. See the *BD Single Cell Genomics Analysis Setup User Guide* (Doc ID: 47383) to install BD Data View v1.1 or later.
2. Run the BD Rhapsody Analysis pipeline. See the *BD Single Cell Genomics Analysis Setup User Guide* to run the appropriate pipeline with FASTQ and FASTA files on the Seven Bridges Genomics platform or on a local installation.
3. From the analysis pipeline, obtain the output files:
   - <sample_name>_DBEC_MolsPerCell.csv
   - <sample_name>_UMI_Adjusted_Stats.csv
   - <sample_name>_bh-tSNEcoordinates.csv
   - <sample_name>_<num_clusters>_Labels.csv
   - <sample_name>_<num_clusters>_cluster_features.csv

Workflow steps

Step	Purpose
1	Load data.
2	View projections of high-dimensional data.
3	Evaluate sequencing depth.
4	View cluster labels.
5	Identify the clusters.
6	Explore the clusters.

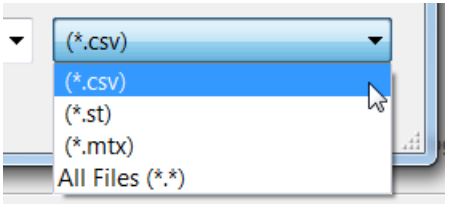
Loading data

1. In the Load data panel, click **Load data**:



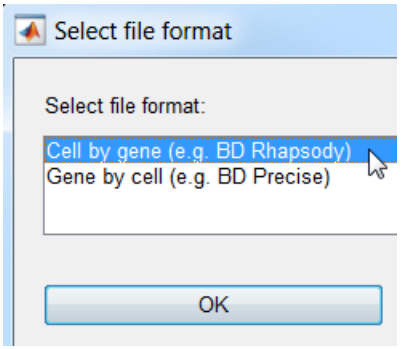
Note: Files formatted as .mat are readable by BD Data View.

2. Browse for the <sample_name>_DBEC_MolsPerCell.csv or <sample_name>_Expression_Data.st file. Ensure that you have specified .csv or .st file types in the explorer window:



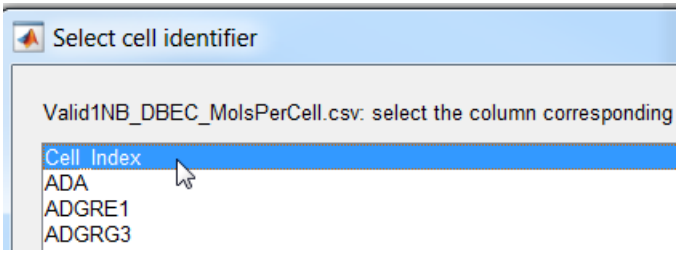
After you select the file and click **Open**, the Select file format window is displayed.

3. Select **Cell by gene (BD Rhapsody)** format:

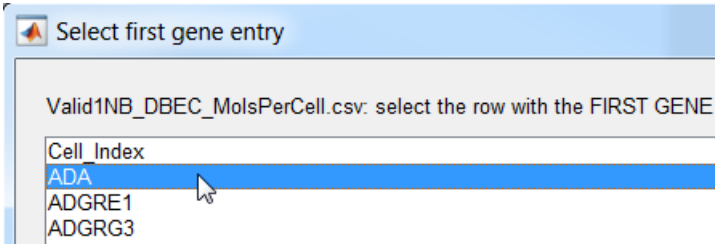


The Select cell identifier window is displayed.

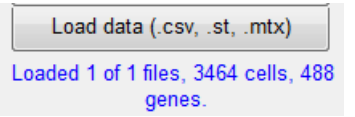
4. Select the column with the cell identifiers (**CELL ID**). For example:



5. Select the row with first gene entry, and then click **SELECT**. For example:



The software indicates that the file was uploaded and displays the number of cells and genes in the file. For example:



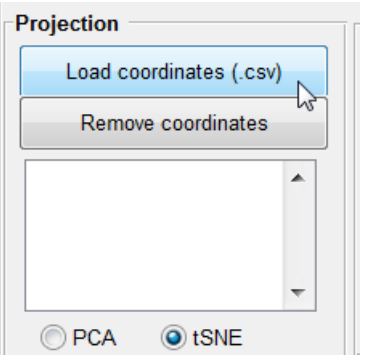
Note:

- To view the name of the file loaded, click **Show file names**.
- To delete an uploaded file, click **Clear**.

Viewing projections of high-dimensional data

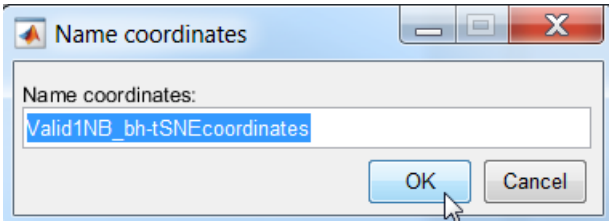
Use output from clustering analysis with the filename <sample_name>_bh-tSNEcoordinates.csv to load coordinates.

1. In the Projection panel, click **Load coordinates**:



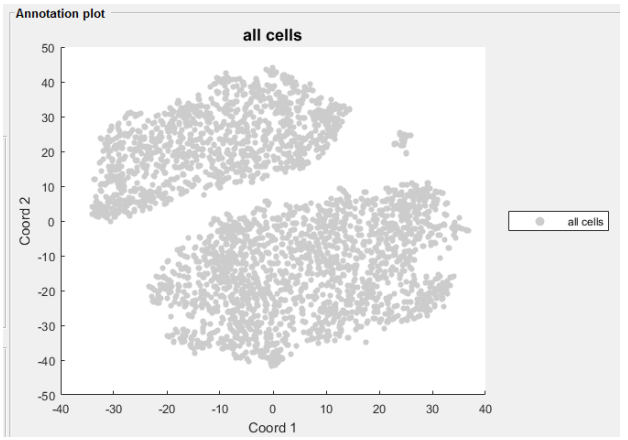
2. Browse to the <sample_name>_bh-tSNEcoordinates.csv file, and then click **Open**.

3. (Optional) Rename the coordinates. For example:



4. Click **OK**.

   The data points are displayed. For example:



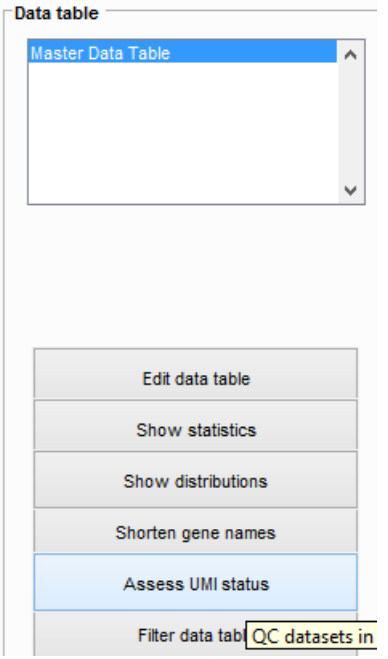
Note: To remove the coordinates, click **Remove coordinates,** and then click **Yes.**

Evaluating sequencing depth

Sequencing depth can be an important consideration in analysis of data with molecular indices. The BD Rhapsody Analysis pipeline outputs a file called <sample_name>_UMI_Adjusted_Stats.csv to document which gene has sufficient depth to undergo DBEC distribution-based error correction.

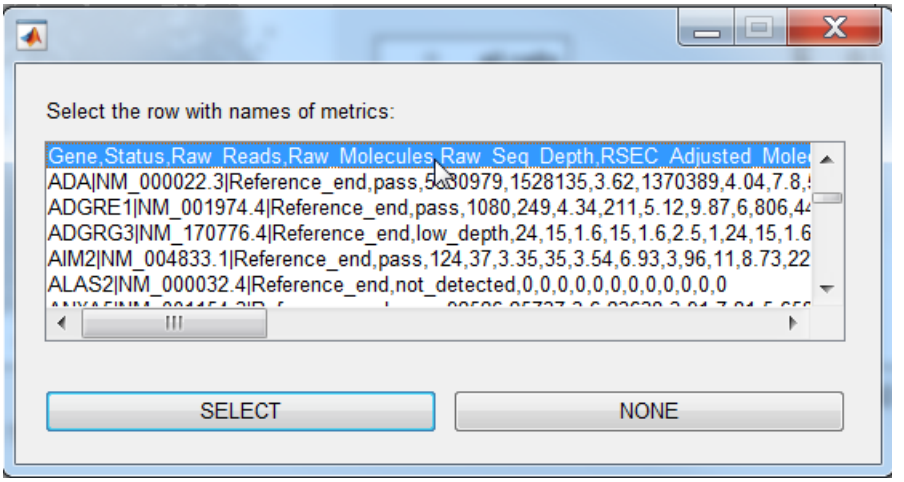
For cell type clustering of a single sample, data quality control can be skipped. Proceed to Viewing cluster labels (page 152). For comparative analysis of multiple samples, perform data quality control.

1. In the Data table panel, click **Assess UMI status**:

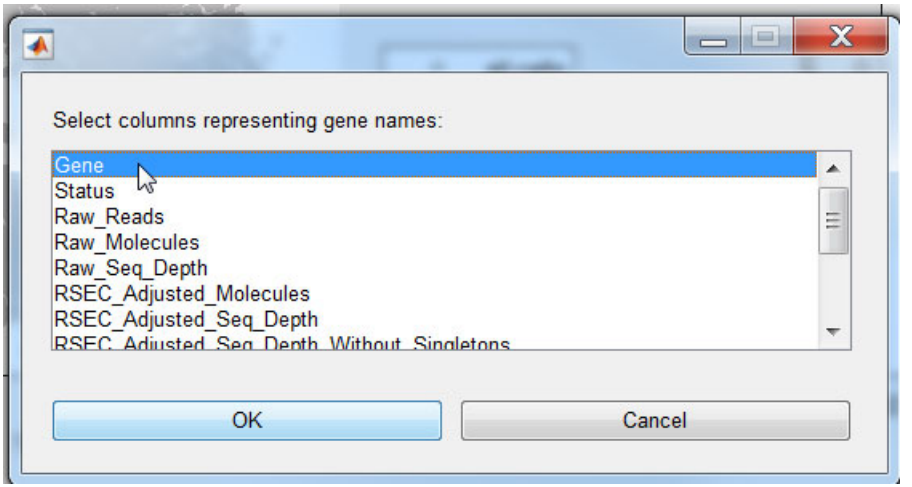


2. Browse to the appropriate file with the UMI adjusted statistics, <sample_name>_UMI_Adjusted_Stats.csv, and then click **Open**. The field names in the .csv file are listed.

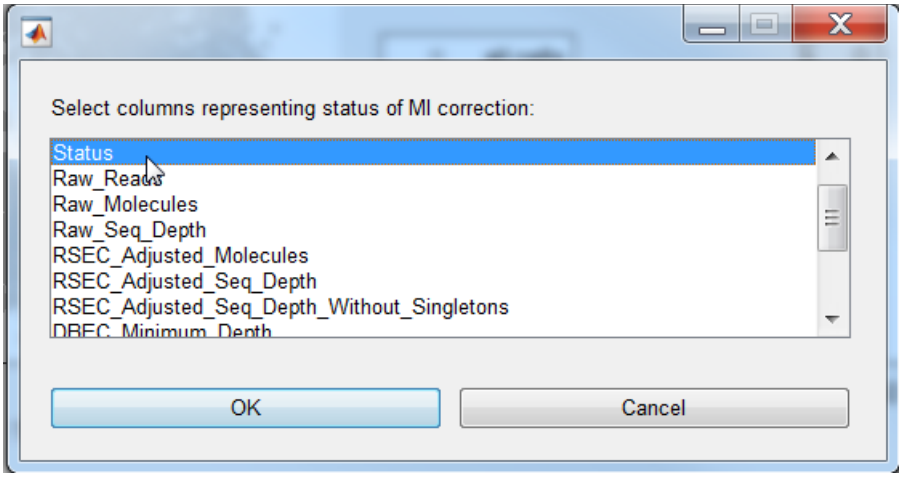
3. Select the row with the names of the metrics, and then click **SELECT**. For example:



4. Select the column name with the gene names, and then click **OK**. For example:



5. Select the column name with the status of MI correction, and then click **OK**. For example:



Three windows are displayed to evaluate the data:

– Figure 1: MI correction status. The application saves the table as CombinedMIAdjustStatus.csv in the current directory. For example:

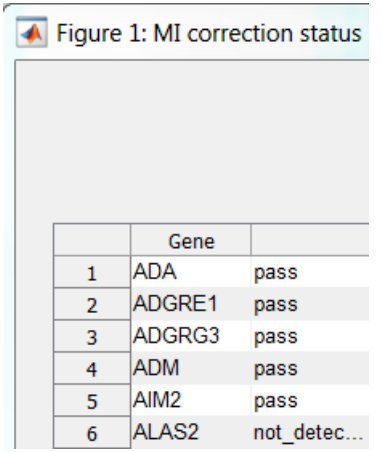


Figure 1: MI correction status

	Gene	
1	ADA	pass
2	ADGRE1	pass
3	ADGRG3	pass
4	ADM	pass
5	AIM2	pass
6	ALAS2	not_detec...

– Figure 2: Percentage counts within each file for non-passing genes. Those genes that have insufficient depth are listed in descending order according to their contribution to the total molecule counts of the entire sample. For example:

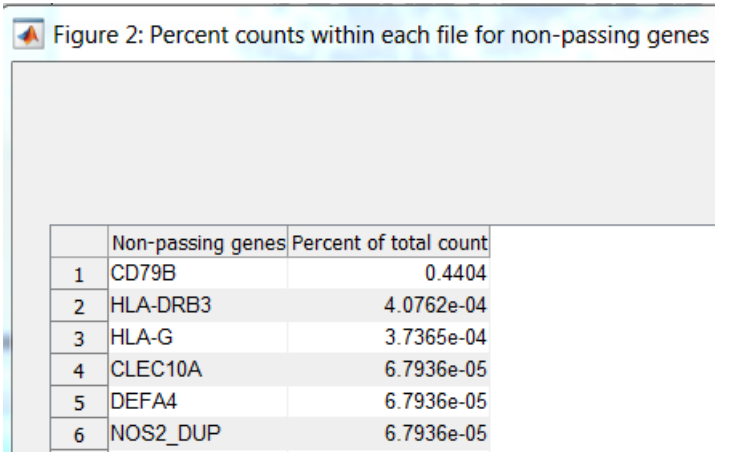
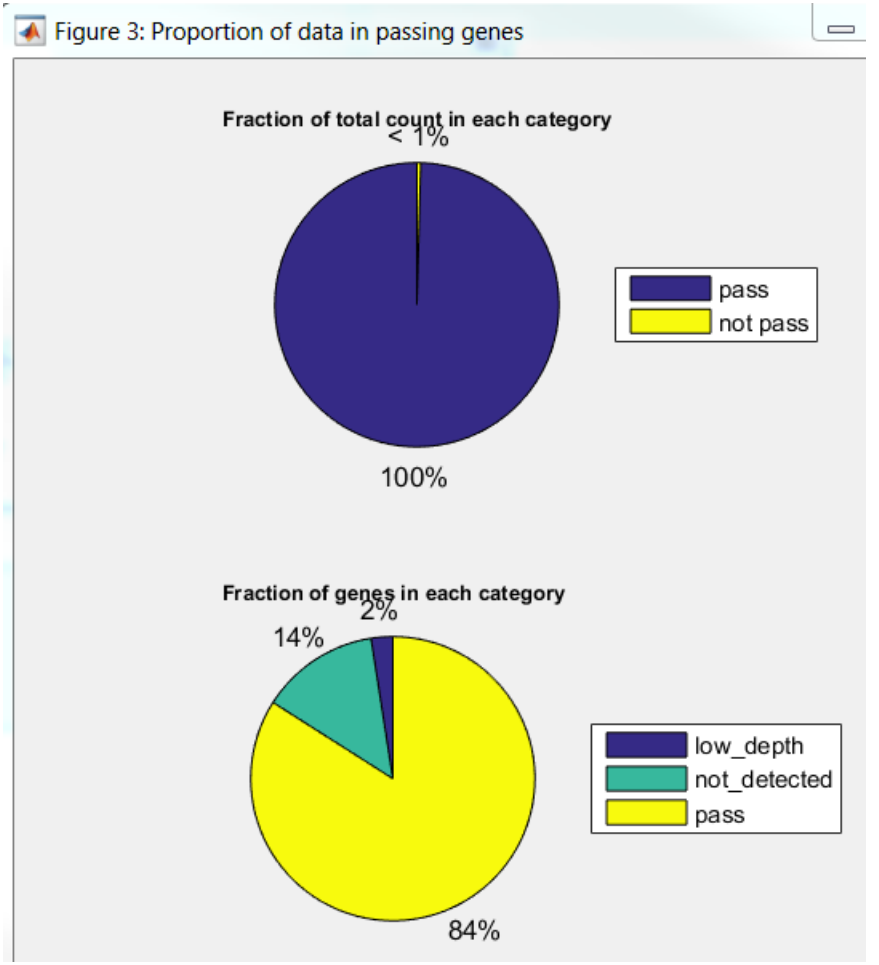


Figure 2: Percent counts within each file for non-passing genes

	Non-passing genes	Percent of total count
1	CD79B	0.4404
2	HLA-DRB3	4.0762e-04
3	HLA-G	3.7365e-04
4	CLEC10A	6.7936e-05
5	DEFA4	6.7936e-05
6	NOS2_DUP	6.7936e-05

– Figure 3: Proportion of data in passing genes. The proportion of each category of genes and the contribution of each category of genes to the total molecules for each loaded file are displayed. For example:

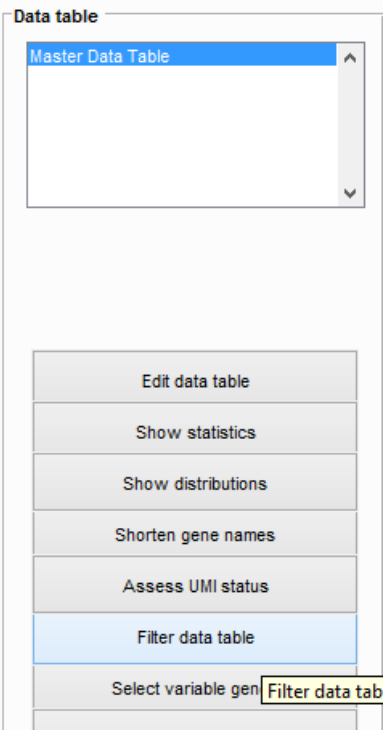


The set of graphs is saved as MIStatus_QC.pdf in the current directory.

6. If the sequencing depth status for all genes is satisfactory, proceed to Viewing cluster labels (page 152). If under-sequenced genes are to be removed, proceed to step 7.

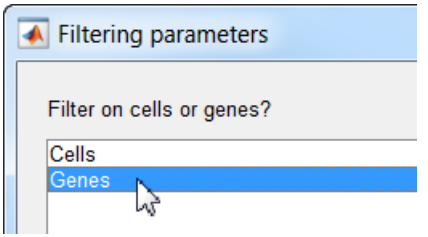
   Note: If the application is for in-depth analysis, many of the genes are low-depth, and the proportion of reads from low-depth genes is considerable (>10–20%), consider sequencing the library further. If the experiment is only to identify distinct types of cells, and one sample is being considered, resequencing might not be necessary. Filter under-sequenced genes by proceeding to step 7.

7. In the Data table panel, click **Filter data table**:



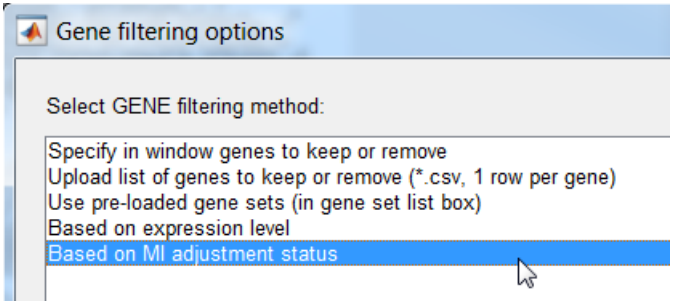
The filtering parameters window is displayed.

8. Click **Genes**, and then click **OK** to filter on genes:



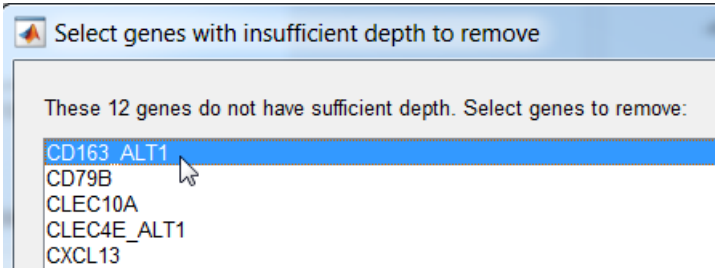
The filtering method window is displayed.

9. Click **Based on MI adjustment status**, and then **Select method(s)**:



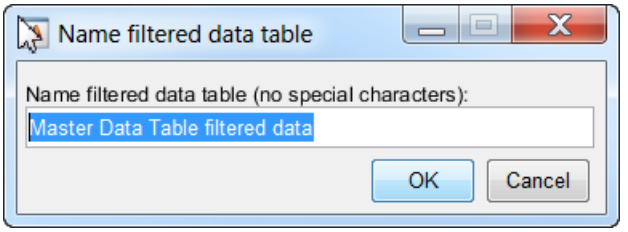
The window Select genes with insufficient depth to remove is displayed.

10. Select to remove genes with insufficient depth or **Select all**, and then click **OK.** For example:

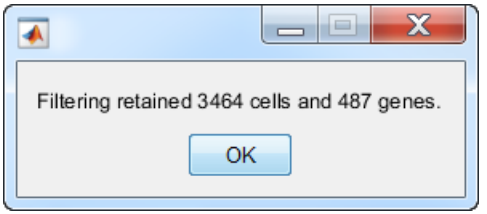


The Name filtered data table window is displayed.

11. If desired, rename the filtered data table. For example:



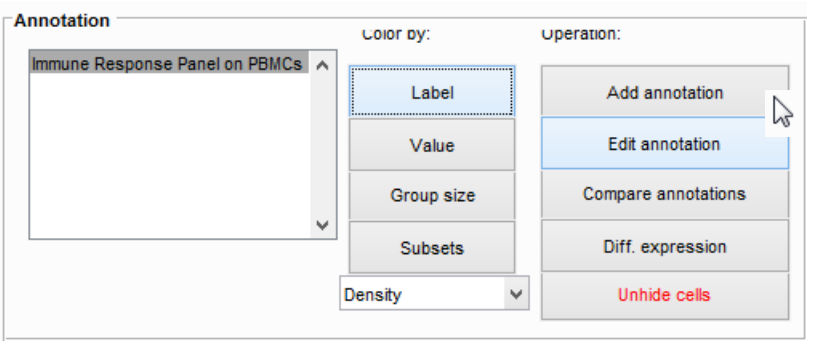
The software acknowledges how many cells and genes were retained. For example:



Viewing cluster labels

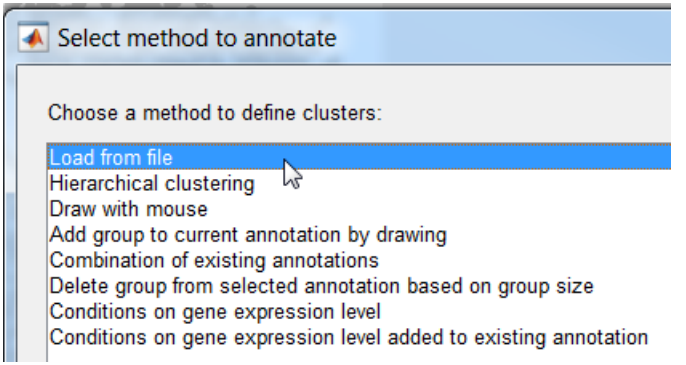
View the cell clusters defined by the clustering analysis pipeline by loading the cluster assignment file <sample_name>_<num_clusters>_Labels.csv.

1. In the Annotation panel, click **Add annotation**:



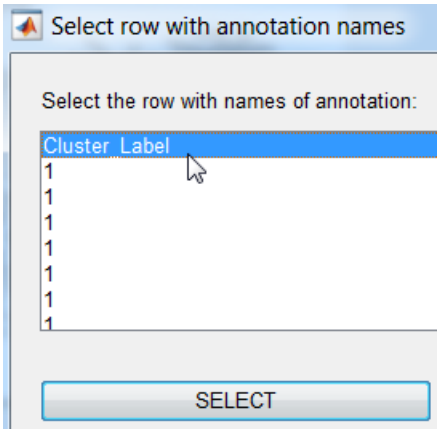
The window Select method to annotate is displayed.

2. Click **Load from file**, and then click **OK**:

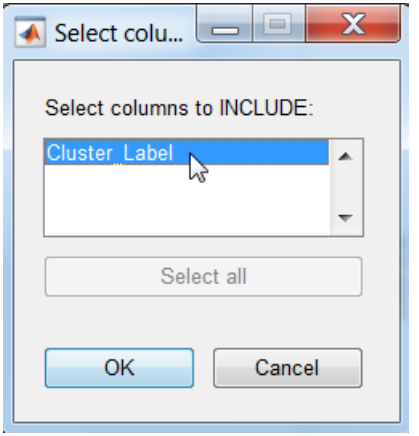


3. Browse for <sample_name>_<num_clusters>_Labels.csv in the cluster analysis output. Click **Open**.

4. Select the row from the file with the names for annotating the clusters, and then click **SELECT**. For example:

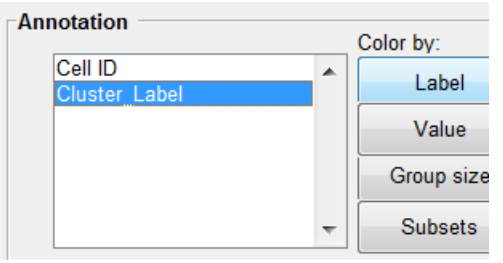


5. Select the columns that have the cluster label, and then click **OK**. For example:

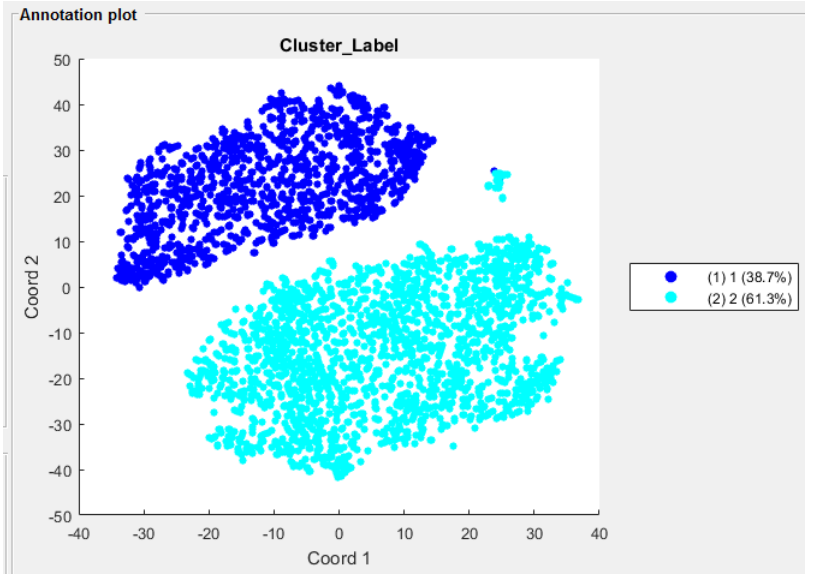


The software lists the .csv file in the Annotation list box.

6. In the Annotation panel, click **Label** to refresh the plot and see the annotations:



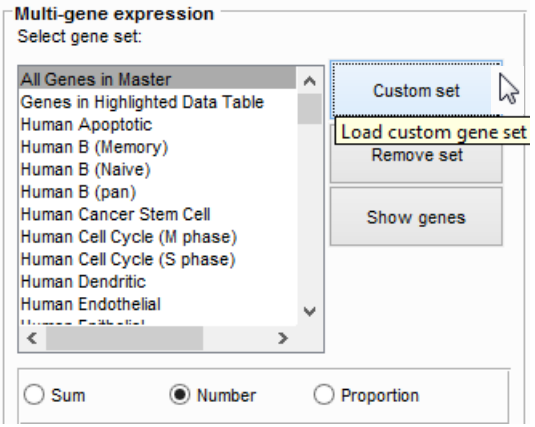
For example:



Identifying the clusters

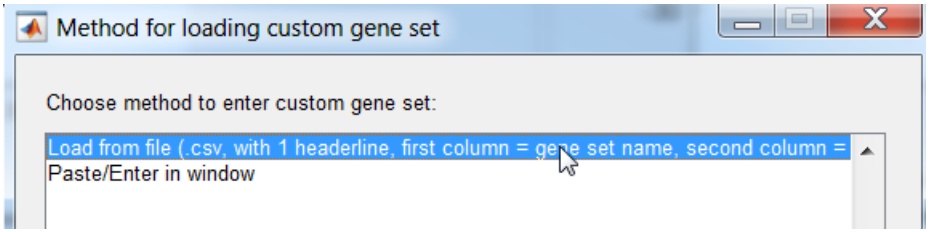
The biological identity of each of the clusters can be determined by the set of genes statistically over-represented by the cluster. A list of over-represented genes in each cluster is in the file <sample_name>_<num_clusters>_Cluster_Features.csv.

1. In the Multi-gene expression panel, click **Custom set**:

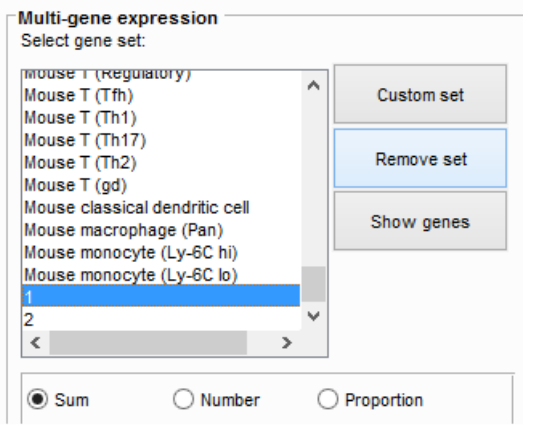


   The window Method for loading custom gene set is displayed.

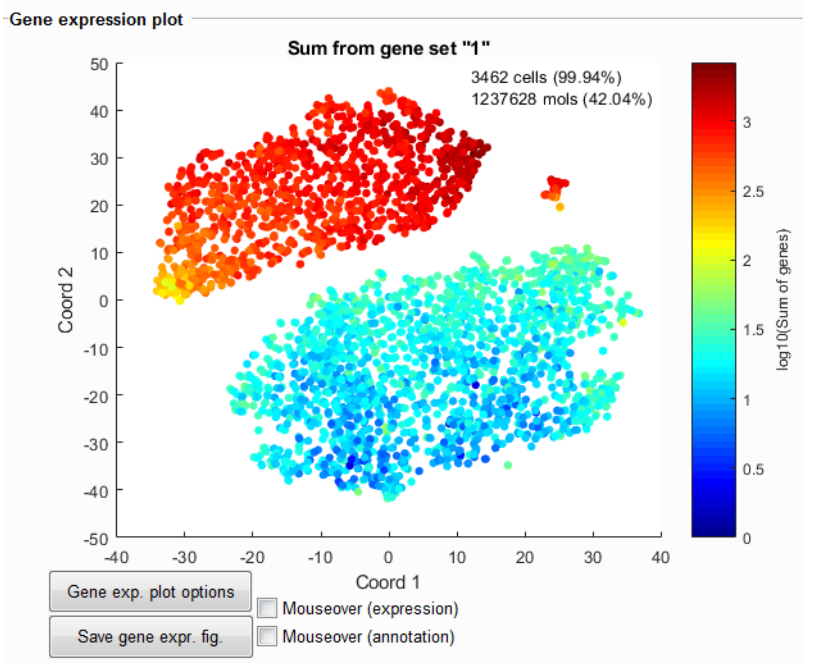
2. Click **Load from file**, and then click **OK** to enter the custom gene set. For example:



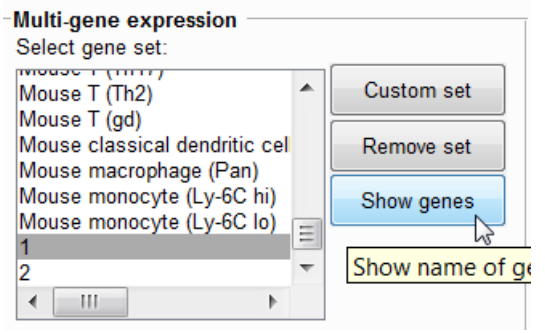
3. Browse to the appropriate file corresponding to the number of clusters, <sample_name>_<num_clusters>_Cluster _Features.csv, and then click **Open**. The field names in the .csv file are listed. The custom gene sets are added in the Multi-gene expression panel. For example:



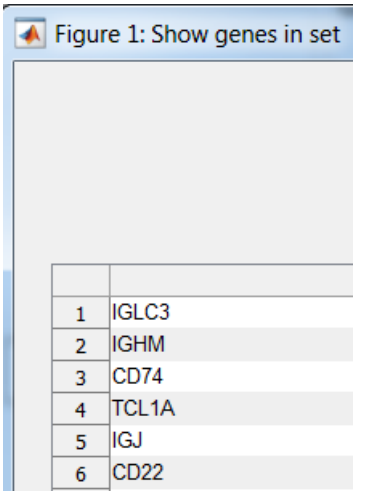
The over-represented genes associated with cluster 1 are displayed by clicking a gene set in the Select gene set box:



4. In the Multi-gene expression panel, click **Show genes** to view the list of over-represented genes associated with a cluster:



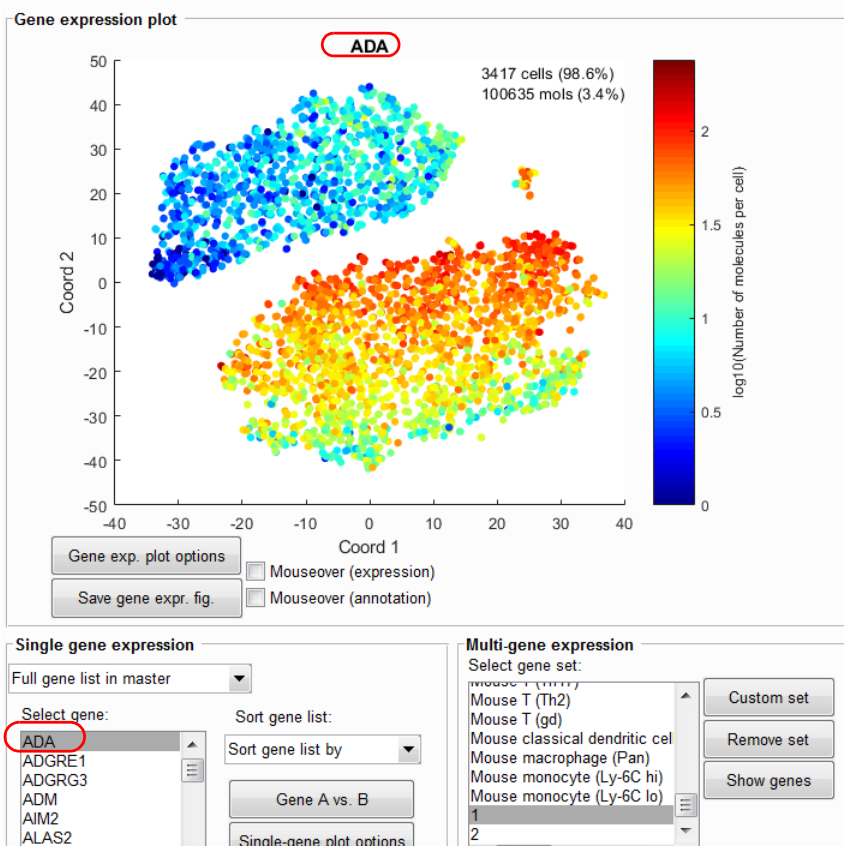
For example, here is a partial list:



This list of genes contains marker genes for B cells, suggesting that the cluster is from the Ramos cell line, which is derived from B cell lymphoma.

Exploring single gene expression

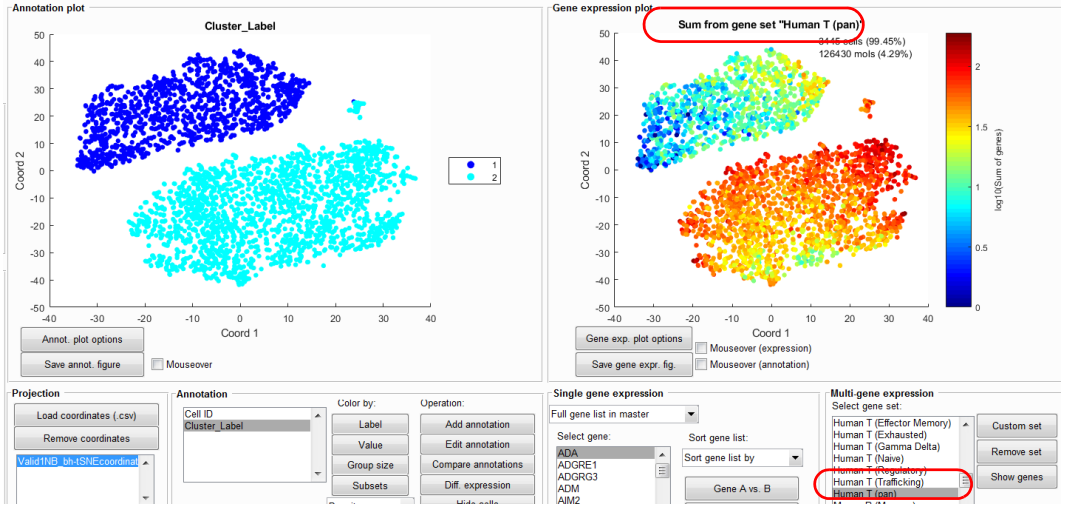
In the Single gene expression panel, click genes to view their detected level (molecules per cell). For example, selecting the ADA gene displays its expression across clusters in the right gene expression plot:



Exploring multi-gene expression

In the Multi-gene expression panel, click gene sets to view their detected level (molecules per cell). Gene sets in the multi-gene expression list have been curated from the scientific literature and are provided to help identify cell type.

In this example, choosing **Sum** (the sum of the number of molecules from all genes in the highlighted gene set detected per cell) and then selecting **Human T (pan)** displays relatively high expression of the gene set in the cluster of Jurkat cells, a T cell-derived cell line:



Analyzing multiple samples with BD Data View

Introduction

This workflow shows how to compare multiple data sets with BD Data View. In this example, BD Data View is used to find any differences in cell type proportion and expression pattern between healthy and diseased peripheral blood mononuclear cells (PBMCs). The application is also used to detect specific gene expression differences between the two cell groups.

Before you begin

1. Install BD Data View, and run the BD Rhapsody Analysis pipeline. See Before you begin (page 138).
2. From the analysis pipeline, obtain the output files:
   - PBMC_Healthy1_DBEC_MolsPerCell.csv
   - PBMC_Disease1_DBEC_MolsPerCell.csv
   - PBMC_Healthy1_UMI_Adjusted_Stats.csv
   - PBMC_Disease1_UMI_Adjusted_Stats.csv
   - <sample_name>_Metrics_Summary.csv

Workflow steps

Step	Purpose
1	Load data.
2	Evaluate sequencing depth.
3	Generate projections of high-dimensional data.
4	Identify doublets clusters.
5	Filter out potential doublets and genes with inconsistent UMI status.
6	Randomly subsample an equal number of cells from each sample.
7	Generate projections of high-dimensional data.
8	Define clusters manually.
9	Rename the clusters.
10	Compare cell type proportions between samples.
11	Find gene expression differences between healthy and diseased samples.
12	Display differential expression in a heatmap.
13	Explore cell subsets between samples with hierarchical clustering.

Loading data

For more detailed instructions on loading data, see Loading data (page 140).

1. In the Load data panel, click **Load data**.
2. Browse for the two files with filename <sample_name>_DBEC_MolsPerCell.csv, and then hold down the **Ctrl** (control) button to select the files:
   - PBMC_Healthy1_DBEC_MolsPerCell.csv
   - PBMC_Disease1_DBEC_MolsPerCell.csv

   After you select the file and click **Open**, the Select file format window is displayed.
3. Select **Cell by gene (BD Rhapsody)** format, and then click **OK**:

   The files are concatenated (*Master Data Table*). The Select cell identifier window is displayed.
4. Select the column with the cell identifiers (**CELL ID**).
5. Select the row with first gene entry, and then click **SELECT**.
6. Enter short names for each file. For example, *Diseased* and *Healthy*.

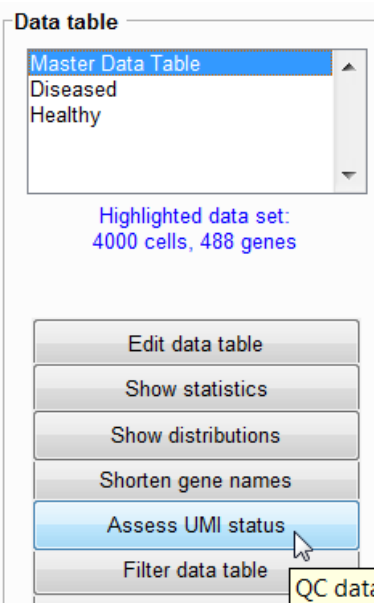
   Note: To view the statistics of the concatenated and loaded files, click **Show statistics** in the Data table panel. In this example, there are Healthy: 12,735 cells and Disease: 7,996 cells.

Evaluating sequencing depth

Sequencing depth is an important consideration in analysis of data with molecular indices of multiple files. The BD Rhapsody Analysis pipeline outputs a file called <sample_name>_UMI_Adjusted_Stats.csv to document which gene has sufficient depth to undergo DBEC.

Another pipeline output, the metrics summary, lists sequencing depth statistics for each of these samples. From the metrics summary, sequencing is not saturated (RSEC depth 2.74, sequencing saturation 81.78% for Healthy, and RSEC depth of 4.79, sequencing saturation of 90.3 for Disease). When RSEC depth is <6, some of the genes might not pass threshold for the DBEC algorithm to be run. If a gene is corrected by DBEC for one sample but not for the other, this can cause an apparent batch effect. To reduce the chance of batch effect, one solution is to remove those genes from consideration when comparing multiple samples.

1. In the Data table panel, click **Assess UMI status:**



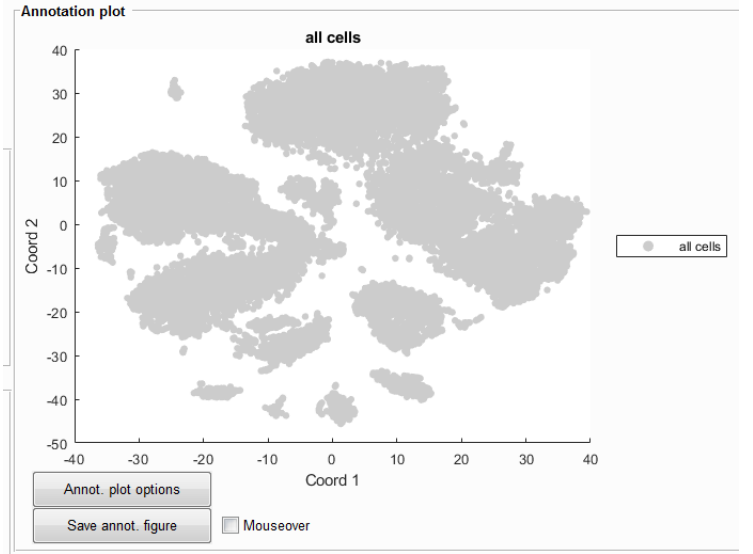
2. Browse to the appropriate two files with the UMI adjusted statistics, <sample_name>_UMI_Adjusted_Stats.csv, select both files; and then click **Open**. The field names in the .csv file are listed.
3. For each file, select the row with the names of the metrics, and then click **SELECT**.
4. For each file, select the column name with the gene names, and then click **OK**.
5. For each file, select the column name with the status of MI correction, and then click **OK**. Three windows are displayed to evaluate sequencing depth: MI correction status, Percentage counts within each file for non-passing genes, and Proportion of data in passing genes.

Note: This analysis showed that a few abundant genes such as LYZ, IGKC, CD74, and HLA-C are not passing threshold in at least one of the samples. BD Biosciences recommends removing them from t-SNE and comparative analysis, because of apparent batch effect caused by inconsistent UMI adjustment status. See Filtering out potential doublets and genes with inconsistent UMI status (page 169).

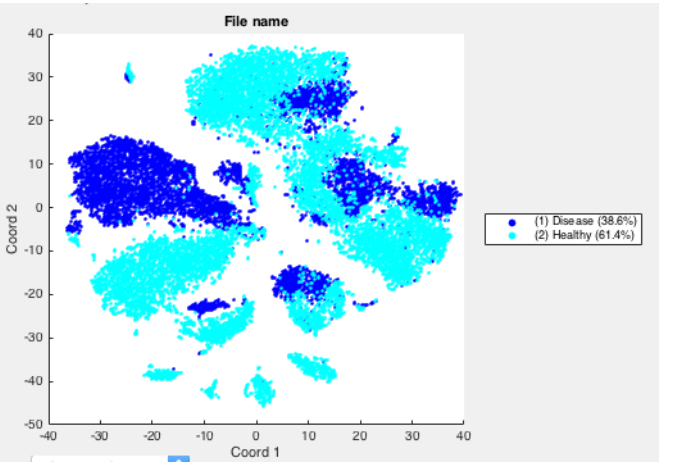
Generating projections of high-dimensional data

You can generate the t-SNE coordinates by selecting **Master Data Table**, **tSNE**, and then clicking **Calculate** in the Projection panel. The coordinates are saved with the session.

The data points are displayed. For example:



To view the data points by cell label, select File name, and then click **Label** in the Annotation panel to refresh the plot. For example:



Note: To remove the coordinates, click **Remove coordinates,** and then click **Yes.**

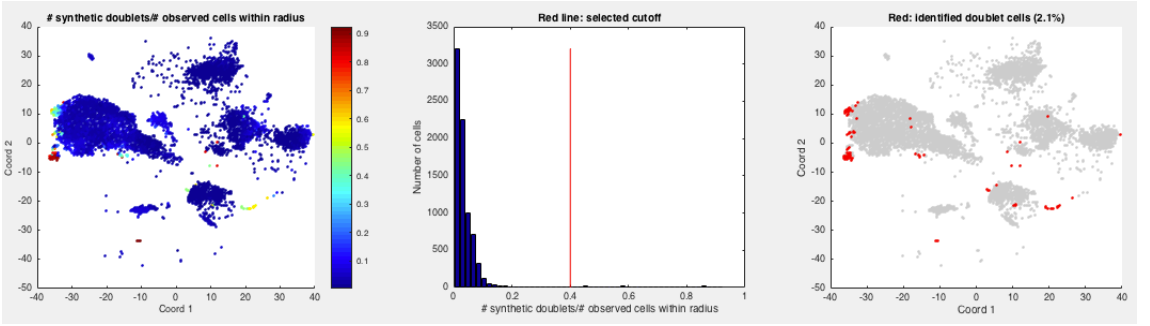
Identifying doublet clusters

Identifying doublet clusters avoids misinterpretation of distinct clusters when they are, in fact, artifacts. See Identify doublet clusters (page 98).

Identify doublet clusters of individual samples, because each library can have different rates of doublet formation depending on cell loading density.

1. In the Data table panel, click to select the data file. For example, the file labelled **Healthy**.

2. In the Data table panel, click **Identify doublet clusters.** For information on parameters, see Identify doublet clusters (page 98). Doublets display in the annotation plot in the Identify doublet cell clusters window. For example:



3. Click **Export doublet annotation** to export the doublet file. For example, Healthy_Doublet.

4. Repeat steps 1–3 for the Diseased file.

5. In the Annotation panel, click **Add Annotation**, and then select **Combination of existing annotations** to combine doublet files. Control-click to select multiple files in the Combine annotations window. For example, combine **Healthy_Doublet** and **Diseased_Doublet**, and name the new annotation **Combined_Healthy_Diseased_Doublets**.

6. In the Annotation panel, click to select **Combined_Healthy_Diseased_Doublets** in the annotation list, and then click **Rename/ reorder**. The combined_healthy_diseased_doublets contain annotation groups: doublet|not_considered, not_considered|doublet, not_considered|singlet, singlet|not_considered. The *not considered* are the cells in Healthy when only the Disease data set is considered for the doublet identification or vice versa (in Diseased but not Healthy). To clean this up, rename any groups with the term "doublet" in it to *doublet*. Rename groups with the term "singlet" in it to *singlet*. Rename the annotation as *BothFiles_Doublet*.

Filtering out potential doublets and genes with inconsistent UMI status

1. In the Data table panel, click **Master Data Table**, and then click **BothFiles_Doublet** in the Annotation list box.
2. Click **Filter data table**, and select filtering by **Cells** and **Genes**. The Cell filtering options window displays.
3. For cell filtering, select **Based on current annotation**, and then click **Filter cells**. Select a group of cells to remove. In this example, select **doublet**.
4. For gene filtering, select **MI adjustment status**, and then select all genes with insufficient depth. Name the new data table **QCed**. In this example, **Select all**.

Randomly subsampling an equal number of cells from each sample

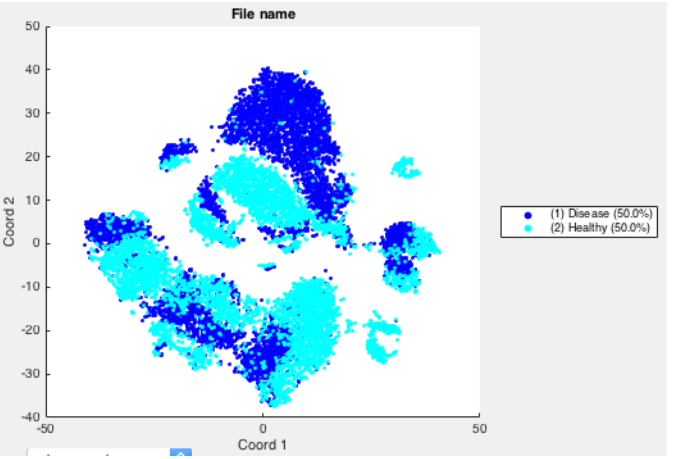
Randomly subsample an equal number of cells from each sample to ensure that the difference in cell number between samples has no affect on differential gene expression and cell type proportion analysis.

1. In the Data table list box, click **QCed**.
2. In the Annotation list box, click **File name**.
3. In the Data table panel, click **Filter data table**, filter data table by **Cells**, and then select **Random subsampling** as the cell filtering method.
4. Select **Groups in highlighted annotation**.

5. In this example, enter **7000** for Disease and Healthy files. BD Data View will randomly subsample 7,000 cells from each file.
6. Rename the data set **QCed_random7000**.

Generating projections of high-dimensional data

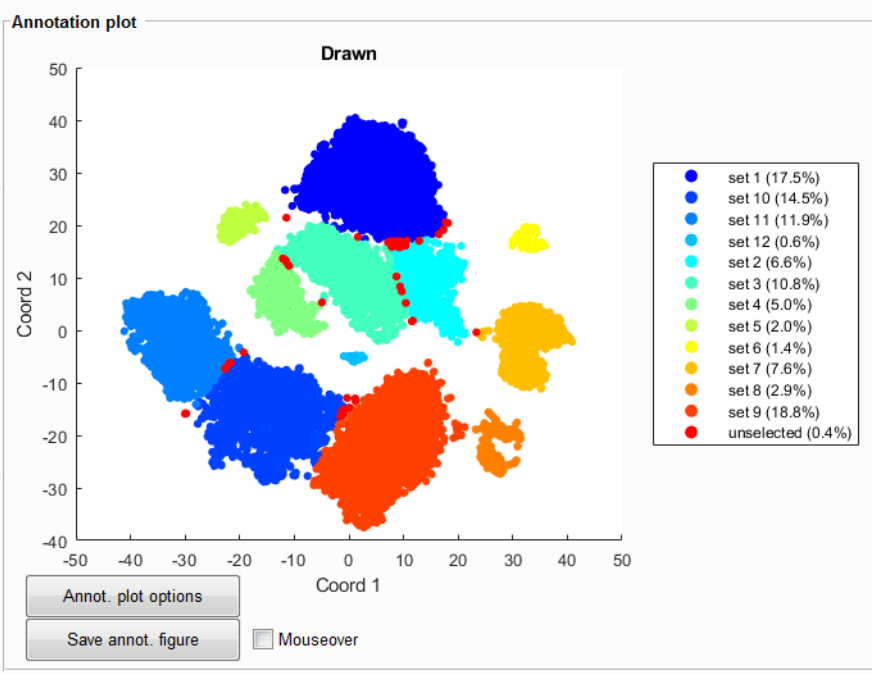
1. In the Data table panel, select **Non-normalized**.
2. In the Data table list box, select **QCed_random7000**.
3. In the Projection panel, select **tSNE**, select **fast tsne** as the implementation, and then click **Calculate**. In this example, the projected QCed_random7000 cell labels:



Defining clusters manually

1. In the Annotation panel, click **Add Annotation**, and then select **Draw with mouse**. Click **OK**.
2. In this example, enter 12 clusters.

3. Click the mouse to draw points around the defined cluster. For example:

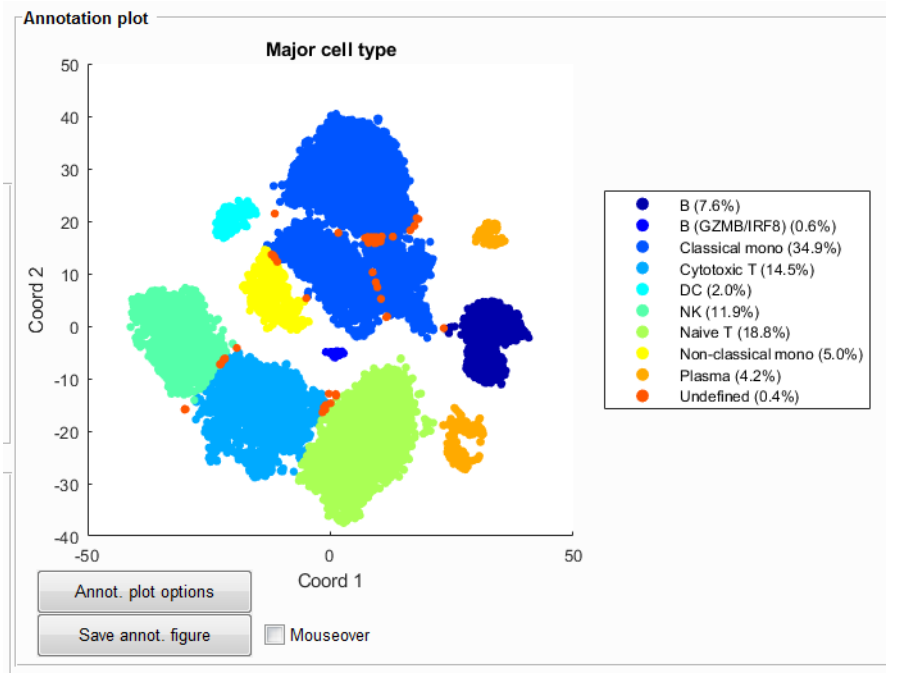


4. Select the file in the Annotation list box and click **Edit annotation** and **Rename Annotations** to rename annotation as **Drawn**.

Renaming the clusters

1. Click **Edit annotation**, and select **Rename groups**.
2. Rename the clusters to major immune cell types by inspecting for marker genes. In the Rename clusters window, select the group name, enter a new name in the text box, and then click **Update cluster name**.
3. Enter a name for the renamed clusters of annotations. In this example, **Major cell type**.

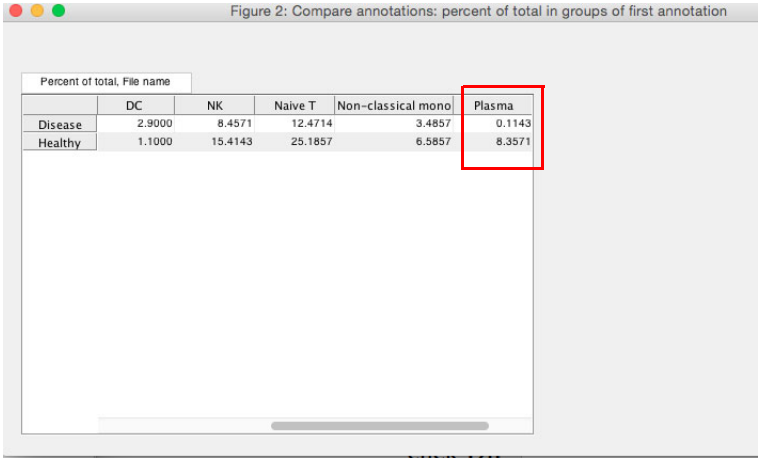
4. Click **Export annotation** to save the renamed clusters. In this example, the exported annotation is displayed this way:



Comparing cell type proportions between samples

In the Annotation panel, click **Compare annotations**, and then select **File Name** and **Major cell type** in the Compare annotations window. The number of cells and percent of total of each cell type in each file is displayed in three windows. You can save the output as a .csv file, and it can be used for plotting.

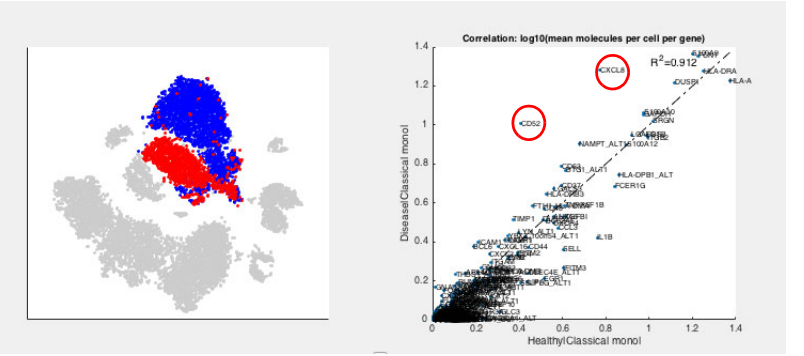
In this example, we observe that Healthy has a much higher proportion of plasma cells than in Disease:



Finding gene expression differences between healthy and diseased samples

1. In the Annotation panel, click **Add annotations**, and then select **Combination of existing annotations**. The Combine annotations window is displayed.
2. Select **File name** and **Major cell type**, and then click **OK**. In this example, groups such as Healthy NK cell and Disease T cell are created.
3. Rename the new annotation *CellType_FileName*.
4. In the Annotation panel, select the new annotation, and then click **Diff. expression**. The Differential expression analysis window is displayed. If there are more than 1,500 genes, a warning displays to save the session and option to proceed.
5. Select two groups to compare.

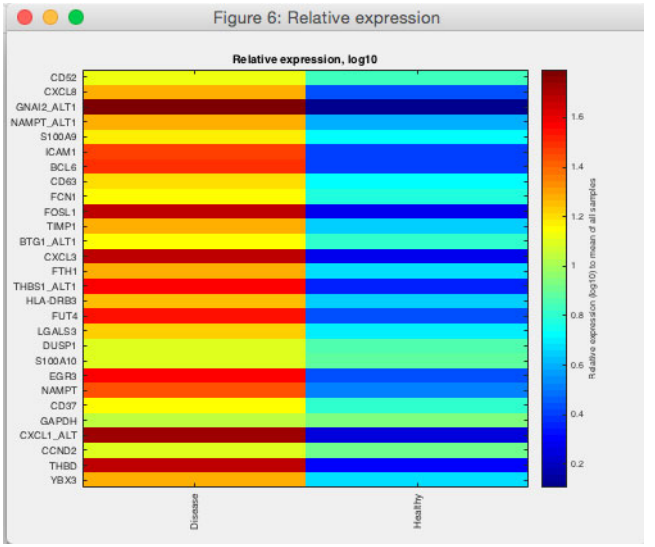
   In this example, on the t-SNE projection, we observed that the classical monocytes of the healthy and disease files occupy distinct space. In order to understand the genes that drive the distinct clustering, we select **Healthy|Classical mono** and **Disease|Classical mono** for comparison. The genes that are highly expressed in Disease|Classical mono include CD52 and CXCL8:



6. Click **Save Results: selected cluster x vs. selected cluster y** to export the list of genes in .csv format that are differentially expressed between the two samples.

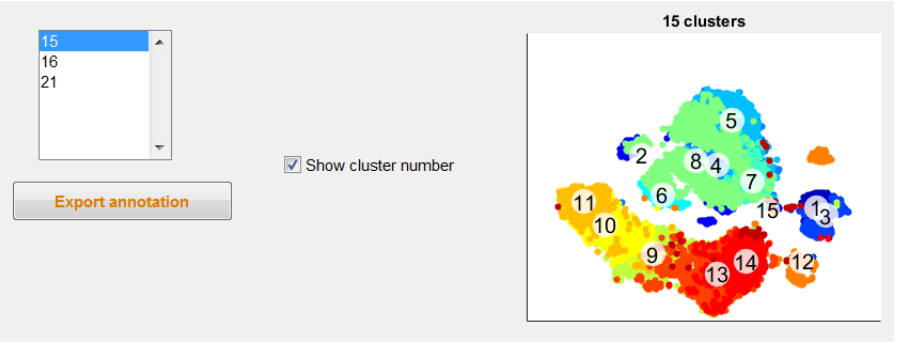
Displaying differential expression in a heatmap

1. In the Multi-gene expression panel, click **Custom set**, and then navigate to the sub-folder **/DiffExp** (**and in this example to QCed_random7000**) to load the **selected cluster x vs. selected cluster y** file (HealthyClassicalmono_vs_DiseaseClassicalmono.csv). The differential expression files are listed in the Multi-gene expression panel. In this example, the two sets of genes are DiseaseClassicalmono_over_HealthyClassicalmono, and HealthyClassicalmono_over_DiseaseClassicalmono.
2. In the Annotation panel list box, select **File Name** and in the Multi-gene expression list box, in this example, select **DiseaseClassicalmono_overHealthyClassicalmono**.
3. In the Multi-gene expression panel, click **Multi-gene plot options > Heatmap (group)**. The Select gene set(s) for heatmap (group) window is displayed.
4. Select **Use current highlighted gene set**.
5. Enter the heatmap name, **Relative expression**. The relative expression of the two gene sets are displayed. In this example:



Exploring cell subsets between samples with hierarchical clustering

1. In the Annotation panel, click **Add annotation**. The window Select method to annotate is displayed.
2. Click **Hierarchical clustering**, and then click **OK**. The Hierarchical clustering window is displayed.
3. Select the **recursive dendrogram split** algorithm. BD Data View runs the algorithm. In this example:



4. Click **Export annotation**, enter in this example, Hierarchical-recursive.
5. Continue to explore genes that drive differences between clusters by repeating the analysis with the differential analysis function in the same window.

Designing a targeted panel from whole transcriptome amplification (WTA) RNA-seq data

Introduction

We analyze whole transcriptome data obtained from the 10x Chromium™ System to design a targeted panel that can be used on the BD Rhapsody Single-Cell Analysis system.

Before you begin

Obtain the sequencing output files:

- barcodes.tsv
- genes.tsv
- matrix.mtx

All three files must be in the same folder.

Workflow steps

Step	Purpose
1	Load data.
2	Shorten gene names.
3	Select the most variable genes.
4	Normalize the data to cell total.
5	Generate projections of reduced data.
6	Perform hierarchical clustering of the cell labels.
7	Find representative genes for each cluster.
8	Find sets of correlating genes.
9	Rename the clusters.
10	Compile a list of genes of interest.
11	Verify that the chosen list of genes produces the desired clustering.

Loading data

Load matrix.mtx. Ensure that you search for the files according to the selected format.

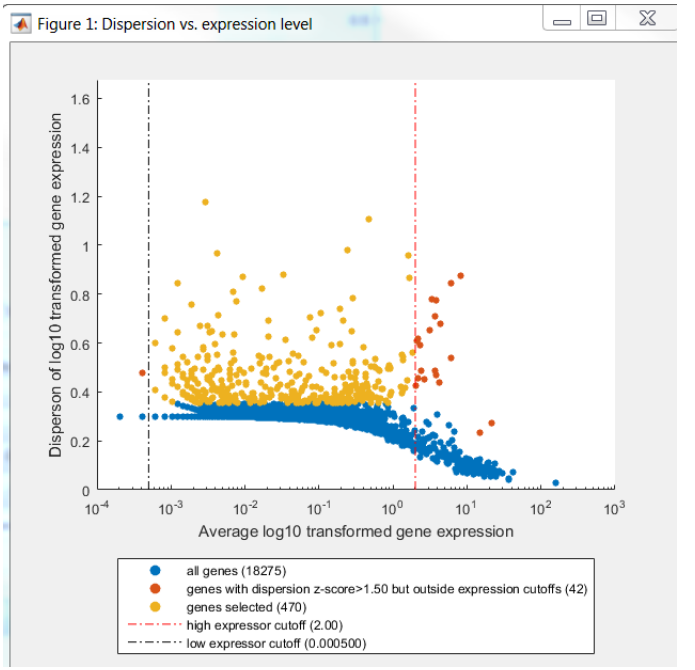
For detailed instructions on loading data, see Loading data (page 140).

Shortening gene names

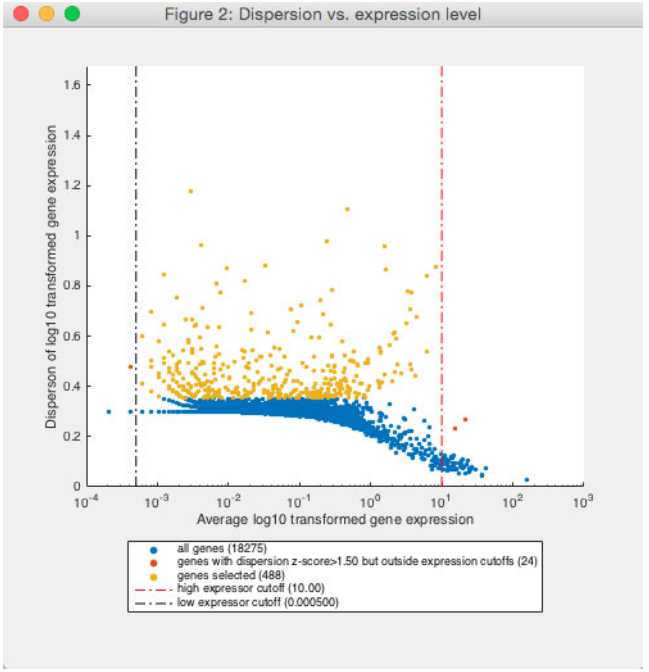
1. In the Data table panel, click **Shorten gene names**. The Select delimited window is displayed.
2. Select **tab** as the delimiter, and then click **Select**. In this example, select to keep **TSPAN6**.

Selecting the most variable genes

1. In the Data table panel, click **Select variable genes**.
2. Select the most variable genes in the data set. To start, select the default parameters to reduce the data set from >10,000 genes to a few hundred. A new data set, *Feature Selection*, is displayed in the Data table list box. The Dispersion vs. expression level window is displayed. For example:



3. (Optional) To capture higher expressor genes, click Select variable genes again, but set the high expression threshold to 10. A new Dispersion vs. expression level window is displayed with more variable genes:



Select variable genes reduced the total genes from 18,275 to 488.

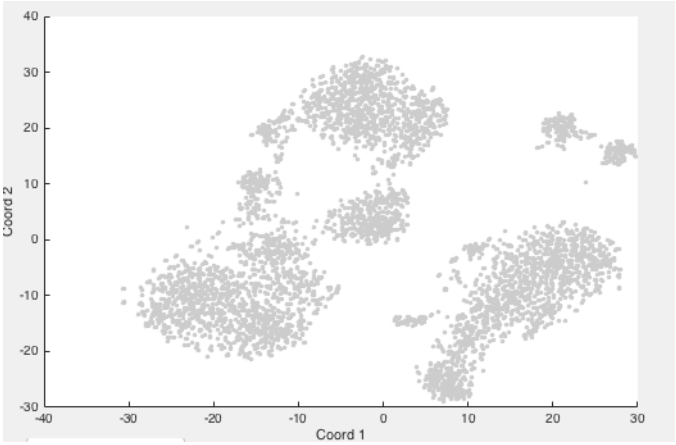
For information on parameters, see Select variable genes (page 97).

Normalizing the data to cell total

In the Data table panel, click the **Normalize to cell total** to normalize to the Master Data Table. Normalization reduces the effect of total molecules per cell on the t-SNE projection. The data is normalized by dividing the count of each gene from each cell to the cell total multiplied by the median of the cell total from all cells.

Generating projections of reduced data

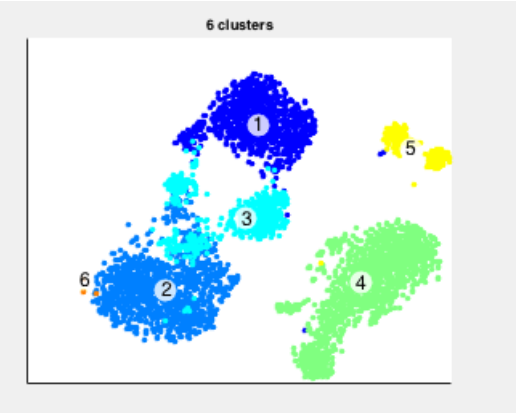
1. In the Data table panel, select **Normalized**
2. In the Data table list box, select **Feature Selection**.
3. Select **tSNE** in the Projection panel, select **fast tsne** as the implementation, and then click **Calculate**. In this example, the cell ID projections look like this:



Performing hierarchical clustering of the cell labels

1. Select **Normalized** in the Data table panel and **Feature Selection** in the Data table list box.
2. In the Annotation panel, click **Add annotation**, and then select **Hierarchical clustering**.
3. Select **Recursive dendrogram split**, an unsupervised method for clustering. The recursive dendrogram split algorithm outputs four sets of cluster assignments: 6, 9, 12, and 14 clusters. In this example, six clusters were chosen:

   Recursive dendrogram split



4. Export the cluster assignment annotation.

   For detailed instructions on performing hierarchical clustering on cell labels, see Exploring cell subsets between samples with hierarchical clustering (page 176).

Finding representative genes for each cluster

1. Select a hierarchical cluster annotation in the Annotation list box.
2. In the Annotation panel, click **Diff. expression**. The Differential expression analysis window is displayed.
3. Start with a default p-value of 1^{-10}.
4. Click **Save Results: every cluster x vs. rest** to export the list of genes in .csv format that are differentially expressed between each cluster versus the rest of the clusters.
5. If necessary, lower the p-value threshold to include more genes. Lists of genes are saved as .csv file in *FeatureSelection/DiffExp*.

Finding sets of correlating genes

This analysis is especially useful to analyze rare cells that do not form obvious clusters on a tSNE projection. The result is groups of genes that are co-expressed by a group of cells.

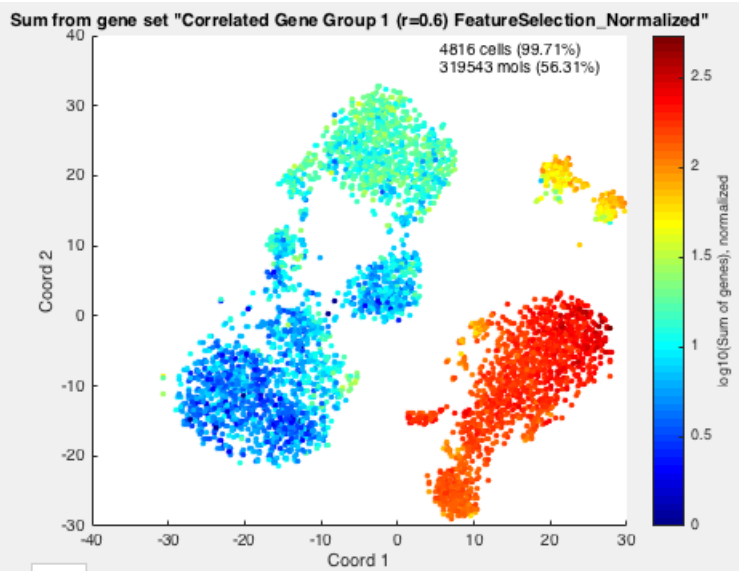
1. Select **Feature Selection** data table in the Data table list box.
2. In the *Multi-gene* expression panel, click **Correlated genes**.
3. Use the default setting of 0.6 for the correlation coefficient threshold. Lower the threshold to include more genes. The correlated gene list is exported as a .csv file to *FeatureSelection/Correlation*, and the file is also listed in the gene set list in the Multi-gene expression panel.
4. Select **Highlighted gene set** from the drop-down menu in the Single gene expression panel.
5. Click each correlated gene group in the Multi-gene expression list box to view correlated genes.

   By clicking a correlated gene group, the gene expression plot is updated to display the total expression of the set of genes per cell.
6. In the Multi-gene expression panel, click **Show genes** to view each group of selected correlated genes.

Note: To understand the biological significance of a group of correlated genes, consider using gene ontology analysis. Copy a list of genes for searching on a gene ontology website such as geneontology.org.

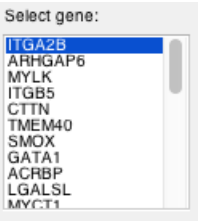
Example

Choose correlated group 1, which shows the genes co-expressed as the combination of genes is heavily detected in one cluster:

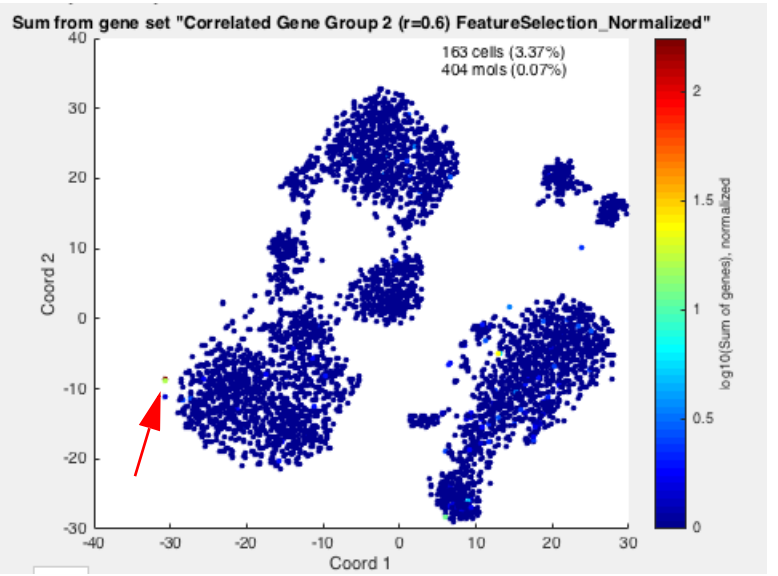


Closer inspection reveals that these genes are specific to monocytes.

Choose correlated group 2:

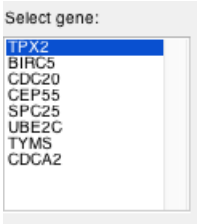


The gene expression plot shows very few cells with expression of this combination of genes:

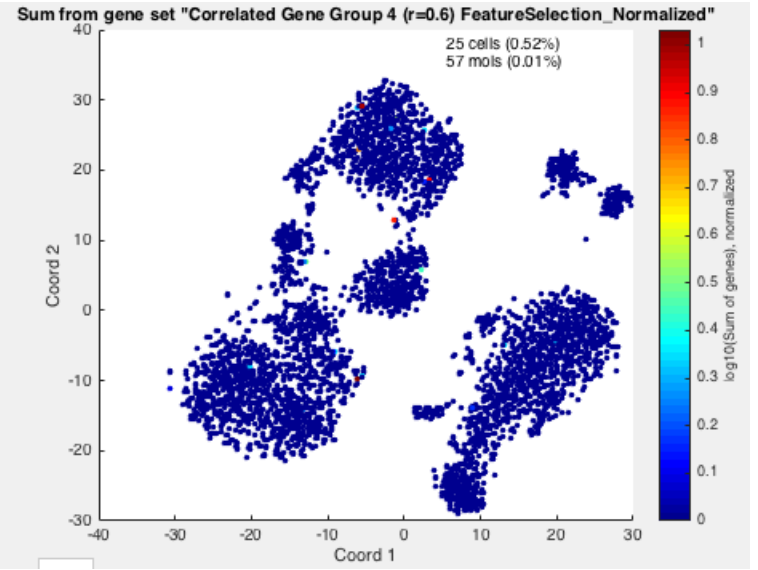


Gene ontology indicates basophil differentiation and platelet formation, suggesting that these cells are likely some type of granulocyte.

Choose correlated group 4:



Co-expressed genes are less obvious on the gene expression plot, because very few cells co-express them, and the cells that do express them also do not belong to one cluster:



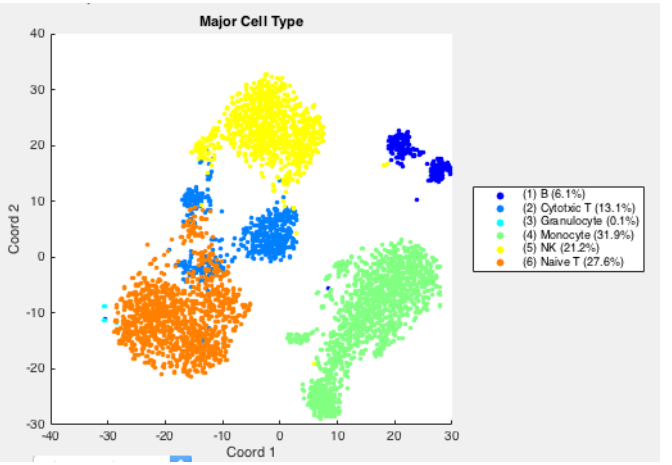
Gene onotology indicates mitotoic spindle assembly, cell division, chromosome segregation, suggesting this group of genes is related to cell cycle, and the cells expressing this combination are undergoing mitosis.

7. (Optional) Using the results from differential analysis, rename the cluster assignment with the cell type name. Click **Edit annotation > Rename groups** to rename clusters.

Renaming the clusters

Using the results from differential analysis and gene correlation, rename the cluster assignment with the cell type name.

1. Click **Edit annotation**, and select **Rename groups**.
2. Rename the clusters to major immune cell types by inspecting for marker genes. In the Rename clusters window, select the group name, enter a new name in the text box, and then click **Update cluster name**.
3. Enter a name for the renamed clusters of annotations. In this example, **Major cell type**.
4. Click **Export annotation** to save the renamed clusters. In this example, the exported annotation is displayed:



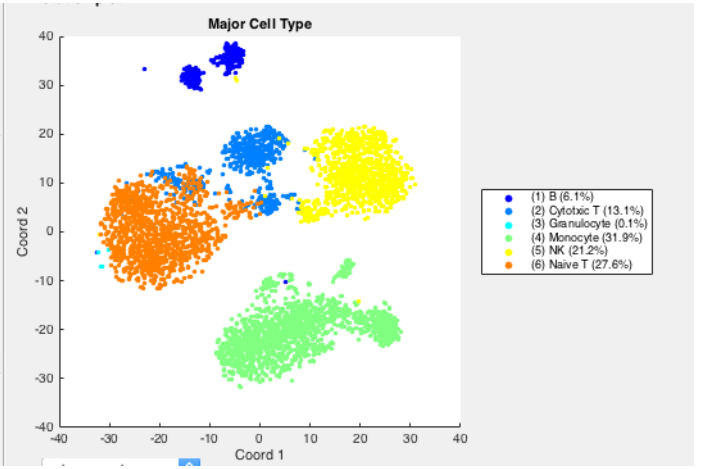
Compiling a list of genes of interest for a targeted gene panel

1. Obtain the differentially expressed gene list (see Finding representative genes for each cluster (page 184) and the correlated gene list (see Finding sets of correlating genes (page 184).
2. Use Microsoft® Excel® to explore the list of representative genes of each cell cluster and the list of correlated genes.
3. Create a list of genes that represents each major cell type in the sample and their association with biological pathways.
4. Combine the list with additional genes from other sources such as the scientific literature and verified data. In this example, we combined all of the representative genes from each cluster as well as correlated gene groups 2 and 4. This became a panel of 272 genes. A fully custom panel, which can be ordered from BD Biosciences, can have up to 500 targets.

Verifying that the chosen list of genes produces the desired clustering

1. Create a list of genes (one row per gene) as a .csv file.
2. Select **Master Data Table** in the Data table list box.
3. Filter the original data with the list of curated genes with **Filter data table**, and then click **Gene filtering**.
4. Load the .csv file, keep the genes, and name the set *Candidate gene list*.
5. Select **Candidate gene list** in the Data table list box, and ensure that it is normalized.
6. Regenerate the t-SNE projection.

7. Verify that the major clusters that you want to identify are present. For example, when selecting **Label** by cell type:



Analyzing a multiplexed sample with BD Data View

Introduction

In this example, we load Sample Tag and sample name annotations to view the Sample Tag calls for a specific multiplexed sample. The sample contains peripheral blood mononuclear cells (PBMCs) from two different donors. Using a BD™ Single-Cell Multiplexing Kit, the cells of each donor are labelled with a unique Sample Tag and then pooled for cell capture, library preparation, and sequencing. Use BD Data View to analyze the pipeline results of the pooled sample and individual PBMC-labelled samples.

Before you begin

1. Install BD Data View and run the BD Rhapsody Analysis pipeline.
2. From the analysis pipeline, obtain the output files:
   - Combined_<sample_multiplex_name>_DBEC_MolsPerCell.csv
   - <sample_name>_Sample_Tag_Calls.csv
   - <sample_multiplex_name>_Metrics_Summary.csv

Workflow steps

Step	Purpose
1	Load the data table.
2	Load the Sample Tag and sample name data.
3	Generate a projection of high-dimensional data.
4	View the Sample Tag and sample name labels.
5	Analyze multiple samples with BD Data View.

Loading the data table

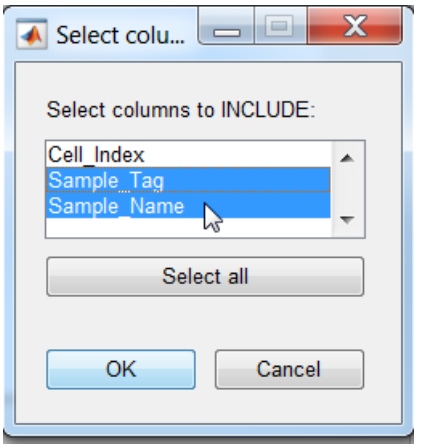
Load Combined_<sample_multiplex_name>_DBEC_MolsPerCell.csv. Ensure that you search for the files according to the selected format.

For detailed instructions on loading data, see Loading data (page 140).

Loading the Sample Tag and sample name data

1. In the Annotation panel, click **Add annotation.**
2. Click **Load from file**, and then click **OK**.
3. Browse for <sample_name>_Sample_Tag_Calls.csv in the sequencing analysis output. Click **Open**.
4. Select the row from the file with the Sample Tag names, and then click **SELECT**.

5. Press **Ctrl+click** to select the columns that include the Sample Tag and sample names that match the specific Sample Tags, as shown here:

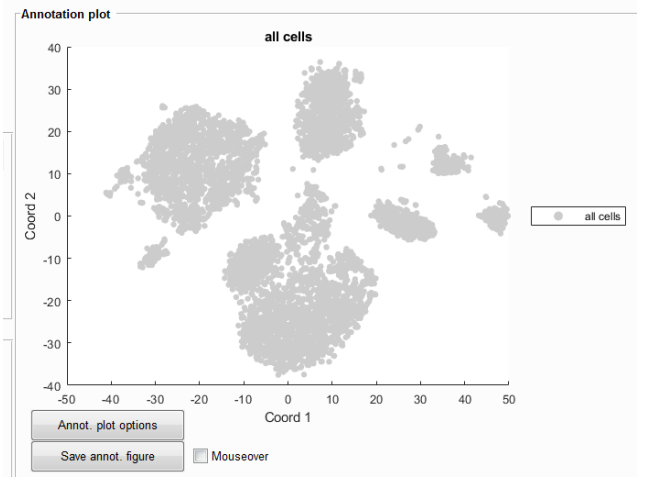


The software lists the columns in the Annotation list box.

Generating a projection of high-dimensional data

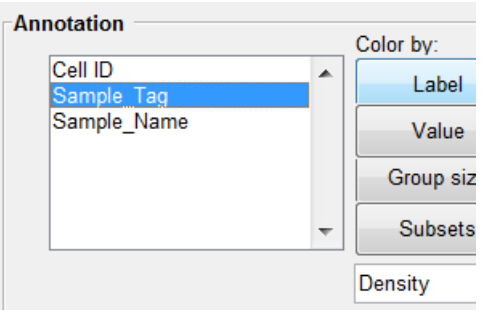
You can generate the t-SNE coordinates by selecting **Master Data Table**, **tSNE**, and then clicking **Calculate** in the Projection panel. Use the default settings. The coordinates are saved with the session.

The data points are displayed. If they are not, click **Cell_ID** and then **Label** in the Annotation panel. For example:

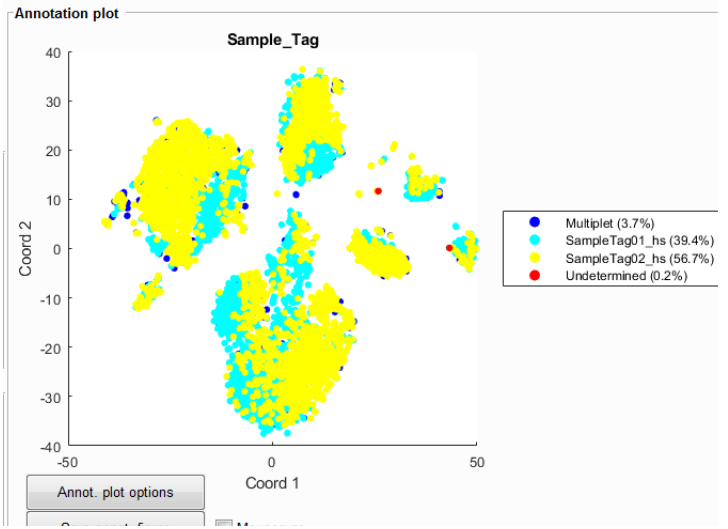


Viewing the Sample Tag and sample name labels

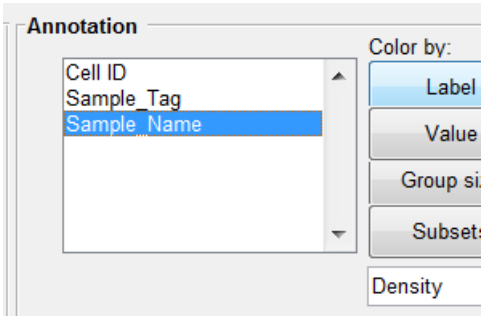
1. In the Annotation panel, click **Sample_Tag,** and then click **Label** to refresh the plot and see the annotations by Sample Tag:



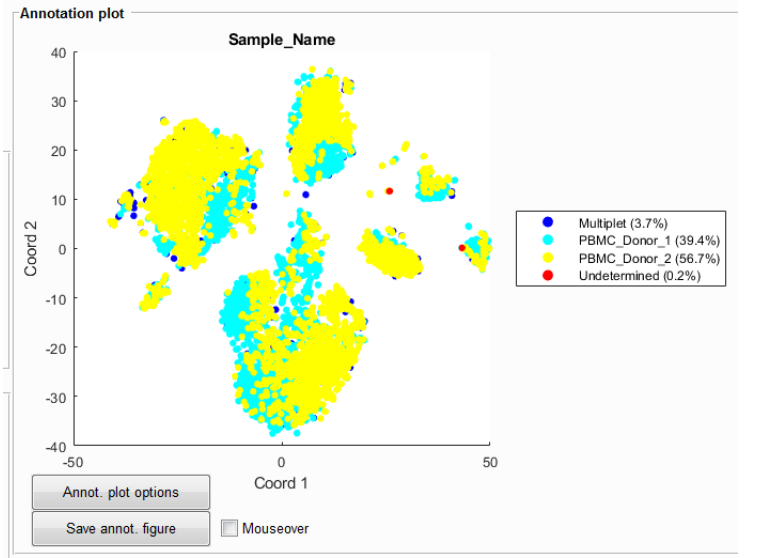
For example:



2. In the Annotation panel, click **Sample_Name,** and then click **Label** to refresh the plot and see the annotations by sample name:



For example:

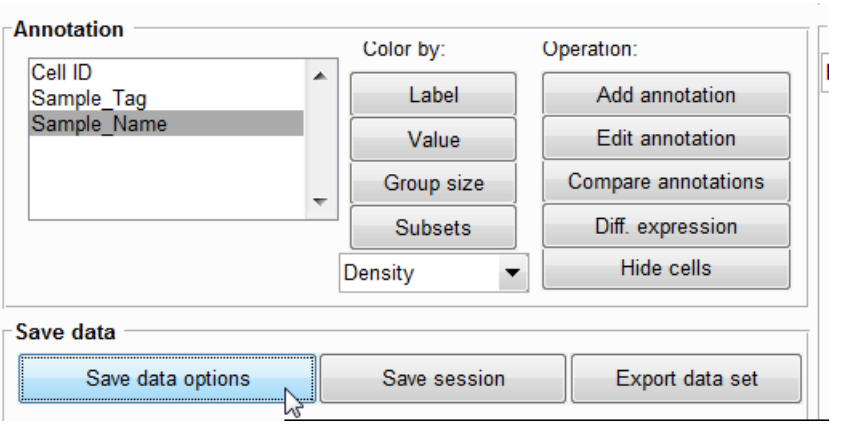


3. Proceed to Analyzing multiple samples with BD Data View (page 161).

Managing sessions

Saving a session

1. In the Save data panel, click **Save data options**:



2. Select a figure format (PDF, SVG, PNG) from the drop-down menu.
3. Click **Select directory** to save output, and then browse to a location to save the session. Click **Select folder**. and close the dialog box.

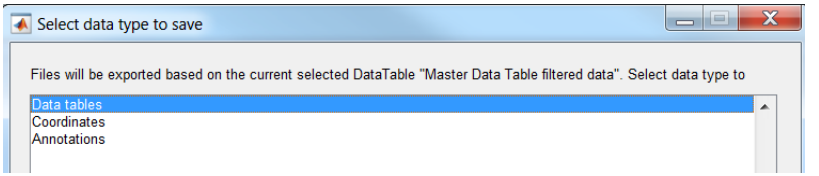
   Note: The default directory for saving the session is the same directory where the data files are stored.
4. Click **Save session**. Rename the session, if desired, and then click **OK**.

Exporting data files in .csv format

You can export the filtered data table where genes with insufficient sequencing depth have been removed.

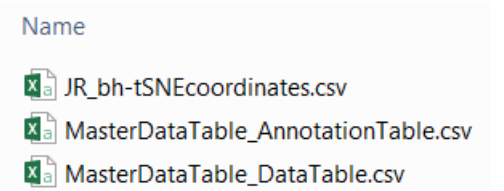
Use alphanumeric characters and the underscore in saved filenames. Do not use special characters.

1. In the Data table panel, select the data table to be saved. In this example, click **Master Data Table filtered data**.
2. In the Save data panel, click **Export data set**. The Select data type to save window is displayed.
3. Select **Data tables**:



4. If you select **Select all**, you can select to save:
   - Files
   - Types of data tables
   - All genes or non-zero genes
   - Sets of coordinates
   - Annotations

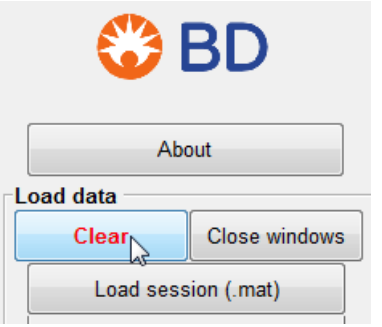
   The data are exported to the selected directory. For example:



Clearing a session

Clearing a session clears all displayed data.

1. In the Load data panel, click **Clear** to clear the data from the session and begin a new session:

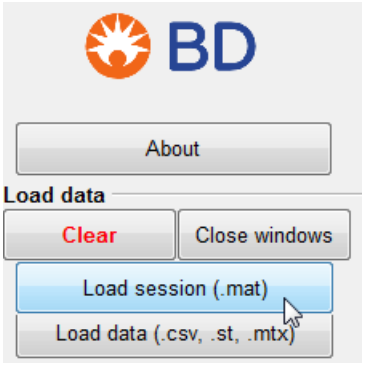


   The Clearing data window displays.

2. Click **Yes** to clear the session.

Loading a session

1. In the Load data panel, click **Load session**:



2. Browse, and select a .mat file, and then click **Open**. The session opens in the same state in which it was saved.

Managing errors encountered with BD Data View

Introduction

In BD Data View, most errors are displayed as warning messages in pop-up windows or within the application. If you encounter an error, and no warning messages display, copy the last error messages listed on the terminal window, and then send them to BD Biosciences technical support at researchapplications@bd.com.

Plots and text do not display properly

Possible causes	Recommended solutions
Conflict between high- and low-resolution monitor settings	1. Save the session. 2. Restart BD Data View with the current monitor resolution setting.

Glossary

B

BAM An alignment file in binary format. A binary SAM file.

C

CIGAR Compact Idiosyncratic Gapped Alignment Report. A sequence of base lengths to indicate base alignments, insertions, and deletions with respect to the reference sequence. See samtools.github.io/hts-specs/SAMv1.pdf.

CLS Cell label sequence.

D

DBEC Distribution-based error correction.

F

FASTA Text-based format that contains one or more DNA or RNA sequences.

FASTQ A file in standardized, text-based format that contains the output of read bases and per-base quality values from a sequencer.

L

L Common sequence.

M

molecule A unique combination of a cell label, UMI sequence, and a gene. Without UMI adjustment methods, it is called *raw molecule*. With RSEC UMI adjustment, it is called *RSEC-adjusted molecule*. With additional DBEC UMI adjustment, it is called *DBEC-adjusted molecule*.

P

PhiX Control library used for sequencing runs.

R

R1 reads Contains information about the cell label and UMI.

R2 reads Contains information about the gene.

RSEC Recursive substitution error correction.

S

SAM Tab-delimited text file with sequence alignment data.

singlet A putative cell where more that 75% of sample tag reads are from a single tag.

singleton Clustering: Cell not assigned to any of the clusters. UMI correction/adjustment: Molecule that is represented by only one read.

U

UMI Unique Molecular Identifier. A string of eight randomers immediately downstream of the cell label sequence (CLS) 3 of the R1 read that is used to uniquely label a molecule.