Small RNA v1.0
BaseSpace App Guide

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

illumına®

Introduction

The BaseSpace® App, Small RNA v1.0, analyzes small RNA samples. The app aligns reads against 4 reference databases (abundant, mature miRNA, other RNA, and genomic) and outputs hits to mature miRNAs, isomiRs, and piRNAs. Optionally, the app performs novel precursor discovery and pairwise differential expression analysis. Pairwise differential expression analysis identifies differentially expressed miRNAs, precursor groups, miRNA families, and piRNAs for each pair of sample groups.

Compatible Libraries

See the BaseSpace support page for a list of library types that are compatible with the Small RNA v1.0 App.

Workflow Requirements

▸ A maximum of 100 samples is possible per analysis.
▸ The maximum size is 200 gigabases per sequence.
▸ The maximum size is 25 gigabases per sample.

Versions

The following components are used in the Small RNA v1.0 App.

Software	Version
Isis (Analysis Software)	2.5.52.11
SAMtools	0.1.19-isis-1.0.2
Bowtie (Aligner)	0.12.8
MiRDeep*	3.2
DESeq2	1.0.17

Workflow Diagram

Figure 1   Small RNA v1.0 App Workflow
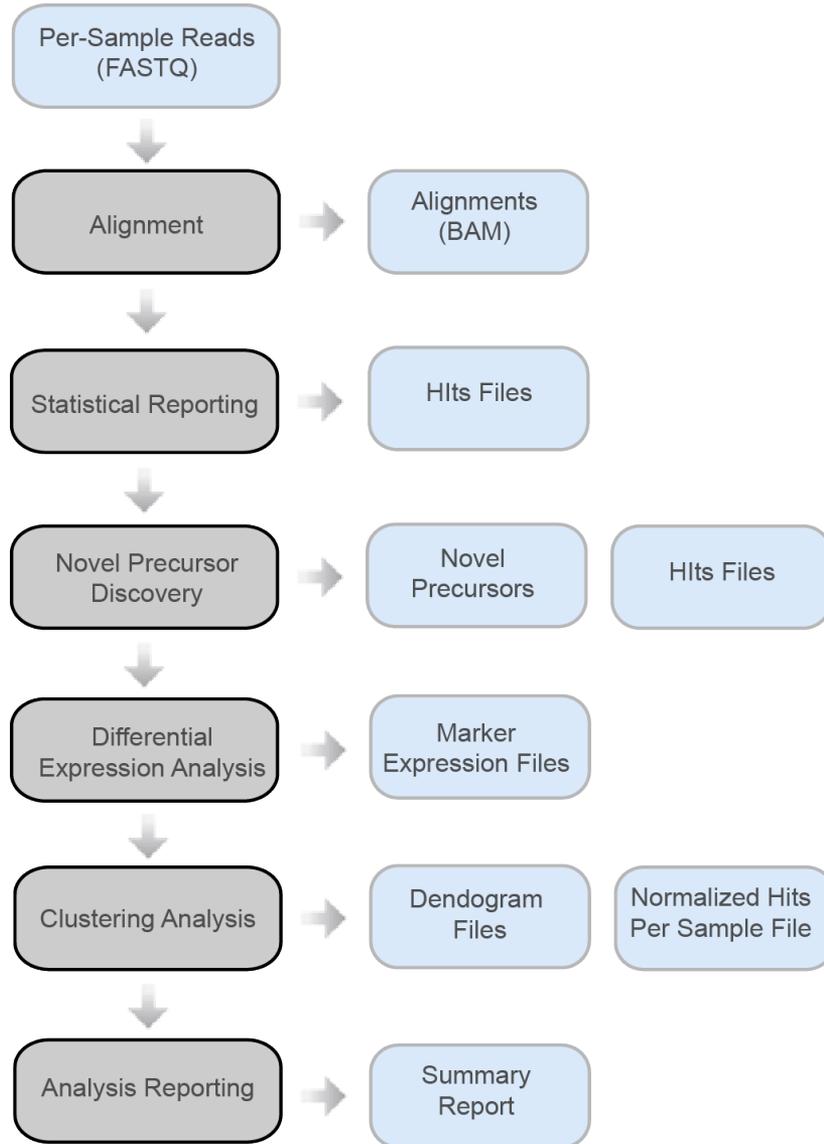
Set Analysis Parameters

1 In BaseSpace, click the **Apps** tab.

2 Click Small RNA.

3 From the drop-down list, select **version 2.0**, and then click **Launch** to open the app.

4 In the **Analysis Name** field on the app input form, enter the analysis name.
By default, the analysis name includes the app name, followed by the date and time that the analysis session starts.

5 From the **Save Results To** field, select the project that stores the app results.

6 [Optional] Select from the following **Options** checkboxes:
  ▸ **Enable Novel Precursor Discovery**—Enable novel precursor prediction using miRDeep*.
  ▸ **Enable Pairwise Differential Expression Analysis**—Enable pairwise differential expression analysis using DESeq2.

7 From the **Species** field, select the species of your samples.

8 From the **MiRBase Version** field, select the microRNA database version.

9 From the **Samples** field, select the samples to analyze. Select the samples as follows.

  a Enter a **Label** for the sample group.
  b Click **Select Sample(s)** to open the Select Sample(s) window.
  c [Optional] Search for the samples you want to analyze.
  d Select the checkbox next to the appropriate samples.
  e Click **Confirm**.

  To set up multiple groups for differential expression analysis and novel precursor discovery, select the samples for the first group, and then click + to add a group. Add a label for each group.
  When pairwise differential expression analysis is enabled, at least 2 groups are required.
  When novel precursor discovery is enabled, novel precursor discovery is performed for each group.

10 If you selected **Enable Pairwise Differential Expression Analysis** in the **Options** drop-down list, click the **Select Pairs of Groups for Differential Expression Analysis** checkbox to select pairs of groups. By default, pairwise differential analysis is performed for all pairs. Select the pairs as follows.

  a In the **Control Group** and the **Comparison Group** fields, enter the label that you assigned to each group in the Samples section.
  b Click + to add more pairs.

11 Click **Continue**.

When analysis is complete, the status of the app session is updated automatically and an email is sent to notify you.

Analysis Methods

The Small RNA v1.0 workflow aligns reads against 4 reference databases (abundant, mature miRNA, other RNA, and genomic) and outputs hits to mature miRNAs, isomiRs, and piRNAs. Optionally, the app performs novel precursor discovery and pairwise differential expression analysis. Pairwise differential expression analysis identifies differentially expressed miRNAs, precursor groups, miRNA families, and piRNAs for each pair of sample groups.

Alignment

The Small RNA v1.0 workflow uses Bowtie v0.12.8 to align each cluster against abundant, mature miRNA, other RNA, and genomic reference databases[1].

Statistical Reporting

Reads are assigned to 1 of the 4 databases in order of precedence.

Figure 2 Read Assignment Order of Precedence
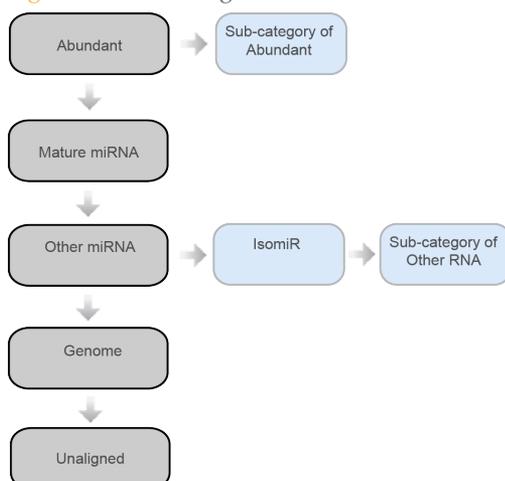
For example, a read that matches the abundant database is counted as abundant, regardless of other matches. Reads that match the mature miRNA database, but not the abundant database, are counted as miRNA.

- **Mature miRNA Hits**—Only exact matches to mature miRNA records are counted. A read must align to the start of a reference sequence on the same strand and have the same length of the reference sequence. No mismatches are allowed.
- **IsomiR Hits**—A read is counted as isomiR if it is a subsequence on the same strand of a precursor sequence, but is not the 5' or 3' mature miRNA sequence. The precursor sequences are obtained from the miRNA database (miRBase) and are part of the other RNA database.
  - An isomiR ID is in the form of "{start position}_{sequence read}_{precursor ID}". The start position is 0-based and relative to the precursor. The precursor ID is the miRBase precursor name.
  - Mismatches are allowed. IsomiR hits are further filtered to remove artifacts of PCR and sequencing errors.
- **piRNA Hits**—A read is counted as piRNA if it aligns to a piRNA sequence, but has not been counted as isomiR. The piRNA sequences are obtained from the NCBI Nucleotide database. For more information, see www.ncbi.nlm.nih.gov/nuccore.

6

▸ **miRNA Family Hits**—A miRNA family consists of miRNA precursors belonging to the same family. The miRNA family information is obtained from miRBase. A read contributes to the count of a miRNA family if it aligns to 1 of the precursors in the family.

▸ **Precursor Group Hits**—A precursor belongs to a group if it produces a mature miRNA produced by another precursor in the group.

Novel Precursor Discovery

Novel precursor discovery is performed for each sample group. Reads that are assigned to the other RNA and genome categories, but not the precursor subcategory, are pooled across samples in the group. Genomic locations for these reads are available through Bowtie alignment. The command-line version of miRDeep* 3.2 is used to predict novel precursor based on these aligned reads[2].

The novel precursor ID format is "{sample group}_{reference sequence ID}_novelMiR_{nonnegative integer}".

Differential Expression Analysis

Differential expression analysis is performed for miRNA families, precursor groups, miRNAs, and piRNAs. If novel precursor discovery is enabled, the precursor groups include both known and novel precursor groups.

MiRNAs include mature miRNAs and isomiRs (know precursor). If novel miRNA prediction is enabled, miRNAs also include novel mature miRNAs and isomiRs (novel precursor).

DESeq2 filters out low expressed markers with normalized mean count < 10 before testing[3]. The DESeq2 variance model is used to detect and exclude outliers based on extreme variation between replicates. The status (filtered or passed) and the result of the analysis (mean expression, fold change, standard error, p-value, q-value, etc.) is reported for each marker.

If novel precursor discovery is enabled, the novel precursors predicted for the 2 sample groups in each pairwise differential expression analysis are merged. The merged novel precursor ID format is "{control group}_vs_{comparison group}-novelMiR-{nonnegative integer}:{ID1}[+ID2[+ID3...]]", where ID1, ID2 and ID3 are the IDs of the merged member novel precursors. Because the first part of the ID uniquely identifies a merged novel precursor, the member novel precursor IDs are omitted in reports.

Similarly, novel precursors are merged across all the sample groups to generate global plots, such as the expression heat map. The merged novel precursor ID format is "global-novelMiR-{nonnegative integer}:{ID1}[+ID2[+ID3...]]". The member precursor IDs are also omitted in reports.

References

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25.

An J, Lai J, Lehman ML, Nelson CC. (2013) miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res 41(2):727-37.

Love MI, Huber W, Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550.

Analysis Output

To view the results, click the **Projects** tab, then the project name, and then the analysis.

Small RNA v1.0 Output Navigation Bar
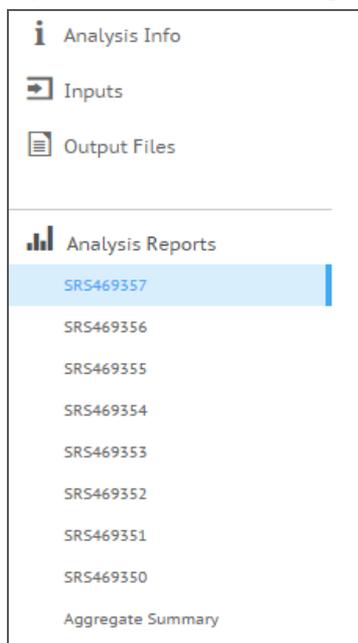
After analysis is complete, access the output through the left navigation bar.

▸ **Analysis Info**—Information about the analysis session, including log files.

▸ **Inputs**—Lists the samples and settings specified for the analysis session.

▸ **Output Files**—Output files for the sample.

▸ **Aggregate Summary Analysis Report**—Analysis metrics for the aggregate results, displayed when multiple samples are analyzed.

▸ **Sample Analysis Reports**—Analysis reports for each sample.

Analysis Info

The Analysis Info page displays the analysis settings and execution details.

Row Heading	Definition
Name	Name of the analysis session.
Application	App that generated this analysis.
Date Started	Date and time the analysis session started.
Date Completed	Date and time the analysis session completed.
Duration	Duration of the analysis.
Session Type	Number of nodes used.
Status	Status of the analysis session. The status shows either Running or Complete.

Log Files

File Name	Description
CompletedJobInfo.xml	Contains information about the completed analysis session.
Logging.zip	Contains all detailed log files for each step of the workflow.
SampleSheet.csv	Sample sheet.
SampleSheetUsed.csv	A copy of the sample sheet, generated at the beginning of the workflow.
SmallRNARunStatistics.xml	Provides methylation-related information.
WorkflowError.txt	Workflow standard error output (contains error messages created when running the workflow).
WorkflowLog.txt	Workflow standard output (contains details about workflow steps, command line calls with parameters, timing, and progress).

Inputs

The Inputs page lists the samples and settings specified for the analysis session.

Output Files

The Output Files page provides access to the output files for each sample analysis.

Sample Output Files

The following files are present in each sample folder.

File Name	Description
{sample group}_S{sample index}.abundant.bam	Alignments against abundant sequences. One alignment per read.
{sample group}_S{sample index}.abundant.bam.bai	BAM index file.
{sample group}_S{sample index}.mirna.bam	Alignments against mature miRNA sequences. One alignment per read.
{sample group}_S{sample index}.mirna.bam.bai	BAM index file.
{sample group}_S{sample index}.rna.bam	Alignments against other RNA sequences. One alignment per read.

File Name	Description
{sample group}_S{sample index}.rna.bam.bai	BAM index file.
{sample group}_S{sample index}.genome.bam	Alignments against the genome. One alignment per read.
{sample group}_S{sample index}.genome.bam.bai	BAM index file.
Hits.txt	Hits to mature miRNAs.
IsoHits.txt	Hits to isomiRs.
FilteredIsoHits.txt	Hits to filtered isomiRs.
PrecursorHits.txt	Hits to precursor groups.
FamilyHits.txt	Hits to miRNA families.
PiRNAHits.txt	Contains the number of reads counted for each piRNA listed, for each sample.
AllHits.txt	A verbose table of hits to mature miRNAs and other RNAs.
HitsByCategory.txt	Breakdown of hits by category.
AbundantHitsByCategory.txt	Breakdown of abundant hits by category.
MiRNAHitsByCategory.txt	Breakdown of miRNA hits by category.
RNAHitsByCategory.txt	Breakdown of other RNA hits by category.
ReadLengthDistribution.txt	Distribution of small RNA sequence length.
Novel—Available when novel precursor discovery is enabled.	
{sample group}_S{sample index}.bam	Alignments against the genome belonging to the sample used to predict novel precursors.
{sample group}_S{sample index}.bam.bai	BAM index file.
{sample group}_result	miRDeep* result file.
FilteredNovelIsoHits.txt	Hits to filtered isomiRs on novel precursors.
NovelHits.txt	Hits to novel mature miRNAs.
NovelIsoHits.txt	Hits to isomiRs on novel precursors.
NovelPrecursorHits.txt	Hits to novel precursor groups.
Report	
{sample group}_S{sample index}.report.pdf	PDF report of the sample analysis.

Aggregate Summary Output Files

An Aggregate Summary folder is available when more than 1 sample is analyzed. The TXT files in the top-level Aggregate Summary folder are similar to the files in the individual sample folders, but contain data for all samples.

File Name	Description
Novel—Available when novel precursor discovery is enabled.	
{sample group}_sorted_bams.zip	Alignments used to predict novel precursors for the sample group.
{sample group}.result	miRDeep* result file.
NovelHits_{sample group}.txt	Hits to novel mature miRNAs for samples in the group.
NovelIsoHits_{sample group}.txt	Hits to isomiRs on novel precursors for samples in the group.
FilteredNovelHits_{sample group}.txt	Hits to filtered isomiRs on novel precursors for samples in the group.
NovelPrecursorHits_{sample group}.txt	Hits to novel precursor groups for samples in the group.
Differential	
global.result	Merged miRDeep* results across all sample groups (available when novel precursor discovery is enabled).
{control group}_vs_{comparison group}.result	Merged miRDeep* results between the control and comparison groups (available when novel precursor discovery and pairwise differential expression analysis are enabled).
[Marker Type]	Family, precursor, miRNA, and piRNA.
Global	
counts.csv	Raw counts.
normalizedcounts.csv	Normalized counts for library size. If normalization fails, this file is not present.
deseq.corr.csv	Spearman sample correlations based on the raw counts.
deseq.corr.pdf	PDF file of the sample correlation plot.
deseq.corr.png	PNG file of the sample correlation plot.
rlog.csv	Regularized log-transformed normalized counts by DESeq2. If DESeq2 fails, this file is not present.

File Name	Description
pca.csv	Principal components based on the regularized log-transformed normalized counts. If rlog.csv is missing, this file is not present.
pca.pdf	PDF file of the scatter plot based on the first 2 principal components. If there are < 2 principal components, this file is not present.
pca.png	PNG file of the scatter plot based on the first 2 principal components. If there are < 2 principal components, this file is not present.
[control group]_vs_[comparison group]—Available when pairwise differential expression analysis is enabled.	
deseq.counts.csv	Raw counts.
deseq.res.csv	CSV file of the DESeq2 results. If DESeq2 fails, this file is not present.
deseq.disp.pdf	PDF file of the DESeq2 dispersion plot. If DESeq2 fails, this file is not present.
deseq.ma.pdf	PDF file of the DESeq2 MA plot. If DESeq2 fails, this file is not present.
deseq.scatter.pdf	PDF file of the scatter plot of genes in the control and comparison groups. The values are back-calculated based on the DESeq2 mean and fold change estimate. If DESeq2 fails, this file is not present.
Report	
aggregate.report.pdf	PDF report of the aggregate summary.

BAM File Format

A BAM file (*.bam) is the compressed binary version of a SAM file that is used to represent aligned sequences up to 128 Mb. SAM and BAM formats are described in detail at https://samtools.github.io/hts-specs/SAMv1.pdf.

BAM files use the file naming format of SampleName_S#.bam, where # is the sample number determined by the order that samples are listed for the run.

BAM files contain a header section and an alignments section:
‣ **Header**—Contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.
‣ **Alignments**—Contains read name, read sequence, read quality, alignment information, and custom tags. The read name includes the chromosome, start coordinate, alignment quality, and the match descriptor string.

The alignments section includes the following information for each or read pair:
‣ **RG:** Read group, which indicates the number of reads for a specific sample.
‣ **BC:** Barcode tag, which indicates the demultiplexed sample ID associated with the read.
‣ **SM:** Single-end alignment quality.

- ▸ **AS:** Paired-end alignment quality.
- ▸ **NM:** Edit distance tag, which records the Levenshtein distance between the read and the reference.
- ▸ **XN:** Amplicon name tag, which records the amplicon tile ID associated with the read.

BAM index files (*.bam.bai) provide an index of the corresponding BAM file.

Sample Analysis Reports

The Small RNA v1.0 App provides an overview of statistics per sample on the Analysis Reports sample pages. To download statistics in the Small RNA Sequencing Report, click **PDF Summary Report**.

Sample Information

Table 1   Sample Information Table

Statistic	Definition
Sample ID	Normalized BaseSpace sample ID to ensure uniqueness.
Sample Group	Sample group specified through the input form.
Number of PF Reads	Number of pass-filter reads.

Quality Control Statistics

Table 2   Quality Control Statistics Table

Statistic	Description
Small RNA Length Distribution	A histogram of the number of reads per small RNA length.
Read Distribution	The number and percentage of reads for the following categories: Abundant, Genome, miRNA, Other RNA, and Unaligned.
Abundant Distribution	The number and percentage of reads for abundant small RNA. The subcategories are species-dependent.
miRNA Distribution	The number and percentage of reads for miRNA with the following subcategories: known mature miRNA, IsomiR (known precursor), novel mature miRNA, and IsomiR (novel precursor). The latter 2 subcategories are only present when novel precursor discovery is enabled.
Other RNA Distribution	The number and percentage of reads for other RNA. The subcategories are species-dependent.

Summary and Top Sequences

The summaries for the following marker types of small RNA are provided in the analysis report:
- ▸ Mature miRNA
- ▸ IsomiR (Known Precursor)
- ▸ Precursor Group

- miRNA Family
- piRNA

The following summaries are available when novel precursor discovery is enabled:
- Novel Mature miRNA
- IsomiR (Novel Precursor)
- Novel Precursor Group

Table 3 Summary Tab

Statistic	Definition
Number of [Marker Type] with Reads	The number of [marker type] that has at least one supporting read.
Total Number of Reads	The total number of reads assigned to [marker type].
Hits File	A filterable list of the sequences and the number of reads per sequence. The PDF report only lists the top 10 sequences with most reads.
miRDeep* Result File	A filterable list of novel precursors predicted by miRDeep*.

Table 4 miRDeep* result file headings

Heading	Definition
ID	Novel precursor identifier
Score	Score of the novel precursor
Chr	Chromosome
+/-	Strand
Hairpin	Precursor location
Mature	Mature miRNA location
Expr	Mature miRNA expression computed by miRDeep*

Visualization by Precursor

Enter a precursor in the **Precursor** field or select a precursor from the dropdown to visualize mature miRNA and isomiR hits. If present, click the blue hyperlink of the precursor name to open the precursor entry in the miRBase database.

The following information about the precursor is displayed.
- 5' Counts—Read count of the 5' mature miRNA.
- 3' Counts—Read count of the 3' mature miRNA.
- 3'/5' ratio—Ratio of 3' to 5' mature miRNA read counts.
- Read counts of mature miRNAs and isomiRs.

If a novel precursor is selected, only one mature miRNA is predicted. Total Counts is provided instead of the 5' Counts and 3' Counts. Total Counts is the total number of reads assigned to the mature miRNA and isomiRs.

This section is not included in the PDF of the analysis report.

Summary Analysis Report

The Small RNA v1.0 App provides an Aggregate Summary for all samples. Statistics are plotted against samples with tables providing additional metrics.

If novel precursor discovery is enabled, the Novel miRNAs Predicted by miRDeep section contains filterable lists of the novel miRNAs predicted by miRDeep* for each sample group.

Differential Expression Analysis

Differential expression analysis is performed for the following marker types:

▸ miRNA Family
▸ Precursor Group
▸ miRNA
▸ piRNA

Click a tab to view the sample correlation, PCA plot, expression heat map, and pairwise analysis. The pairwise analysis tab is available when pairwise differential expression analysis is enabled.

Sample Correlation

A heat map showing the relative similarity between all samples in the analysis. Each row and column represents 1 sample, ordered by similarity (hierarchical clustering). The color of each field indicates the Spearman Rho correlation between these replicates.

PCA Plot

A scatter plot of samples using PC1 and PC2, the first and second principal components computed by principal component analysis, respectively.

The plot points are colored according to the sample groups, which match the legend in the upper right corner. Place your cursor over a plot point to view a tooltip with the coordinates and the sample ID.

Expression Heat Map

An expression heat map visualizing the marker by sample expression matrix. The matrix contains regularized log-transformed counts produced by DESeq2. The marker and sample dendrograms show the hierarchical clustering of markers and samples, respectively.

Select a group or sample using the dropdowns.

The expression heat map is not included in the PDF report.

Pairwise Analysis

Select a pair of sample groups in the dropdown for a summary of the analysis.

Table 5   Pairwise Analysis

Statistic	Definition
Total [Marker Type]	Total number of [marker type] in the analysis.
Tested [Marker Type]	Number of [marker type] tested for statistical significance.

Statistic	Definition
Differentially Expressed [Marker Type]	Number of differentially expressed [marker type] with adjusted p-value of 0.05 or lower.

Pairwise Analysis Browser

The browser shows an interactive scatter plot of the log2 fold change against the normalized mean count for a marker. Filter the results using the following metrics.

Filter	Definition
Test Status	• OK—Test successful. • Low—Low average expression across samples (mean normalized count across all samples < 10). • Outlier—A single replicate strongly affects the result.
Significance	Select **True** for markers with a multiple-testing adjusted p-value (q-value) for differential expression of < 0.05. Otherwise, select **False**.
Absolute log2(fold change)	Drag the left or right end of the bar to filter according to the absolute log2(fold change) value.
[Marker Type]	Search for a particular marker in the plot and the marker table below the scatter plot.

The marker table below the scatter plot shows the metrics for individual marker results. Click a marker to circle the corresponding dot in the scatter plot. Similarly, click a dot in the scatter plot to highlight the corresponding marker in the marker table.

The following additional metrics are reported in the marker table:

▸ Std. err. log(fold change)—Standard error of the fold change estimate.

▸ q value—Multiple-testing adjusted p-value for differential expression (used for the significance filter).

Revision History

Document	Date	Description of Change
Document # 15067960 v01	January 2016	Reorganized topics, updated writing style.

Notes

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 6   Illumina General Contact Information

Website	www.illumina.com
Email	techsupport@illumina.com

Table 7   Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Japan	0800.111.5011
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
China	400.635.9898	Singapore	1.800.579.2745
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	Taiwan	00806651752
Hong Kong	800960230	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000
Italy	800.874909		

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download in PDF from the Illumina website. Go to support.illumina.com, select a product, then select **Documentation & Literature**.