ScreenSeq Pipeline Results Documentation

Overview

The ScreenSeq pipeline is designed for analyzing CRISPR screen sequencing data. This document describes the output files and results generated by the pipeline, including quality control metrics, raw counts, and differential analysis results.

Version Information

- **Pipeline Version**: v1.1.0
- **Code Repository**: ScreenSeq GitHub Repository

Output Files Overview

The ScreenSeq pipeline generates several types of output files:

1. **Quality Control Files**: Statistics and metrics for data quality assessment
2. **Count Files**: Raw and processed count data for each sample
3. **Differential Analysis Files**: Statistical comparisons between experimental groups

Counts and Quality Control Statistics

Quality Control Statistics File: `<ProjectNo>____STATS.xlsx`

This file contains overall quality control statistics for the entire run. It provides metrics to assess data quality and identify potential issues.

Column	Description
Sample	Sample identifier
Total	Total number of reads
Num.Processed	Number of reads with valid sgRNA sequences
Num.Library	Number of reads found in the sgRNA library
PCT.Useable	Percentage of usable reads (Num.Library/Total)

File Columns

Quality Assessment The `PCT.Useable` column provides a critical measure of library quality: - **High values** (>80%) indicate good library quality - **Low values** (<50%) may indicate sequencing or library preparation issues - **Inconsistent values** across samples may indicate batch effects or technical problems

Recommendation: Investigate samples with PCT.Useable values significantly lower than the group average.

Raw Counts File: `<ProjectNo>____COUNTS.xlsx`

This file contains the raw (unnormalized) count data for each sample and sgRNA.

Column	Description
sgRNA	sgRNA sequence
Gene	Target gene name
ProbeID	Unique probe identifier
LibName	Library name
Samp1	Raw counts for Sample 1
...	...
SampN	Raw counts for Sample N

File Columns

Differential Analysis Results

For each statistical comparison, the pipeline generates two output files:

- **PDF Report**: `<ProjectNo>_DiffAnalysis_<SetName>_.pdf`
- **Excel Results**: `<ProjectNo>_DiffAnalysis_<SetName>_.xlsx`

Where: - `<ProjectNo>` = Project number identifier - `<SetName>` = Specific dataset name for the comparison

PDF Report Contents

The PDF file contains four quality control and analysis plots:

1. Boxplot of Normalized Log2 Data

- **Purpose**: Visualize data distribution across samples
- **What to look for**:
- Outlier samples within replicate groups
- Batch effects or systematic biases
- Overall data quality and consistency

2. Multidimensional Scaling (MDS) Plot

- **Method**: Uses `plotMDS` from Bioconductor's edgeR package
- **Purpose**: Dimensionality reduction to visualize sample relationships
- **What to look for**:
- Sample clustering by experimental groups
- Outlier samples
- Clear separation between different sample groups

3. Scatter Plot: Log Average Intensity vs Log Fold Change

- **Purpose**: Visualize differential expression results
- **Features**:
- Each point represents a probe
- Significant probes highlighted in red
- Shows relationship between expression level and fold change

4. Volcano Plot: Gene-Level Significance Analysis

- **Purpose**: Visualize gene-level differential expression results
- **Axes**:
- X-axis: Log fold change
- Y-axis: P-value (log scale)
- **Features**: Significant genes highlighted in red

Excel Results File

The Excel file contains two sheets with detailed statistical results:

Sheet 1: ProbeLevel Contains probe-level differential analysis results:

Column	Description
ProbeID	Unique probe identifier
SEQ	Probe sequence
LIB	RNA library identifier
FC	Fold change in natural units (FC=Group1/Group2)
logFC	Log2 fold change
PValue	Raw p-value
FDR	False discovery rate (multiple test corrected)
avgAll	Average counts across all samples
avg.Group1	Average counts in Group 1
avg.Group2	Average counts in Group 2

Sheet 2: GeneLevel Contains gene-level differential analysis results using the `camera` function from edgeR:

Column	Description
NGenes	Number of probes for this gene
Direction	Aggregate direction of change
PValue	Raw p-value

Column	Description
FDR	False discovery rate (corrected p-value)
logFC	Log2 fold change

Last updated: [2025-07-19] Pipeline Version: v1.1.0