nCounter Advanced Analysis 2.0

Plugin for nSolver Software

User Manual

NanoString Technologies®, Inc.

530 Fairview Ave N
Seattle, Washington 98109

www.nanostring.com

T: 206.378.6266
888.358.6266

E: info@nanostring.com

MAN-10030-03, January 2018

Intellectual Property Rights

This nSolver™ Analysis Software user manual and its contents are the property of NanoString Technologies, Inc. ("NanoString"), and is intended for the use of NanoString customers solely in connection with their operation of the nCounter® Analysis System. The nCounter Analysis System (including both its software and hardware components) and this User Manual and any other documentation provided to you by NanoString in connection therewith are subject to patents, copyright, trade secret rights, and other intellectual property rights owned by or licensed to NanoString. No part of the software or hardware may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into other languages without the prior written consent of NanoString. For a list of patents, see www.nanostring.com/company/patents.

Limited License

Subject to the terms and conditions of sale of the nCounter Analysis System, NanoString grants you a limited, non-exclusive, non-transferable, non- sublicensable, research use only license to use this proprietary nSolver software with the nCounter Analysis System only in accordance with this manual, the manual for the nCounter Analysis System, and other written instructions provided by NanoString. Except as expressly set forth in the terms and conditions, no right or license, whether express, implied, or statutory, is granted by NanoString under any intellectual property right owned by or licensed to NanoString by virtue of the supply of this software or the proprietary nCounter Analysis System. Without limiting the foregoing, no right or license, whether express, implied, or statutory, is granted by NanoString to use the nSolver Analysis Software or nCounter Analysis System with any third-party product not supplied or licensed to you by NanoString, or recommended for use by NanoString in a manual or other written instruction provided by NanoString.

Trademarks

NanoString Technologies, NanoString, the NanoString logo, nCounter, nSolver, PlexSet and Plex^{2}are registered trademarks or trademarks of NanoString Technologies, Inc., in the United States and/or other countries. All other trademarks and/or service marks not owned by NanoString that appear in this manual are the property of their respective owners.

Copyright

Contents

Introduction

Advanced Analysis 2.0 Basics

NanoString Technologies' nCounter assays are designed to provide a single-tube, ultra-sensitive, reproducible, and highly-multiplexed method for detecting nucleic acid targets across all levels of biological expression. These assays provide direct detection of targets using molecular barcodes, most without the necessity of reverse transcription or amplification. nCounter assays are processed on the fully-automated Prep Station followed by data collection on a Digital Analyzer; alternatively, processing and data collection may be accomplished together on the SPRINT instrument. The nSolver 4.0 Software Analysis System is provided to organize, view, and prepare your data for statistical interpretation.

Advanced Analysis 2.0 is conveniently provided as a link from the nSolver dashboard. It draws from powerful academic open-source analysis tools, provides a simple interface to guide you through analysis, and displays the results in an interactive HTML document. Each Advanced Analysis is performed using R, a powerful statistical software program. Familiarity with R is not required as users only need to interact with a simple wizard within nSolver 4.0.

The basic steps needed to prepare your data in nSolver 4.0 are covered in this manual (see the *nSolver 4.0 Data Preparation* section of this manual); for more information on this process, see the *nSolver 4.0 User Manual* (MAN-C0019).

Changes from Advanced Analysis 1.1 to 2.0

Advanced Analysis 2.0 is keeping pace with the rapidly expanding nCounter technology. In this version, data analysis becomes more **data-focused** and less analyte-restricted. Single Nucleotide Variance (**SNV**) analysis is supported, as is **Fusion** data analysis.

Workflow

The following steps are common to all Advanced Analyses. The *Advanced Analysis 2.0 Quick Start Guide* in the next section leads you through these basic steps.

For more details on a subject:

- Click the relevant step in the workflow below.
- Follow the hyperlinks in the Quick Start Guide.
- Navigate the manual using the Table of Contents and relevant links.

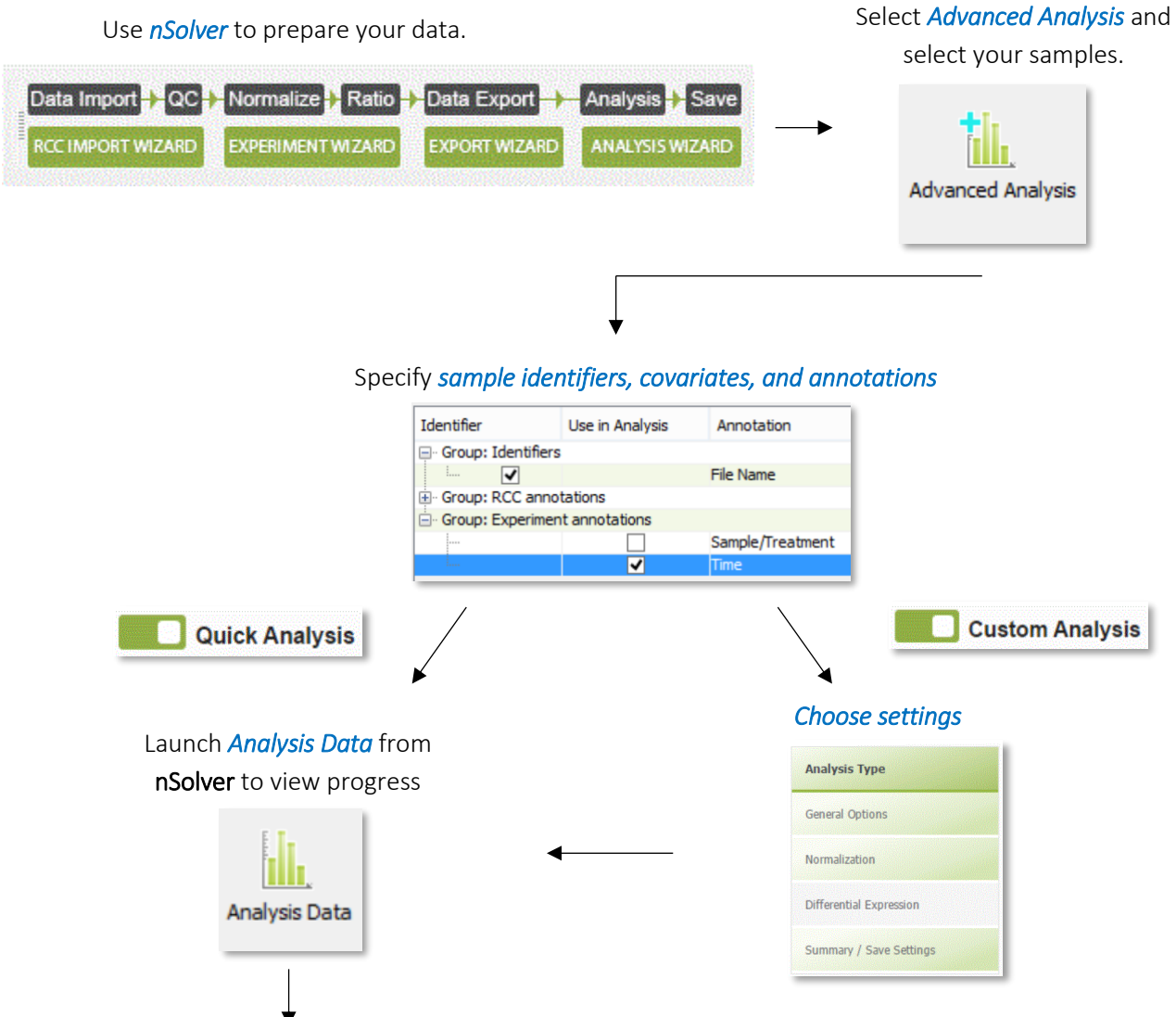
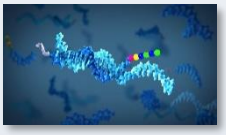


Figure 1: Advanced Analysis workflow

Analyte Types

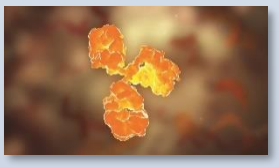
Advanced Analysis 2.0, in conjunction with nSolver 4.0, is designed to identify and support the analyte types listed below, either alone or in any combination with each other. At this time, it is not designed to analyze Plex^2, PlexSet, CNV, or miRNA data. Most commercial NanoString panels are supported and this list may be updated periodically. Contact support@nanostring.com with questions.



Messenger RNA (mRNA) – A mRNA molecule is a nucleic acid of 400-10,000 bases which serves as a template for protein synthesis (translation). mRNA panels are offered stand-alone, in Gene Expression Panels, and with miRNA panels in the miRGE Assay kits.

Single Nucleotide Variance (SNV) – SNV refers to a single- or multi-base change of up to 20 bases, which may exist as an insertion or deletion, occurring in human genomic DNA. Vantage 3D DNA SNV assays and the Vantage 3D DNA solid tumor panel are designed to detect such sequence variations at specific positions at levels as low as 5% allele frequency, thereby permitting the detection of somatic mutations commonly seen in cancer.

Fusion – A gene fusion event, which results in a hybrid gene formed from two previously distinct genes, happens through translocation, chromosomal inversion or interstitial deletion. Fusions are often used as prognostic markers in cancer diagnosis. NanoString offers direct detection and counting of fusion events in two customizable Lung and Leukemia gene fusion panels: the nCounter Vantage 3D Gene Fusion panels and nCounter Gene fusion Panels (Ex US).



Protein – Proteins are translated from mRNA producing polypeptides which perform the majority of active function within biological systems. Vantage 3D Protein Panels target proteins and phospho-proteins in a variety of cell types with the Immune Cell Profiling, Immune Cell Signaling, and Solid Tumor Lysate and FFPE Panels.

Figure 2: Analyte Types that can be analyzed using Advanced Analysis 2.0

Installation – nSolver 4.0, Advanced Analysis, & R

Requirements

Before running Advanced Analysis for the first time, ensure you have the following:

- A reliable internet connection which allows the download and installation of R libraries.
- Permissive firewall settings which allow R Script to write files to the home directory and that allow access to the websites necessary for full functionality.
- Adequate time to allow R library downloads; this can take up to one hour. This requirement is for first-time Advanced Analysis users, only.
- Practice data. NanoString strongly recommends practicing with sample data before using Advanced Analysis on experimental/clinical data. Contact support@nanostring.com.

Downloads

Advanced Analysis must be separately downloaded from the NanoString website and imported into the nSolver 4.0 application. All Advanced analysis plugins distributed by NanoString depend on a specific R version. Refer to the instruction manual of the specific Advanced Analysis plug-in you intend to use to ensure you have the correct R version installed.

Instructions for the following software downloads are listed individually below: **nSolver 4.0 Analysis Software**, **R 3.3.2**, and the **Advanced Analysis 2.0** plugin.

Downloading nSolver 4.0 Analysis Software

If you have been using another version of **nSolver 4.0 alpha**, you will need to back up your database and start with a **clean or blank nSolver 4.0 database**. Then, download and install the software.

Windows users:

- Navigate to c:\users\<username>\appdata\roaming\. Rename your nSolver4 folder to *nSolver4_old* (or similar). You may need to *show hidden files* in order to see the *appdata* folder.
- **Download** and extract **nSolver 4.0** from https://www.nanostring.com/products/analysis-software/nsolver. **Install** the nSolver 4.0 application.
- When prompted to **Install R**, select **Yes** (see next section).

Mac users:

- From your home directory, make sure your hidden files are shown so you can see your nSolver4 folder. Rename it *nSolver4_old* (or similar).
- **Download** and extract **nSolver 4.0** from https://www.nanostring.com/products/analysis-software/nsolver. **Install** the nSolver 4.0 application.

Downloading R 3.3.2

R 3.3.2 is required for version 2.0 of the Advanced Analysis.

Windows users:

- You will be given the option to download R 3.3.2 when you install **nSolver 4.0**. If you did not, go to https://cran.r-project.org/bin/windows/base/old/3.3.2/.
- If you've previously used a different version of R with Advanced Analysis and are updating to a new version of R, you will need to **change the R home path in nSolver**. Select **Analysis** on the top toolbar in nSolver and select **Change R Home Path** to the R 3.3.2 installation folder. Browse to the desired directory and then select **Ok**.

Mac users:

- Install R separately. Go to https://cran.r-project.org/bin/macosx/old/R-3.3.2.pkg.
- Install XQuartz if you use Mac OS X 10.10 or higher. Go to https://www.xquartz.org/.
- You may need to download **R Switch** or a similar app to replace your current version of R with 3.3.2. Alternatively, you may uninstall all other R versions.

When initiating an analysis in Advanced Analysis 2.0, nSolver 4.0 will check the version of R you have installed and will issue a warning if it is a version incompatible with the program.

Downloading Advanced Analysis 2.0

You will find the most recent version of Advanced Analysis on https://www.nanostring.com/products/analysis-software/nsolver. Save this to your computer as a compressed .zip file. *Do not extract the files before uploading them to nSolver.*

In **nSolver 4.0,** select **Analysis** on the top toolbar (see Figure 3) and select **Advanced Analysis Manager**. Any previously-installed versions of Advanced Analysis will be displayed. You can **Remove** them or simply **Import** the current version. To import, select the **Import New Advanced Analysis** button and navigate to the .zip file with the current Advanced Analysis version. This version will be added to the list within the Advanced Analysis Manager. Select **OK**.

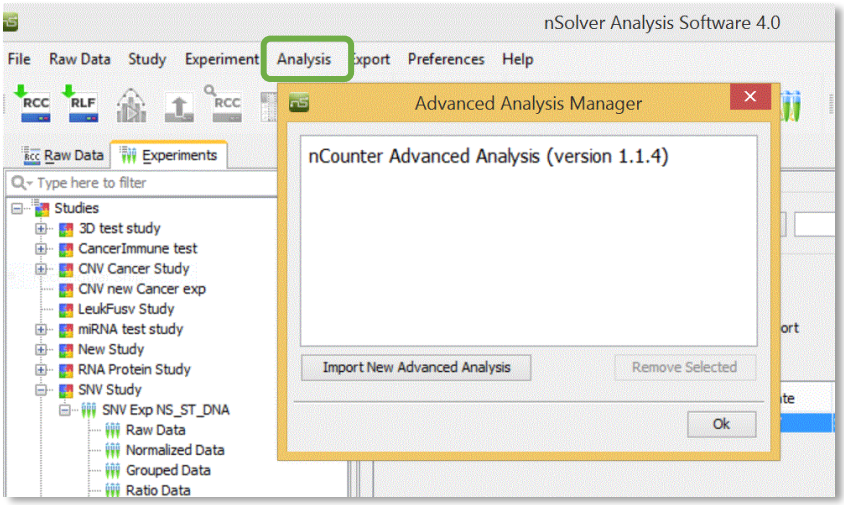


Figure 3: Importing Advanced Analysis or changing the analysis version

Advanced Analysis 2.0 Quick Start Guide

The Advanced Analysis software plugin provides a number of R-based statistical tools with minimal input from the user. Before beginning, ensure you have a reliable internet connection and security settings that allow pop-ups. First-time Advanced Analysis users should ensure they allow adequate time to download required R libraries (this ~750 MB file may take up to 1 hour to download). See the *Installation* section for download and installation instructions.

1. Experimental Design & nSolver 4.0 Data Preparation: **Import** your **RCC** and **RLF files** to nSolver 4.0 and create an **Experiment**. For more on this topic, see the *nSolver 4.0 User Manual* (MAN-C0019) or the *nSolver 4.0 Quick Start Guide* (MAN-10049). Annotate samples, bearing in mind that the annotations will be used as variables in Advanced Analysis. On the **Experiments** tab, highlight the raw or normalized data and select **Advanced Analysis**.
2. Creating an Advanced Analysis: Highlight the desired **Advanced Analysis** version (if more than one installed), choose a **Name** for the analysis, and **Browse** for the location in which you would like it saved. Select an **Identifier** that is unique to each sample (including SNV references) and one or more **Covariate** by checking appropriate boxes. Use the drop-down menu in the **Categorical Reference** column to set a reference group as your baseline. Selecting **Quick Analysis** will result in Overview, Normalization, and Differential Expression analyses for expression data and variant call detection analyses for SNV and Fusion data. **Custom** may be selected when wanting to customize analysis settings; these settings are addressed on pages 3-4 of this guide.
3. Viewing Analysis: Return to the nSolver 4.0 dashboard, select your experiment, and expand the navigation tree. Highlight the Analysis data level and find your most recent analysis on the list. Highlight it and select **Analysis Data**. This will open a window in your browser; you may need to **Allow Blocked Content**, depending on your internet security settings. Libraries will load, status messages will dynamically appear in the browser, and ultimately, an analysis screen will appear. See next page for descriptions of plots and options available.

Workflow

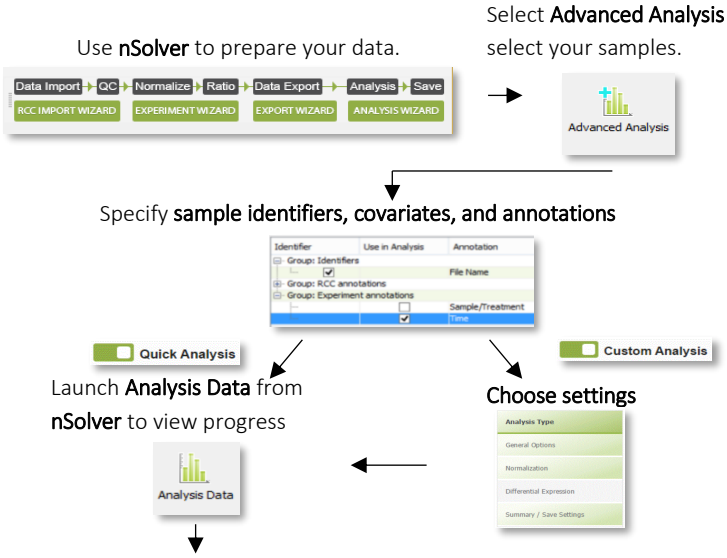


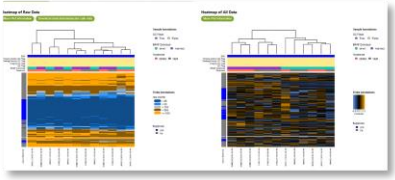
Figure 4: Advanced Analysis workflow

View **QC** results and **data analysis** in pathways of interest (this step continued from previous page).

Overview | Normalization | Diff Expr | GSA | PathView | Analysis Parameters | Share

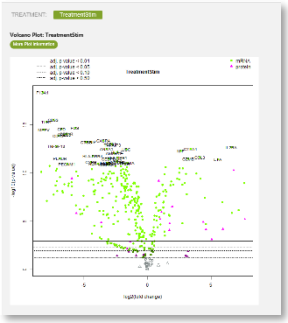
Overview

Overview heatmaps display raw data, allowing you to identify gene sets with low counts and normalized data clusters, which gives you a high-level view of possible associations within the data. Choose to view only genes in particular gene sets along the left side of the window and choose to view Principal Component analysis, study design, and QC data along the top of the window.



Differential Expression

This module isolates the effect of each variable on the data. It displays a linear regression of the differential gene expression for each variable as a volcano plot.

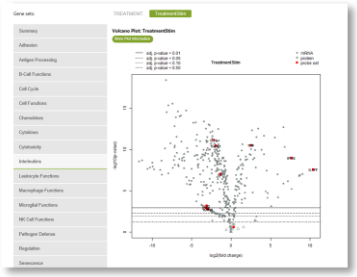


Share

This allows you to access the Advanced Analysis report as a sharable zip file. Once it is saved to your computer, extract AdvAnalysisReport.zip and view the HTML report outside of nSolver. This folder also contains all the analysis output images and data files.

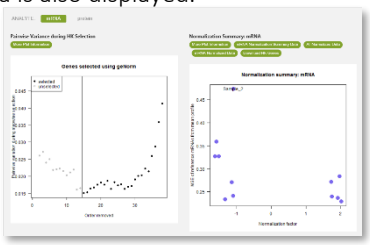
Gene Set Analysis (GSA)

GSA overlays differential expression data for sets of genes grouped by biological function, considering the covariates and relative to the baseline.



Normalization

This module allows you to normalize mRNA and protein data separately. It uses the geNorm algorithm for mRNA, choosing only the most stable housekeeping genes. Scatter plots display the effect of the chosen normalization settings on the data. Protein expression data is also displayed.



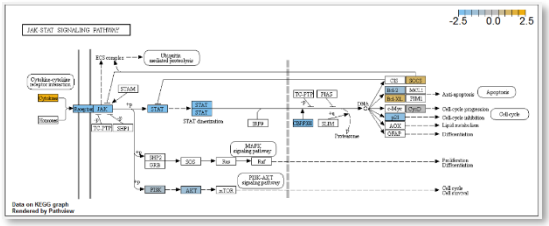
Analysis Parameters

Under this tab, you may view all analysis settings and details. You may also review the reasons behind any aborted analyses.

SNV | Fusion

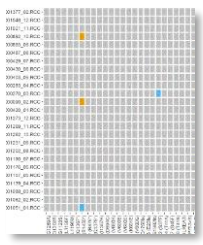
PathView

This module displays different KEGG pathways and highlights pathway members most differentially expressed in your data.



SNV- & Fusion-specific plots

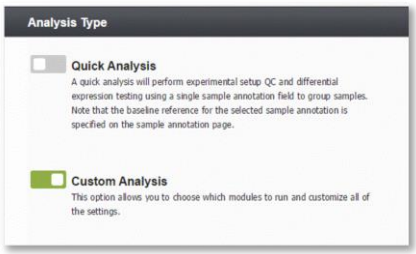
SNV and Fusion variant detection call summaries can be found on these tabs. QC metrics specific to these assays can also be found in this section.

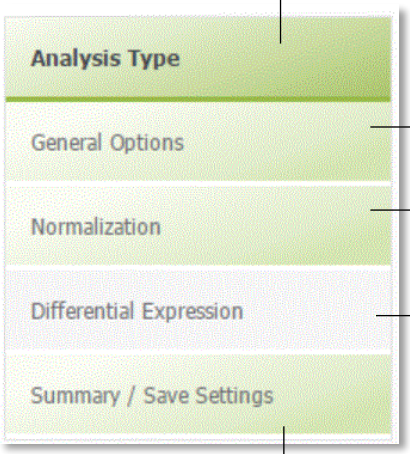


Custom Advanced Analysis Settings & Plots

Analysis Type

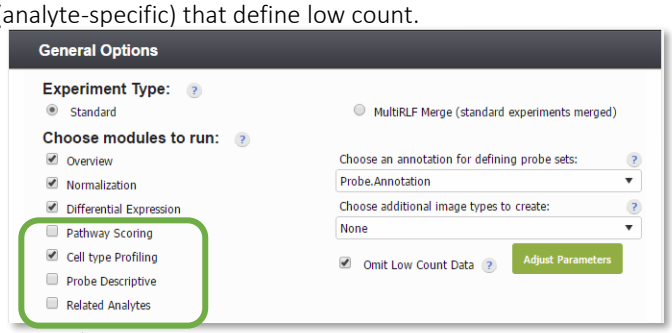
Here, you choose between Quick and Custom Analysis.





General Options

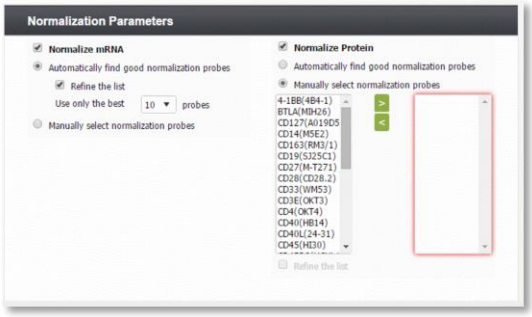
On this menu, choose the modules to run, confirm the experiment type, choose a probe annotation file, and determine any additional image types (.pdf, .jpg, etc.) to create. Use the check box to omit low count data and then Adjust Parameters to alter the thresholds (analyte-specific) that define low count.



Selecting these modules adds them to the menu. See following page.

Normalization

Advanced Analysis allows you to normalize each analyte type with its own custom settings. Manually select probes or allow the software to automatically select the best performing probes. It can also refine the list of probes to the top 10 (or other number of your choice). See previous page for resulting plot.



Summary/Save Settings

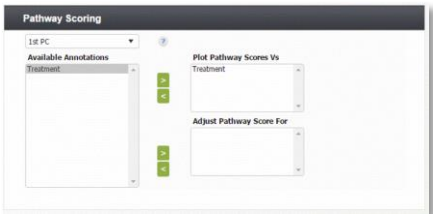
This displays a summary of your settings for the current analysis and allows you to save them for a future application.

Differential Expression

On this menu, choose one or more variables to include in your differential expression model. Predictors and confounders are treated equally in this model, but results will only be shown for predictors. Choose to run DE using the **Optimal** or **Fast/Approximate** method. The Optimal method is robust for estimating differential expression when probe counts are low or near background but computationally demanding. The Fast/Approximate method works well for probe counts observed significantly above noise. The PathView plots can be colored by either t-statistic or fold-change. See previous page for resulting plot.

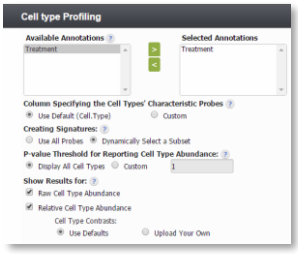
Pathway Scoring

Use the green arrows to select variables to plot against pathway scores and variables to adjust for before calculating pathway scores. See below for plot.



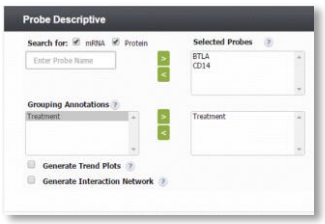
Cell Type Profiling

Use the green arrows to move at least one covariate from **Available Annotations** to **Selected** to analyze cell population abundance.
See below for resulting plot.



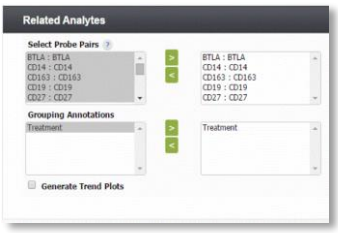
Probe Descriptive

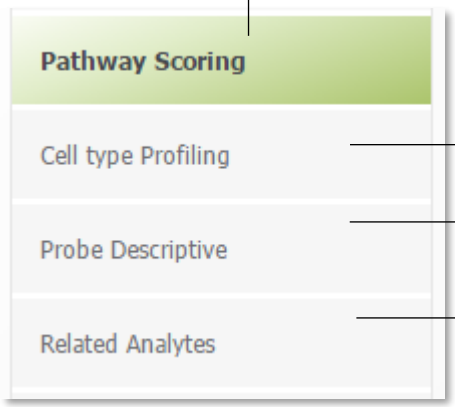
Search for probe names to calculate detailed metrics on a smaller subset of genes. At least 5 genes need to be entered for Principal Component Analysis biplots. See below for resulting plot.

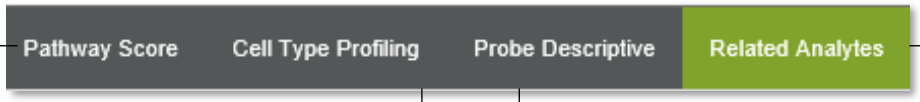


Related Analytes

Related probes for different analytes will be listed. Use the green arrows to move the probe pairs of interest. See below for resulting plot.

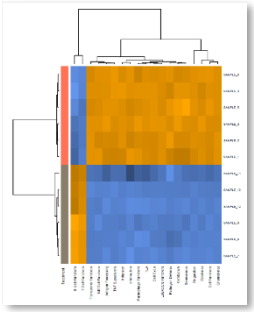






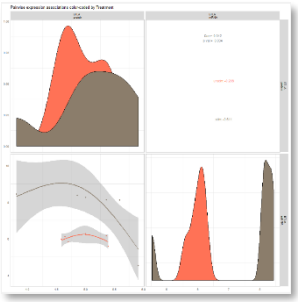
Pathway Score

The Pathway Score summarizes the data from a pathway's genes with a single score. The heatmap of Pathway scores shows a high-level overview of how the pathway scores change across samples.



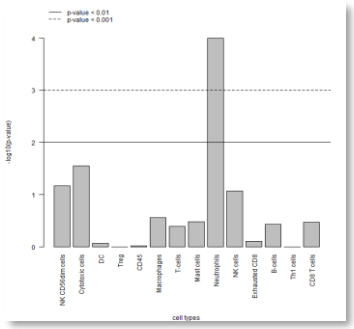
Related Analytes

This module compares the expression levels of multiple analytes when they have been linked in the probe annotation file. It applies all the tools of the Probe Descriptive module to each pair of related analytes.



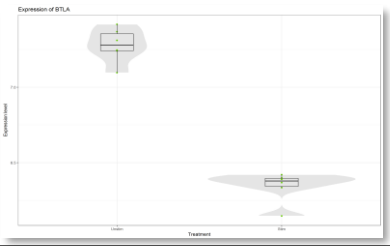
Cell Type Profiling

This module quantifies various cell types using cell type-specific marker genes.



Probe Descriptive

This module provides detailed descriptive analysis of 1–15 genes selected using univariate and correlation plots. When at least 5 probes are selected, PCA biplots and parallel coordinate plots will also be generated.



What to Do Before Performing Advanced Analysis

Meaningful and effective Advanced Analysis outputs rely on properly-designed experiments and well-prepared data.

Experimental Design

Experimental design drives the quality and clarity of downstream analysis results. Considering the number of samples, replicates, and variables ahead of time is essential.

- If working with categorical variables, arrange to run **at least three *biological* replicates** in each category.
- **Probe annotations** document the biological significance of the probes and link them to the pathways with which they are associated. Check your probe annotation file to ensure the fields you need are filled. For help with your annotation file see the *Managing Probe Annotations* section.
- The default unique **Identifier** for your samples is the file name. This may appear long or complicated in visualizations, so you may consider creating a simpler identifier using the Description column in nSolver. Ensure that any SNV references you may incorporate utilize the same identifier category.
- Use your **sample annotations** to label both confounders and predictors. These will become your **covariates** to choose for analysis.
    - A **Confounder** is a variable which affects your data but which is not scientifically relevant. **Technical confounders** are variables such as run date or cartridge lot. **Experimental confounders** are variables such as patient body mass index or age. You will want to investigate each confounder's effect on your data in such a way that it does not complicate the effect of any predictors included for analysis. Ensure that any confounding variables do not wholly overlap with predictors.
    - A **Predictor** is a variable which affects your data and which is scientifically relevant. Examples include treatment type, treatment time, and cell line. You will want to investigate each predictor's effect on your data in such a way that it is not complicated by the effect of any confounders included for analysis. Sample annotations (established during experiment-creation in nSolver) can be used to distinguish predictors and test for their effect on the data in Advanced Analysis.

Tips: recommended workflow

It is helpful to run Advanced Analysis through a multiple passes process.

First pass: include all samples and all possible predicting and confounding variables. Run the **Quick Analysis** and view the **Overview**, **Normalization**, and **Differential Expression** modules and check plots on these tabs for clustering and bias; these indicate variables which are impacting your data. Use this information to determine which samples and covariates to choose for your next pass.

Second pass: remove samples that failed QC in the first pass. Choose the covariate that is most scientifically relevant to your project and set up a **Custom Analysis**. Choose analysis modules and parameters that fit your experimental design.

More passes: analysis can be further modified after reviewing analysis results from previous passes. This includes removing outlier samples, using a different covariate(s) and applying different parameters to analysis.

nSolver 4.0 Data Preparation

Advanced Analysis requires either raw or normalized data from an **nSolver experiment** as well as **the appropriate RLF**. Below is an abbreviated description of the nSolver 4.0 workflow required to prepare your data for Advanced Analysis. Refer to the *nSolver 4.0 User Manual* (MAN-C0019) for more details.

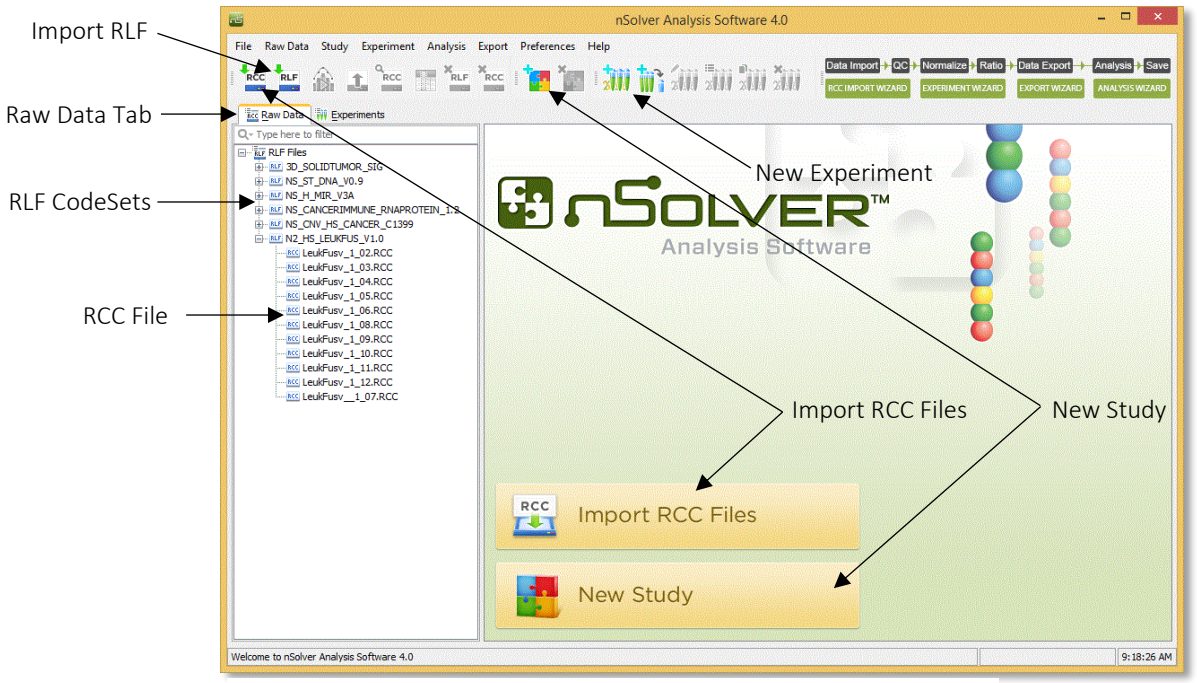


Figure 5: nSolver dashboard – raw data tab

Import Files & Explore Raw Data

A **Reporter Code Count (RCC) file** is an output file generated by nCounter instruments. One RCC file is produced for each sample tested; this one file contains the barcode counts from each gene and control in the CodeSet. RCC data files should be saved on your computer or USB drive. **Open your data folder** and **unzip** data files using right click and **Extract All**. Open **nSolver 4.0™** and select **Import RCC Files**. Browse for and select your samples of interest. Select **Data has fusion probes** if working with fusion data (this allows you to designate fusion probes). Select **Open**. Review **QC** parameters, then select **Import**.

A **Reporter Library File (RLF)** is a file specific to your CodeSet. It provides nCounter instruments and the nSolver 4.0 software application with valuable information about the CodeSet, such as the assignment of probe to gene. To import, select the **Import RLF File** icon on the toolbar at the top of the page. **Browse** to navigate to the folder in which your RLF file is stored and select **Import**. Importing the RLF is required for Advanced Analysis. SNV references may be run on a separate RLF; this should be imported, as well.

After importing to nSolver 4.0, your RCC data files will be stored under the corresponding RLF file CodeSet on the **Raw Data** tab. Selecting the CodeSet name allows you to view all RCC files in a table format. Scroll to check for QC flags. Use Description column to create shortened sample identifiers.

Create a Study and Experiment

Select the **New Study** button to create a study, then the **New Experiment** button to create an experiment under that study. Follow the prompts to select the samples to include in your experiment.

Annotations assigned here can be selected later as covariates (variables) in Advanced Analysis (see the *Identifiers and Covariates* section). Carefully consider your experimental design at this stage (see the *Experimental Design* section), as it can have an impact on visualizations created later in Advanced Analysis.

Background correction and **Normalization** steps are not necessary if the data will be processed by Advanced Analysis; the plug-in will perform its own thresholding and normalization and will pull the raw data from nSolver. You will be prompted to select **SNV reference samples** (RCCs and RLF must already be imported into nSolver) if working with SNV data. Fold Changes (**Ratios**) can be calculated by specifying the sample(s) that represent the baseline of your experiment.

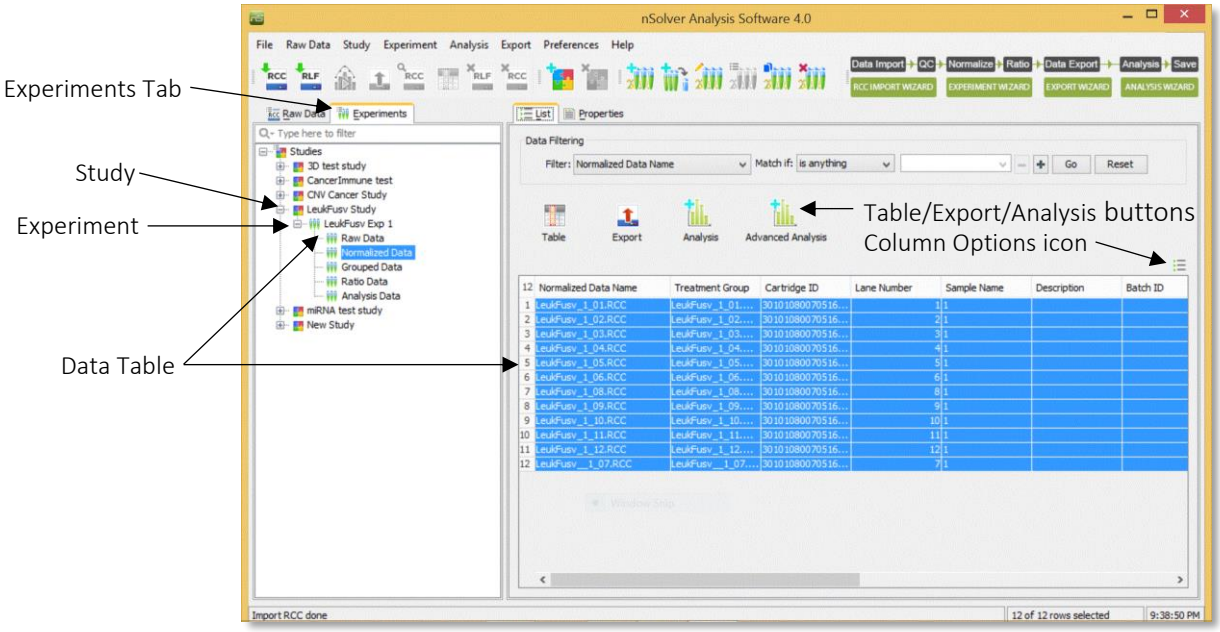


Figure 6: nSolver dashboard - experiments tab

Creating an Advanced Analysis

Select Data

Once you have created an experiment in nSolver 4.0, expand the navigation tree on the Experiments tab (by clicking on the **+** sign) and highlight either the **Raw** or **Normalized** data level of your experiment.

- Raw data is typically used for most single-RLF experiments since the QC processes in Advanced Analysis are more sophisticated than those in nSolver.
- Normalized data should be selected for any multi-RLF experiment.

Highlight the samples you want in your experiment, utilizing the **Exclude Selected** and/or **Keep Selected** buttons, if desired. The **Filter** tool is available, as well.

Select **Advanced Analysis**. Select the version of Advanced Analysis you'd like to use (if more than one is installed) choose a **Name** for the analysis, and **Browse** for the location in which you would like the resulting file saved.

Select **Next**. A warning will appear if nSolver detects a version of R which is incompatible with the program (R version 3.3.2 is required for Advanced Analysis 2.0). See the *Downloading R 3.3.2* section.

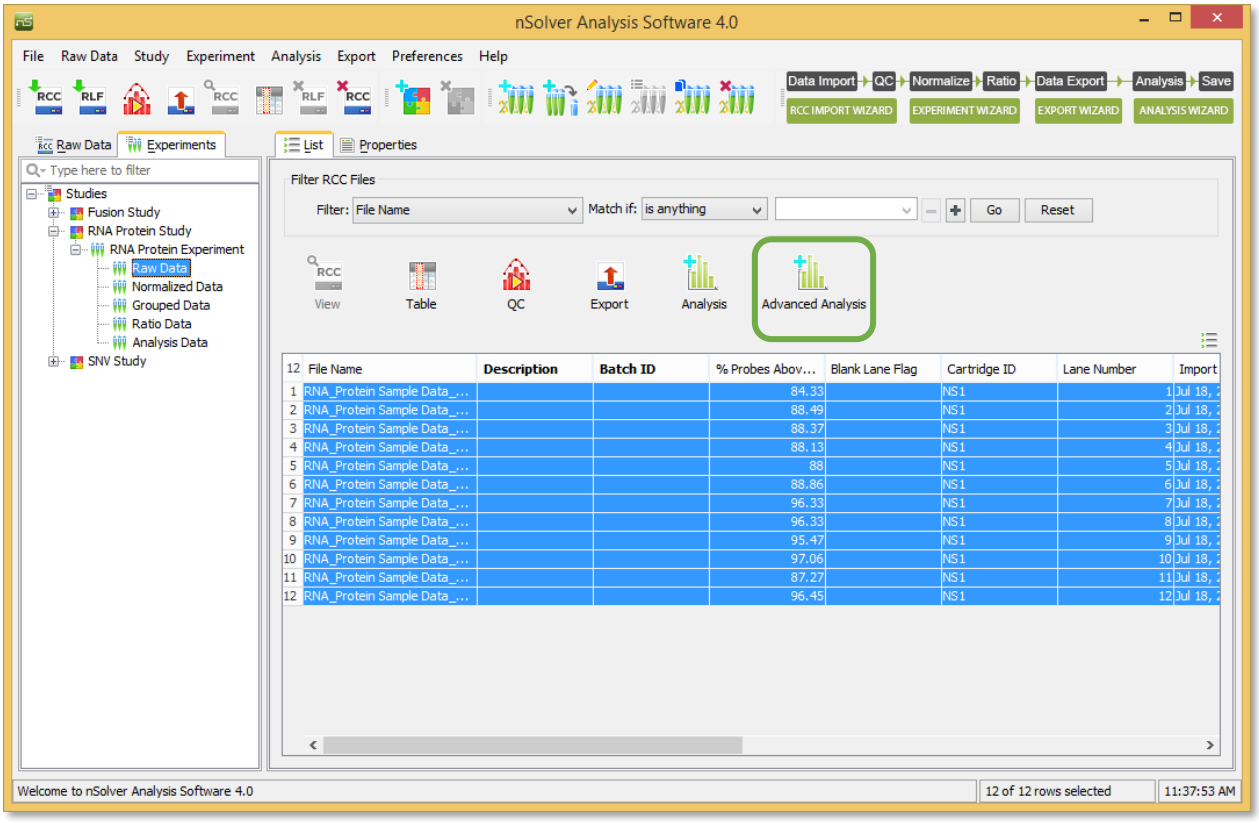


Figure 7: starting an Advanced Analysis

Identifiers and Covariates

Identifiers are unique names that differentiate every sample from the others. The Sample File Name will always be unique, but can be long, so you may choose another type of identifier for this reason. Any sample attributes that are unique from sample-to-sample will have a check box in the Identifier column and will be available for use. Only one box may be checked for the Identifier category.

Covariates are variables which the Advanced Analysis tool can isolate and assess the effect of. At least one covariate must be selected by checking a **Use in Analysis** box. Multiple covariate options are available, including:

- Any RCC file attributes, including Cartridge ID, Lane Number, Assay type, Scanned date, Comments, FOV Count, and Binding Density. Note that these technical covariates are useful for QC purposes only (e.g. assessing batch effects).
- Any sample annotations added to the lanes in the nSolver experiment wizard during the creation of the experiment (see the *nSolver Data Preparation* section).
- Any additional sample annotations imported from an external text file in this dialog box.

Too many covariates selected in one analysis can complicate matters; it is often wise to consider which variables are potential confounders and which are potential predictors and run multiple analyses, selecting different combinations of covariates in each analysis.

Select the type of identifier and covariate you have using the **Choose Type** column; you may choose categorical, continuous, or True/False. If categorical, you will need to select a **Categorical Reference**, which will serve as a baseline sample for analysis.

Select one check box in the **Identifier** column and at least one in the **Use for Analysis** (covariate selection) column and select **Next.**

Use the **Import** or **View Annotations** buttons to import new or view existing sample annotations, respectively.

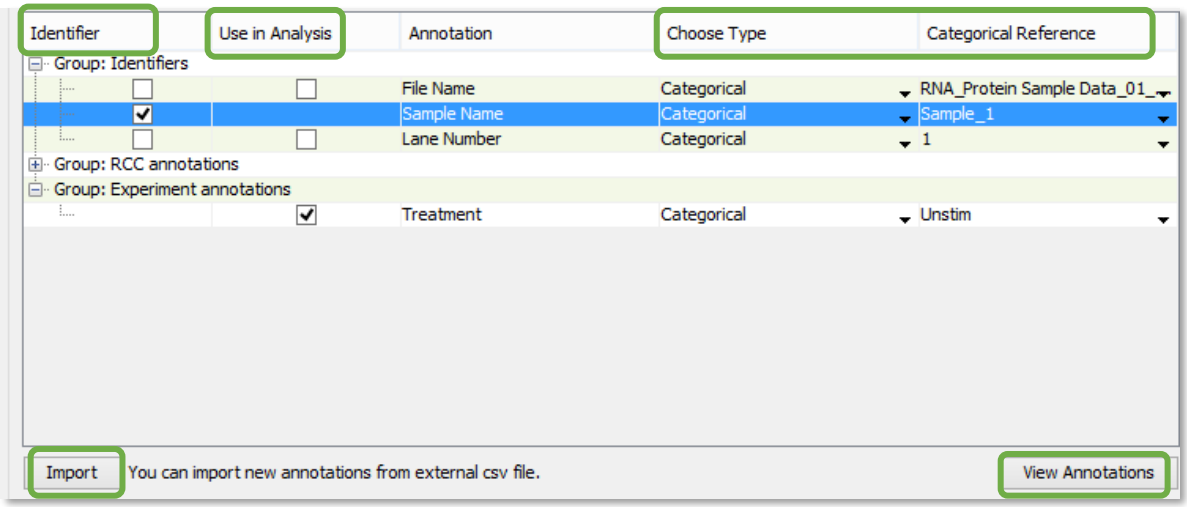


Figure 8: selecting identifiers and covariates

Analysis Type

In a **Quick Analysis**:

- The analysis is performed with default parameters.
- Probe annotations are not required for mRNA and Protein analyses.
- The core modules are preselected – Overview, Normalization, Differential Expression, GSA, and PathView.
- Only a *single* covariate is used for differential expression (DE) analysis.

In a **Custom Analysis**:

- Multiple menu tabs appear to the left of the screen (see the *General Options Custom Analysis Menu* section, as well as the *Custom Options* section found in each respective module section), allowing you to customize parameters.
- Probe Annotations are required for GSA, PathView, Cell Type Profiling, and Pathway Scoring.
- In addition to the core modules, Overview, Normalization, Differential Expression, GSA, and PathView, you have access to Related Analytes, Probe Descriptive, Cell Type Profiling, and Pathway Scoring. You can select or deselect these to customize your analysis.
- You may select multiple covariates for analysis.

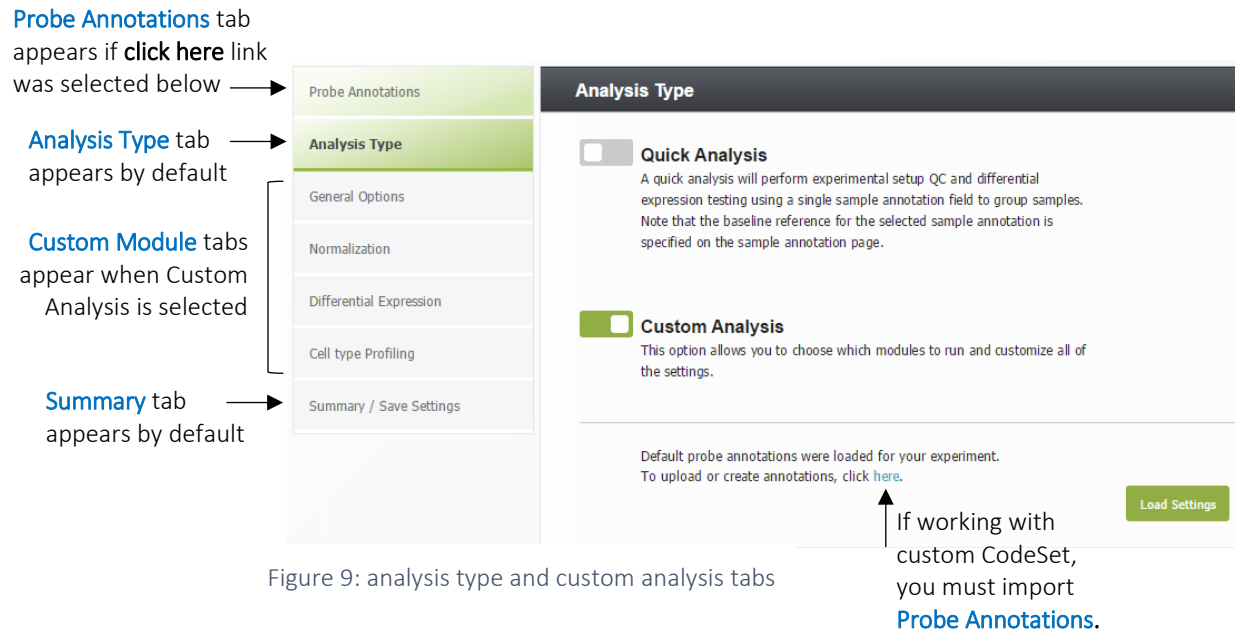


Figure 9: analysis type and custom analysis tabs

Probe Annotations

A message indicating the status of your probe annotations will appear at the bottom of the *Analysis Type* window. Most commercial panels come with probe annotations pre-loaded; you may use these, replace them with you own file, or customize them using the **click here** link. See the *Managing Probe Annotations* section.

Load Settings

Select this button to browse for a saved settings file from a previous analysis from a common CodeSet. This will load the saved settings as well as the probe annotations. Covariates may need to be re-selected for analysis; navigate to the module menus to reselect or confirm covariates for analysis.

Manage Probe Annotations

The nCounter Advanced Analysis plugin uses probe annotations to define the biological functions of the respective probes present. For most commercial CodeSets, the Probe IDs are collectively imported through the RLF. If working with a custom-designed nCounter CodeSet, however, you may either request probe annotations from NanoString or create a probe annotation file using a template (see sections below).

Probe annotations:

- Define KEGG IDs that associate pathway membership of the target gene or expression characteristics of a cell type to perform cell type profiling of the data.
- Assign Gene Set membership where a 'set' identifies a broad biological function category such as 'Adhesion'.
- Identify Related probe-pairs such as mRNA and Protein counterparts of a target gene (sharing the same NCBI gene ID).

If You Do Not Have Default Probe Annotations

For Custom CodeSets and a subset of NanoString Panels, Probe Annotation files are not automatically uploaded by the software; in these cases, an alert at the bottom of the *Analysis Type* screen will be displayed: **Default probes could not be loaded for some or all of the probes in your experiment**. In this case, you may do one of the following:

- Run a Quick Analysis with no probe annotations (mRNA and Protein).
- Request a probe annotation file from NanoString (see the *Requesting Probe Annotations from NanoString* section).
- Create a custom probe annotation file from a template (see the *Creating Probe Annotations for Custom CodeSet Data* section).

Creating Probe Annotations for Custom CodeSet Data

In the *Analysis Type* window, select the **click here** link to open the *Probe Annotations* window. You may also access this window by selecting the *Probe Annotations* tab, if visible. Select the **Download CSV** button and save the **ProbeAnnotations.csv** file to your computer. Modify this template file to include annotations that suit your analysis needs. The properties of each of the columns are explained in Table 1. See the *Importing Probe Annotations Files* section for next steps.

Requesting Probe Annotations from NanoString

If you would like to request a probe annotation file from NanoString, **send an email** to bioinformatics@nanostring.com with a request for probe annotations, including the following information:

- The **name of the RLF** for the nCounter data that you wish to analyze.
- The **annotation database** that you would prefer to employ - GO molecular function, GO cellular component, GO biological process, KEGG BRITE, KEGG Pathway, or Reactome.
- If working with a multiRLF Merge Experiment, include the **nSolver experiment report for the multiRLF Merge** experiment as an attachment.
- If the data is from a **CodeSet Plus RLF, send the RLF file** that was used to scan the nCounter cartridge to generate your data.

Save the .csv file for probe annotations that you receive from NanoString to your computer. See the *Importing Probe Annotations Files* section for next steps.

Importing Probe Annotation files

In the *Analysis Type* window, select the **click here** link to open the *Probe Annotations* window. You may also access this window by selecting the *Probe Annotations* tab, if visible. Select the **Import CSV** button. Browse to the desired probe annotation file and select **Open**.

Scroll through the preview of the annotations displayed in the screen to confirm that your custom annotations have been applied. When you are satisfied, select the *Analysis Type* tab again and confirm that the message at the bottom of the window now indicates that **Probe Annotations were loaded for your experiment**.

Proceed to the *General Options* section if running a Custom Analysis or the *Back to the nSolver Dashboard* section if running a Quick Analysis.

Table 1: Probe Annotations file format

Column Number	Column Name	Description
1	ProbeID	The Probe ID must be unique within the file
2	CodeSet.Name	RLF
3	Probe.Label	Generic name for the target mRNA or Protein
4	Analyte.Type	Indicates whether the probe detects an RNA or Protein target
5	Is.Control	Boolean. TRUE or FALSE
6	Control.Type	Indicates whether a Control probe target has Exogenous (e.g., ERCC probes), Endogenous (Housekeeping), or Negative controls
7	Related.Probes	Semicolon-delimited list of Probe IDs. Identifies probes for mRNA and Protein counterparts of a gene. Probes for splice isoforms or phosphorylated vs. non-phosphorylated counterparts. Values in this column are necessary to run the Related Analytes module
8	Probe.Annotation	Semicolon-delimited list of annotations. Identifies Probe sets characteristic of a biological function. By default, this column defines the annotations for grouping probes for Gene Set analysis and Pathway scoring Modules.
9	KEGG.Pathways	Semicolon-delimited list of KEGG Pathway IDs; values in this column are necessary to run the PathView Module
10	Cell.Type	Identifies cell types in which target genes have known characteristic expression. By default, this column defines the annotations for running the Cell Type Profiling module.
11	Official.Gene.Name	HUGO gene name http://www.genenames.org (I'd replace this entirely with: "Official gene symbol per NCBI". That happens to be HUGO for human but MGI for mouse…
12	Fusion.probe.type	Denotes whether the probe identifies a junction (fusion), 5' expression (5p) or 3' expression (3p)
13	Fusion.base	The group of fusion products to which the probe relates. For instance, the oncogene related to an imbalance probe or set of fusions (like BCR-ABL) to a specific fusion junction
14	SNV.probe.type	Denotes whether the probe identifies reference or variant bases
15	SNV.LocID	An identifier linking all related variant and reference probes for analysis purposes
16	SNV.annot	Name for each probe to be displayed, intended to be more reader-friendly than the name in the RLF.
17		Add additional columns starting at column 12. When present, these will be available as a custom annotation column for specifying the column defining probe sets for Cell Type Profiling, Gene Set Analysis, and Pathway Scoring.

General Options Menu

If you've chosen to run a Custom Analysis, you can start customizing on this tab.

Choose the appropriate **Experiment Type. Standard** refers to a single-RLF experiment. A **MultiRLF Merge** is a multi-RLF experiment created by combining more than one pre-created experiment in nSolver.

Choose an annotation for defining probe sets using the dropdown menu. This will impact the Gene Set options available in some modules. By default, *Probe.Annotation* is selected.

The question mark button reveals additional information.

The exclamation mark button reveals an alert and brief explanation as to why an option may be unavailable (greyed out).

Choose additional image types to create. The default image format is .png. Use the dropdown menu to specify additional image formats for each image.

The **Omit Low Count Data** checkbox permits the software to remove genes that fall below a given low count level. You can use the **Adjust Parameters** button to change the threshold options for the different analyte types. The Overview heatmaps depict the probes pruned from analysis with a blue *below threshold* bar (see the *Overview Module* section).

Choose modules to run. Click the module check boxes to display the corresponding tabs on the left under *Analysis Type*. Click the appropriate tab to review settings and options. Some options may not be available (may be dimmed) due to incompatibility with analyte types detected in the data and/or limitations in the probe annotations.

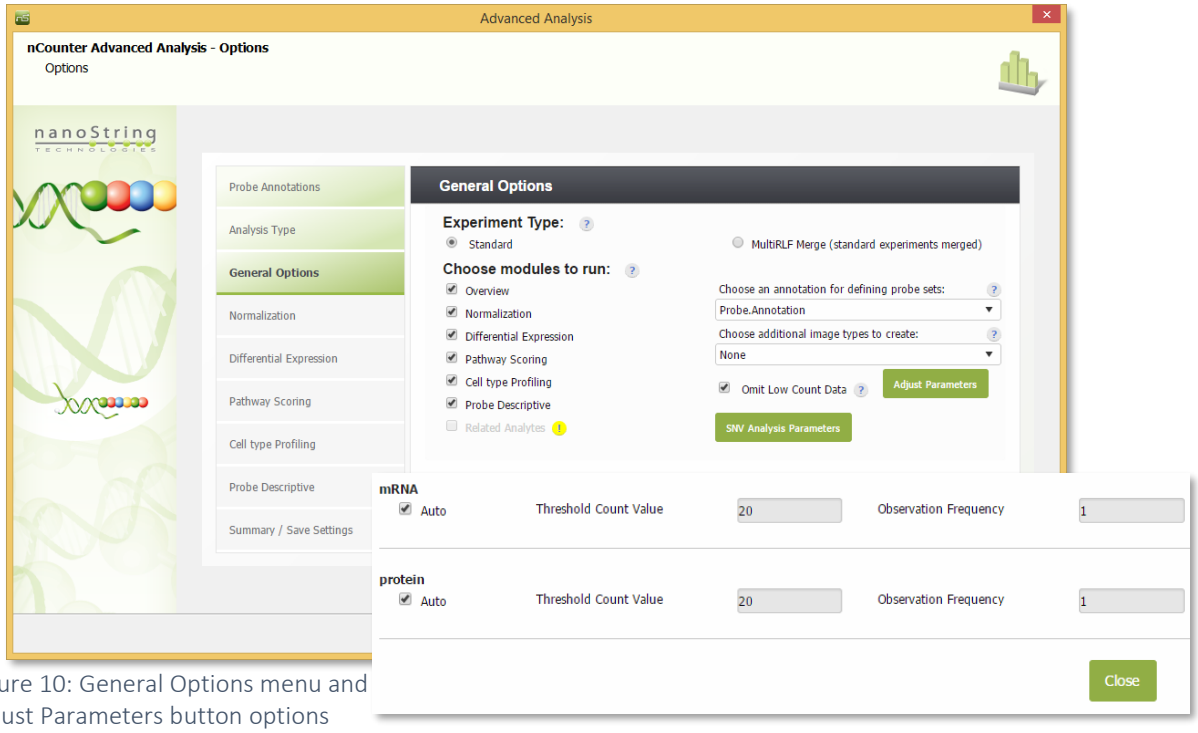


Figure 10: General Options menu and Adjust Parameters button options

Custom Analysis Module Menus

See individual module sections for information on Custom Analysis menu options. They are listed here: *Normalization*, *Differential Expression* (includes GSA and PathView), *Pathway Scoring*, *Probe Descriptive*, *Cell Type Profiling*, and *Related Analytes*. Custom Options for *SNV* and *Fusion* are included on the General Options tab.

Summary/Save Settings

The Summary/Save Settings tab provides information about the current analysis and allows you to save the settings and apply them to a subsequent analysis for data derived from an identical CodeSet. This is especially useful when looking at the effects of different annotations on analysis.

To save the settings for a subsequent analysis with a common CodeSet, select the **Save Settings** button on this tab.

To use these settings in a subsequent analysis with a common CodeSet, use the **Load Settings** button in the *Analysis Type* window (see the *Analysis Type* section).

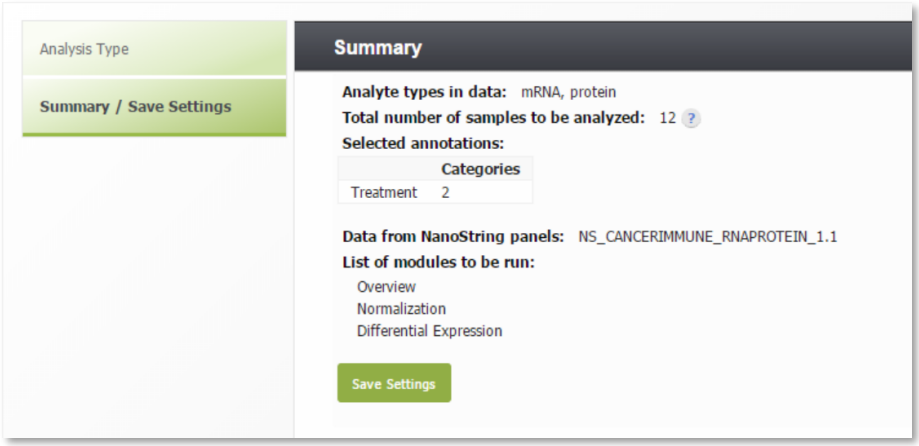


Figure 11: Summary / Save Settings tab

Back to the nSolver dashboard

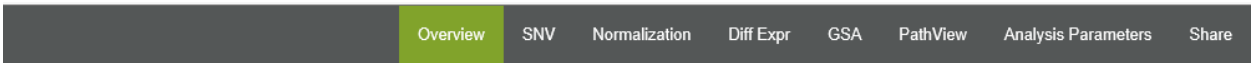
Select **Finish.**

You will be retuned to the nSolver dashboard. Highlight your analysis in the list and select **Analysis Data** to view your plots and options.

This will open an HTML window and dynamically display the progam's status. When complete, a summary screen will appear. Click through the different plots and options for viewing data.

Overview Module

The Overview module provides a general overview of the data through descriptive plots, organized into four categories: **Heatmaps, PCA (principal component analysis), Study Design,** and **Other QC.** Heatmap and



PCA plots can be drawn as a summary of all probes or in just the specific **gene set** of interest. Note: Fusion and SNV data do not produce an overview module.

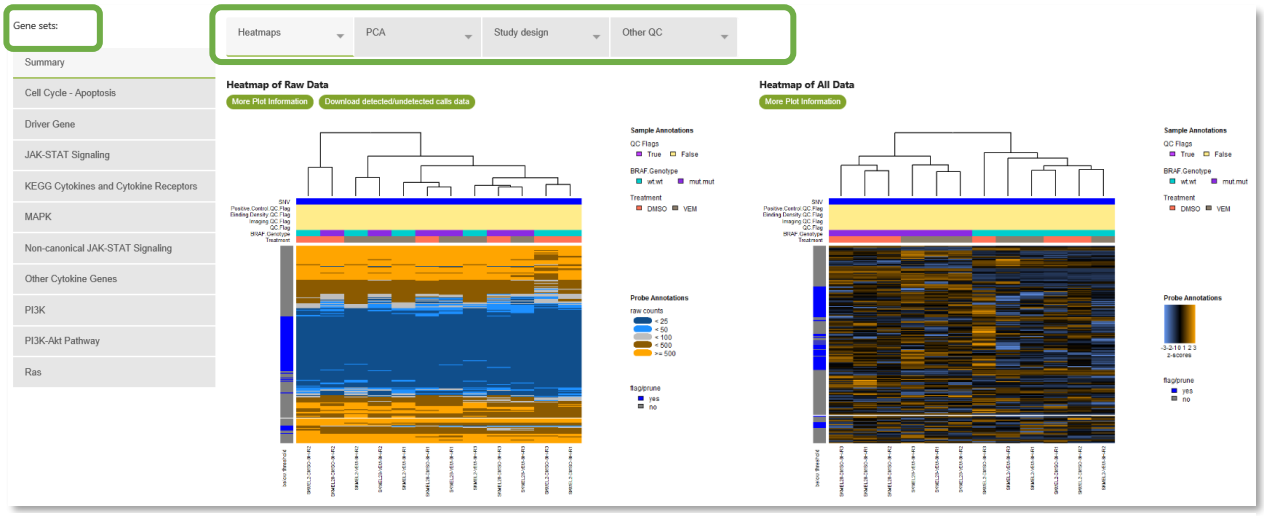


Figure 12: Overview module window and options

Before You Start Overview

This module is not intended to be used in in-depth analyses; it should be used as a QC tool and way to get a general impression of your data.

Designations for SNV and Fusion variant status as well as **covariates** will appear at the top of the heatmap. Some covariates will be used to perform principal component analysis. Consider what covariates you want to investigate and how your covariate conditions overlap with each other. A potential confounder which overlaps with a potential predictor should be analyzed separately. All factors to be investigated in the present study should be **annotated** and **selected for analysis**.

The **More Plot Information** button provides a description of the plot.

The **Detected/Undetected Calls** button opens a .csv data table that can be viewed, edited, printed, and saved. You may also Save or Save as without opening. **0** indicates data below threshold and **1** indicates data above threshold.

Interpreting Results of Overview Plots

Raw Data Heatmap

This heatmap is generated from raw data and allows quick identification of samples and gene sets with low signal. Each row of the heatmap is a single probe, and each column is a single sample. Colored horizontal bars along the top of the plot identify SNV or Fusion variant status, if applicable, as well as QC flag status and covariate categorization. The blue bar labeled **below threshold** on the left indicates probes whose counts have fallen below threshold in all samples (see the *General Options Menu* section) and will be trimmed out of further analysis for all modules except Probe Descriptive and Related Analytes. Unlike other plots, clicking anywhere on this image will initiate the interactive heatmap in a browser window; clicking again will return you to the original view.

- Dark Blue bars: counts < background (25)
- Light Blue bars: counts < 50
- Grey bars: counts < 100
- Brown bars: counts < 500
- Tan bars: counts ≥ 500

Datasets with exclusively low raw counts (e.g., counts < 100) may arise from experimental failure or low input. Data with expressions near background must be interpreted carefully. You may consider using a higher effective amount of input target.

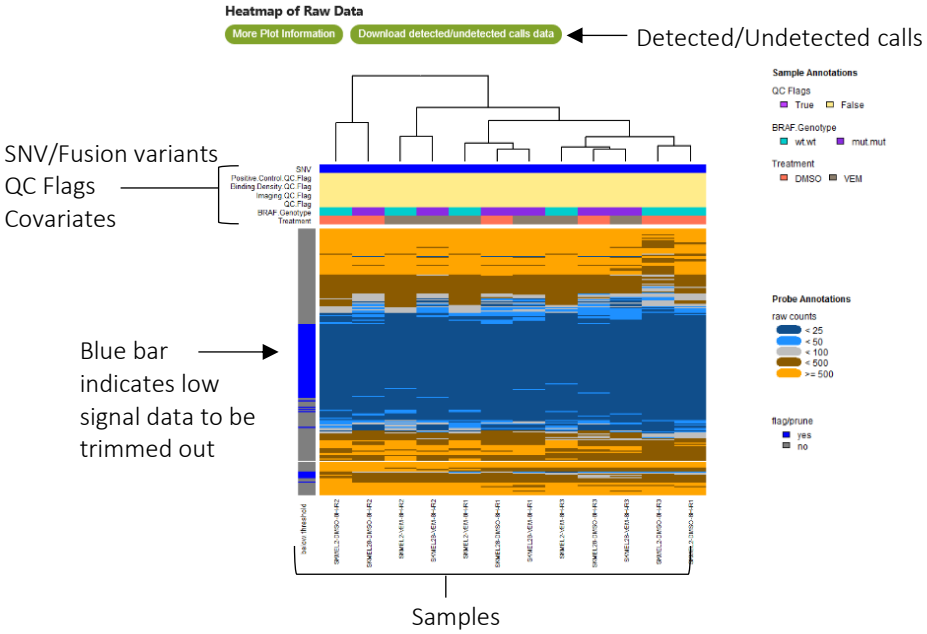


Figure 13: Overview - Heatmap of Raw Data

The **detected/undetected calls** button links to a .csv file stating whether each probe is above background, with 0 indicating below and 1 indicating above background. If you did not specify a detection threshold (see the *General Options Menu* section), probes for mRNA will be called detected if they have more than double the counts of the median negative control.

	A	B	C	D	E	F	G
1		TP53	IL22RA2	IL2	CCR5	PRLR	LIF
2		mRNA	mRNA	mRNA	mRNA	mRNA	mRNA
3	SKMEL2-DMSO-8h-R1_04.RCC	1	0	0	0	0	1
4	SKMEL2-DMSO-8h-R2_04.RCC	1	0	0	0	0	1
5	SKMEL2-DMSO-8h-R3_04.RCC	1	0	0	0	0	1
6	SKMEL2-VEM-8h-R1_10.RCC	1	0	0	0	0	1
7	SKMEL2-VEM-8h-R2_10.RCC	1	0	0	0	1	1
8	SKMEL2-VEM-8h-R3_10.RCC	1	0	0	0	0	1
9	SKMEL28-DMSO-8h-R1_04.RCC	1	0	0	0	1	1
10	SKMEL28-DMSO-8h-R2_04.RCC	1	0	0	0	0	1
11	SKMEL28-DMSO-8h-R3_04.RCC	1	0	0	0	1	1
12	SKMEL28-VEM-8h-R1_10.RCC	1	0	0	0	0	0

above background detection call

Figure 14 :Overview - detected/undetected calls table

Heatmap of All Data

This is a heatmap of the normalized data. This data is plotted by **z-score** and is meant to provide a high-level view of the data and possible associations to covariates of interest. Each row of the heatmap is a single probe, and each column is a single sample. Colored horizontal bars along the top of the plot identify SNV or Fusion variant status, if applicable, as well as QC flag status and covariate categorization. The blue bar labeled **below threshold** on the left indicates probes whose counts have fallen below threshold in all samples (see the *General Options Menu* section) and will be trimmed out of further analysis for all modules except Probe Descriptive and Related Analytes. Clicking anywhere on this plot results in a zoomed-in image; clicking again returns you to the original view.

This plot is scaled with relation to the average probe performance across samples to give all genes equal mean and variance. Hierarchical clustering is used to generate dendrograms.

- Blue: low expression
- Black: average expression
- Orange: high expression

Click anywhere on the normalized heatmap to open an **interactive view**.

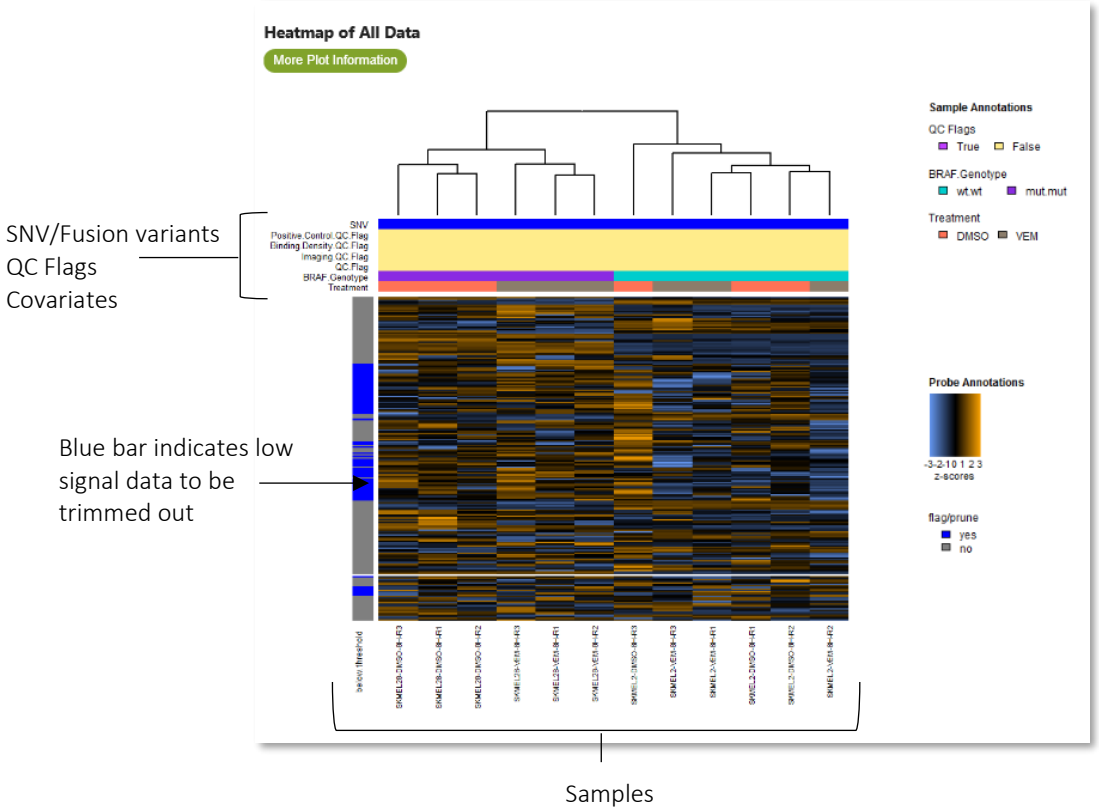


Figure 15: Overview – Heatmap of All Data

Highlight an area on the **selection bars** to the right and below to zoom in. Click on the main image to zoom back out. Click within the main image to open a window which allows you to adjust the plot and label settings. Right-click to save the HTML file or page. Use the **X** in the upper-right corner of the window to return to the original view.

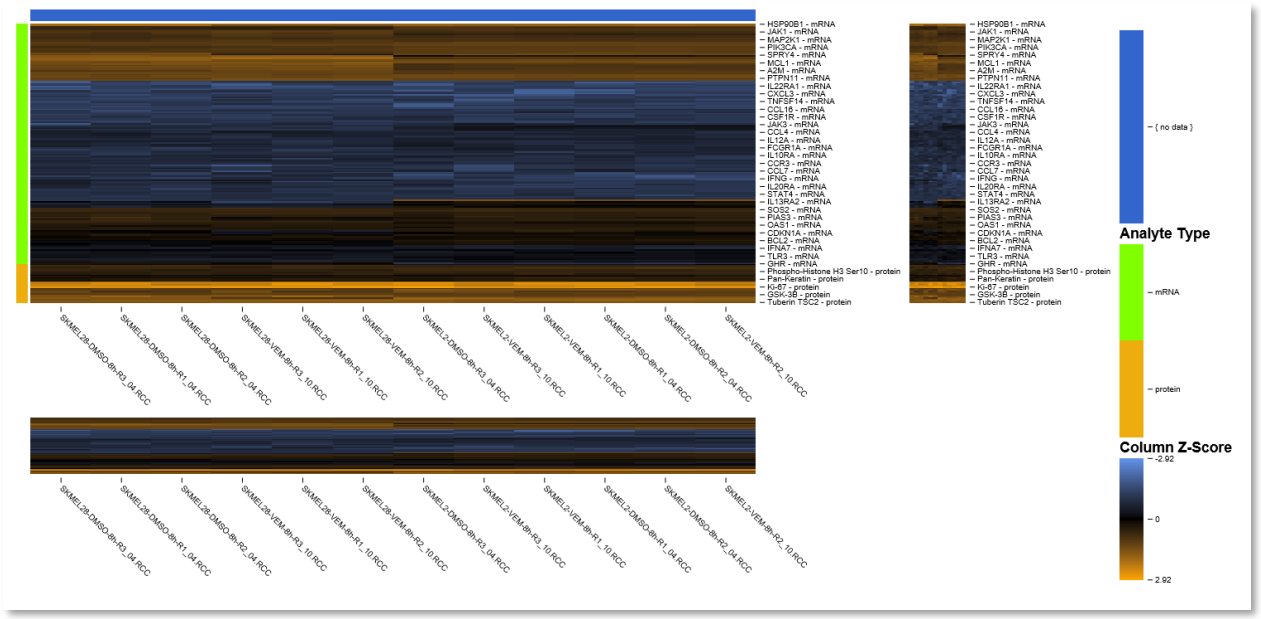


Figure 16: Overview - Heatmap of All Data - zoom view

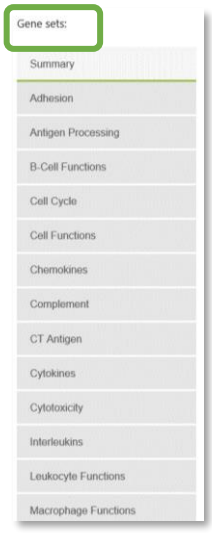


Figure 17: Overview - gene set list

Select one of the gene sets along the left side of the window to view a normalized data heatmap specific to that set of genes.

When you select a particular gene set from the left-hand tabs, a heatmap of normalized data for just the genes in that gene set is displayed. Expression values are centered and scaled. Orange indicates high expression; blue indicates low expression.

.

Principal Component Analysis (PCA)

Use the dropdown menu in the **PCA:** header to choose which covariate to analyze.

Principal component analysis transforms data with multiple variables into a linear set of principal components. Principal component 1 (PC1) captures the highest level of variance, PC2 the next highest, PC3 next, and so on. The resulting image (see Figure 18) plots each PC vs. another twice and colors the points by the selected covariate (once on one side of the diagonal and once on the other). The boxes on the diagonal each contain a PC name; all plots in the same row will have this PC on their y-axis and all plots in the same column will have this PC on their x-axis. Viewing the PCA plot for one covariate and then toggling to another may help identify clusters in the data associated with a covariate.

Figure 18: Overview – PCA plots with different covariates selected

In the **3D Bio Data Example** (see *Appendix A*), a clear separation of **BRAF.Genotype** data points can be seen in the PC1 vs. PC2 results, meaning that changes in this variable cause clear, consistent changes in the data. **Treatment** does not have the same effect.

In the example shown in Figure 18, the first three principal components identify the variability in the data associated with BRAF genotype status. Once you have reviewed the PCA plots for your data, you should then review the covariates plot under Study Design (see below) to recognize any covariates that are highly correlated with the biological covariate. Then, review the PCA plots by that confounding variable. Identified confounding variables can be adjusted for in DE analysis or pathway scoring analysis. If specific gene sets are selected from the **Gene set** tabs along the left side of the window, the same plots will be shown, but only the genes defined in the specific gene set will be used in the analysis. A gene can occur in multiple gene sets.

Outliers may be biologically interesting or caused by technical artifacts such as failed reactions. Samples that were initially flagged by nSolver and now appear as outliers in Advanced Analysis should be treated with caution. Repeat the analysis after excluding outliers and confirm that any important analysis results hold even when these samples are removed.

Study Design

The Study Design tab allows you to look at all the covariates and their relationships.

Examine these plots before viewing the main analysis results. Compare some of the technical covariates (Binding Density, for example) to biological annotations (Subtype, for example). Seeing the distribution of samples among the covariates and conditions may give context to an observed result or suggest changes needed to the experimental design. If one covariate wholly overlaps with another, it will be difficult to discern if one is a predictor and the other a confounder. For example, in an experiment testing different subtypes, if each subtype was scanned on a different date, the scanned date covariate could confound the effect that subtype has on the samples. As an additional example, if samples belonging to different genotypes were correlated with binding density (which is a surrogate for sample input quantity), any conclusions drawn should be based upon adjusting for binding density as a confounder.

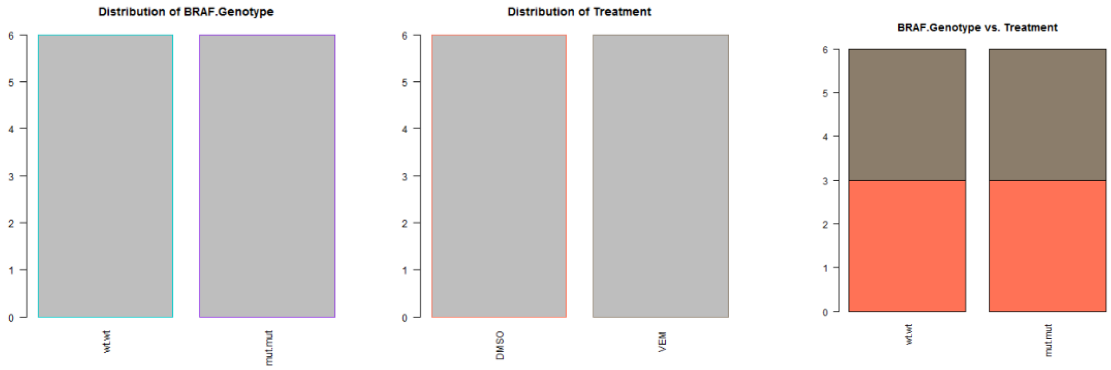


Figure 19: Overview - Study Design

Other QC

Other QC provides two types of analysis graphs, **histograms of p-values** and **mean and variance scatter plots.** The histograms provide a good way to see what variables are having a big impact on your data (see below). This knowledge, combined with what the experimental covariates tell you about what variables are correlated, allows you to separate these variables in Differential Expression analysis and avoid confounding.

Histograms

Covariates with no association with gene expression display mostly flat histograms, and covariates with widespread effects on gene expression have peaks near zero. Technical covariates with such left-weighted histograms may have biological relevance, and it is sometimes advisable to adjust for them in differential expression analyses to avoid confounding.

In some cases, a covariate with no effect will be correlated with a covariate with a powerful effect, producing a left-weighted histogram. In datasets with larger sample sizes, there is little harm in adjusting differential expression analyses for likely unimportant technical variables like Scanned Date, but in smaller datasets, including irrelevant variables will reduce statistical power.

In the **3D Bio Data Example** (see *Appendix A*), the conclusions drawn from the PCA (above) are reinforced by the p-value histograms under the **Other QC** tab, which shows a clear left-weighted plot for **BRAF.Genotype** samples, meaning there are a number of p-vaules in the significant range, close to zero. The **Treatment** p-values are more evenly distributed, indicating lower significance.

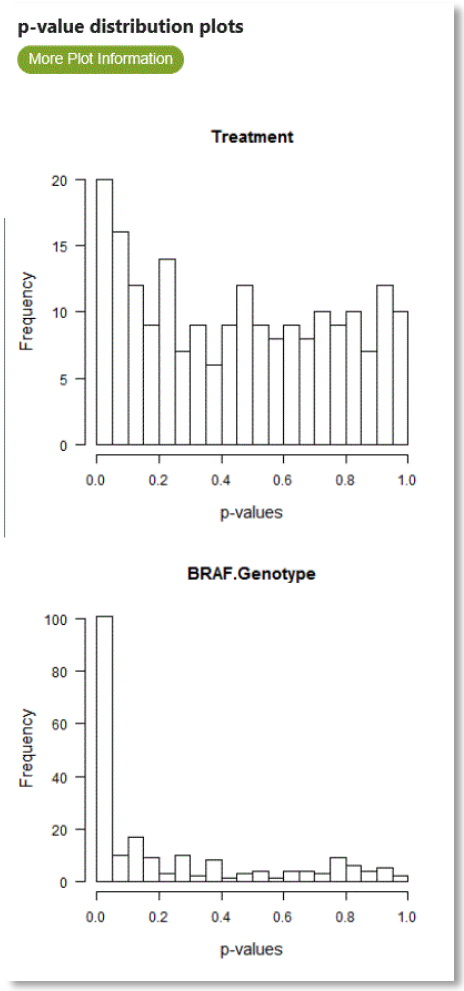


Figure 20: Overview - Other QC p-value Distribution Plots

The Mean and Variance Scatter plot

This plot shows each gene's variance in the log-scaled, normalized data against its mean. Highly variable genes are indicated by gene name. Housekeeping genes are color coded according to their use in (or omission from) normalization. The plot confirms that the selected housekeeping genes are stable (given their low variability and moderate expression levels). This plot highlights genes that are expressed at moderate to high levels and show great variability; these may be of interest for further study.

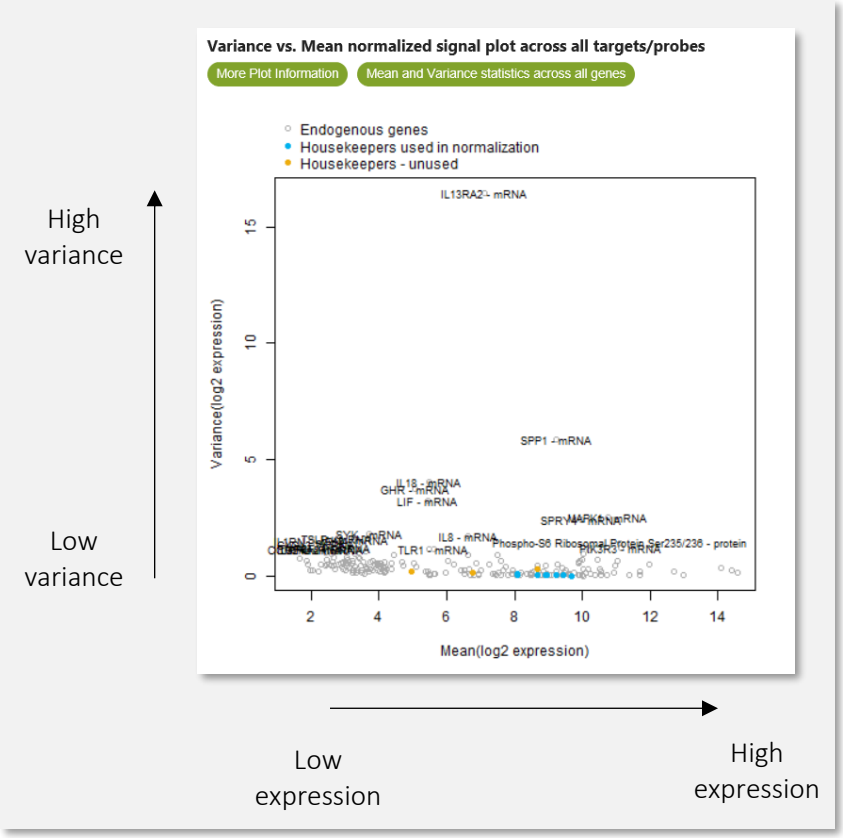
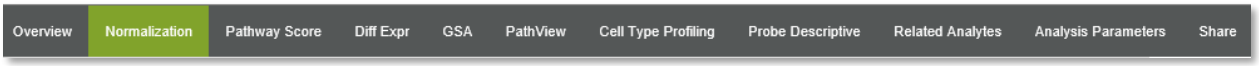


Figure 21: Overview - Other QC - Mean and Variance Scatter Plot

The **More Plot Information** button provides a description of the plot.

The **Mean and Variance statistics across all genes** button opens a .csv data table that can be viewed, edited, printed, and saved. You may also Save or Save as without opening. It provides the average normalized count and the variance normalized count for each probe.

Normalization Module



Data normalization seeks to eliminate run-to-run and sample-to-sample technical variability in the raw counts, which arises from inconsistencies in effective sample input and fluctuations in the overall efficiency in capturing and counting target molecules. This module normalizes each analyte-type separately, resulting in clickable analyte-type tabs which reveal respective plots.

For **mRNA data**, the Normalization module displays two plots: the **Pairwise Variance During HK Selection** plot, detailing the selection process of the geNorm algorithm (Vandesompele, 2002), and the **Normalization Summary**, which summarizes the performance of these chosen normalization genes.

For **Protein data**, the Normalization module displays three plots. The first is the **Probe Stability** plot, which ranks the stability of all proteins in the dataset and selects the 15 most stable probes for normalization. Second, is the **Normalization Summary**, which summarizes the performance of these chosen normalization genes. The third plot is the **Protein Expression Threshold** plot, which visualizes the background-subtracted normalized counts for each of the analyzed proteins (counts lower than zero after background subtraction are thresholded to zero).

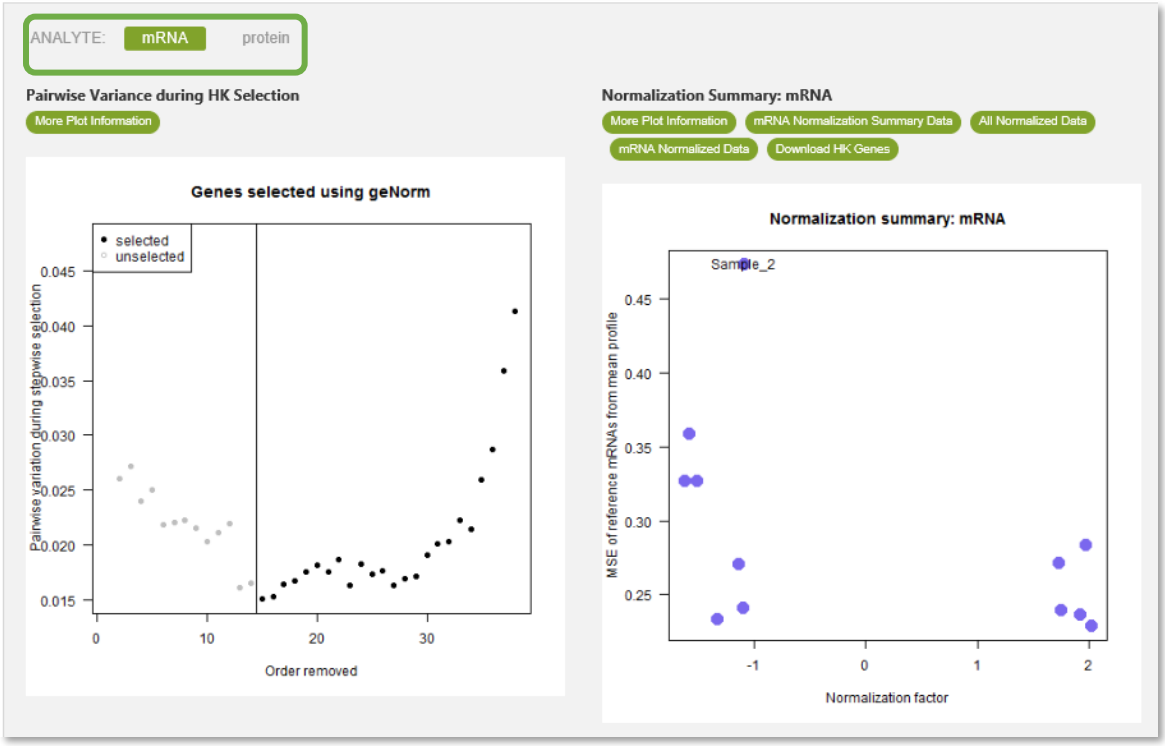


Figure 22: Normalization module window and options

Before You Start Normalization

Advanced Analysis does not automatically detect whether input data is **raw or normalized.** Raw data is usually the preferred selection, as the built-in algorithms help to determine the best normalization probes. Please note that normalization performed using the Advanced Analysis module will override any normalization previously performed in nSolver.

Because **multiRLF Merge experiments** originate from multiple nSolver experiments (whose data has presumably already been normalized), the Normalization module will not be available (will appear greyed out).

Most commercial CodeSets come with pre-identified potential reference genes. The built-in geNorm algorithm will determine the best performing of those reference genes and use them for normalization.

Custom Options for Normalization

The Normalization Parameters tab allows you to specify parameters to **normalize mRNA and Protein probes independently**.

For each analyte type detected in the data, select **automatic** or **manual** methods for choosing Normalization/Reference genes. Automatic normalization is the default; check the **Refine the list** box to customize this list. Manual normalization allows you to specify candidate normalization probes and refine it to a list consisting of at least 5 normalization probes

The question mark button reveals additional information.

The exclamation mark button reveals an alert and brief explanation as to why an option may be unavailable (greyed out).

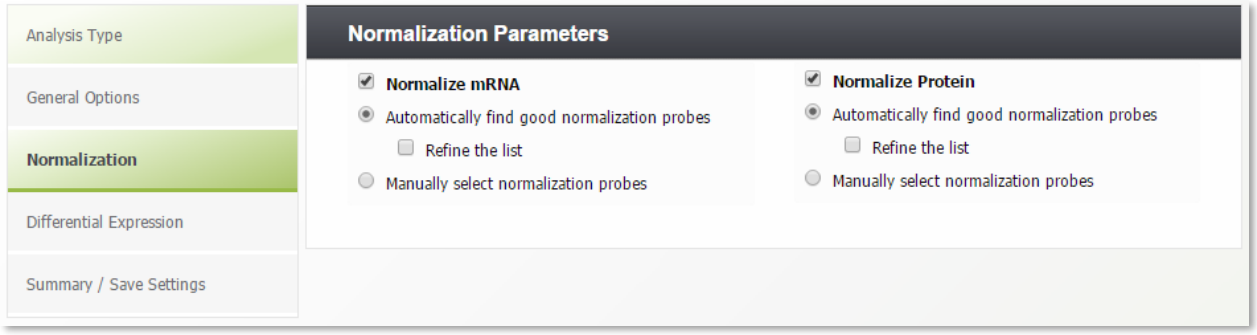


Figure 23: Normalization Custom Analysis menu

Interpreting Results of Normalization Plots

mRNA Plots

For **mRNA**, in the **Pairwise Variance during HK Selection** plot, the ideal normalization genes are determined by selecting those that minimize the pairwise variation statistic. We see the order in which genes have been eliminated from the list of stable housekeepers (HK) as we travel along the x-axis. The y-axis depicts the measure of pairwise variation, which is re-calculated as the housekeeper pool gets smaller and smaller. The final two genes are not displayed, since the statistic can no longer be calculated. Pairwise variation will drop as the less-stable reference genes are removed. At a certain point, the program will determine that removing any more reference genes will begin to increase pairwise variation again; this signifies that it has reached the most optimal arrangement and that the most stable reference genes have been identified.

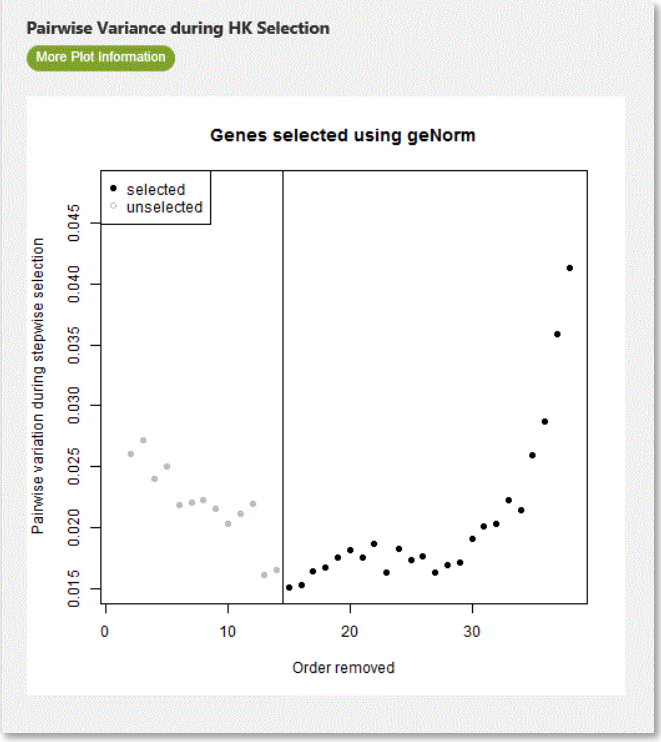
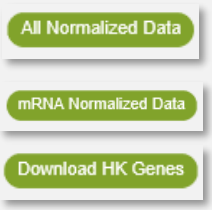


Figure 24: Normalization - mRNA Pairwise Variance plot

The **More Plot Information** button provides a description of the plot.



Each button opens a .csv data table that can be viewed, edited, printed, and saved. You may also Save or Save as without opening. Respectively, they provide all normalized data, mRNA normalized data, and a list of the housekeeping genes and the order in which they were chosen by geNorm.

For **mRNA**, the **Normalization Summary** depicts samples by their normalization factor on the x-axis and their Mean Squared Error (MSE) on the y-axis. As the normalization factor for a sample increases on the x-axis, the standard error of the reference genes decreases. Samples with lower counts will therefore have noisier data.

The overall quality of the normalization decreases as the MSE increases on the y-axis.

Samples with MSE values far outlying the other samples are designated with their sample names on the plot. For these samples, the chosen reference genes might not be effective in their normalization. The list of selected housekeepers can be downloaded by selecting the **Download HK Genes** button.

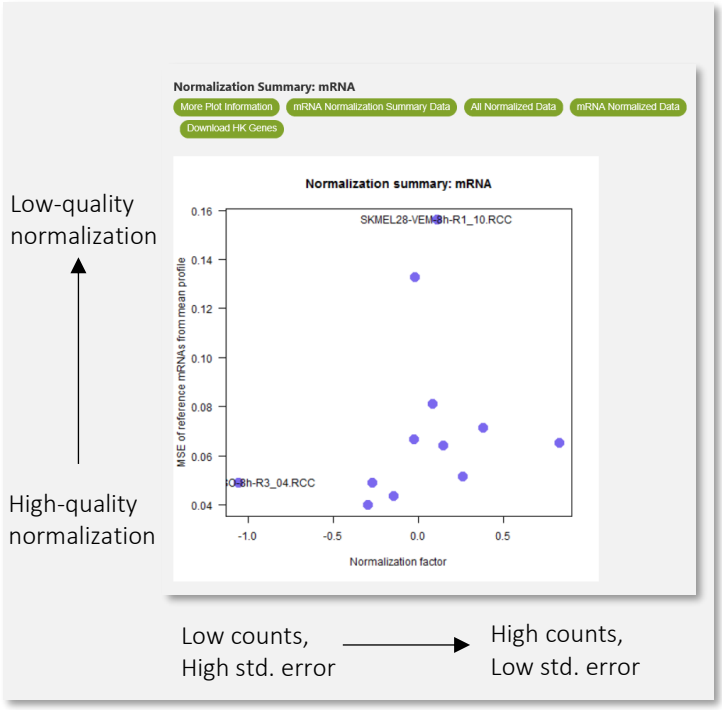


Figure 25: Normalization - mRNA Normalization Summary plot

Protein Plots

In the **Protein Stability** plot, a measure of stability, the Mean Absolute Deviance (MAD), is determined. Ideal proteins for normalization have low MAD. Stability in this context is defined in terms of how closely each probe follows the sample average fold change. The intuition behind this method is that the average (across all Protein probes) up/down fold change for a sample relative to the median expression profile (across all samples) roughly estimates the normalization factor. In this setting, an ideal normalizer probe in any sample shows a small deviation from the average fold change of that sample, and this is used to rank Protein probes on how closely they resemble ideal normalization candidate probes.

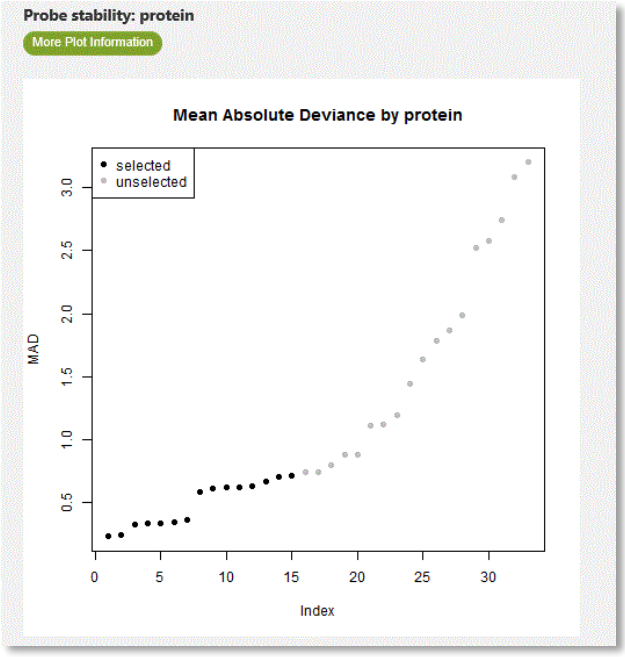


Figure 26: Normalization - Protein Stability plot

For **protein**, the **Normalization Summary** depicts samples by their normalization factor on the x-axis and their Mean Squared Error (MSE) on the y-axis. As the normalization factor for a sample increases on the x-axis, the standard error of the reference genes decreases. Samples with lower counts will therefore have noisier data. The overall quality of the normalization decreases as the MSE increases on the y-axis.

Samples with MSE values far outlying the other samples are designated with their sample names on the plot. For these samples, the chosen reference genes are not effective in their normalization. The list of selected normalizers can be downloaded by selecting the **Download Normalizer Proteins** button.

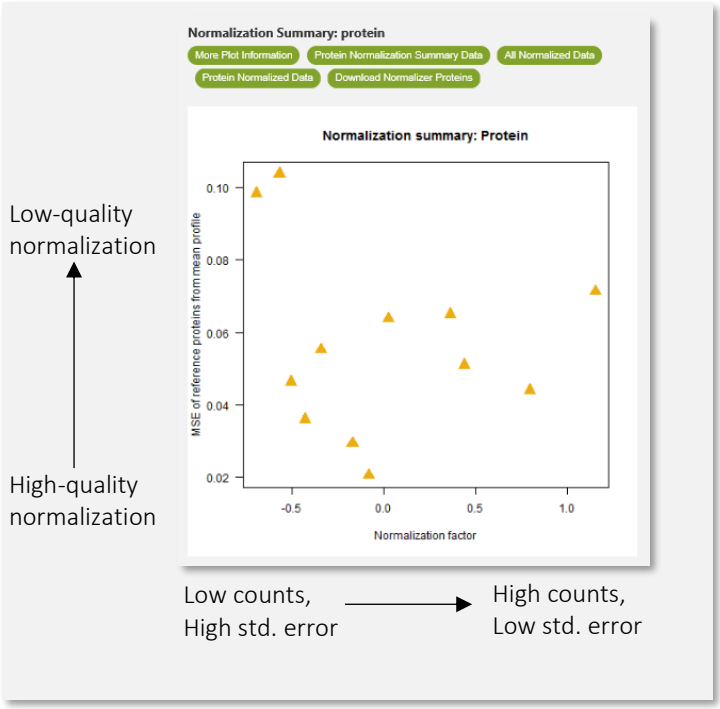


Figure 27: Normalization - Protein Normalization Summary plot

For **protein**, the **Expression Threshold plot** depicts log-normalized expression thresholded to zero based on the by-lane background level. This plot relies on the mean of the negative antibody results and an estimation of error. To estimate error, the module uses either the standard deviation (if there are three antibodies to work with) or the deviation from the PC1 best fit line (if there are only two antibodies to work with).

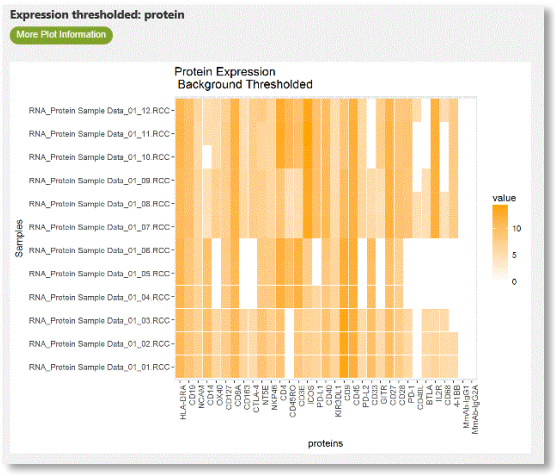


Figure 28: Normalization - Protein Expression Threshold plot

Normalization Algorithm Details

As both sample input and reaction efficiency are expected to affect all probes uniformly[1], normalization for run-to-run and sample-to-sample variability is done by dividing counts within a lane by the geometric mean of the reference/normalizer probes from the same lane (i.e., all probes/count levels within a lane are adjusted by the same factor) [2].

geNorm selection of housekeeping genes

Normalizer probes can be specified by the user. When not specified by user (default), normalizer probes are selected using the widely used geNorm algorithm (Vandesompele, 2002) as implemented in the Bioconductor package NormqPCR. While expression of a good housekeeping gene may vary between samples in non-normalized data, the ratio between two good housekeepers should be very stable. In other words, good housekeepers are expected to rise and fall together and at the same rate. geNorm relies on this behavior to iteratively remove candidate housekeepers with the least stable expression relative to other candidates. geNorm is implemented in the Advanced Analysis module through the function *selectHKs* in the NormqPCR package. This function, using the geNorm algorithm, ranks genes on the V number (variation between successive norm factors as reference genes are removed). Genes are excluded when their V number is equal to or less than the smallest V number for all the genes plus one (Vmin + 1).

To understand the how geNorm is implemented, consider the case where we have:

- **n samples**
- **p candidate housekeeping genes**
- **$gene_i$ and $gene_j$**, which are the raw expression of any pair of genes, i and j, respectively, from this set of p genes.

For each of the n samples we could compute: **$\log_2(gene_i/gene_j)$.** Taking the standard deviation of the n log ratio values gives a statistic that captures how these two genes in our sample set deviate from perfect co-expression (as perfect co-expression across sample set would result in the standard deviation of zero). We can call this value **V_{ij}**.

For $gene_i$, we can calculate **Vi_1 through Vi_p** and then take their average value to represent the average degree of dissimilarity in expression pattern between $gene_i$ and all the other genes in the set of candidate genes. This value was called **gene stability measure, M_i**, by Vandesompele and colleages.

[1] This assumption holds true from empirical observations when expressions are not near the background counts. Deviation from this assumption becomes stronger for expression nearing the background.

[2] This normalization does not account for any batch effect that may exist if data from multiple CodeSet batches are being analyzed together in the same study. In the case of multiple batches, we recommend the use of reference or calibration samples to quantify and adjust for variability in probe efficiency across batches of CodeSet before any subsequent analysis is performed. Some of the modules (e.g., DE) allow adjustment for technical variables such as batch effect, however, when the experimental conditions and batch effect are confounded, we cannot correct for the batch effect and use of a reference sample is needed.

- The larger M_i, the more dissimilar the pattern of expression of gene i to the other candidate housekeeper genes in the set.
- M can thus be used to rank the candidate housekeepers from the **best (lowest M)** to **worst (highest M)** in terms of their similarity (co-expression) to other candidate housekeepers.
- We could subsequently choose the top 5 or 10 or however many housekeeper genes we believe will be optimal.

If the optimal number of genes to be selected is not known, we can use an iterative process to select the optimal number of housekeepers from among the candidates. Tracing the variation statistic as genes are removed iteratively can allow us to find the point at which variation is minimized and relatively stable; this is the point with the optimal number genes. Given **n** x **p** matrix of n samples and p candidate housekeeper genes, the housekeeper selection proceeds as follows:

- Compute the normalization factor (NF) as the log geomean of the p genes for each sample to get NF_p.
- Compute M_1 through M_p.
- Remove the gene with highest value of M.
- Re-compute the normalization factor as the log geomean of the remaining p-1 genes for each sample to get NF_{p-1}.
- Evaluate the stability of the new normalization factor, NF_{p-2}, by quantifying the change between NF_p and NF_{p-1}. This is by variation statistic: $Vp/p-1 = \text{standard deviation } (NF_p - NF_{p-1})$.
- Re-compute M for each of the remaining p-1 genes.
- Remove the gene with the highest M.
- Repeat until all but the last two genes are removed.

Normalization

Consider the graph plotting the (raw) log count of normalizing probe in each of the samples against the (raw) log geometric mean of those probes. In this context, points corresponding to each sample follow a trend line with slope 1 and an intercept that captures run-to-run variability. In this setting, adjusting for run-to-run variability simply involves subtraction of the intercept, which should bring the values corresponding to any of the normalizing probes across all samples very close to one another (with some variability due to noise). This is what we may expect when normalization is working well.

If there is large deviation from the expected line of slope 1, substantial variability will remain even after subtraction of the intercept. This can be an indication of poor normalization quality. In Figure 29, sample A has a positive normalization factor (indicating larger-than-average expression levels of normalizing genes) and sample B has a negative normalization factor (indicating smaller-than-average expression levels of normalizing genes). Additionally, we observe that the mean squared error, MSE, of sample A is larger than the MSE of sample B, which may be taken as sample A not conforming to the described normalization adjustment model as well as sample B.

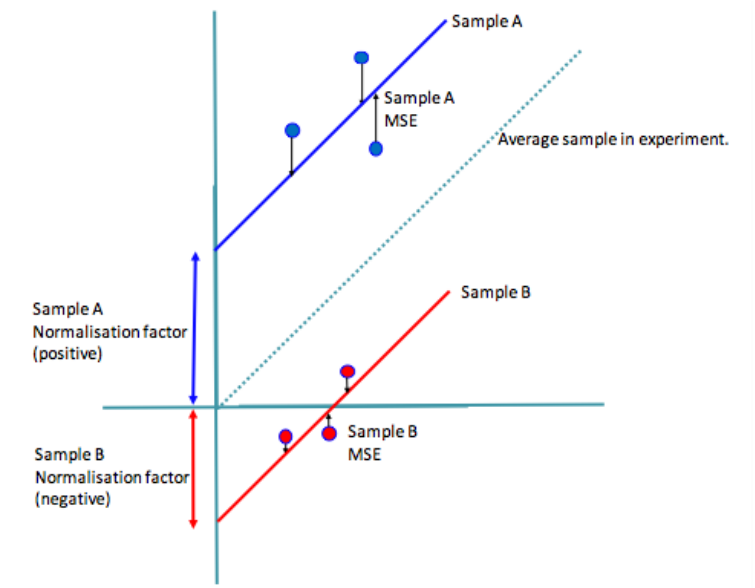


Figure 29: Diagrammatic representation showing how the values for the Normalization summary are generated

Protein Expression Threshold

With the mean and error known, the background threshold for each line of the Protein Expression Threshold plot is equal to the mean of negative antibodies for the lane + 1.96* estimated error. In the heatmap, any antibody value below its estimated background is set to zero.

Differential Expression Module

The Differential Expression (DE) module is used to identify the specific targets which exhibit significantly increased or decreased expression in response to the chosen covariate. This module provides the basis for the Gene Set Analysis (GSA) and PathView modules and should be viewed prior to both. GSA explores differential expression of particular pathways and PathView shows that differential expression in the context of pathway figures. The DE module lays the ground work for this, which can be seen in the volcano plot and the significant gene table.

The **DE volcano plot** displays each target's -$\log_{10}$ (p-value) and $\log_2$ fold change with respect to the selected covariate. Highly **statistically significant targets** fall at the top of the plot, and highly **differentially expressed genes** fall to either side. Green point colors and horizontal lines indicate various False Discovery Rate (FDR) thresholds.

The **40 most statistically significant targets** are named in the accompanying chart.

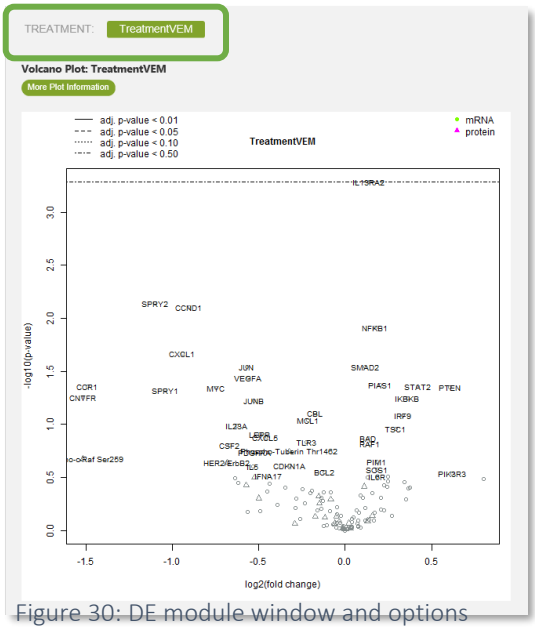


Figure 30: DE module window and options

The **More Plot Information** button provides a description of the plot.

The **Download CSV Data** button opens a .csv data table that can be viewed, edited, printed, and saved. You may also Save or Save as without opening. The name of this button will vary with the type of data available for download.

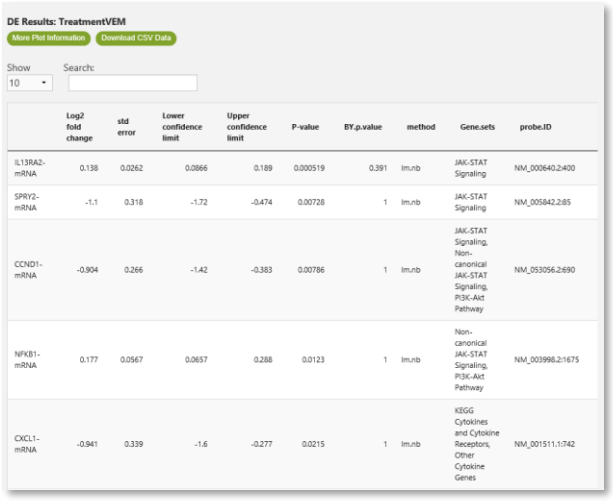


Figure 31: DE - statistically significant targets table

Before You Start Differential Expression

Multivariate DE analysis requires thoughtful setup. Sample covariates include predictors, variables that are scientifically interesting or at the heart of the study, confounders, technical variables that impact expression but are of no interest to the study, and uninteresting variables that do not impact expression. The linear regressions treat predictors and confounders identically, but results are only reported for predictors. It is recommended that you **base your covariates on factors that are scientifically believed to account for (explain) gene expression** in your system. In addition:

- Ensure that any **variables you include do not strongly correlate with each other**, and similarly, ensure two or more categorical variables don't have redundant categories (see Figure 32). This essentially nullifies the effect of both variables and the DE analysis will randomly drop one or both from the model. Correlation and level-redundancy can be detected using the Study Design tab of the Overview module. See the *Study Design* section.
- **At least one variable needs to be chosen as a predictor** (if using Custom Analysis); additional variables may be selected as predictors or confounders. See the *Custom Options for Differential Expression* section.
- **Models with fewer variables are preferable.** Generally, linear regression weakens as the ratio of variables to the number of samples grows since including too many covariates in a model can diminish its ability to detect the effects of the variable you care most about.
- Similarly, when working with categorical variables, **models with fewer categories are preferable.** Comparison of each category to the reference category is treated as another variable; adding categories is equivalent to adding additional variables, weakening the ability of the model to determine the effect on expression.

In the example in Figure 32, the *normal* category in the *Type* column overlaps completely with the *normal* category in the *Subtype* column. Not only is the *Type* annotation less informative than the *Subtype* annotation, but DE module may have a difficult time with this. To remedy this, the *Type* column should be dropped.

	Type	Subtype
1	cancer	type1
2	cancer	type1
3	cancer	type1
4	cancer	type2
5	cancer	type2
6	cancer	type2
7	normal	normal
8	normal	normal
9	normal	normal
10	normal	normal
11	normal	normal
12	normal	normal

Figure 32: Annotation example - redundant variables may cause DE analysis to fail

Custom Options for Differential Expression

Custom Analysis can effectively isolate the effect of multiple covariates on gene expression and avoid confounding due to technical variables by allowing multiple predictors and confounders to be included in the multiple regression model.

Highlight the **Available Annotations** of choice and move at least one to the **Selected Predictors** window with the green arrows. You may designate **Selected Confounders**, as well, if desired.

The **Fast/Approximate** method for estimating DE can be used for most datasets, but **Optimal**, although more time consuming, is more accurate for low count data and should be used for datasets with low input samples or a high degree of low count targets.

The question mark button reveals additional information.

The exclamation mark button reveals an alert and brief explanation as to why an option may be unavailable (greyed out).

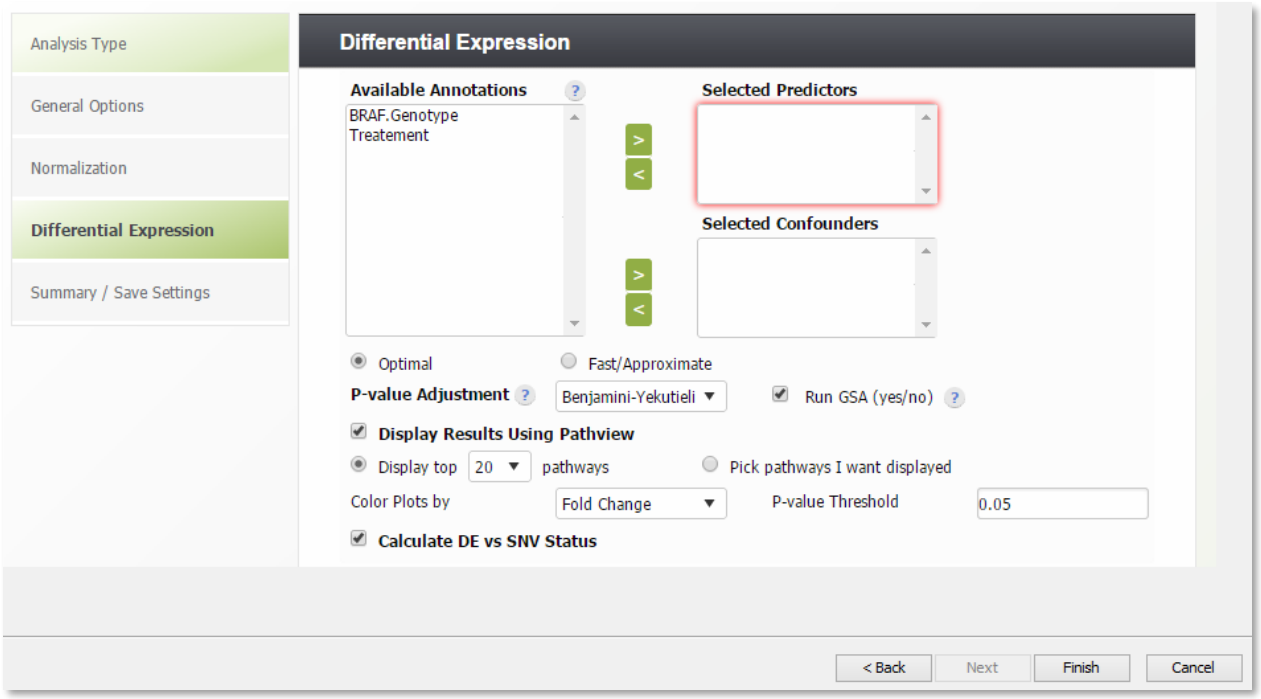


Figure 33: DE Custom Analysis menu

Since nCounter data is multiplex in nature, we provide the option to apply an adjustment to the p-values before plotting them in DE to correct for the high number of comparisons. You can select **none** if you would prefer raw p-value thresholds throughout DE plots. There are three methods for **P-value Adjustment**:

- The **Bonferroni** correction is a very conservative approach to multiple testing: it multiplies each p-value by the number of genes tested. Although genes with low Bonferroni-corrected p-values have very strong evidence for differential expression, many genes worth consideration may be ruled out by this method.
- The **Benjamini-Yekutieli** method returns moderately conservative estimates of false discovery rate (FDR), but, importantly, makes the assumption there may be some biological connection between

genes. FDR is the proportion of genes with equal or greater evidence for differential expression (i.e. equal or lower raw p value) that are expected to be "false discoveries" due to chance. For example, if a gene has p = 0.02 and FDR = 0.25, then 25% of the genes with p ≤ 0.02 are expected to be false discoveries.

- **Benjamini-Hochberg** is a method of estimating FDR that assumes that the genes and variable studied do not have an impact on each other. This would be the best choice when it can be assumed that the majority of targets and covariates studied don't have a common biological/ functional focus.

As introduced earlier, the DE results can be viewed through the optional Gene Set Analysis (GSA) and PathView modules. To run those, select the **Run GSA** and **Display Results Using PathView** boxes.

- GSA will result in summary heatmaps as well as labeling of the DE volcano plot, such that the genes of each pathway are highlighted. See the *Gene Set Analysis (GSA)* section.
- Selecting to display results using **PathView** will then allow you to **display the top 20 pathways** or choose a different number (the analysis time will increase with the number of pathways requested). You may also choose to display a **hand-picked selection of pathways** to view. The software will overlay DE information over each pathway figure.
- You may choose to **Color Plots by** fold change or T-statistic, and choose a **P-value Threshold**.

Interpreting Results of Differential Expression Plots

Volcano Plots

The differential expression results are displayed as a volcano plot for each variable chosen as a predictor in the regression analysis and table. A volcano plot visualizes the results for the chosen covariate in the DE model. If using Quick Analysis, you will only have one covariate's analysis to view. If you chose Custom Analysis and chose multiple covariates, you can click on the buttons above the plots to choose which covariate's analysis to view.

The genes of greatest interest will be both high in the graph (corresponding to a very small p-value) and at either the right or left side (corresponding to greatly increased or decreased expression). mRNA probes will be displayed as solid circles and Protein probes as triangles. Note the following:

- Note where the p-value thresholds lie and how much of your data is above the significance threshold for your study (and is therefore appearing significant). If you selected a p-value adjustment on the Custom Analysis menu, your thresholds will reflect the adjusted p-value, whereas the axes will be based on raw p-values.
- Points above your p-value threshold will be shown in color (mRNA in purple, protein in gold). See Figure 34a. If all points are uncolored and there are no thresholds on the plot (as in Figure 34b), this indicates that none your data points have p-values at a significant level.
- Data points should often be fairly spread across the plot (and not clustered to one side, for example); if not, check normalization settings and explore if there is a biological reasons for this skewing.

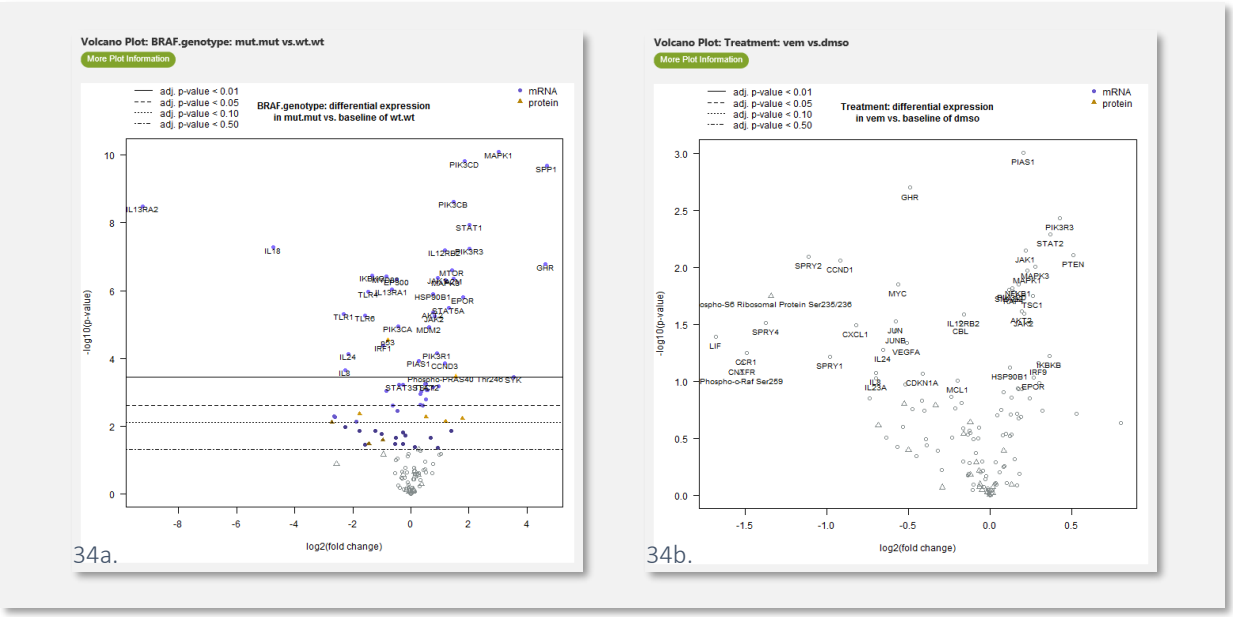


Figure 34: DE - Volcano Plots

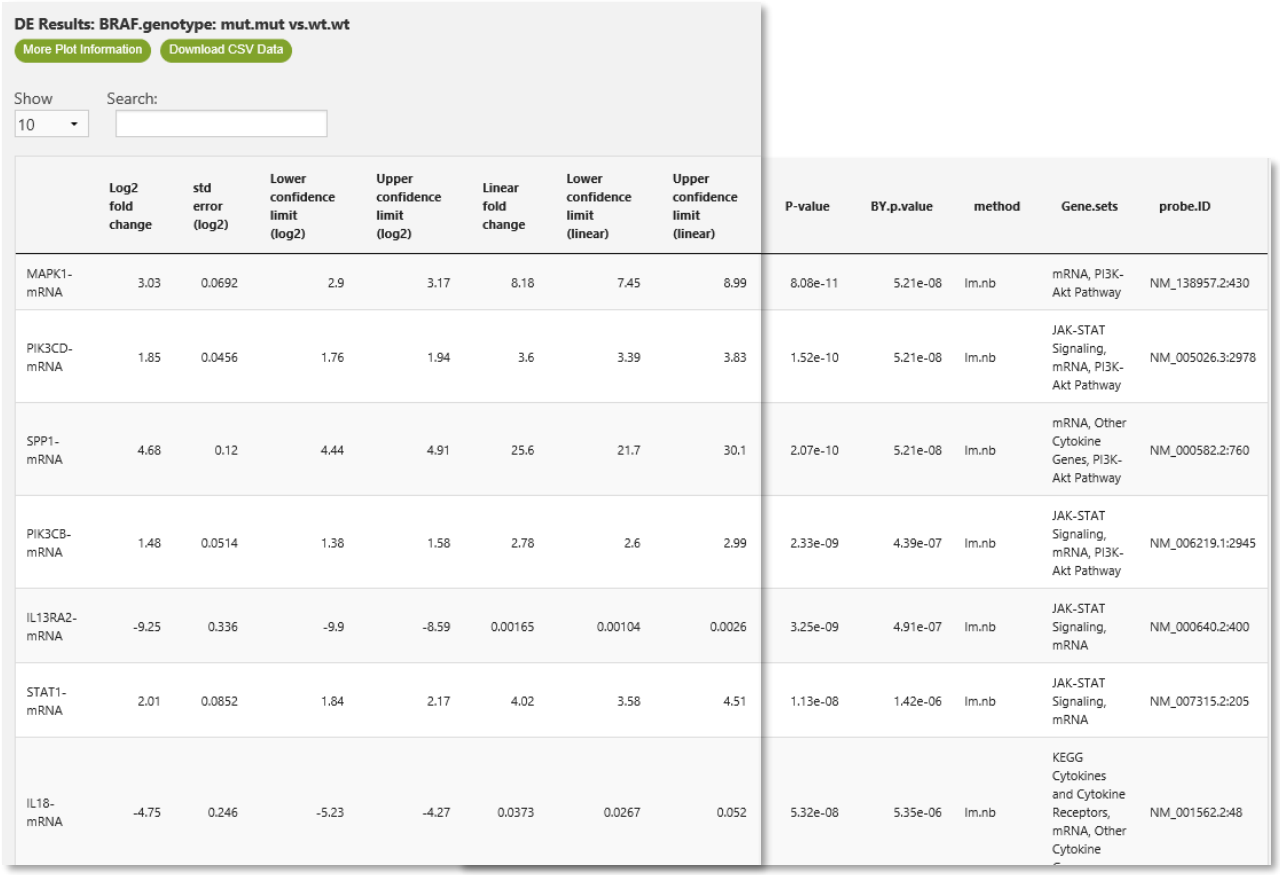
In the **3D Bio Data Example** (see *Appendix A*), the **Volcano Plot** for the covariate **BRAF.Genotype** depicts the differential expression of genes in mut/mut samples relative to the wt/wt samples. It shows multiple p-vaule (significance level) thresholds (Figure 34a, above). Only probes with p-values in the significant range

are colored and named. Viewing this plot under the Treatment tab shows a colorless plot with no p-value thresholds (Figure 34b), indicating Treatment did not result in signficant gene expression changes.

Significant Genes Table

The corresponding table presents the genes with the lowest p-values for differential expression with respect to the selected covariate. The estimated log fold-change represents the average magnitude of a gene's differential expression. Note the following:

- For **categorical** covariates, a gene is estimated to have $2^{\text{log fold-change}}$ times its expression in baseline samples, holding all other variables in the analysis constant.
- The 95% confidence interval for the log fold-change is also presented, along with a p-value and an adjusted p-value or FDR if requested.
- For **continuous** covariates, for each unit increase in the selected covariate, a gene's expression is estimated to increase by $2^{\text{log fold-change}}$ fold, holding all other variables in the analysis constant.
- Log fold-change values have a slightly different interpretation for continuous variables. For continuous variables, the fold-change must be read in the context of the range of the variable. If the variable has a small range, a unit increase is a huge difference, and large log fold-changes are to be expected. In contrast, if we studied the covariate “drug dose in milligrams,” we would expect very small estimated log fold-changes, not because the drug has a small effect but because an extra 1 mg of the drug has a small effect.



DE Results: BRAF.genotype: mut.mut vs.wt.wt

More Plot Information | Download CSV Data

Show 10 | Search:

	Log2 fold change	std error (log2)	Lower confidence limit (log2)	Upper confidence limit (log2)	Linear fold change	Lower confidence limit (linear)	Upper confidence limit (linear)	P-value	BY.p.value	method	Gene.sets	probe.ID
MAPK1-mRNA	3.03	0.0692	2.9	3.17	8.18	7.45	8.99	8.08e-11	5.21e-08	lm.nb	mRNA, PI3K-Akt Pathway	NM_138957.2:430
PIK3CD-mRNA	1.85	0.0456	1.76	1.94	3.6	3.39	3.83	1.52e-10	5.21e-08	lm.nb	JAK-STAT Signaling, mRNA, PI3K-Akt Pathway	NM_005026.3:2978
SPP1-mRNA	4.68	0.12	4.44	4.91	25.6	21.7	30.1	2.07e-10	5.21e-08	lm.nb	mRNA, Other Cytokine Genes, PI3K-Akt Pathway	NM_000582.2:760
PIK3CB-mRNA	1.48	0.0514	1.38	1.58	2.78	2.6	2.99	2.33e-09	4.39e-07	lm.nb	JAK-STAT Signaling, mRNA, PI3K-Akt Pathway	NM_006219.1:2945
IL13RA2-mRNA	-9.25	0.336	-9.9	-8.59	0.00165	0.00104	0.0026	3.25e-09	4.91e-07	lm.nb	JAK-STAT Signaling, mRNA	NM_000640.2:400
STAT1-mRNA	2.01	0.0852	1.84	2.17	4.02	3.58	4.51	1.13e-08	1.42e-06	lm.nb	JAK-STAT Signaling, mRNA	NM_007315.2:205
IL18-mRNA	-4.75	0.246	-5.23	-4.27	0.0373	0.0267	0.052	5.32e-08	5.35e-06	lm.nb	KEGG Cytokines and Cytokine Receptors, mRNA, Other Cytokine	NM_001562.2:48

Figure 35: DE - Significant Genes Table

Differential Expression Algorithm Details

Data model

Let y_j be the count of probe in sample *j* after normalization to the housekeeping probes, j = 1,…,*J*. We assume y_j is the sum of the background noise z_j and the true expression x_j, where z_j and x_j follows negative binomial (NB) distribution:

$z_j \sim NB(\lambda_b, \phi_b)$,

$x_j \sim NB(\mu_j, \phi)$,

$\log(\mu_j) = X_j^T \beta$.

The negative binomial model contains a dispersion parameter ϕ. It accommodates the variance of probe expression within biological replicates, which is not of interest in differential expression (DE) analysis. When $\phi = 0$, the negative binomial model reduces to the Poisson model.

Notations

λ_b: mean of background noise. Estimated using all negative controls in all samples.

ϕ_b: dispersion of background noise. Estimated using all negative controls in all samples.

μ_j: mean expression in sample *j*.

ϕ: dispersion.

X^T: *J*×*P* matrix for the sample annotation. *P* is the number of covariates including the intercept term. *J* is the number of samples.

X_j^T: the j^{th} row of X^T, annotation of sample *j*.

β: *P*×1 matrix for the parameter.

Estimation of model coefficients

Model 1: Mixture negative binomial model

- **Parameter estimation:** Let $f(x|\mu, \phi)$ be the probability mass function (PMF) of the negative binomial distribution with mean parameter μ and dispersion parameter ϕ. The marginal probability mass function for y_j can be derived as

$$\mathrm{p}(y_j|X_j, \beta, \phi) = \sum_{x=0}^{y_j} f\left(x\middle|e^{X_j^T\beta}, \phi\right) \cdot f(y_j - x|\lambda_b, \phi_b)$$

  The log likelihood function is

$$L = \sum_j \log \mathrm{p}(y_j|X_j, \beta, \phi)$$

  The parameter μ and ϕ are estimated by maximum likelihood method:

$$\hat{\beta}, \hat{\phi} = argmax_{\beta,\phi} L$$

  This can be obtained via the **MLE** function in **R/stats4**.

- **Inference and p-value calculation:** If the MLE exists, the hessian matrix at MLE is evaluated:

$$H = \frac{\partial^2 L}{\partial \beta \, \partial \beta^T} | \beta = \hat{\beta}, \phi = \hat{\phi},$$

  and the variance-covariance matrix is H^{-1}.

  The test hypothesis is:

$$H_0: \beta_p = 0 \text{ vs } H_1: \beta_p \neq 0,$$

where *p* is the index of the covariate in the design matrix.

The Wald test is conducted and the test statistic is:

$$S = \frac{\hat{\beta}_p}{\sqrt{H_{pp}^{-1}}}$$

where H_{pp}^{-1} is the p^{th} element on the diagonal of H^{-1} matrix.

Model 2: Simplified negative binomial model

The complexity of algorithm in model 1 is proportional to the total count of the probe, which can result in long computation time for probes with large counts. Model 1 can be simplified to the following form when y_j is significantly greater than the background mean λ_b:

$$p(y_j|\beta, \phi) = f\left(y_j - \lambda_b \middle| e^{X_j^T \beta}, \phi\right)$$

The maximum likelihood estimate of β is then obtained using the **glm.nb** function in **R/MASS**.

Model 3: Log-linear model (linear regression)

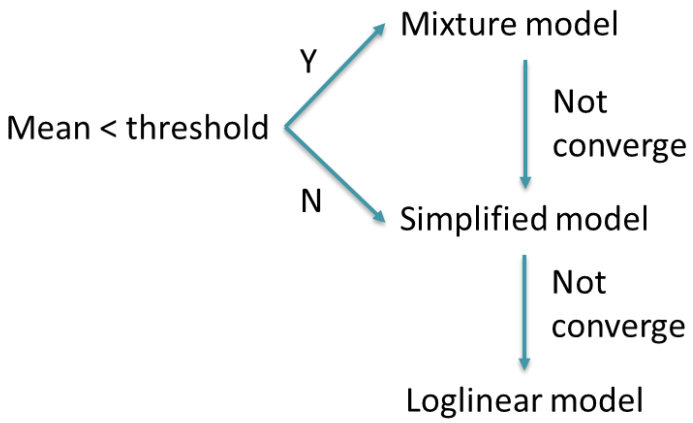
In case the algorithms in 1 and 2 fail to converge and lead to unstable estimate of the parameters, log transformation is taken on the counts. Assume normal distribution of the log transformed data:

$$\log(y_j - \lambda_b) \sim N(X_j^T \beta, \sigma^2)$$

The maximum likelihood estimate of β is obtained via **lm** function in **R**.

Flow of algorithm

The flow of the algorithm works as follows: the mean of the gene across all samples is compared against the threshold, where the threshold is 10-fold of background noise. If the gene mean is above the threshold, the mixture model in 1 is simplified to 2. If mixed model in 1 does not converge, the simplified model in 2 is applied instead. If model 2 does not converge, the loglinear model in 3 is used.



Variables

If your variable is categorical, you will be asked to assign one level of the category as the baseline or reference level. If a category has a reference level (normal) and levels A, B, C and D as well as another covariate, Binding Density (whether confounder or predictor), a linear regression will be run for each gene using the following model:

$$E \log_2(\text{expression}) = ß_0 + ß_1(I_A) + ß_2(I_B) + ß_3(I_C) + ß_4(I_D) + ß_5(\text{binding.density})$$

Depending on each sample label (whether Normal, A, B, C or D), only one of I_{Normal}, I_A, I_B, I_C, I_D will take on value 1 and the rest will be 0. (Note that I_{Normal} is not in the model and its coefficient value is absorbed by $ß_0$ term). binding.density here is a continuous variable.

Optimal Method

For each gene, the Optimal method (see the *Custom Options for Differential Expression* section) infers differential expression with respect to specified covariate(s) using a negative binomial mixture model for low expression probes or a simplified negative binomial model for high expression probes. The Fast method uses the simplified negative binomial model for all probes. In situations of algorithm not converging, the linear regression method will be used instead. High or low expression is determined by how high the probe mean is across all samples relative to the negative controls. At least one covariate must be selected as the predictor. Analysis will take into account the selected confounders but results will only be displayed for covariates designated as predictors.

Running the Optimal model is computationally intensive and run time is proportionate to data size and number low expression probes. It may take several 10s of minutes depending on the data size and count distribution.

Gene Set Analysis Module

Gene set analysis (GSA) summarizes the change in regulation within each defined gene set (selected along the left side of the window) relative to the baseline (or in the case of continuous variable, per unit change in variable). The values calculated are the global significance score and the directed global significance score and are expressed in heatmaps and/or a data table.

Before You Start GSA

Since much of GSA originates from Differential Expression Analysis, see the *Before You Start Differential Expression* section.

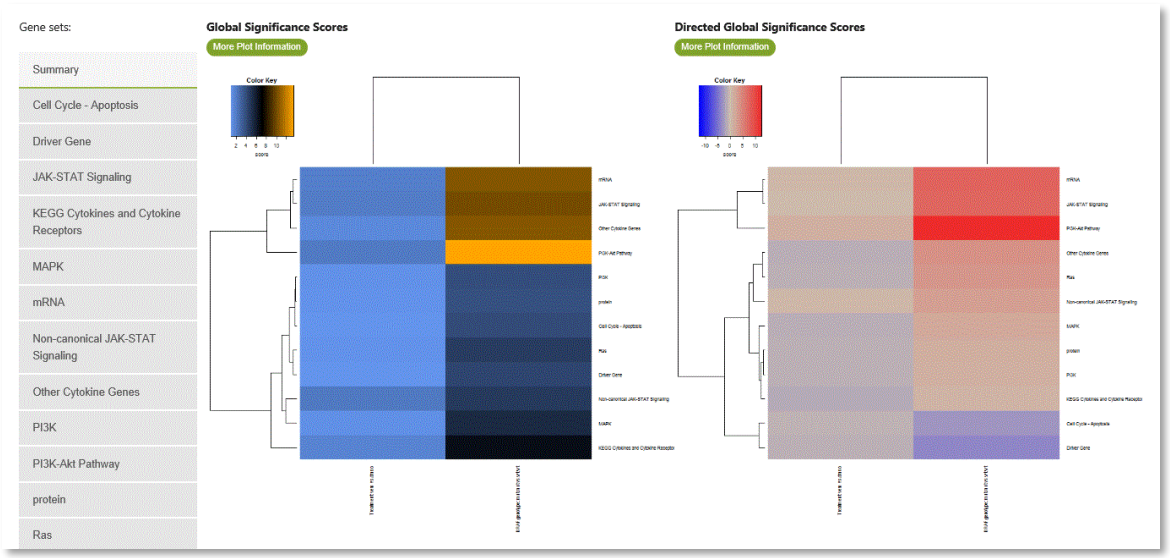


Figure 36: GSA module window and options

Custom Options for GSA

There is no custom menu for GSA. The Differential Expression menu, however, features a checkbox indicating whether to run GSA (see the *Custom Options for Differential Expression* section).

Interpreting Results of GSA Plots

Summary - Global Significance Scores

Global significance scores (also called undirected global significance scores) measure the overall differential expression of the selected gene set relative to selected covariates, ignoring whether each gene is up- or down-regulated.

The chosen covariates are listed along the bottom of the heatmap and the various genesets are listed along the right side.

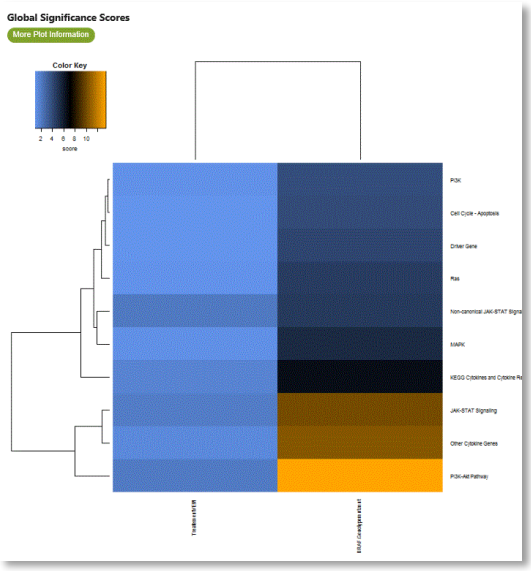


Figure 37: GSA - Undirected Global Significance Scores plot

Summary – Directed Global Significance Scores

Directed global significance scores measure the extent to which a given gene set is up- or down-regulated relative to a given covariate. It is calculated similarly to the undirected global significance score, but it takes the sign of the t-statistics into account.

The chosen covariates are listed along the bottom of the heatmap and the various genesets are listed along the right side.

In the **3D Bio Data Example** (see Appendix A), we see that the **BRAF.Genotype** is associated with more variable results among the gene sets than **Treatment** (Figure 37). We can see from the **Directed Global Significance Scores** plot (Figure 38) that the **P13K-Alt Pathway** gene set has the highest score in the BRAF.Genotype category.

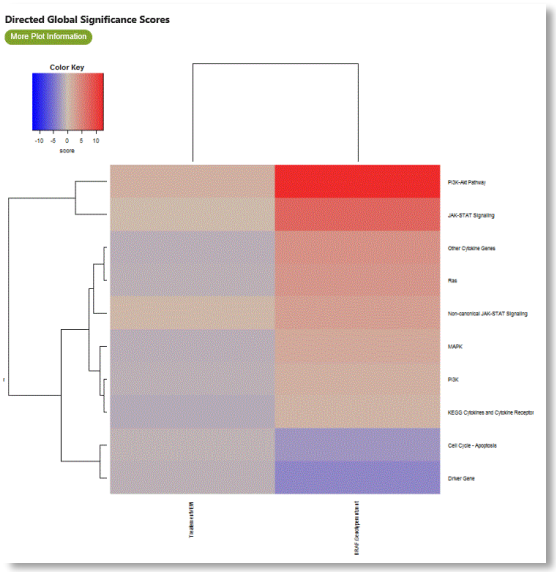


Figure 38: GSA - Directed Global Significance Scores plot

If only a single variable is chosen as a predictor, then a table will take the place of a heatmap, showing values for directed and undirected global significance.

Undirected and Directed global significance scores table

More Plot Information | Download CSV Data

Show 10 | Search:

	Undirected BRAF.Genotypemut.mut	Undirected TreatementVEM	Directed BRAF.Genotypemut.mut	Directed TreatementVEM
Cell Cycle - Apoptosis	4.015	1.015	-2.672	-0.867
Driver Gene	4.298	0.986	-3.672	-0.979
JAK-STAT Signaling	10.133	1.97	7.442	0.371
KEGG Cytokines and Cytokine Receptors	6.637	1.702	0.967	-1.383
MAPK	5.454	1.173	1.936	-1.169
Non-canonical JAK-STAT Signaling	4.766	2.141	2.692	0.613

Figure 39: Global Significance Scores Table

Gene set of choice – covariate of choice

Selecting a pathway along the left side of the window results in volcano plot and table of values. The volcano plot is a replicate of that drawn in the Differential Expression module, but all data points are greyed except those in the selected pathway (those data points are colored).

Selecting the **P13K-Alt Pathway** gene set results in the Differential Expression volcano plot, overlaid with colored points which reflect the probes in that gene set. We can see that there are a number of probes from this gene set with significant results.

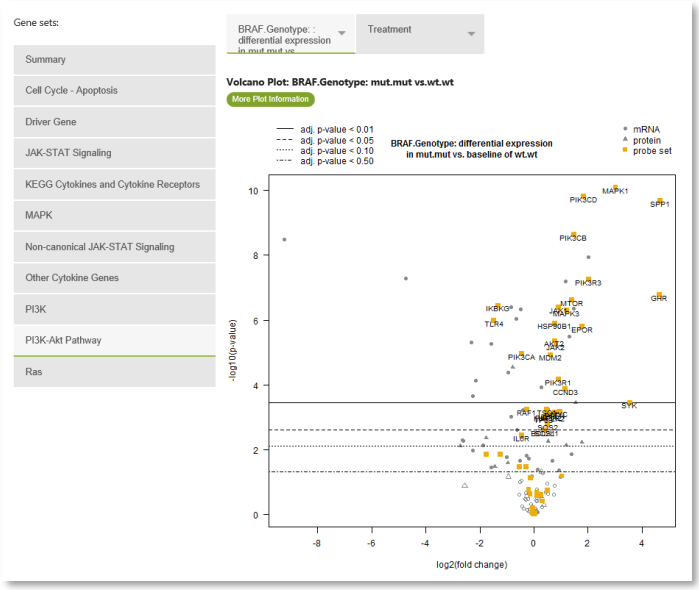


Figure 40: GSA - Volcano plot by gene set and covariate

GSA Algorithm Details

Differential expression analysis calculates a t-statistic for each gene against each covariate in the model. A gene set's global significance score for a covariate measures the cumulative evidence for the differential expression of genes in a pathway and is calculated as the square root of the mean squared t-statistic of genes.

$$global\ significance\ statistic = \left(\frac{1}{p}\sum_{i=1}^{p} t_i^2\right)^{1/2},$$

where t_i is the t-statistic from the i^{th} pathway gene.

The directed global significance statistic is similar to the global significance statistic, but rather than measuring the tendency of a pathway to have differentially expressed genes, it measures the tendency to have over- or under-expressed genes. It is calculated similarly to the undirected global significance score, but it takes the sign of the t-statistics into account:

$$directed\ global\ significance\ statistic = sign(U)|U|^{1/2}$$

where $U = \left(\frac{1}{p}\sum_{i=1}^{p} sign(t_i) \cdot t_i^2\right)$

and where sign(U) equals -1 if U is negative and 1 if U is positive.

A pathway with both highly up-regulated and highly down-regulated genes can have a very high global significance statistic, but a directed global significance statistic that is relatively close to zero. The two statistics will be equal in a pathway that contains genes regulated in only one direction.
For each gene set, the volcano plot is redrawn and table produced as described in the DE module, with the exception that the genes in that pathway are highlighted on the plot and displayed in the table

PathView Module

The PathView module overlays the Differential Expression analysis results with various KEGG pathways. Elements that are over-expressed in this pathway are colored gold, those that are under-expressed are colored blue, and those that are unchanged are gray.

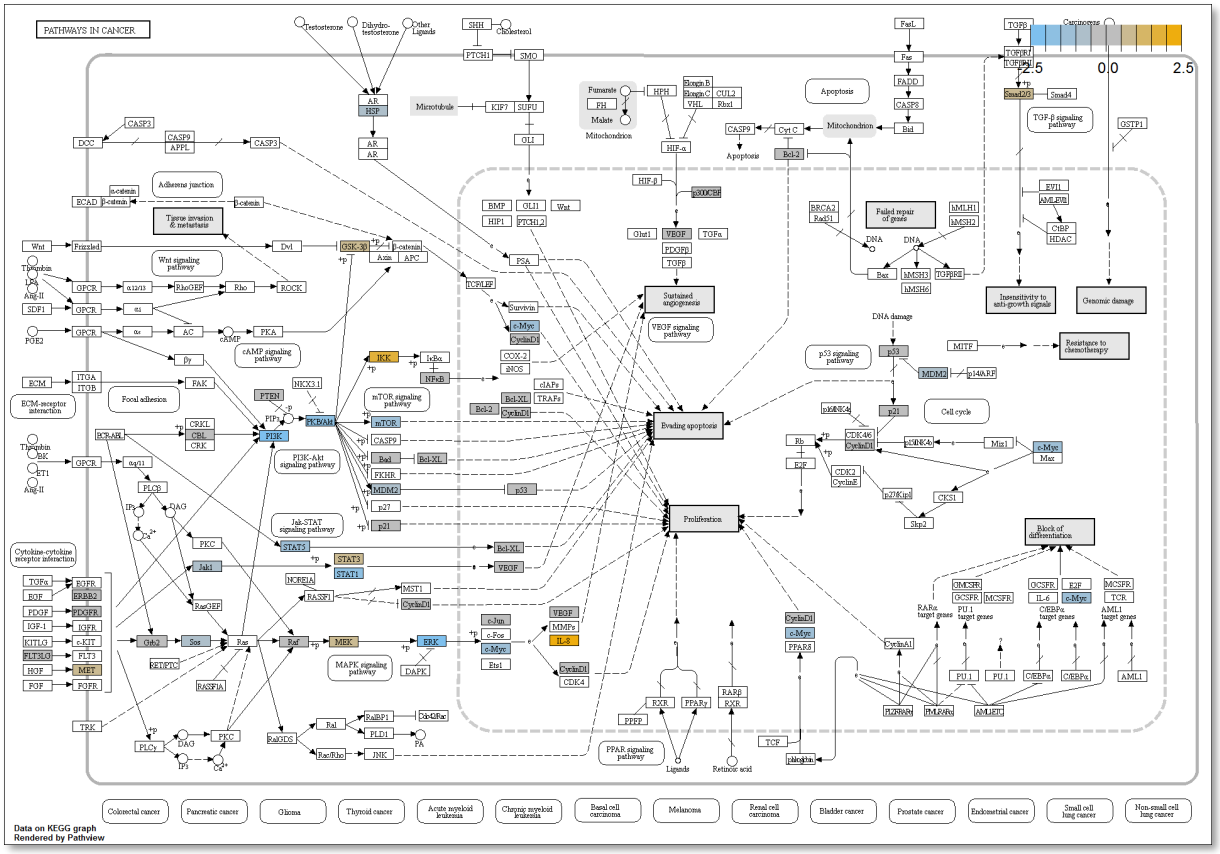


Figure 41: PathView module view

Before You Start PathView

PathView plots are simply DE results overlaid with KEGG pathways. See the *Before You Start Differential Expression* section.

Custom Options for PathView

There is no custom menu for PathView. The Differential Expression menu, however, features a checkbox to choose whether to display PathView, dropdowns for how many pathways to offer and whether to color plots by fold-change or t-statistic, as well as a box to enter a p-value threshold for plotting. See the *Custom Options for Differential Expression* section.

Interpreting Results of PathView Plots

Throughout each of the KEGG pathways offered along the left side of the window, nodes associated with genes are colored **blue** if the data suggests that they are down-regulated or **gold** if it suggests that they are up-regulated. The KEGG pathways listed are those with the highest level of differential expression for your dataset; the number of top pathways offered depends on the number chosen on the Custom Analysis menu (20 is default; see the *Custom Options for Differential Expression* section) and how many probes in your dataset map to those pathways.

Note that the default p-value threshold is an un-adjusted p-value, so some colored nodes may represent false positives. Also, before inferring significance from the abundance or paucity of differentially expressed genes in a particular pathway, consider the percentage of genes from that pathway that are actually represented in the CodeSet. Studying the impact of the data on the overall pathway in addition to its effect on the individual parts results in a more holistic analysis.

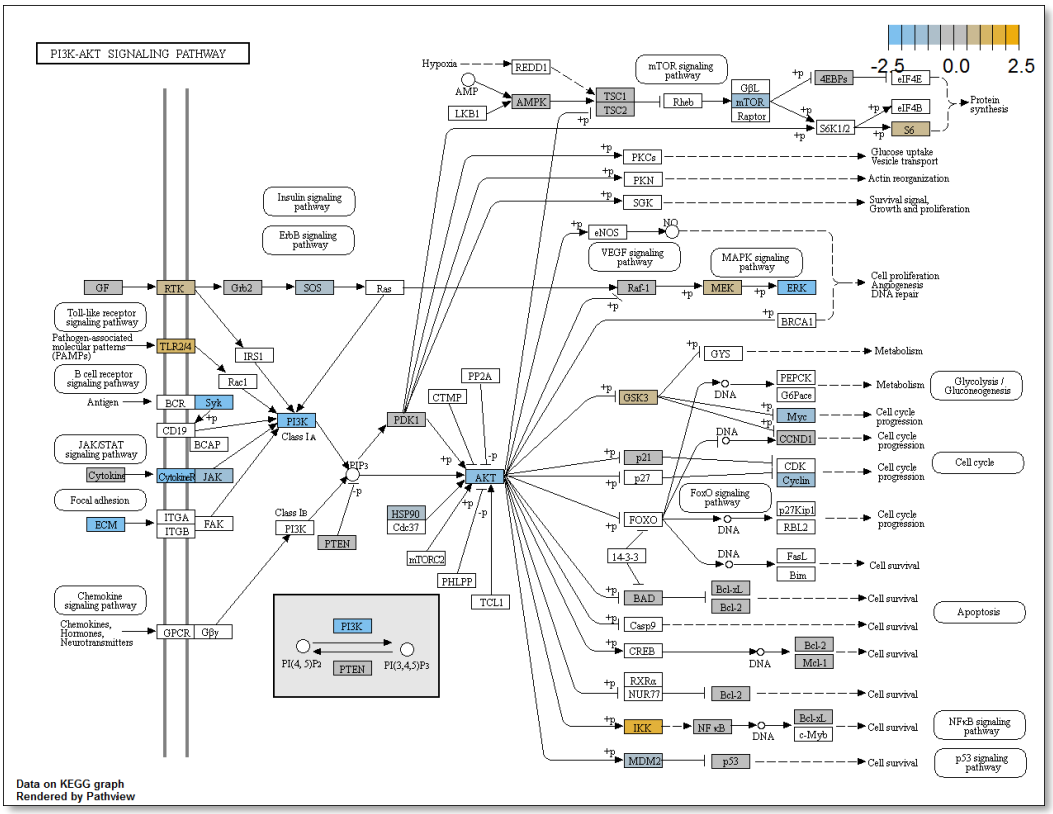


Figure 42: PathView plot

As a next step to the GSA analysis in the **3D Bio Data Example** (see *Appendix A*), we can view the pathways that include our gene set(s) of interest in the PathView module. Here, we select the **P13K-Alt Pathway** (Figure 42) to see where our genes of interest lie in this particular pathway. Colored boxes show the specific elements of the pathway that were differentially expressed and whether they are up- or down-regulated in our data. If we decided to later run the Probe Descriptive module, we might enter these genes for analysis.

Pathway Scoring Module

Just as Differential Expression analysis of individual genes or gene sets is used to research the effect of covariates on a dataset, the Pathway Score can be used to summarize the data from a pathway's genes into a single score.

At least one covariate must be chosen against which to plot the scores, while the effects of other variables that may be highly correlated with gene expression can be removed from the analysis by adjusting the score with respect to those variables (see the *Custom Options for Pathway Scoring* section). Pathway scores are calculated as the first principal component of the pathway genes' normalized expression.

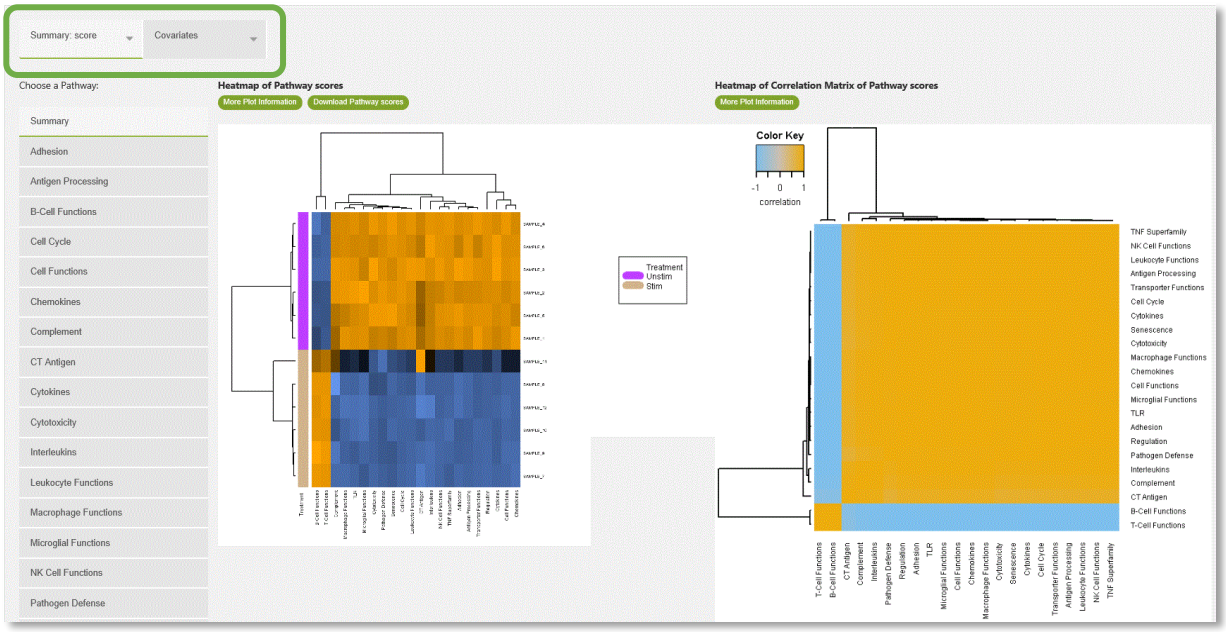


Figure 43: Pathway Scoring window and options

When the software generates pathway scores, there can be some ambiguity in the directionality of those scores. The software will attempt to orient them such that increased score corresponds with increased expression in a majority of the pathway genes. In pathways where the first PC is somewhat balanced between up-regulated and down-regulated genes, however, the direction of the pathway score can be somewhat unpredictable.

Like any complex statistical metric, Pathway Scores should be interpreted with caution. Although the first principal component of a gene set may reflect pathway activity or deregulation in some settings, the scores may be confounded by biological effects (i.e.,

The **More Plot Information** button provides a description of the plot.

The **Download Pathway Scores** button opens a .csv data table that can be viewed, edited, printed, and saved. You may also Save or Save as without opening. It contains pathway scores for all samples and pathways.

proliferation or immune cell abundance) or technical effects (i.e., sample input or preparation) unrelated to the pathway activity. For these reasons, pathway scores can be a useful tool for understanding your data in some settings, but misleading or meaningless in others. Interpretation of scores should never be performed without correlating them to other analysis results (such as differential expression testing), to ensure that they are placed in the correct biological context.

Before You Start Pathway Scoring

To run the Pathway Scoring Module, you must choose **Custom Analysis** as your **Analysis Type** and check the appropriate box on the **General Options** tab. Once you have done that, the Pathway Scoring tab will appear in the list and you will be able to select it for customization (see the *Custom Options for Pathway Scoring* section).

Custom Options for Pathway Scoring

Your available annotations will appear on the Pathway Scoring tab. Use the green arrows to move over those annotations with which you would like to plot the Pathway Score (to the **Plot Pathway Score Vs** field) and those for which you would like to adjust it (to the **Adjust Pathway Score For** field).

Adjusting for covariates removes their signal from the data before pathway scoring is performed. To be precise, when this option is selected, each gene will be regressed against the selected covariates and pathway scoring will be performed on the residuals of these regressions.

The question mark button reveals additional information.

The exclamation mark button reveals an alert and a brief explanation as to why an option may be unavailable (greyed out).

It is usually advisable to **Adjust Pathway Score For** various **technical variables** that are suspected to influence gene expression. These may be needed to account for (e.g.) data generated by different operators, from different labs, or using different lots of NanoString reagents. Adjusting for **biological variables** is a more difficult decision. In some cases, you may want to score pathway status independent of one biological variable to isolate the effect of another biological variable. For example, in data with multiple subtypes and multiple treatment groups, the signal from a subtype may exceed the signal from a treatment group. In this case, adjusting for subtype will help the pathway scores capture the effects of the treatment group. Even if there is only one biological variable, it can sometimes make sense to adjust for it. For example, adjusting for the treatment group can encourage pathway scores to reflect treatment-independent tumor state, which could be desirable depending on the biological question of interest.

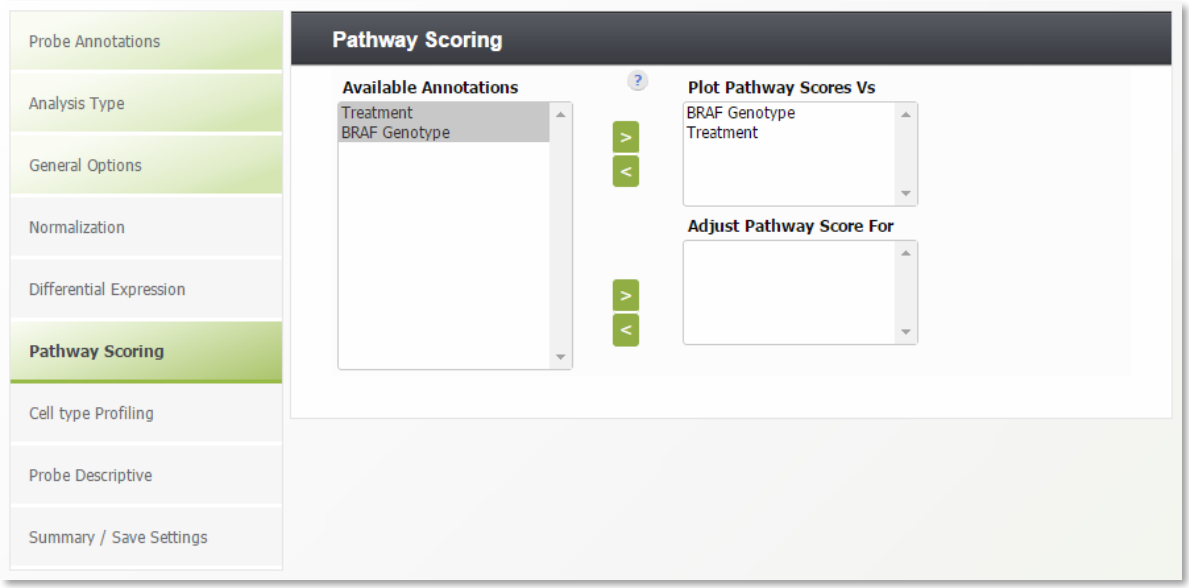


Figure 44: Pathway Scoring Custom Analysis menu

Interpreting Results of Pathway Scoring Plots

For a given pathway, PC analysis scores each sample using a linear combination (a weighted average) of its gene expression values, weighing specific genes to capture the greatest possible variability in the data. Thus, the first PC will reflect whatever factor(s) emerge as the main driving force of variability in gene expression for that dataset.

Summary Plots

The Heatmap of Pathway Scores is a high-level overview of how the pathway scores change across samples. Pathways are listed on the horizontal axis and samples are listed vertically. Using this plot, you may begin to understand how pathway scores cluster together and which samples exhibit similar pathway score profiles. Orange indicates high scores; blue indicates low scores. Scores are displayed on the same scale via a Z-transformation.

In the RNA-Protein dataset used in this example (Figure 45), we can see that the six samples in the unstimulated group exhibit high scores with *T-cells* and *B-cells*, but low scores with all others. The Stimulated group tested opposite these results.

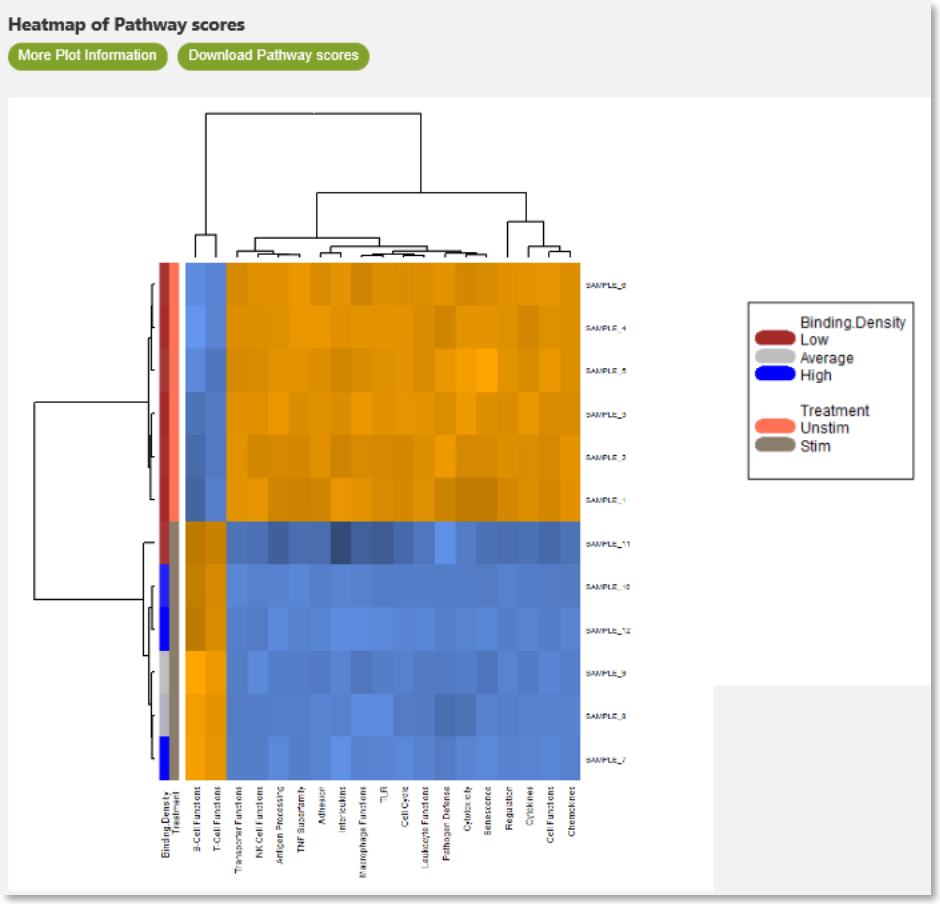


Figure 45: Pathway Scoring - Heatmap of Pathway Scores

The **Heatmap of Correlation Matrix of Pathway Scores** is a heatmap showing the correlation matrix of pathway scores. Pathways are listed on both the horizontal and vertical axes. Orange indicates positive correlation, while blue indicates negative correlation. Since the values are mirrored across the diagonal, you may limit your observations to either the upper or lower triangular matrix.

Similar to the previous heatmap, the RNA-Protein dataset used in this example (Figure 46), shows that *T-cell Function* and *B-cell Function have* negative correlations with all other pathways but positive correlation with each other.

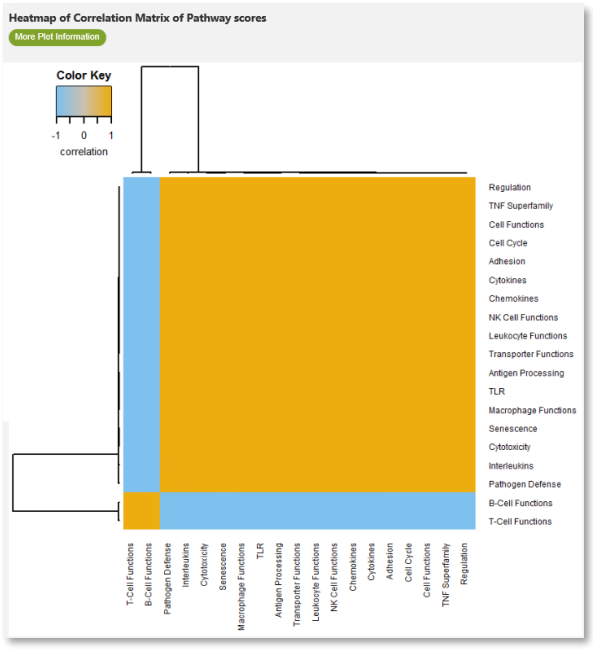


Figure 46: Pathway Scoring module - Correlation matrix of pathway scores

Pathway Measurements vs. Other Pathway Scores

On the *Summary* tab, you may select an individual pathway of interest along the left side of the window. This creates a collection of scatter plots, each with the selected pathway of interest on the x-axis. On each scatter plot's y-axis is an alternative pathway. If you have more than one covariate, you will see a scatter plot collection for each covariate.

This combined view allows you to see how the scores for each pathway compare to scores for other pathways and how the different experimental conditions are distributed across each comparison. You may identify pathways with highly correlated scores in this plot, which may indicate that these are driven by the same underlying factor(s). Others may be almost completely uncorrelated, indicating that they reflect very different biological events.

In the RNA-Protein dataset used in this example (Figure 47), we select *B-Cell Functions* from the list to see this pathway's correlation with other pathways. As in the heatmaps, above, it shows positive correlation with *T-Cell Functions* and negative correlation with all other pathways.

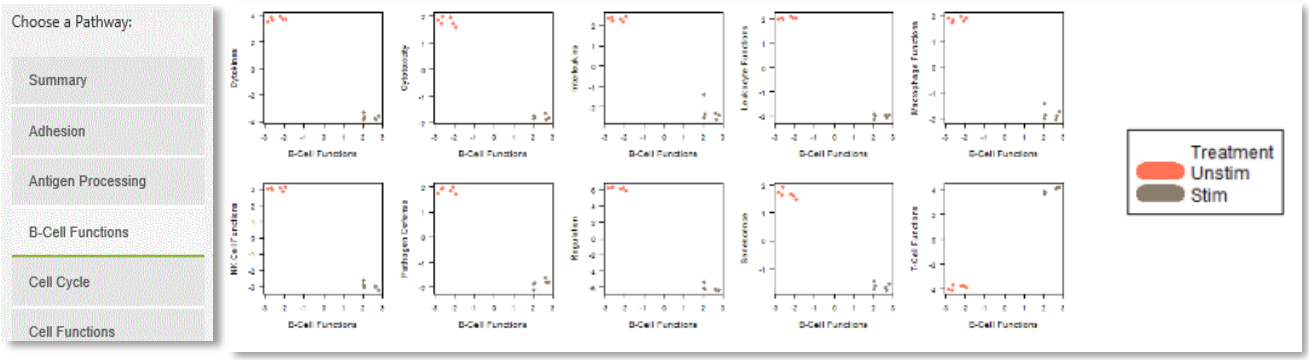


Figure 47: Pathway Measurement vs. Other Pathway Scores

Covariates Plots

Pathway Scores vs. Covariate

Selecting the *Covariate* tab and the *Summary* of all pathways results in a plot of all pathway scores against the covariate chosen earlier on the Custom Analysis menu (see the *Custom Options for Pathway Scoring* section). There is a separate graph for each covariate; pathway scores are plotted to show how they vary across different experimental conditions.

In the RNA-Protein dataset used in this example (Figure 48), we see that scores for *T-Cell Function* and *B-Cell Function* increase between the unstimulated and stimulated groups, while others decrease.

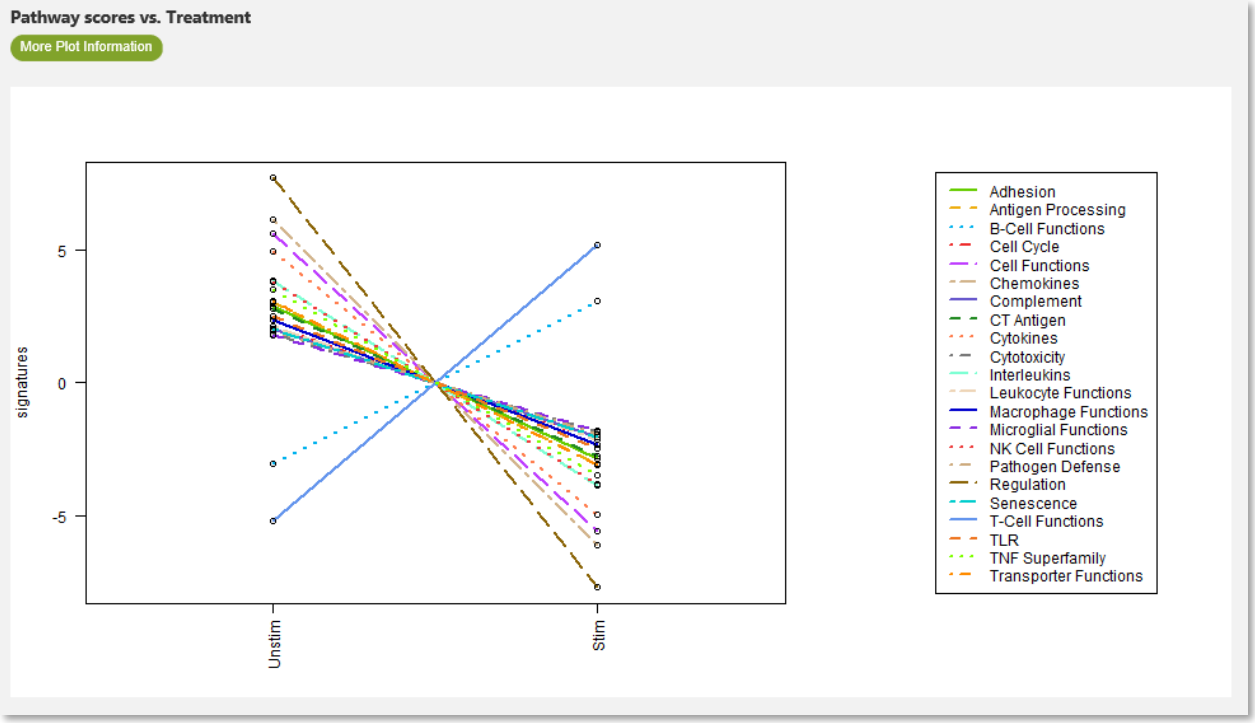


Figure 48: Pathway Scoring module - All Pathway Scores vs. covariate

Pathway of choice vs. Covariate

Selecting the *Covariate* tab and a specific pathway along the left side of the window results in a separate box plot for each experimental condition. Each depicts pathway scores on the y-axis vs. the experimental conditions for the covariate on the x-axis.

In the RNA-Protein dataset used in this example (Figure 49), we select *B-Cell Functions* from the list and see, again, that the unstimulated treatment group exhibits low pathway scores (which often indicates down-regulation of the pathway) while the stimulated treatment group exhibits elevated pathway scores (which often indicates up-regulation).

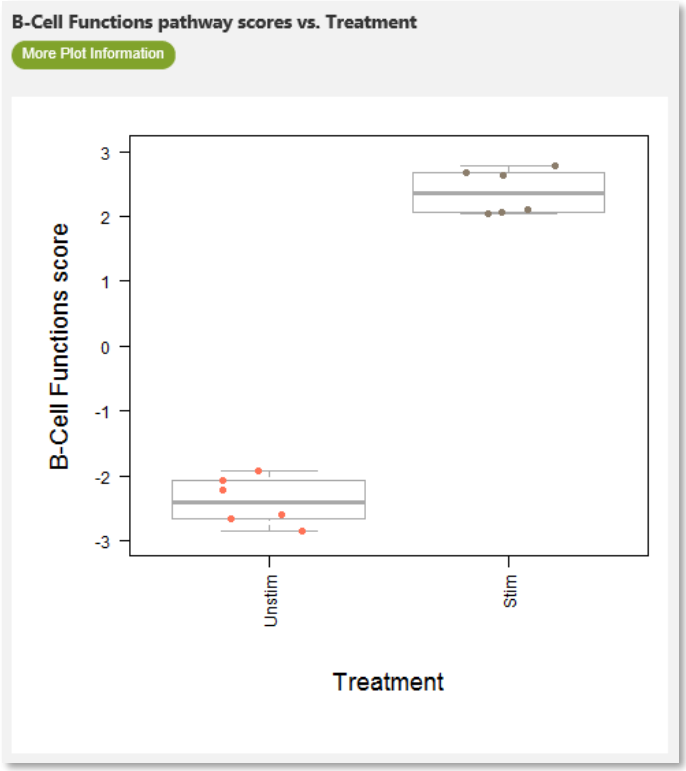


Figure 49: Pathway Scoring module - Pathway of choice vs. Covariate

Pathway Scoring Algorithm Details

This approach of extracting pathway-level information from a group of genes using the first principal component (PC) of their expression data was established by Tomfohr, Lu, & Kepler in 2005.

Probe Descriptive Module

The Probe Descriptive tab provides multiple plots which are focused just on the probes of interest, which you designate on the Custom Analysis menu. **Univariate plots** show the distribution of the probe results according to the variable of choice. **Correlation plots** illustrate the relationship between the probes of interest. **PCA Biplots** display the impact of the expression of probes of interest on the clustering of samples, contrasting principal components (PCs) two at a time, for the variable of choice. **Parallel Coordinate Plots** allow you to view the expression levels of the probes of choice; the experimental group's results overlay each other, each displayed in a different color. The Interaction network plot visualizes a conditional dependency network among the selected probes that best describes the observed data. The Trend Plot visualizes the expression trajectory of a trending variable (e.g. a patient ID, a cancer subtype) typically across an ordinal variable (e.g. time).

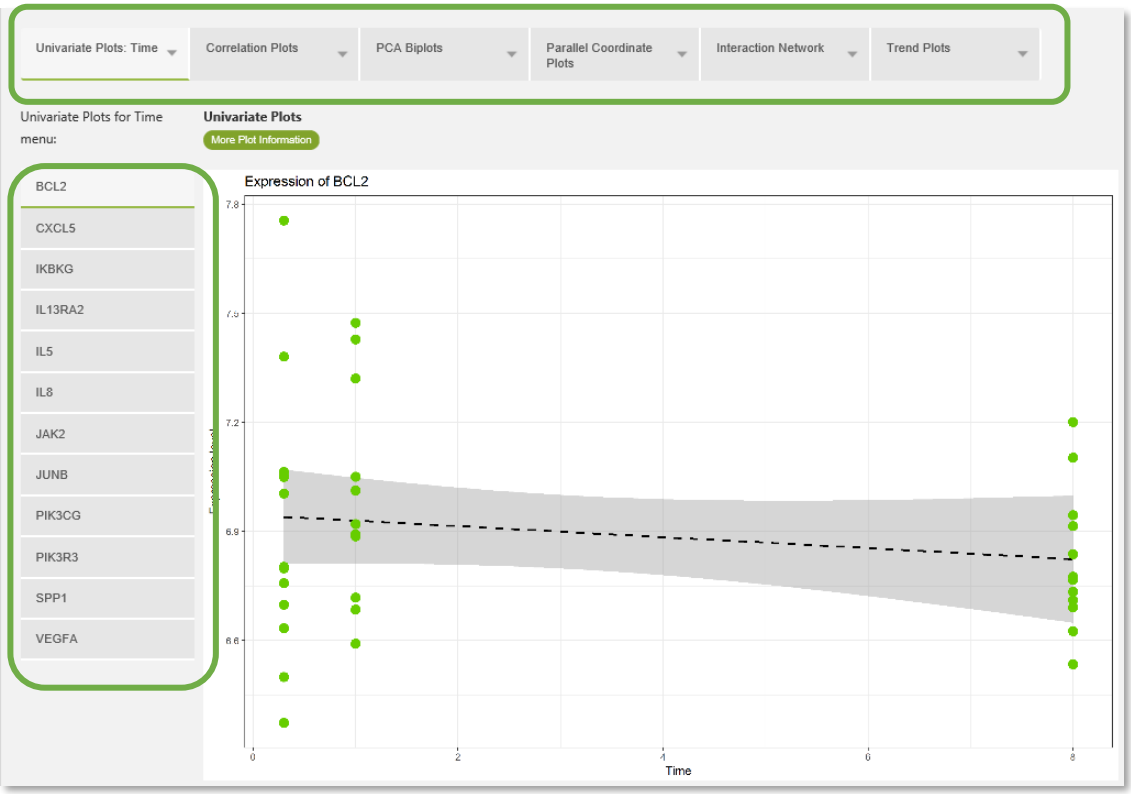


Figure 50: Probe Descriptive window and options

Before You Start Probe Descriptive

Due to the highly descriptive nature of this analysis, you may consider adding **one or only a few covariates** to this analysis.

To choose probes to use for descriptive analysis:

- Run an initial Advanced Analysis *without* the Probe Descriptive module so that you can identify **the most differentially expressed probes** from the DE module plots and tables.
- List 5-15 probes which appeardifferentially expressed across the different groups belonging to the annotation that you wish to analyze. Make sure your list includes genes that are both induced and repressed.

Return to the nSolver dashboard and run a second Advanced Analysis, selecting Custom Analysis. This time, on the General Options tab, select the **Probe Descriptive** box. In the Probe Descriptive module menu, enter the probes that you identified in the first analysis and move them over to the **Selected Probes** window. Select the grouping annotation(s) that you used to identify differentially expressed genes.

Custom Options for Probe Descriptive

Use the checkboxes to search for probes of the analyte-type of interest (RNA, protein, etc). Probe names will dynamically appear as they are typed and can be moved over to the **Selected Probes** field using the green arrow buttons. You may enter up to 15 probes (if five or more are entered, PCA plots will also be generated). Probes used as housekeepers or removed from the analysis via a low count threshold will not be included in the output. To identify probes of interest, consult the *Before You Start Probe Descriptive* section.

Move any annotations to be used in **Grouping** the expression data using the green arrow button. At least one annotation must be selected.

You can check the box to **Generate Trend Plots** if you have covariates to designate as **Interval ID** and as **Series ID**. The interval ID can be an ordered categorical or continuous variable. Additionally, trends across distinct sample annotation groups can be examined by specifying an optional stratifying annotation.

- **Interval ID** is the variable that defines how the data points are ordered along the trend (horizontal axis in plots). Typical covariates that would be specified as Interval IDs are Time (as in the example below – Figure 51), Concentration, and Dosage; there should be three or more groups in this variable.
- **Series ID** defines the groups into which we wish to separate the samples (for example, patient cohorts). In general, the definition of group could extend to the case where each group consists of only one observed entity (for example, one patient). The example below uses BRAF Genotype.
- **Stratifying Annotation** allows you to separate the series ID into groups to see a trend. Since we are interested in how Treatment affects each BRAF genotype (chosen as Series ID, below), we will select it as our stratifying annotation.

Selecting the **Generate Interaction Network** box generates a network that best describes the conditional relationship between your selected probes. You can adjust for a covariate that is expected to influence these probes. In this context, the relation between two probes is defined as their statistical dependence on one another after accounting for their dependence on other probes.

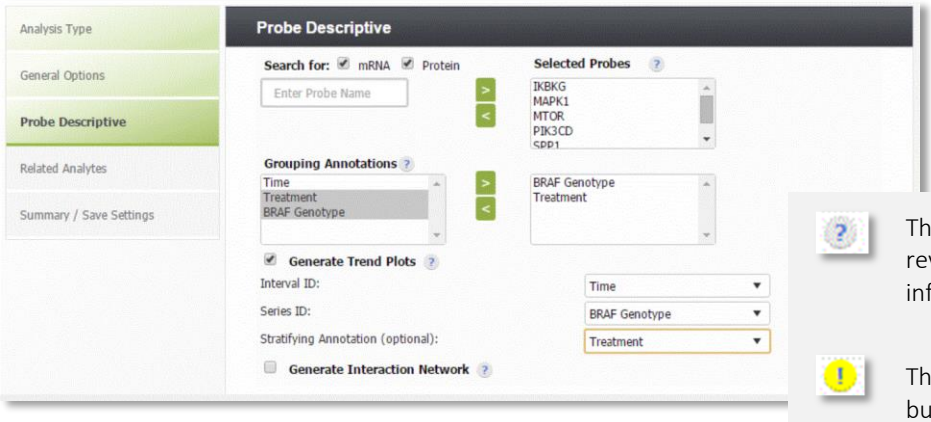


Figure 51: Probe Descriptive module Custom Analysis menu

The question mark button reveals additional information.

The exclamation mark button reveals an alert and brief explanation as to why an option may be unavailable (greyed out).

Interpreting Results of Probe Descriptive Plots

This module provides detailed descriptive analyses of the genes of your choice. The analysis will always include univariate plots and correlation plots. When at least 5 probes are selected, PCA biplots and parallel coordinate plots will also be generated. Interaction network plots will be generated, if selected. Additionally, when trending parameters (Series ID and Interval ID) are defined, you may generate trend plots.

Univariate Plots

For **categorical variables**, a box plot is overlaid with a violin plot providing information on both the $\log_2$ expression quartiles as well as the estimated expression distributions for each level of the categorical variable(s) of interest. For each box in the boxplot:

- The horizontal black line on the box plot represents the median expression.
- The box depicts the 2nd quartile of expression.
- The green dots display each sample's $\log_2$ expression for the specific gene selected (on the left).
- The grey shading represents the estimated distribution of the expression values.

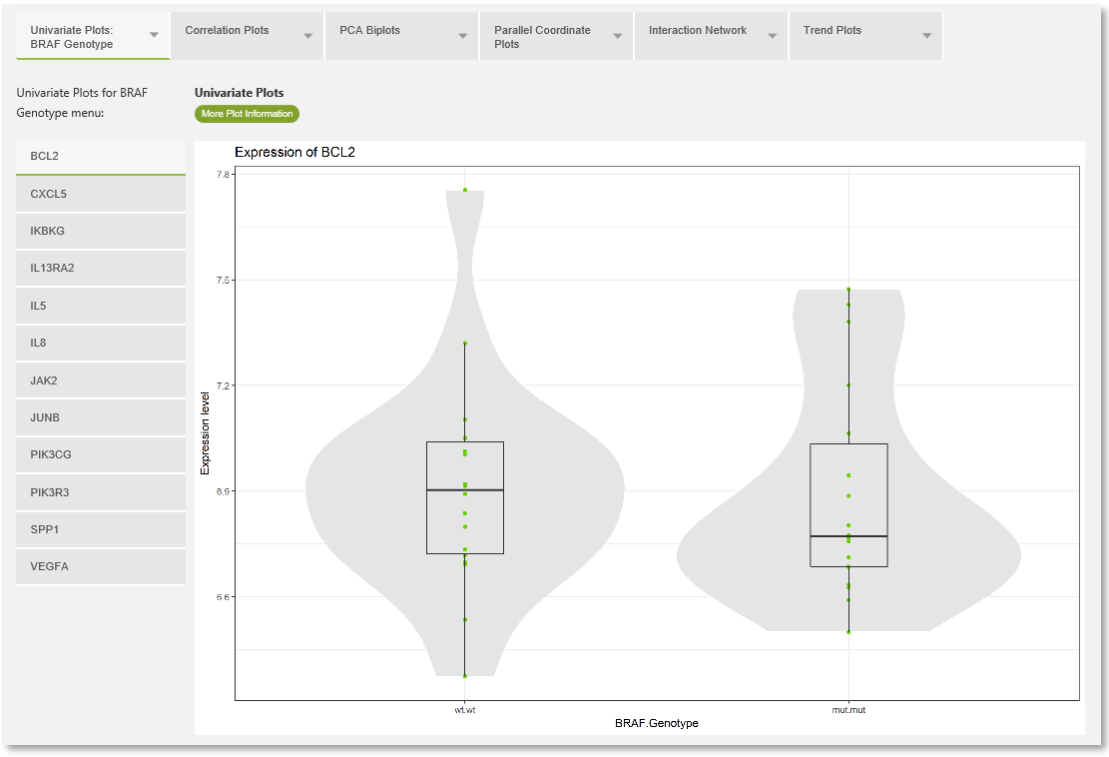


Figure 52: Probe Descriptive module - Univariate biplot for categorical variable

For a **continuous covariate**, a scatter plot is generated, showing each sample's normalized $\log_2$ expression level plotted relative to the continuous variable.

- The dotted line represents the least squares fit, drawn along with the 95% confidence interval (CI).
- The green dots display each sample's expression for the specific gene selected (on the left).
- The grey shading represents the estimated distribution of the expression values.

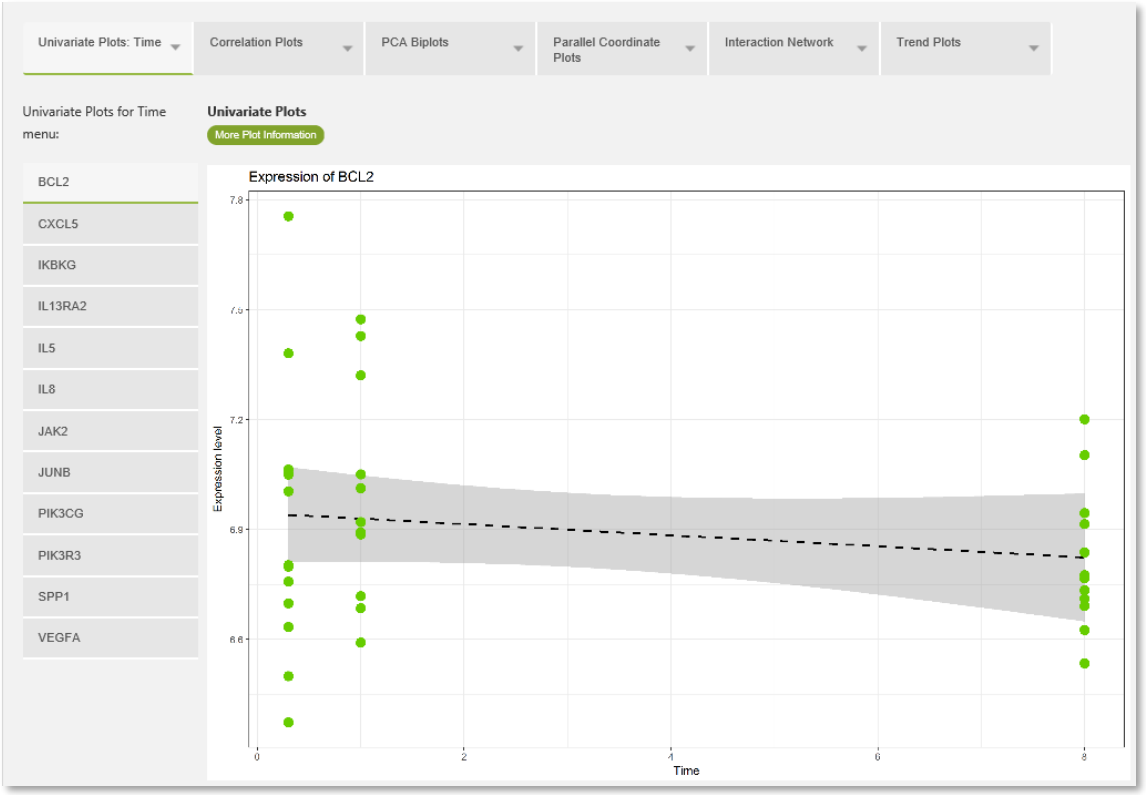


Figure 53: Probe Descriptive module - Univariate plot for continuous variable

Correlation Plots

The correlation plot allows visualization of two sets of information: **distribution of gene expression** and **correlation of gene expression**. When the covariate of interest is continuous, the values are categorized into low, average and high. Each field belongs to the gene listed at the top of its column and the gene listed on the right side of its row.

- The **distribution of expression** for each gene is drawn on the diagonal (note this effectively replicates the violin plot from the univariate analysis), segregating experimental groups belonging to the chosen covariate by color.
- The **correlation of gene expression** for each pair of genes is expressed **numerically** in the top right fields as the overall Pearson correlation coefficient and corresponding p-value. Pearson values of correlation of gene expression segregating covariate groups is also given; groups are separated by color.
- The **correlation of gene expression** for each pair of genes is expressed **graphically** in the lower left fields, plotting the expression values and separating the groups by color.

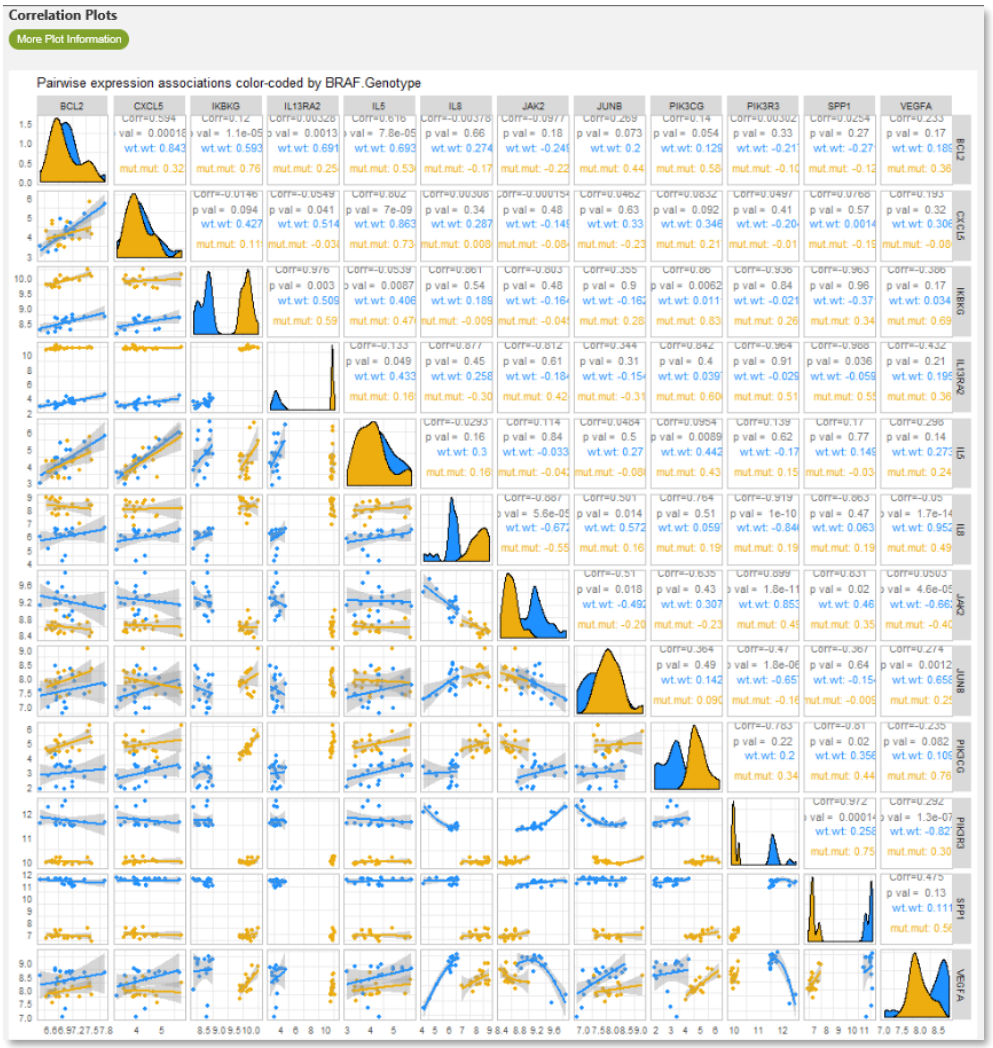


Figure 54: Probe Descriptive module - correlation plots

PCA Biplots

PCA biplots also allow visualization of the relationship between the genes chosen for probe descriptive analysis.

Each biplot shows the spread of the probe data along a pair of principal component (PC) axes. You may choose the PC's of interest from the **PCA Biplots menu** on the left side of the window. The plot includes:

- **Samples:** each point in the PCA biplot corresponds to one sample. The coordinates of each point indicate the sample's Principal Component scores. Samples with similar Principal Component scores have similar gene expression profiles and cluster together. Points are colored by covariate.
- **Ellipses:** each category of the chosen covariate is represented by a colored ellipse. This represents the estimated region where the majority of the samples (68%) of that category type would be expected to fall, assuming the analyzed samples represent the population well. The extent to which the ellipses overlap indicate that gene expression differences are not enough to differentiate among categories of the covariate. When ellipses are non-overlapping, the different categories of the covariate of interest have distinctly different PC scores and gene expression profiles cluster the categories apart.
- **Vectors:** each vector in the biplot corresponds to one gene. The direction and length of the vector indicate how each gene contributes to the principal component. Vectors pointing in the same direction indicate co-regulated genes.

In Figure 55, we can see that PC1 clusters the WT (blue ellipse) and MUT (gold ellipse) categories apart. IL-5, CXCL5, and BCL2 display long projections on PC2 and short projections on PC1, toward either BRAF genotype group,indicating these genes do not have a major impact on the differences between WT and MUT groups.
VEGFA (left) and JUNB (right) display projections in opposite directions, indicating that VEGFA is upregulated when JUNB is downregulated, and vice versa (see the patterns of gene expression on the diagonal of the correlation plot).
JACK2, PICK3RR and SPP1 have long negative projections on PC1, while IKBKG, IL13RA2, IL8 and PIK3CG have long positive projections on PC1, meaning that they contribute in high degree to the clustering apart of WT and MUT samples. Because they display projections in opposite directions, they have opposite patterns of regulation (see the patterns of gene expression on the diagonal of the correlation plot).

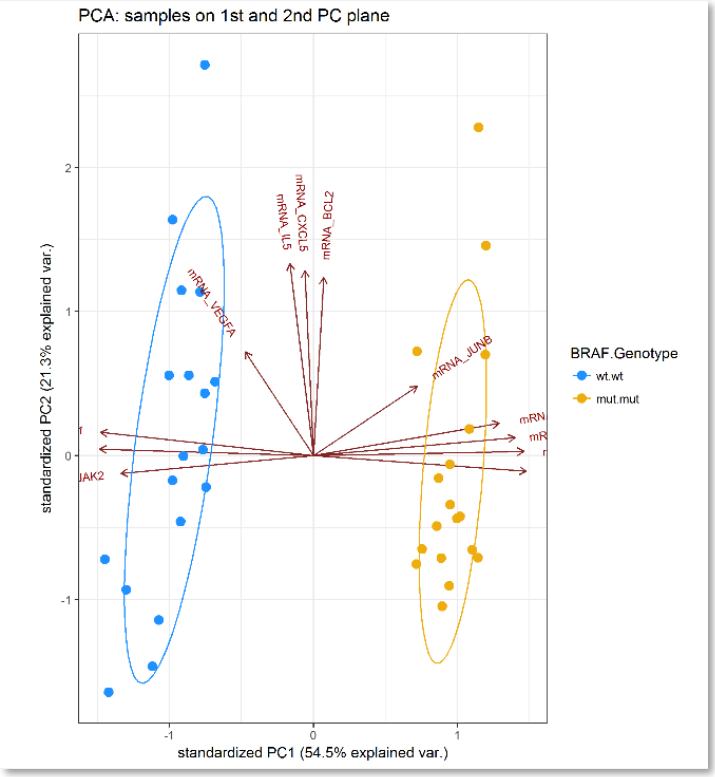


Figure 55: Probe Descriptive module - PCA Biplot

Parallel Coordinate Plots

These plots provide a simple way to see up/down regulation of each gene relative to the covariate of interest. The expression is scaled for each gene across all samples.

This view lets you compare the patterns of gene expression among the different categories of the covariate of interest. When a continuous variable is selected, its values are split into average, high and low.

Figure 56: Probe Descriptive module - Parallel Coordinate Plot

Trend Plot

This plot is designed to enable visualization of the change in expression levels as a function of a variable of interest, the **Interval ID**, grouping your samples by a category, or **Series ID** (see the *Custom Options for Probe Descriptive* section). You can further stratify your samples using a **Stratifying Annotation**.

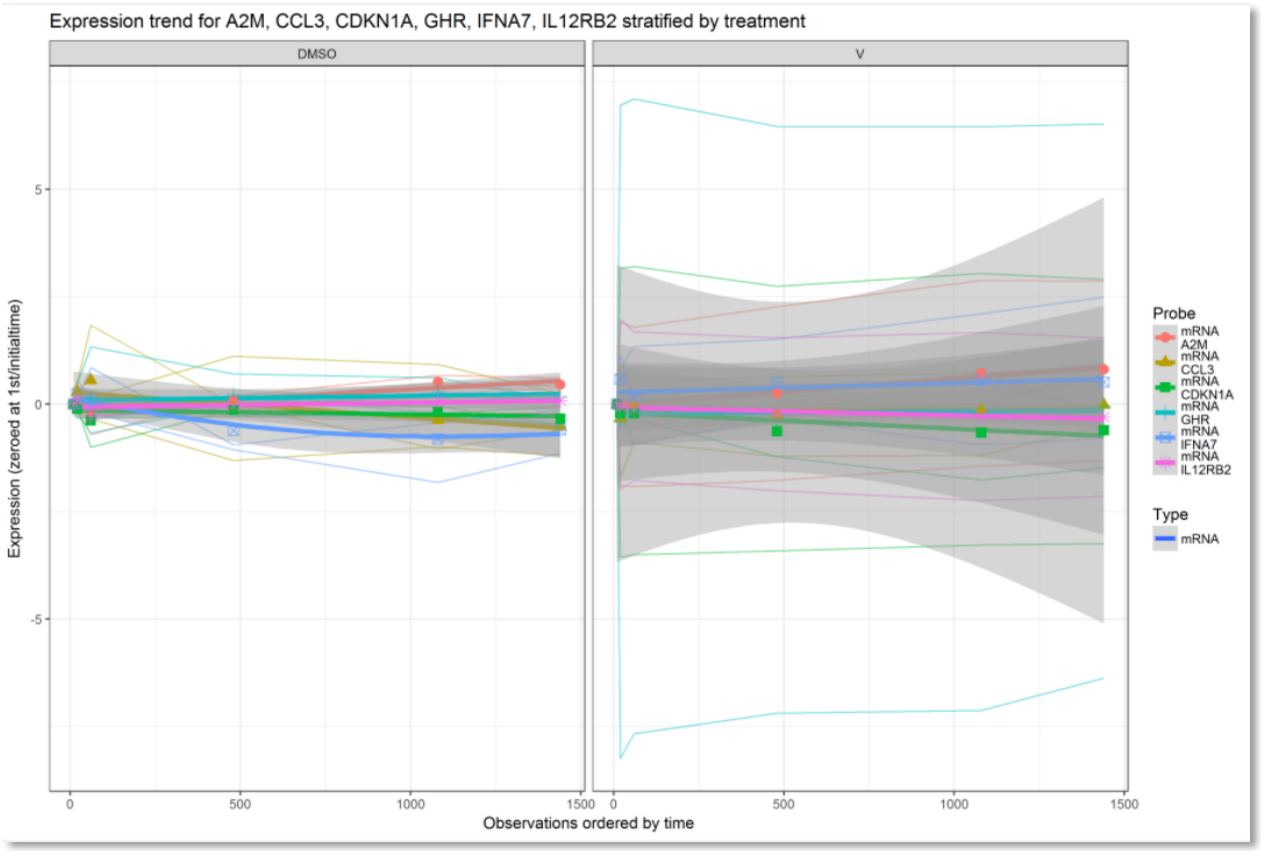


Figure 57: Probe Descriptive module - Trend Plot

The **Interval ID** is the variable of interest; it is typically a continuous variable such as time, concentration, or dosage, and is plotted on the x-axis of the plot. The example in Figure 57 uses time as Interval ID (variable designated continuous).

The **Series ID** defines groups such as patient cohorts (treated vs untreated) or cell lines. The example in Figure 57 uses BRAF genotype (which correlates with cell line) as series ID.

The **Stratifying Annotation** separates the trend plot information into plots for each category of the stratifying variable. The example in Figure 57 uses Treatment (DMSO as vehicle and VEM as treatment) as the stratifying annotation.

For the settings used to create this plot, see the *Custom Options for Probe Descriptive* section. Each probe's gene expression is plotted in a single color over the interval ID variable (time) in three narrow lines, corresponding to the three BRAF genotypes (the series ID variable). The thick line of the same color represents the average of the expression values of the three BRAF genotypes.

Interaction Network Plot

The interaction network plot shows the conditional dependency network among the selected probes, as suggested by the data. This analysis is highly exploratory and is meant, primarily, to aid hypothesis generation. Although the network inferred is the most likely network under the modeling assumptions and based on the data provided, it may not reflect a real biological network.

An edge between two nodes implies an association between them, after accounting for the variability of all other nodes. As with many other analyses, the inference here can be performed on data adjusted for selected variables when the effects of those variables are to be removed from the inference. The thickness of the edges denotes significance or confidence in the inferred edge. The color of edges captures the direction of the effect (i.e., if the nodes have positive or negative conditional dependence). It suggests this is the baseline interaction and matches up to known pathways.

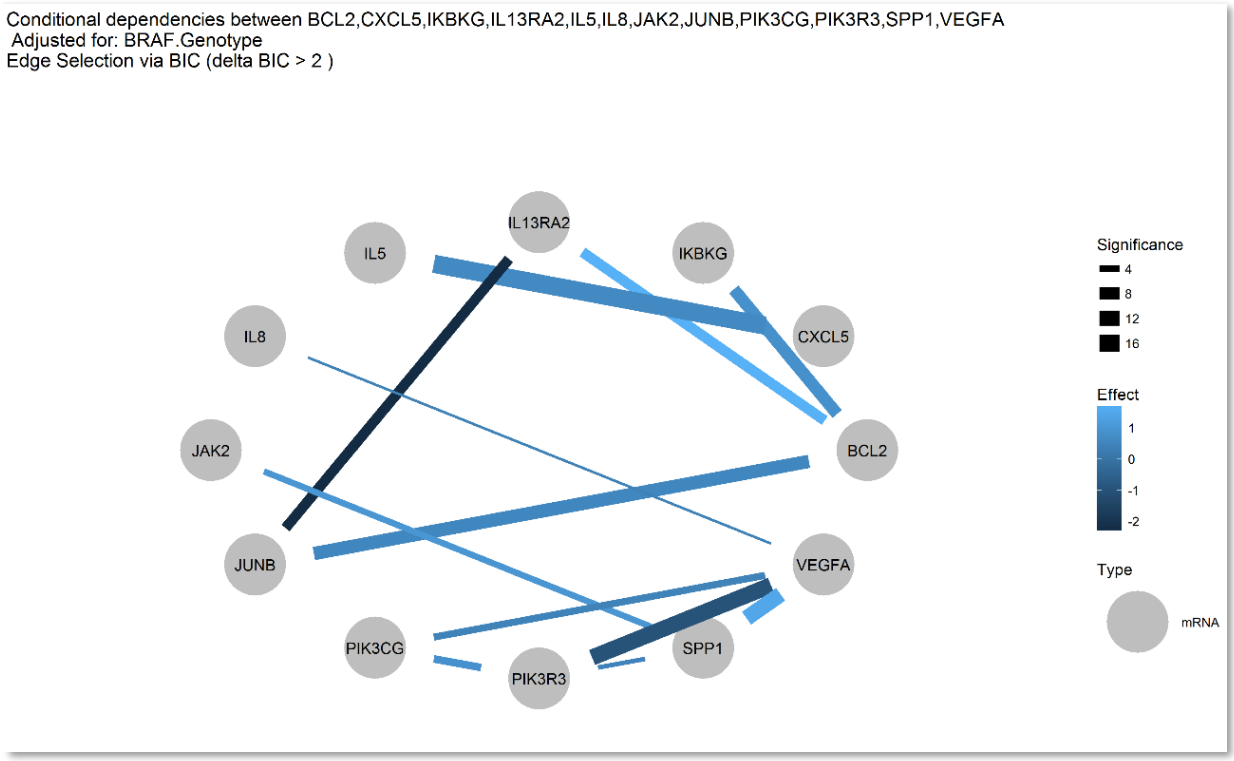


Figure 58: Probe Descriptive module - Interaction Network plot

Cell Type Profiling Module

This Advanced Analysis module uses the method described by Danaher (2017) to measure the abundance of various cell populations. The method quantifies cell populations using marker genes which are expressed stably and specifically in given cell types. These marker genes act as reference genes specific to individual cell types, as they are expressed only in their nominal cell type, at the same level in each cell. The closer the biomarker genes defined in the probe annotation are to this ideal scenario, the more reliable the scores.

Plots are categorized in three tabs along the top of the window: **QC**, **Summary**, and **Covariates**. Each tab's plots can be further categorized, on the left side of the window, as either a **Summary** or by each **Cell Type**.

The **QC** tab within this module displays p-values for correlation of marker gene expression. These p-values should be reviewed before examining the main cell type results. Cell types with high p-values and uncorrelated genes may still produce useful measurements, but will require more skepticism than other cell types.

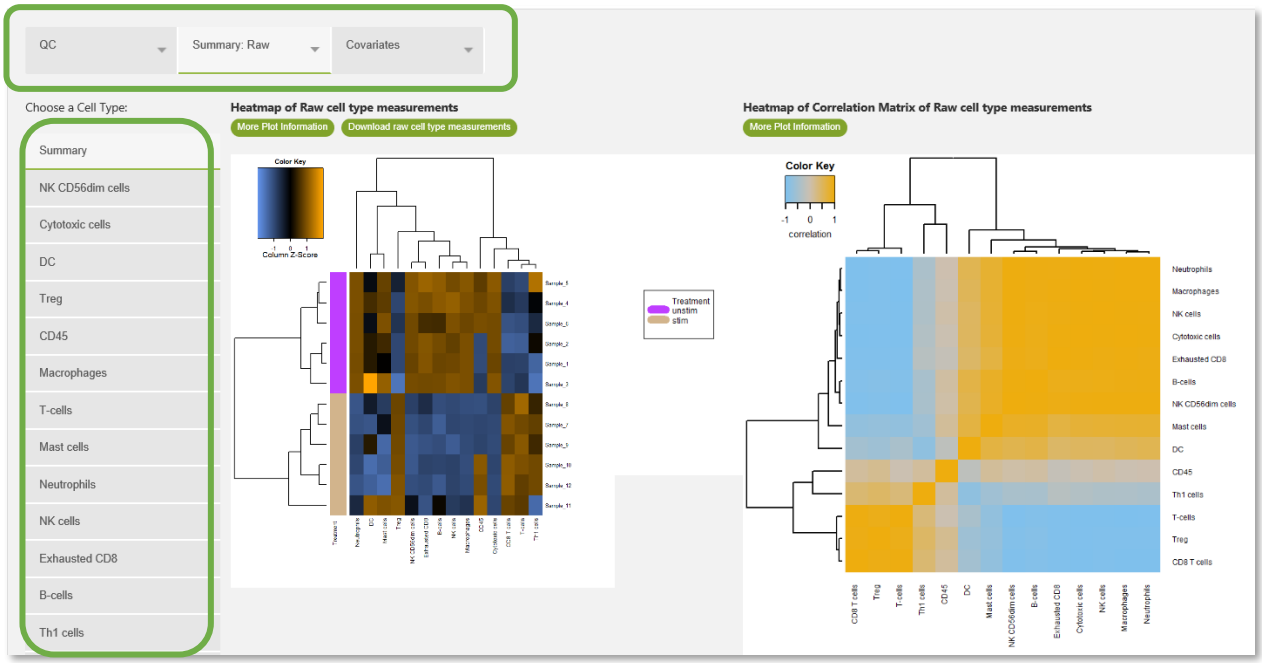


Figure 59: Cell Type Profiling window and options

Once the cell type QC plots have been reviewed, look at the cell type scores on the other tabs. Table 2 lists ways in which to use this data and ways in which it is somewhat limited. Note that because the scores are simple averages of marker gene expression, they convey no information about the absolute number of cells in a sample.

Table 2: Options and limitations of Cell Type Profiling

Comparison/question	Allowed
Calculate the number of cells in sample A	NO – Cell Profile is average of expression levels, and the number of transcripts per cell is unknown.
Compare a cell type's abundance between samples A & B	YES – If a cell type abundance measurement is increased by 1 between two samples, then there is a two-fold increase in the number of the cells present (abundance measurements are in the $\log_2$ space).
Compare the profiles of two cell types in sample A	NO – Cell Profile is average of expression levels for the selected genes, so a difference in values within a sample does not necessarily represent a difference in cell numbers.
Compare the ratio between two cell types in sample A & B	YES – We can claim, for example, that the number of T-cells relative to NK cells in sample A is twice that in sample B.
Compare profile for a cell type between two samples when one sample is from a different dataset	YES – The underlying assumption is that these are cell type-specific reference genes

The **Summary** and **Covariates** tabs allow you to analyze both Raw and Relative cell type abundance estimates.

- Raw cell type measurements are simple averages of the $\log_2$ expression of each cell type's marker genes.
- Relative measurements are calculated as contrasts between raw measurements. This may be useful since the abundance of most cell types might be highly correlated with the tumors' variable amounts of total infiltrate. Relative profiles better reveal differences in the composition of that infiltrate. Also, in PBMCs and other samples where tumor cells do not provide the majority of RNA, relative measurements can be much cleaner and easier to interpret than raw measurements.

Before You Start Cell Type Profiling

This module will only run with CodeSets in which a significant proportion of their genes are cell-type genes.

To run the Cell Type Profiling module, you must choose **Custom Analysis** as your **Analysis Type** and check the appropriate box on the **General Options** tab. Once you have done that, the Cell Type Profiling tab will appear in the list and you will be able to select it for customization (see the *Custom Options for Cell Type Profiling* section).

Custom Options for Cell Type Profiling

Your available sample **annotations (covariates)** will appear in the Available Annotations field. Use the green arrows to move over those annotations (at least one is required) that you want to examine in Cell Type Profiling.

The CodeSet's probe annotation file will designate which genes are cell type specific markers using the column with header "Cell.type". In addition, gene lists can be created by modifying the probe annotation file. To specify the cell types' characteristic probes (markers) select either **Use Default(cell.type)** or designate a **Custom** column. See the *Creating Probe Annotations for Custom CodeSet Data* section.

The question mark button reveals additional information.

The exclamation mark button reveals an alert and brief explanation as to why an option may be unavailable (greyed out).

In **Creating Signatures**, the **Dynamically Select a Subset** option will reject genes that do not behave like marker genes (genes which are poorly correlated with the other markers for the cell type; see Danaher 2017 for details). These will appear with the word "discarded" underneath in some plots (see the *QC Plot for Cell Type Measurements of Choice* section). The **Use All Probes** setting bypasses this QC step and retains all genes, regardless of whether they display cell type specific correlated expression.

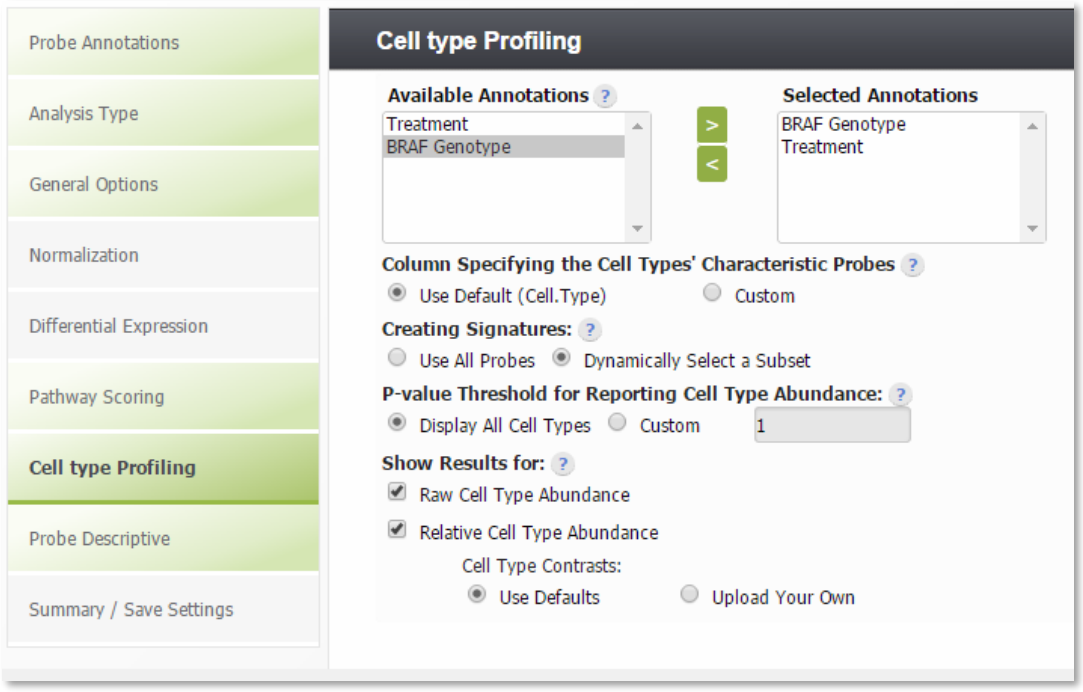


Figure 60: Cell Type Profiling custom options menu

The software tests each cell type's marker genes for better-than-random marker-like co-expression (Danaher 2017) and returns a p-value for each cell type score. The **P-value Threshold** defines the significance threshold for reporting a cell type abundance estimate.

- By default, the module will **display all**, returning results for all cell types regardless of their p-value. This setting may be desirable since gene sets with high p-values may still be useful: even if your dataset does not provide high confidence values, the results of previous authors provide enough evidence to make their use a reasonable choice.
- Alternatively, you may choose **Custom** and enter a value of 0.05 or lower to see results (and calculate relative cell scores, see below) only for cell types whose quantification is further supported by your data. Cell types whose evidence for cell type-specific expression does not meet this level of confidence will be discarded.

Show Results for allows choices in how results are displayed:

- Raw cell type abundance shows the estimated abundances of each individual cell type. Abundance estimates are given on the $\log_2$ scale, so a unit increase in score corresponds to a doubling of a cell type's abundance. As each abundance estimate is simply the average of the $\log_2$ counts of chosen characteristic genes (cell.type genes), these estimates do not support claims about whether one cell type is more abundant than another. Rather, they permit claims that a cell type is more abundant in one sample than in another.

- Relative cell type abundances show *contrasts* between pairs of cell types. For example, rather than measuring CD8 T cell abundance, a relative cell type score measures CD8 abundance relative to overall T cell abundance. A relative abundance measurement is especially useful in a sample comprised of a heterogeneous mix of cell types such as PBMCs.

- Contrasts are ratios of the cell type scores in the form of cell type 1/cell type 2. They will only be displayed if a cell type profile is generated for both the numerator and the denominator. If you wish to upload your own cell type contrasts, you can generate a contrast matrix using a template similar to that shown in Figure 61 and save it as a .csv. You can then select the **Upload Your Own** option on the Custom Analysis menu and **Choose File**. See *Cell Type Profiling Algorithm Details* for more information.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Total TILs	B-cells vs. TILs	Cytotoxic cells vs. TILs	DC vs. TILs	Exhausted CD8 vs. TILs	Macrophages vs. TILs	Mast cells vs. TILs	Neutrophils vs. TILs	NK CD56dim cells vs. TILs	NK cells vs. TILs	T-cells vs. TILs	Th1 cells vs. TILs
2	**B-cells**	0.2	0.8	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.
3	**CD45**	0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.
4	**Cytotoxic**	0.2	-0.2	0.8	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.
5	DC	0	0	0	1	0	0	0	0	0	0	0	
6	Exhaustec	0	0	0	0	1	0	0	0	0	0	0	
7	**Macropha**	0.2	-0.2	-0.2	-0.2	-0.2	0.8	-0.2	-0.2	-0.2	-0.2	-0.2	-0.
8	Mast cells	0	0	0	0	0	0	1	0	0	0	0	
9	Neutroph	0	0	0	0	0	0	0	1	0	0	0	
10	NK CD56d	0	0	0	0	0	0	0	0	1	0	0	
11	NK cells	0	0	0	0	0	0	0	0	0	1	0	
12	**T-cells**	0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.8	-0.
13	Th1 cells	0	0	0	0	0	0	0	0	0	0	0	

Figure 61: Custom cell type contrast matrix file

Interpreting Results of Cell Type Profiling Plots

QC Plots

Barplot of p-values across cell types

p-values from the test for marker-like co-expression are $-\log_{10}$ transformed. Bars above the solid black line indicate statistically significant cell types at a p-value threshold of 0.01. Bars above the dashed black line indicate statistically significant cell types at a p-value threshold of 0.001.

In the RNA-Protein dataset used in this example (Figure 62), Neutrophils and Cytotoxic Cells are the cell types with the most significant p-values.

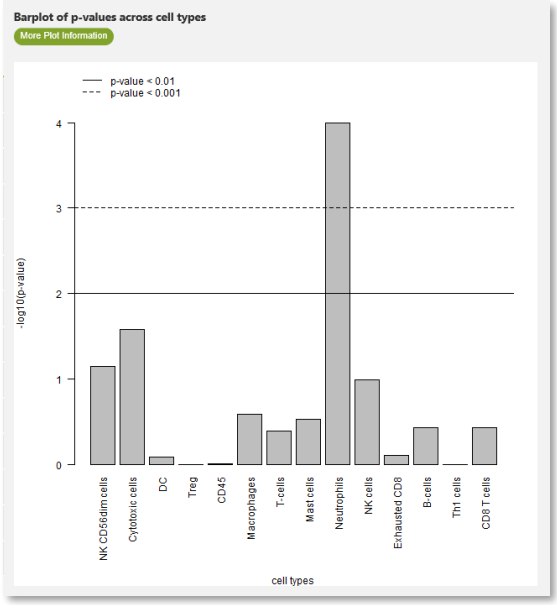


Figure 62: Cell Type Profiling module - QC barplot

QC Plot for Cell Type Measurements of choice

Stable cell type-specific expression of biomarkers allows us to score the cell type's abundances simply by taking the average $\log_2$ expression of its characteristic genes. Selecting the cell type of choice from the left side of the window allows you to view the normalized expression of the genes found to be characteristic of that cell type. If a cell type's characteristic genes are specific to the cell type and stably expressed within it, they will be strongly correlated with a slope of 1. Substantial departures from this pattern indicate noisier quantification of cell type abundance. The resulting image (Figure 63) plots each cell's results vs. another twice (once on one side of the diagonal and once on the other). The boxes on the diagonal each contain a cell name; all plots in the same row will have this cell on their y-axis and all plots in the same column will have this cell on their x-axis. "Discarded" under the cell name indicates that the correlation between cell types was so poor that they qualified to be dropped.

In the RNA-Protein dataset used in this example (Figure 63), SH2D1A and CD3D have the best correlation; CD3E (both mRNA and protein) are so poor, they are discarded.

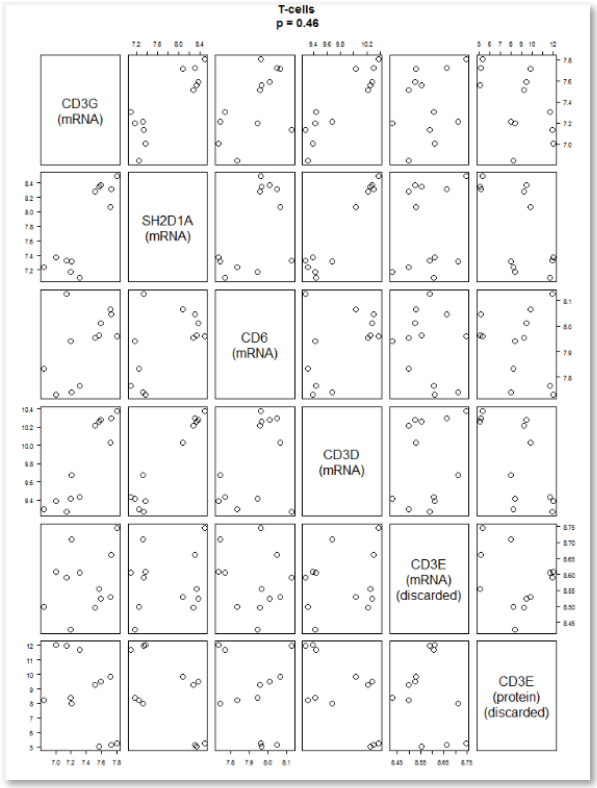


Figure 63: Cell Type Profiling module - QC cell type

Summary Plots

The summary plots can be viewed for **raw** or **relative** data. The summary plots can be viewed for **raw** or **relative** data. Each relative abundance score gives a contrast between cell types' measurements. The **Total TILs** score is defined as the average of the B cell, T cell, CD45, Macrophage and Cytotoxic cell scores. Other relative abundance scores are calculated by subtracting the total TILs score from a single cell type score. For example, the NK cells vs. TILs score is the NK cell score minus the total TILs score.

The **heatmap of raw (or relative) cell type measurements** is a descriptive plot showing the abundance of different cell types. Cell types are listed on the horizontal axis and samples are listed vertically. Orange indicates high abundance whereas blue indicates low abundance. This method does not support comparisons of different cell types. Rather, it supports comparisons of abundance of the same cell type between samples.

In the RNA-Protein dataset used in this example, we can see that the first six samples (the unstimulated group) display a high abundance of several cell type groups where samples 7-12 (the stimulated group) have a low abundance. T-Cells and CD8 T-Cells show the opposite: these groups had low results in the unstimulated samples and higher in the stimulated group.

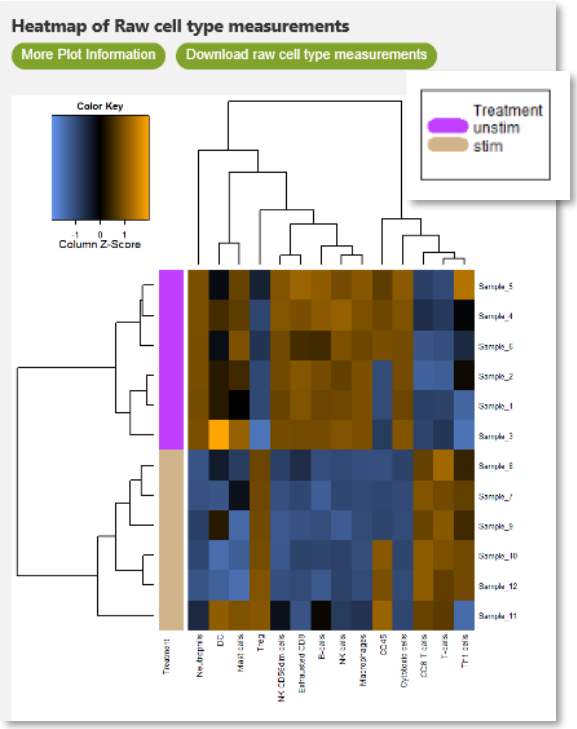


Figure 64: Cell Type Profiling module - cell type heatmap

The **heatmap of correlation matrix of raw (or relative) cell type measurements** shows the correlation between different cell types. Cell types are listed on both the horizontal and vertical axes. Gold shows highly correlated cell types and blue shows highly anti-correlated cell types.

In the RNA-Protein dataset used in this example, we can see that while most of the cell types are highly correlated, T-Cell groups are anti-correlated with them.

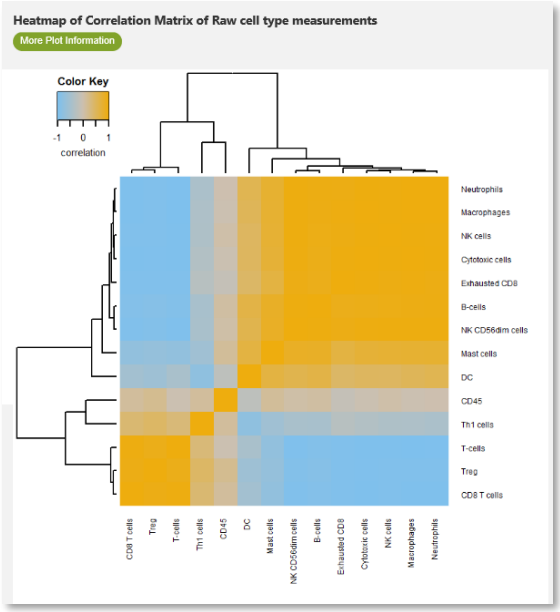


Figure 65: Cell Type Profiling module - cell type correlation

Raw (or relative) cell type measurements vs. other cell type measurements, by covariate

By clicking on a tab for a specific cell type, we can more closely examine its behavior relative to other cell types or cell type ratios.

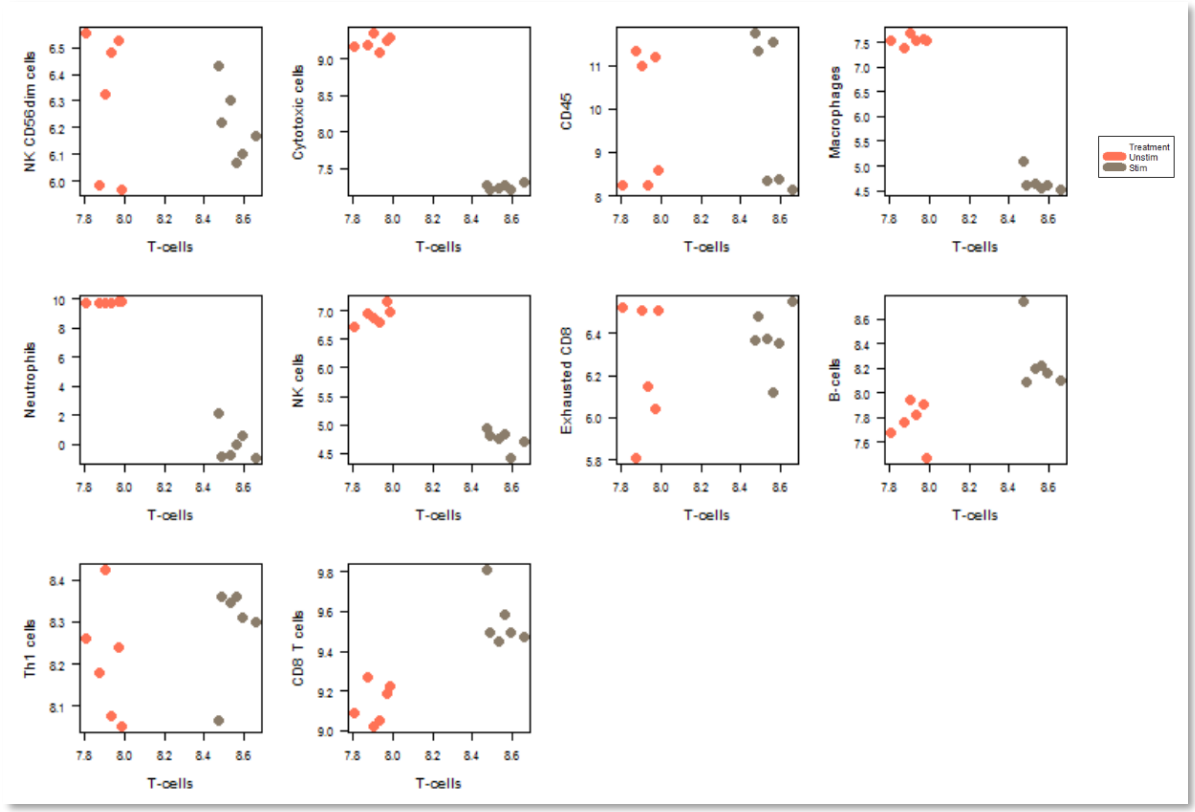


Figure 66: Cell Type Profiling module - cell type measurements, compared

In the RNA-Protein dataset used in this example, we have chosen T-Cells from the tab from the left side of the window. These cells' scores are plotted against each other cell score, colored by treatment. We can see separation between stimulated and unstimulated samples in all plots.

Covariates Plots

The **raw (or relative) cell type measurements vs. covariate** plots the cell type abundance measurements against each selected covariate.

In the RNA-Protein dataset used in this example, we can visualize the change in expression between one experimental group and the other for each cell type. As we saw in the QC barplot, Neutrophils had the biggest change between the unstimulated and the stimulated states.

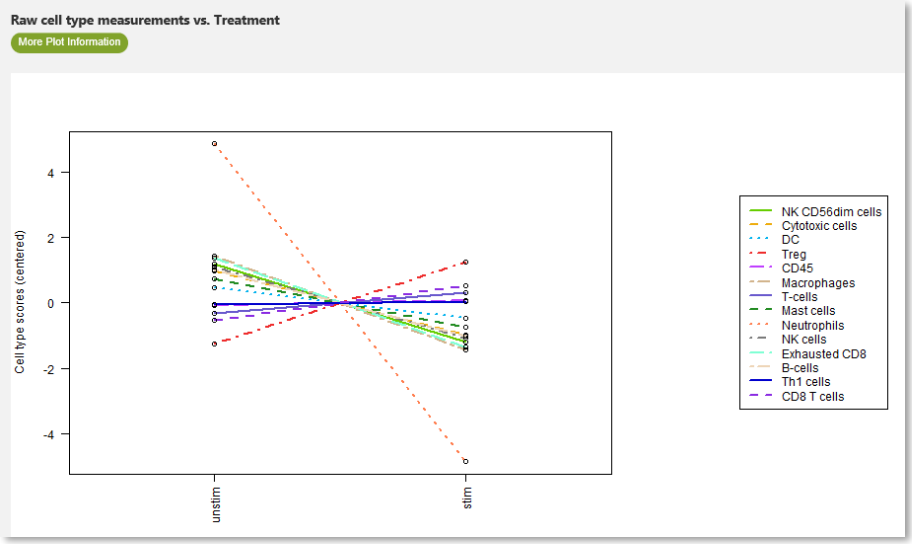


Figure 67: Cell Type Profiling module - cell type measurement by covariate

With **raw (or relative) cell type measurements vs. covariate,** we can examine the relationship between cell populations and selected covariates. Each cell type's score has been centered to have mean 0. As abundance estimates (cell type scores) are calculated in $\log_2$ scale, an increase of 1 on the vertical axis corresponds to a doubling in abundance.

In the RNA-Protein dataset used in this example, we can see the difference between the unstimulated samples' T-cell scores and those of the stimulated samples.

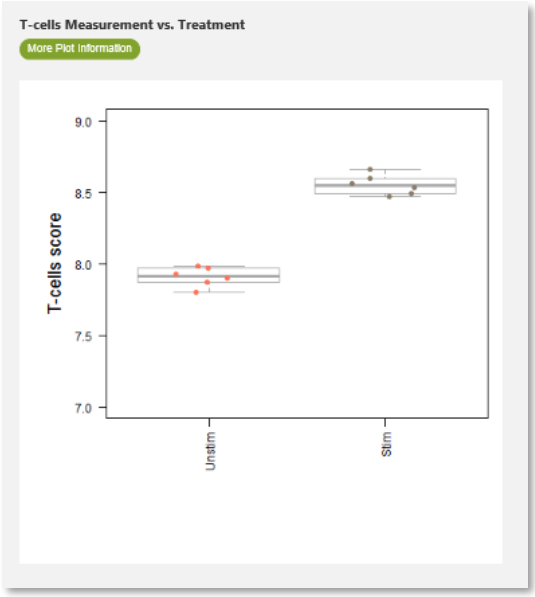


Figure 68: Cell Type Profiling module - cell type vs covariate

Cell Type Profiling Algorithm Details

A cell type's abundance can be measured as the average log-scale expression of its characteristic genes. The algorithm used to identify appropriate marker genes and exclude badly-behaving cell type-specific genes from estimates of cell type abundance is detailed below, as is the permutation test used to derive a p-value assessing a cell type's marker genes. Automatic Screening of Failed

Cell Type-specific Genes

First, we define a similarity metric between two candidate cell type-specific genes. Under the assumption that both genes are specific to the same cell type and consistently expressed within it, they will be highly correlated with a slope of 1. To measure two gene's adherence to this pattern, we employ a slightly modified version of Pearson's correlation metric:

$$similarity(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\frac{(n-1)}{2}(var(x)+var(y))},$$

where *x* and *y* are the vectors of log-transformed, normalized expression values of the two genes, $\bar{x}$ and $\bar{y}$ are their sample means, and var(x) and var(y) are their sample variances. The *similarity()* function equals 1 when the two genes are perfectly correlated with slope of 1 and decreases for gene pairs with low correlation or slope diverging from 1. Since many biologically related genes will exhibit correlation unrelated to a shared cell type, it is important to apply a more stringent measure of similarity than mere correlation.

Our gene selection algorithm is as follows. Assume there are *p* genes and *n* samples.

1. Use the similarity function to compute a *p*p* similarity matrix amongst the genes. Each gene has similarity of 1 with itself.
2. Label all gene pairs with similarity below 0.2 as "discordant."
3. Iteratively remove genes: while there are more than 2 genes remaining and while at least one discordant pair of genes remains:
    a. Count the number of discordant pairs each gene participates in. Call the maximum of these counts *n_discord*.
    b. Identify the genes with n_discord instances of discordance with another gene. Of these genes, remove the single gene with the lowest average similarity to the other remaining genes.

The above process is similar to the geNorm algorithm. This similarity is not a coincidence, as cell type markers genes can be thought of cell type-specific "housekeeper" genes.

Calculation of p-values for Cell Type Gene Sets

We test the null hypothesis that the given gene set exhibits no greater cell type-specific-like behavior than a randomly selected gene set of similar size.

First, we require a metric of a gene set's adherence to the assumption of cell type-specific and consistent expression.

$$concordance(X) = \frac{1}{trace(Cov(X))}\left(p^{-\frac{1}{2}}, \ldots, p^{-\frac{1}{2}}\right) Cov(X) \left(p^{-\frac{1}{2}}, \ldots, p^{-\frac{1}{2}}\right)^{T},$$

where X is the matrix of log-transformed, normalized expression values of the gene set, and where p is the number of genes. The concordance() function evaluates at 1 if all genes are perfectly correlated with a slope of 1, and degrades to 0 as this pattern weakens.

We perform our permutation test as follows. Assume the given gene set has p genes, of which p0 survived the iterative gene selection procedure. Call the data from the gene set X, and the data from the reduced gene set X0.

1. Compute *concordance(X_0)*.
2. Choose 1000 random genes sets of size *p*. Denote the data from a random gene set *X'*.
3. For each gene set, apply the criteria of the gene selection algorithm to reduce *X'* to only its best p_0 genes. Call the data from this reduced random gene set X_0', and compute *concordance(X_0')*.
4. Return a p-value equal to the proportion of *concordance(X_0')* values greater than *concordance(X_0)*.

Also note that there are 3 single-gene cell type scores. These scores cannot be tested with this method; however, the genes in question (CD45 for CD45 cells, Tbx21 (T-bet) for Th1 cells, and FOXP3 for Tregs) are well-characterized.

Cell Type Contrast Matrix File

To generate TIL cell scores using a custom matrix file, the Advanced Analysis Cell Type Profiling module does the following:

1. It calculates raw cell scores as mean $\log_2$ normalized counts of cell type specific markers. Genes included to generate the scores must be those that correlate and behave as cell markers (Danaher, 2017).
2. It correlates all cell type scores vs CD45 scores using the PEARSON equation. The cell scores with correlation coefficients >0.6 will be used to average and get TIL scores.
3. It normalizes raw cell scores to TIL scores.

Here is an example of steps 2 and 3 using 5 cell populations to generate TIL scores: B cells, CD45+ lymphocytes, Cytotoxic T cells, Macrophages and T cells:

Column B in the contrast matrix will be used by nSolver to generate a TIL cell score as the average of (in this example) 5 cell populations highlighted in red (B cells, CD45+ lymphocytes, Cytotoxic T cells, Macrophages and T cells).

$$\mathrm{TIL}_{\mathrm{score}} = \frac{\mathrm{B}_{\mathrm{cell\ score}} + \mathrm{CD45}^{+}_{\mathrm{cell\ score}} + \mathrm{Cytotoxic}_{\mathrm{cell\ score}} + \mathrm{Macrophage}_{\mathrm{cell\ score}} + \mathrm{T}_{\mathrm{cell\ score}}}{5} \quad \text{(Eq 1)}$$

This equation is equivalent to:

$$\mathrm{TIL}_{\mathrm{score}} = 0.2\mathrm{xB}_{\mathrm{cell\ score}} + 0.2\mathrm{xCD45}^{+}_{\mathrm{cell\ score}} + 0.2\mathrm{xCytotoxic}_{\mathrm{cell\ score}} + 0.2\mathrm{xMacrophage}_{\mathrm{cell\ score}} + 0.2\mathrm{xT}_{\mathrm{cell\ score}} \text{ (Eq 2)}$$

The coefficients of the scores in the second equation are annotated in the table (see the Custom Options for Cell Type Profiling section, Figure 61), column B (in red).

Column C is used by Advanced Analysis to generate a contrast ratio of B cells, relative to TIL. In linear space :

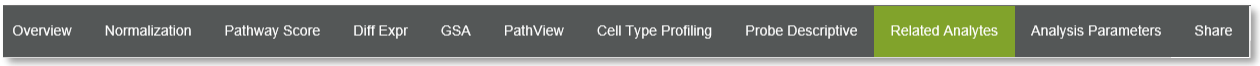
$$\text{Bcells relative to TIL} = \frac{\mathrm{B}_{\mathrm{cell\ score}}}{\mathrm{TIL}_{\mathrm{score}}} \text{ (Eq 3)}$$

This equation is equivalent to:

$$\frac{1\mathrm{xB}_{\mathrm{cell\ score}}}{0.2\mathrm{xB}_{\mathrm{cell\ score}} + 0.2\mathrm{xCD45}^{+}_{\mathrm{cell\ score}} + 0.2\mathrm{xCytotoxic}_{\mathrm{cell\ score}} + 0.2\mathrm{xMacrophage}_{\mathrm{cell\ score}} + 0.2\mathrm{xT}_{\mathrm{cell\ score}}}$$

To annotate this equation on the contrast matrix, we use positive coefficients for scores in the numerator, and negative coefficients for scores in the denominator. Since B cell scores appear in both the numerator and denominator, the coefficient will be the sum of the coefficients. For this example: $1 - 0.2 = 0.8$.

Related Analytes Module



This module enables comparison of expression levels within pairs of probes that have been linked in the probe annotation file. Pairs such as mRNA-Protein or Total Protein-Phosphorylated Protein may already be linked in some files. This module applies all the tools of the *Probe Descriptive Module* to each pair of related analytes. It is especially useful for describing the co-regulation of Protein and mRNA counterparts or of Phosphorylated isoforms.

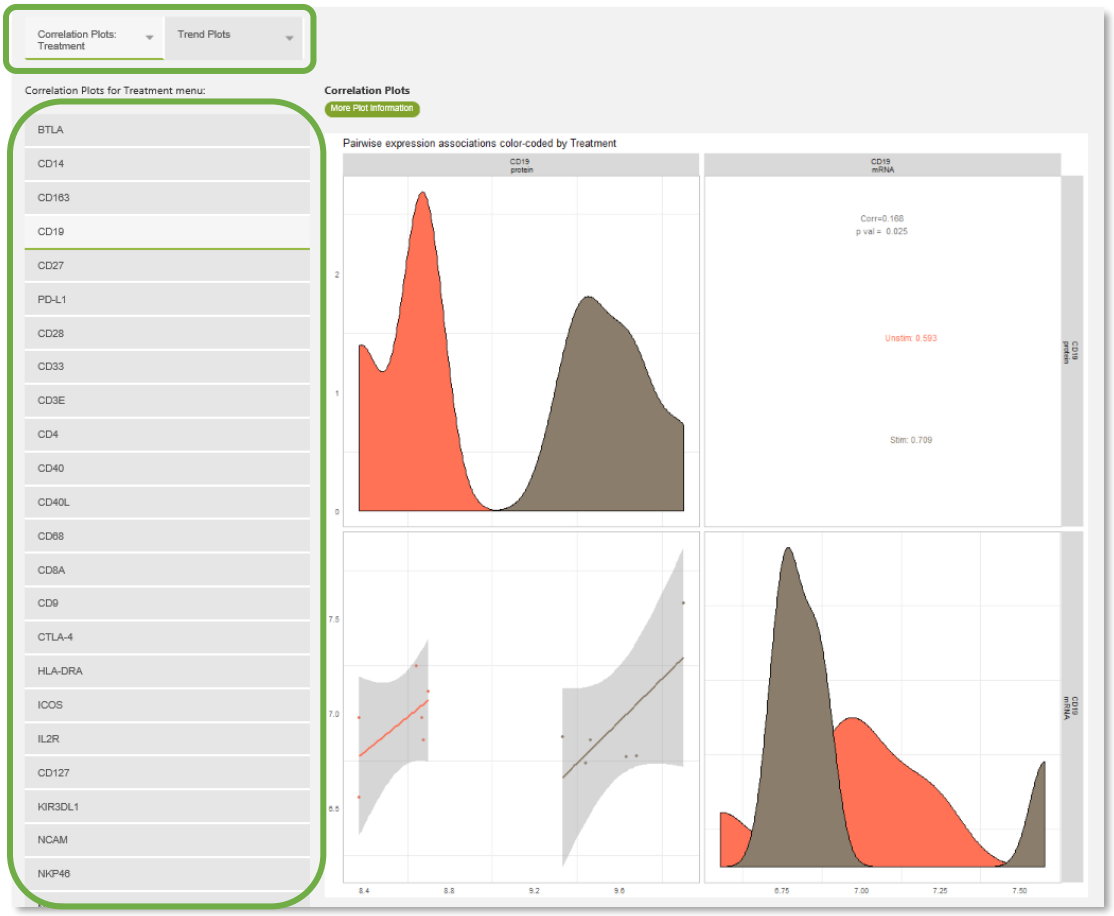


Figure 69: Related Analytes view and options

Before You Start Related Analytes

The CodeSet's probe annotation file will designate which mRNA markers are related to which proteins. To specify or modify these relationships, use the **Related.Probes** column in the probe annotation file. For any probe of interest, enter the probe ID of its related counterpart in the Related.Probes column, and vice versa. In this way, you can link any two probes to look at pairs of mRNA probes for splice variants or pairs of Protein probes (phosphorylated:non-phosphorylated). See the *Creating Probe Annotations for Custom CodeSet Data* section.

Custom Options for Related Analytes

The Related Analytes menu allows you to **Select Probe Pairs** for analysis. The options provided are mRNA and Protein probe pairs that have been defined as related in the Related.Analyte column of Probe Annotations file. Select the pairs of interest from the field on the left and move them to the field on the right with the green arrow button. At least one must be selected

The output graphs will be colored by the categories selected in the **Grouping Annotations** field (for continuous variables, Low, Medium, and High subsets will be computed). Move the annotation(s) desired from the left field to the right using the green arrow button.

The question mark button reveals additional information.

The exclamation mark button reveals an alert and brief explanation as to why an option may be unavailable (greyed out).

You can check the box to **Generate Trend Plots** if you have covariates to designate as **Interval ID** and as **Series ID**. The interval ID can be an ordered categorical or continuous variable. Additionally, trends across distinct sample annotation groups can be examined by specifying an optional stratifying annotation.

- **Interval ID** is the variable that defines how the data points are ordered along the trend (horizontal axis in plots). Typical covariates that would be specified as Interval IDs are Time (as in the example below – Figure 70), Concentration, and Dosage; there should be three or more groups in this variable.
- **Series ID** defines the groups into which we wish to separate the samples (for example, patient cohorts). In general, the definition of group could extend to the case where each group consists of only one observed entity (for example, one patient). The example below uses BRAF Genotype.
- **Stratifying Annotation** allows you to separate the series ID into groups to see a trend. Since we are interested in how Treatment affects each BRAF genotype (chosen as Series ID, below), we will select it as our stratifying annotation.

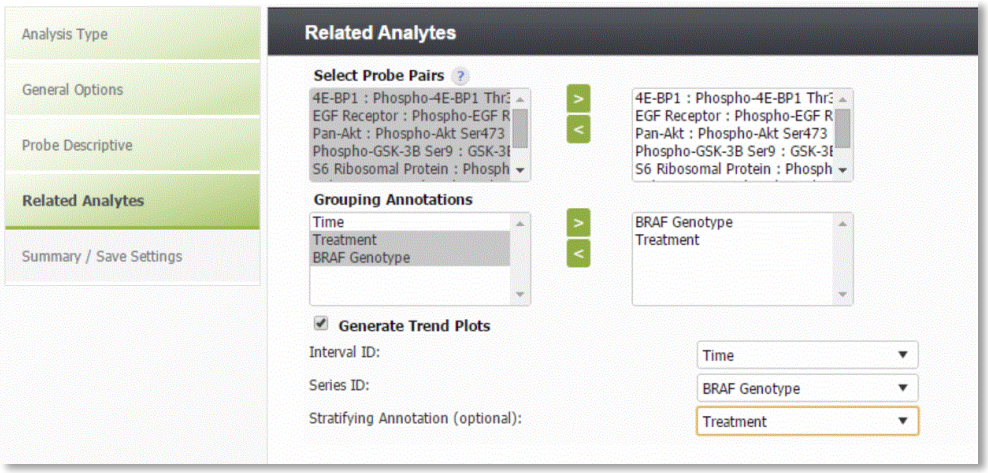


Figure 70: Related Analytes custom analysis menu

Interpreting Results of Related Analytes Plots

Correlation Plots

The correlation plot allows visualization of two sets of information: **distribution of gene expression** and **correlation of gene expression**. When the covariate of interest is continuous, the values are categorized into low, average and high. Each field belongs to the gene listed at the top of its column and the gene listed on the right side of its row.

- The **distribution of expression** for each gene is drawn on the diagonal, segregating experimental groups belonging to the chosen covariate by color.
- The **correlation of gene expression** for each pair of genes is expressed **numerically** in the top right field as the overall Pearson correlation coefficient and the p-value. Correlation of gene expression segregating covariate groups is also given; groups are separated by color.
- The **correlation of gene expression** for each pair of genes is expressed **graphically** in the lower left field, plotting the expression values and separating the groups by color.

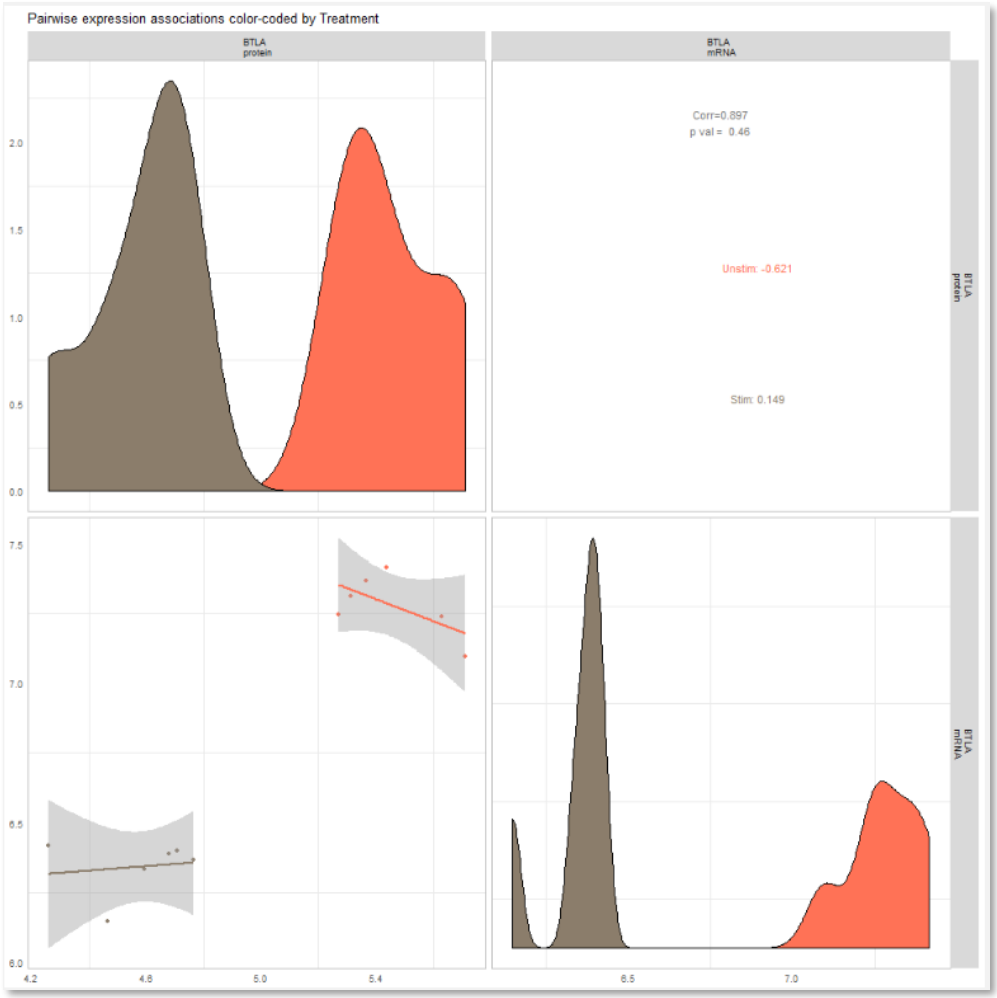


Figure 71: Related Analytes module – Correlations Plot

Trend Plots

This plot is designed to enable tracing of the change in expression levels of the related probes of a sub-category of the variable of interest. Examples of this sub-category could be individual patients, a cell line, or a patient cohort and the variable of interest is (typically) time, concentration, dosage, or order of observation. Choose the probe-pair of interest from the tabs on the left.

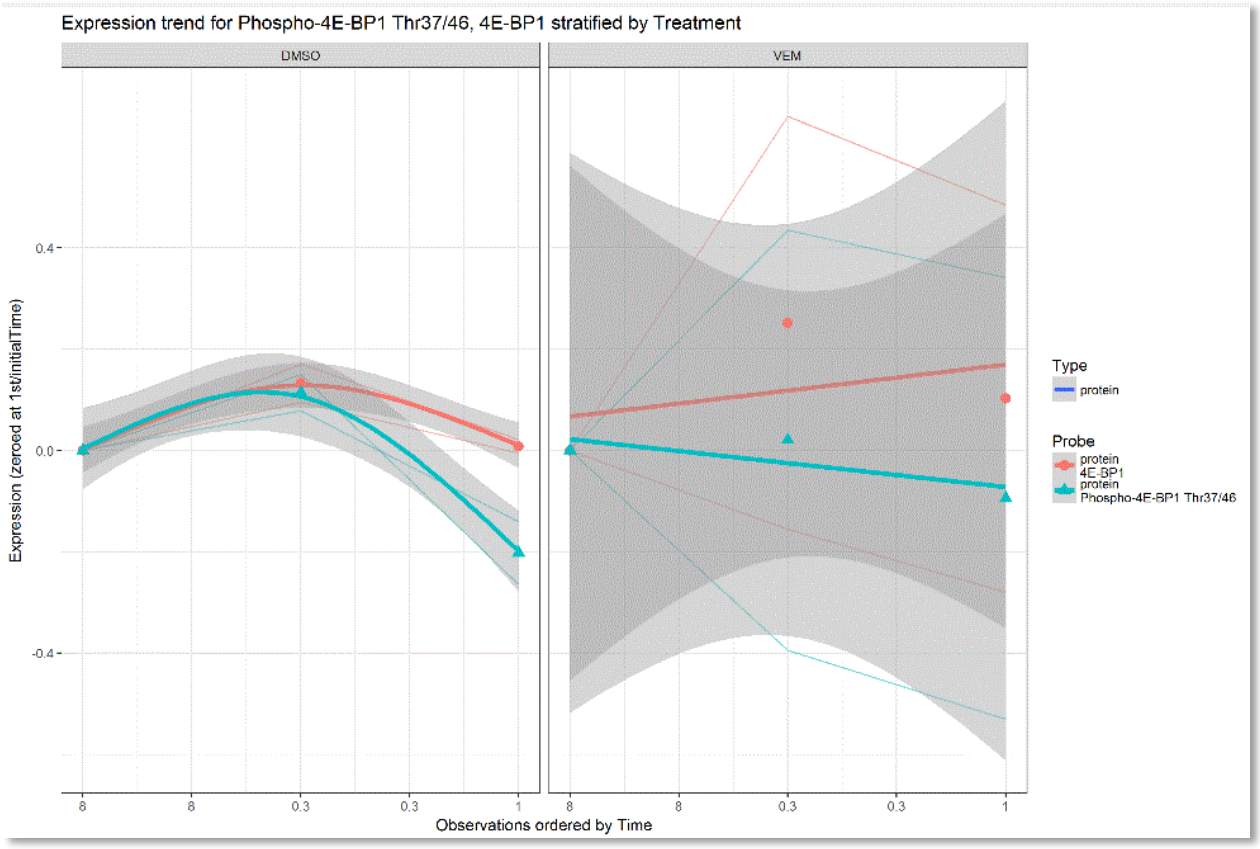


Figure 72: Related Analytes module - Trend Plot

In the example in Figure 72, we see Trend plots, stratified by Treatment type. For the settings used to create this plot, see the *Custom Options for Related Analytes* section. Time is the Interval ID, which establishes the x-axis. Each probe's expression is plotted in a single color over time in three lines. The narrow lines represent each category of the variable set as the Series ID (in this case, BRAF genotype); the thick line represents the average of these values. We can see that all probes responded similarly over time in the DMSO (control) group.

SNV Module

The SNV Module summarizes SNV variant events detected in the data through three different types of plots.

From the Overview tab, you can choose between the **Sample Probe Matrix** or the **Quality Map**. The **Sample Probe Matrix** indicates sites which have tested positive for a variant. The **Quality map** displays the raw counts of the control probes for each sample in the dataset.

The **By Samples** tab allows you to view sample-specific boxplots of the $\log_2$ ratio of counts for each variant probe compared to reference data.

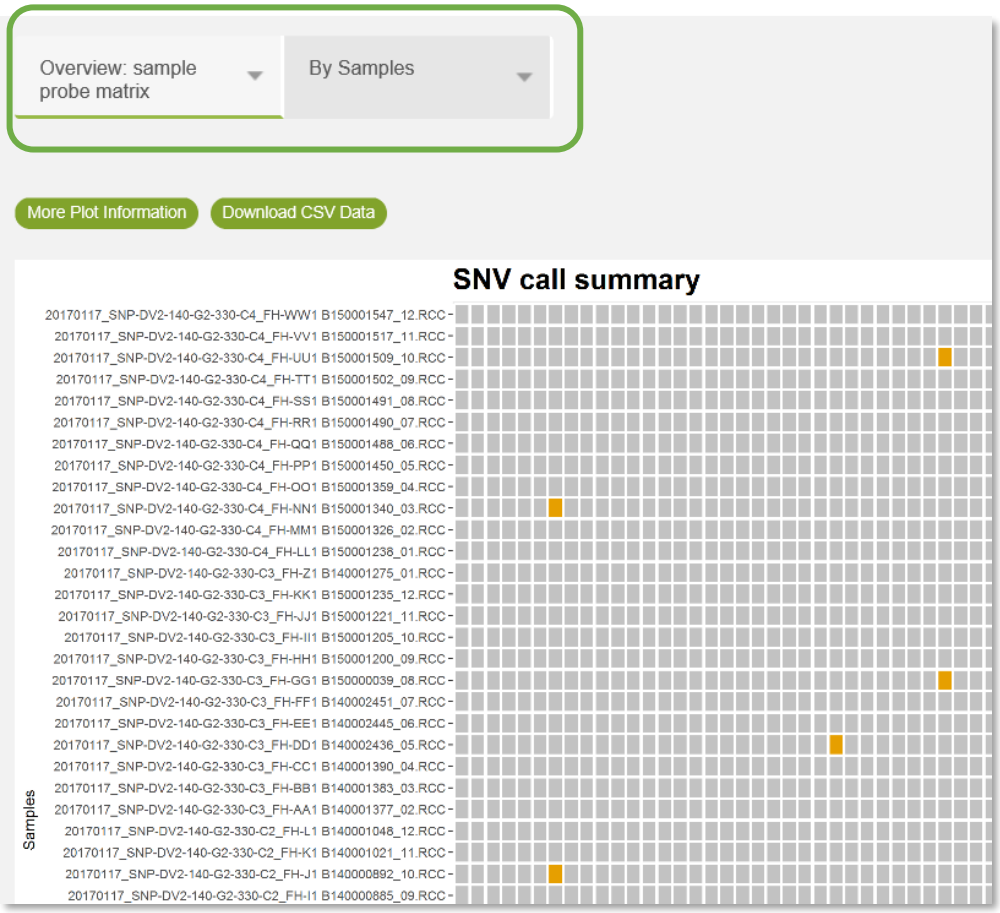


Figure 73: SNV module view and options

Before You Start SNV

SNV analysis requires a set of SNV references to establish a baseline for variant calls. This reference set can be run on a different RLF but should use the same type of unique identifier as the sample set (i.e., if the Description column is used to document shorter sample names for the sample set, you must enter unique same names in the SNV reference set's Description column, as well).

Potential Cross-Hybridization Interactions

Due to the complex, competitive hybridizations that form the foundation of SNV chemistry, there are certain assays that, in order to ensure sensitivity down to 5% allele frequency, may also have affinity for other variant sequences in the assay. These interactions can result in false-positive calls among related probes assaying the same hotspot regions in the genome. Known potential variant cross-hybs are listed in the tables below. Exercise caution when analyzing data that shows positive results in these pairs of assays. The strongest call will likely be the assay listed in the "…When True Positive Present" column, and a weaker, secondary call may appear for the assay listed in the column "Putative False Positive…"

For example, in the Heme panel, when CSF1R COSM947 (Y969C) is present, you have a low chance that CSF1R COSM948 (Y969F) calls will be falsely elevated.

Table 3: Heme Panel Potential Hybridization Pairs

Putative False Positive…	…When True Positive Present	Probability
CSF1R COSM948 (Y969F)	CSF1R COSM947 (Y969C)	Low
DNMT3A COSM52944 (R882H)	DNMT3A COSM99740 (R882P)	Low
FLT3 COSM27650 (D835A)	FLT3 COSM784 (D835V)	Low
IDH1 COSM28748 (R132S)	IDH1 COSM28749 (R132G)	Medium
IDH2 COSM41875 (R140L)	IDH2 COSM41590 (R140Q)	Medium
KIT COSM1310 (D816Y)	KIT COSM1311 (D816H)	High
KIT COSM1311 (D816H)	KIT COSM1310 (D816Y)	Medium
KRAS COSM512 (G12F)	KRAS COSM516 (G12C)	Medium
KRAS COSM512 (G12F)	KRAS COSM520 (G12V)	Medium

Table 4: Solid Tumor Panel Potential Hybridization Pairs

Putative False Positive…	…When True Positive Present	Probability
BRAF COSM473 (V600K)	BRAF COSM476 (V600E)	Low
BRAF COSM475 (V600E)	BRAF COSM476 (V600E)	Low
EGFR COSM12370 (L747_P753>S)	EGFR COSM12369 (L747_T751delLREAT)	High
EGFR COSM12370 (L747_P753>S)	EGFR COSM6255 (L747_S752delLREATS)	High
EGFR COSM12384 (E746_S752>V)	EGFR COSM12416 (E746_T751>VA)	High
EGFR COSM6223 (E746_A750delELREA)	EGFR COSM6225 (E746_A750delELREA)	High
EGFR COSM6255 (L747_S752delLREATS)	EGFR COSM12382 (L747_A750>P)	High
KRAS COSM549 (Q61K)	KRAS COSM550 (Q61E)	Low
KRAS COSM555 (Q61H)	KRAS COSM554 (Q61H)	Low
NRAS COSM585 (Q61H)	NRAS COSM586 (Q61H)	Low

Custom Options for SNV

There is no SNV custom options menu, however, the General Options menu will include a **Specify SNV Parameters** button if it detects SNV data in the set. This button allows you to designate the reference samples, adjust the minimum fold change, and adjust the p-value to modulate SNV calling stringency.

You may select **Quick Analysis** and choose one covariate from the dropdown for analysis.

Alternatively, you may select **Custom Analysis** if you would like to run a multi-RLF analysis, choose multiple covariates, or customize your analysis in another way. The General Options tab will appear (see Figure 74).

Select the **SNV Analysis Parameters** button to reassign SNV References by covariate type (the default is *Is Reference*, referring to the assignment made during nSolver experiment creation) or by file name (manual selection). You may adjust the parameters defining the reference thresholds (*Detected* and *Not Detected*), based on number of $\log_2$ fold changes ($\log_2$FC) and p-value, however, ensure these values are not identical. The algorithm is not designed for a dichotomous output and will reset to default values if it does not detect a lower $\log_2$FC threshold for the *Not detected* setting.

By default, the **EM** (expectation maximization) and **Debias** boxes are checked. In most circumstances, you won't need to deviate from this default. If, however, you are troubleshooting unexpected results, check one of these boxes off at a time and view this effect on your data. EM facilitates the borrowing of information from sample- to-sample; this is a useful model, but if one of the samples is of poor quality, you may need to check this option off to keep this sample from impacting the others. Debiasing is a by-sample bias removal procedure.

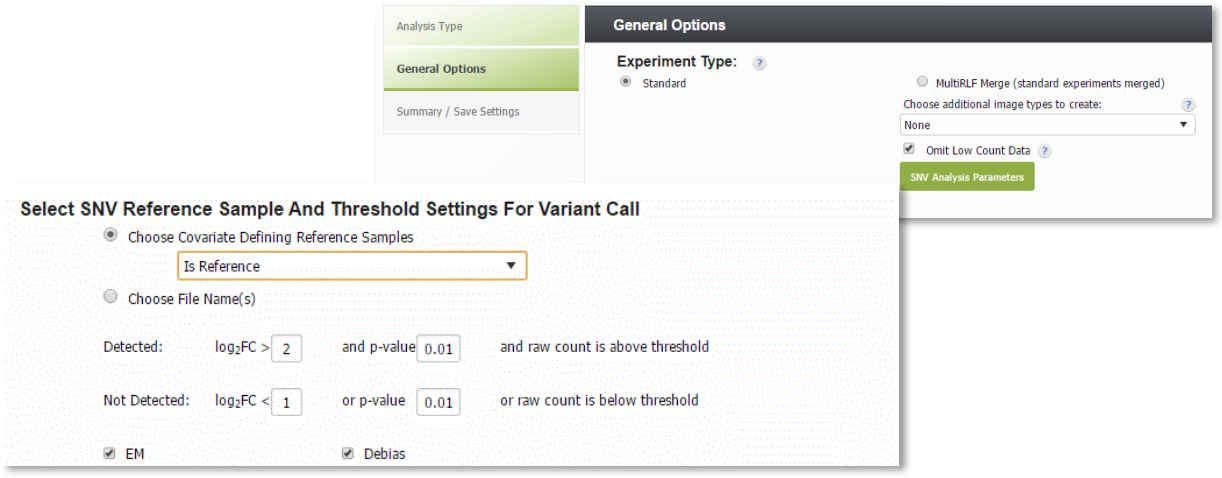


Figure 74: Windows associated with SNV custom analysis options - General Options menu

Interpreting Results of SNV Plots

Overview Sample Probe matrix

This plot shows the samples and probes where SNVs were called. The legend below the plot details the rules for calling SNVs from raw counts, p-values and fold-changes above expected background.

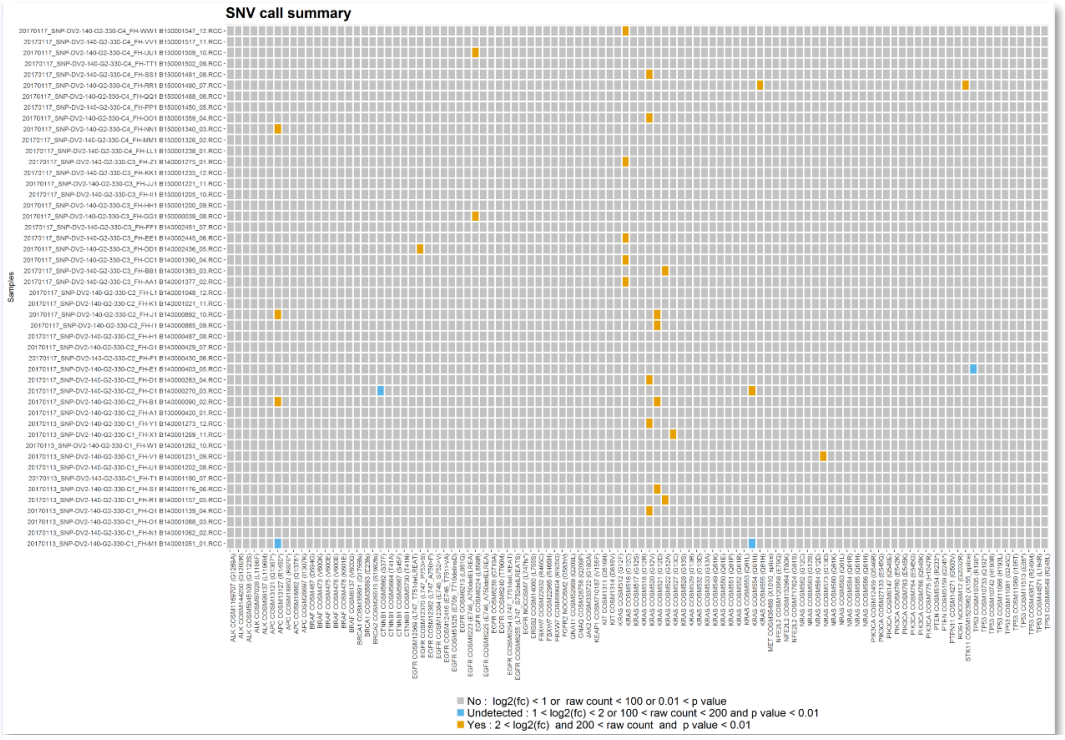


Figure 75: SNV module - Overview plot

In the **3D Bio Data Example** (see *Appendix A*), the **SNV call summary** gives a clear depiction of the SNV calls made in this data. Results are as expected: SKMEL28 samples all exhibited variant calls in the BRAF gene, while SKMEL2 samples all exhibited variant calls in the NRAS gene.

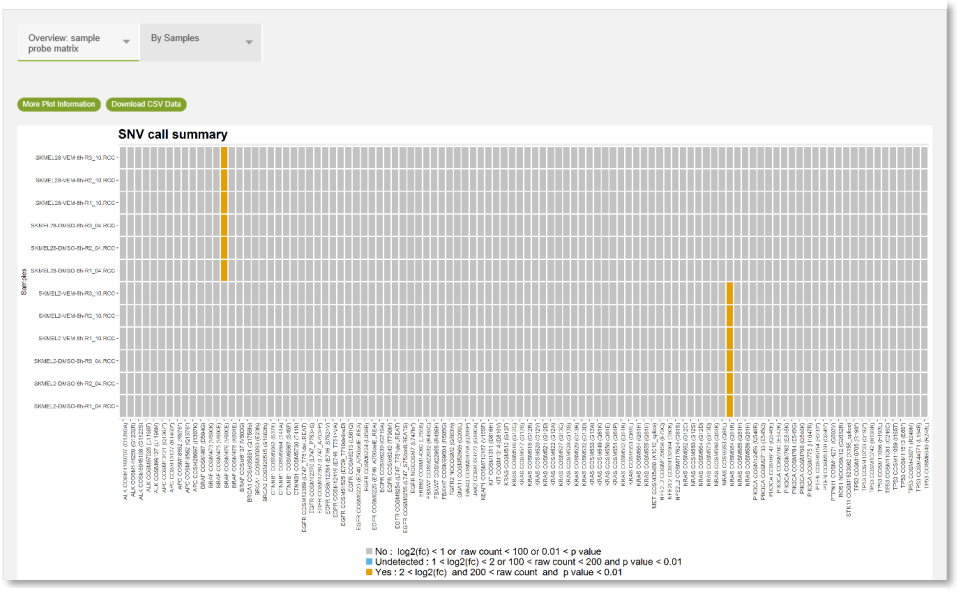


Figure 76: SNV module - Overview plot from 3D Bio Data Example

Overview Quality Map

This heatmap shows the $\log_{10}$ raw SNV control probes values for each (reference and test) sample in the dataset. Using this plot, you can not only detect poor quality samples, but can use the vertical color bar on the left to determine the cause and effect of a possible sample failure. High expression is displayed in red, average in orange, and low in blue.

- The **SNV_INPUT_CTL** class (pink bar in Figure 77) contains probes for amplified, endogenous genes. These demonstrate input gDNA sample quality and amplification success. Relatively low counts here may indicate, for example, an FFPE sample of compromised quality or a suboptimal PCR amplification.
- Probes belonging to **SNV_UDG_CTL** (green bar) are the UDG control probes. Low expression in this class indicates that UDG digestion was successful and your sample is not suspected to be susceptible to cross-contamination.
- Probes belonging to **SNV_PCR_CTL** (orange bar) capture the quality of PCR for each sample. Any sample in this class with significantly low counts may have experienced suboptimal PCR amplification.
- **SNV_NEG** (teal bar) and **SNV_POS** (coral bar) are exogenous assays using two-armed probes. SNV_POS probes have template present and should exhibit high expression; SNV_NEG probes lack a template and should exhibit low expression. Strong signal from the SNV_POS will signify a successful hybridization. A subset of these probes is used to test lane temperature and represent background signal.

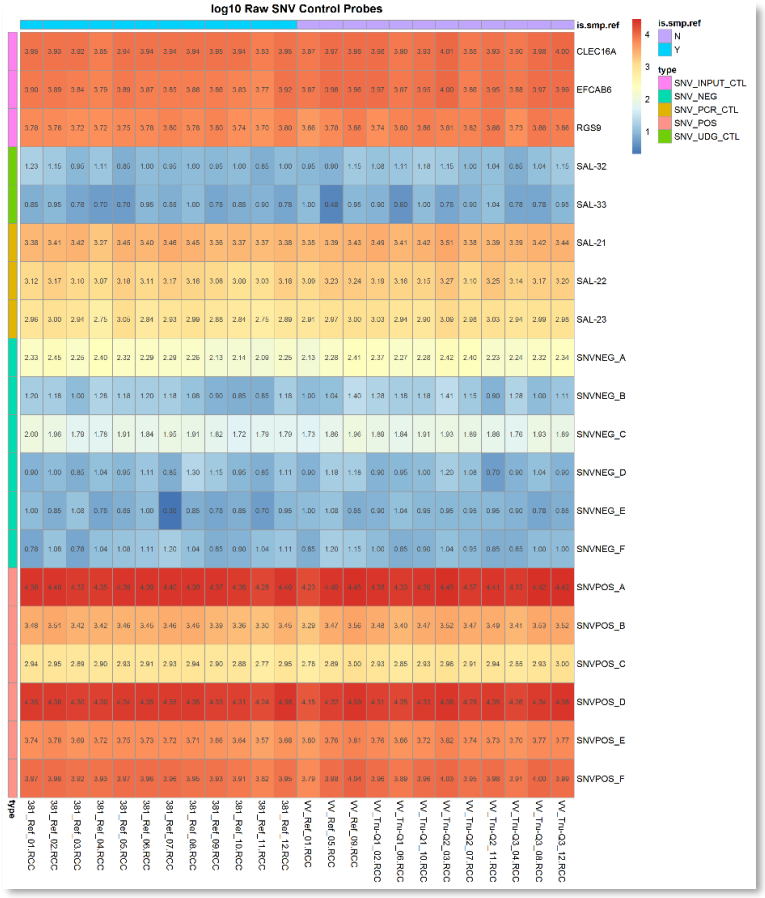


Figure 77: SNV module - Quality map

By Samples Variant plot

This plot displays a single sample's detection results for each probe. Vertical bars show the estimates and confidence intervals for each probe's $\log_2$ fold-change relative to its expected value in reference samples.

- Probes for which an SNV variant was called are highlighted in **gold**. The dashed gold line marks the fold-change threshold required for a SNV call.
- Vertical bars in grey, overlapping the black line at 0, indicate probes that are not statistically significantly above their expected reference level.
- Very low-frequency SNVs may manifest as vertical bars above 0 but with estimates below the fold-change threshold. These are highlighted in **blue**.

The legend below the plot details the rule for calling SNVs from raw counts, p-values and fold-changes above expected reference value.

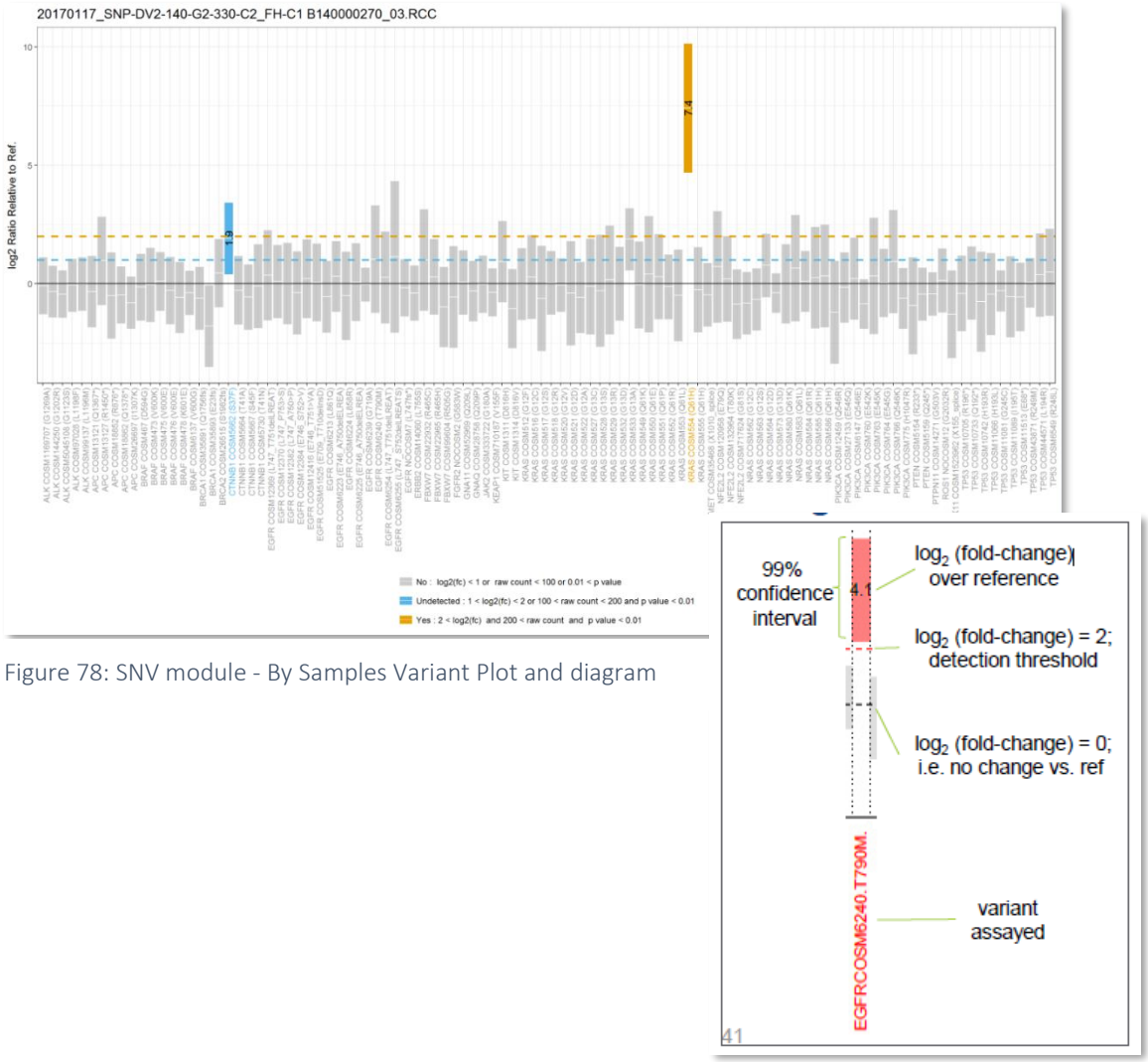


Figure 78: SNV module - By Samples Variant Plot and diagram

SNV Algorithm Details

The primary task of the SNV evaluation algorithm is to make presence/absence calls for mutations detectable in the panel as well as to provide statistics that quantify our confidence in the calls. To this end the algorithm first seeks to characterize the expected distribution of counts for each probe in the panel when no mutation is present. This is learned from reference sample data. Once this distribution is characterized for each probe, the algorithm can perform hypothesis tests evaluating whether a count corresponding to a mutation in a test sample is improbably high when assuming wild type status for the sample. The algorithm consists of three stages. **Preprocessing**, **initial estimation**, and **post estimation refinement**.

Preprocessing

This stage involves data normalization and calculation of data attributes required for estimation and post estimation stage. These attributes include:

- The temperature adjustment factor ($\hat{t}$) is a metric empirically shown to be able to serve as a surrogate for temperature. Its value is:

$$\hat{t} = \log_2\left(\frac{POS_A + POS_D}{POS_B + POS_E}\right) \text{ (Eq.1)}$$

- The normalization factor used to adjust for input/reaction efficiency normalization is the centered mean $\log_2$ SNV_INPUT_CTL probe counts for each lane.
- The estimated background for probes with sufficiently high background is a function of $\hat{t}$ and the wild type count corresponding to the same locus. The following steps are repeated for each probe. Ensure that:
    1. Median raw count WT reference samples > min count threshold (default is 50). If not, set background estimate to 0.
    2. Test sample counts are within +/-1.3 z units of the distribution of counts for that probe's WT reference sample. If not, set the background estimate to 0. Mean and standard deviation for z transformation is computed based on the counts from the reference samples.
    3. The adjusted R^2 is >

$$\widehat{Bg}_{ij} = \beta_0 + \beta_t \hat{t} + \beta_w \dot{W}_{loc.j} + \varepsilon \text{ (Eq.2)}$$

    Where $\dot{W}_{loc.j}$ is the normalized $\log_2$ wild type count at loci j centered by sample set id with sample sets consisting of either reference or test samples. The default adjusted R^2 threshold = 0.6.

Initial Estimation

For each mutant probe p_j, probe j, and sample i, we fit the following model:

$$\log_2(nCount_{ij}) - \widehat{Bg}_{ij} = \beta_j + \beta_{tj}\hat{t}_i + \beta_{ij} S_{ij} + \varepsilon \text{ (Eq.3)}$$

- $nCount_{ij}$ is the normalized count for sample i, probe j.

- S_{ij} is a categorical variable which takes on the value "reference" when the i[th] sample is a reference sample. For a dataset with n test samples and m reference samples, S has n+1 levels and the model has residual degrees of freedom = n + m – (n+2) = m-2.
- β_{ij} is the $\log_2$ fold change normalized count in sample i for probe j relative to the average $\log_2$ count for the same probe in the reference sample. These $\log_2$ fold change values and their corresponding user-specified standard error (e.g. 0.95 or 0.99) are computed and stored. Subsequently, using these estimates as well as the raw counts values calls are made for each mutant probe j. A mutation is called [3]when:
    1. p value < p.threshold (default 0.05)
    2. $\log_2$ Fold change > $\log_2$ fold change threshold (default is 1 i.e. 2 folds or more up)
    3. Raw count > (min count threshold) * $2^{(\log 2 \text{ fold change threshold})}$

Post Estimation Refinement

In a typical dataset, the many tested samples for which no SNV calls were made with high confidence represent data points with wild type status. These can be used to increase our power to characterize the wild type distribution of counts.

- When test sample i and probe j are consistent with wild type status, we replace the sample id in variable S_{ij} with the value *reference*. Completing this exercise for all samples i and probe j, we then re-run the model in Eq.3. This time, the model benefits from 2k increase in residual degrees of freedom (if k non-reference samples are re-designated as *reference*). Following the inference, calls can be made and the iteration can be repeated until changes to the results are minimal (by default the algorithm allows for maximum of 20 cycles).
- The iterative process involves the following steps:
    1. Refit the model in Eq.3 (unless there were no mutant call, in which case term the Sij term is dropped).
    2. Ensure there is no major outlier in the fitted model (this is done based on evaluation of Cooks distance as well as z score of the fitted values). If so, exclude and refit.
    3. Store the adjusted R^2 value of the model.
    4. Fit the model for reference status and compute the corresponding SE.
    5. For any non-reference sample, compute t value by subtracting the expected value of reference and dividing the estimate by prediction standard deviation. Using the same statistic, update the corresponding p value and confidence intervals.
    6. Make the calls as before.

[3] In the current version, the user interfaces allow for specification of two calling categories by the user corresponding to two level of stringencies. The outlined steps for calling corresponds to a case where only the highest stringency is considered a call but the principles of how calls are made are the same in general given a set of user-specified cut offs.

7. Repeat until convergence. Convergence is defined to happen when the median shift in t-values < t.convergence (by default 0.1) or if no change in call status is made. The model with the highest adjusted R^2 is selected as the best solution.

Debiasing

The last refinement involves by-sample bias removal. Specifically, the mean $\log_2$ fold change for wild type calls in each sample is calculated and the value is subtracted from all fold changes estimated for that sample. Subsequent to this adjustment, the t and p values are also adjusted accordingly and the call procedure is repeated.

Fusion Module

nCounter chemistry provides two different approaches to assist in the detection of gene fusions. **Junction probes** are specific to exon-exon pairs and target the sequence of the actual fusion junction (breakpoint). This provides a direct measurement of the fusion event; if a Junction probe has counts above background, then the targeted fusion is present. In contrast, groups of imbalanced, or **End probes,** are used to detect a fusion independent of its exact splice junction. A typical target gene has at least three End probes at each its 5' and 3' ends, well away from known fusion breakpoints. If the 3' probes exhibit higher counts than the 5' probes (meaning that there are abundant 3'end transcripts or a display of imbalanced expression) it indicates that the 3' end has been fused with some unknown 5' partner.

The Advanced Analysis Fusion module summarizes fusion events detected in the data through three different types of plots. It does not require any user input to make fusion calls, however, you may specify a statistical significance level for detection by Junction probe and / or End probes (see the *Custom Options for Fusion* section).

From the Fusion Overview tab, you can choose from the **Detection Summary**, a color-coded sample-gene matrix summarizing fusion calls, or the **Heatmap**, which visualizes raw counts for each probe. The **By Samples** tab provides a more detailed look at the data behind fusion calls, allowing you to view histograms of the $\log_2$ counts from each probe of any sample in your dataset. The **Fusion Summary Report** provides specific details on the probe results used to make fusion calls for each gene.

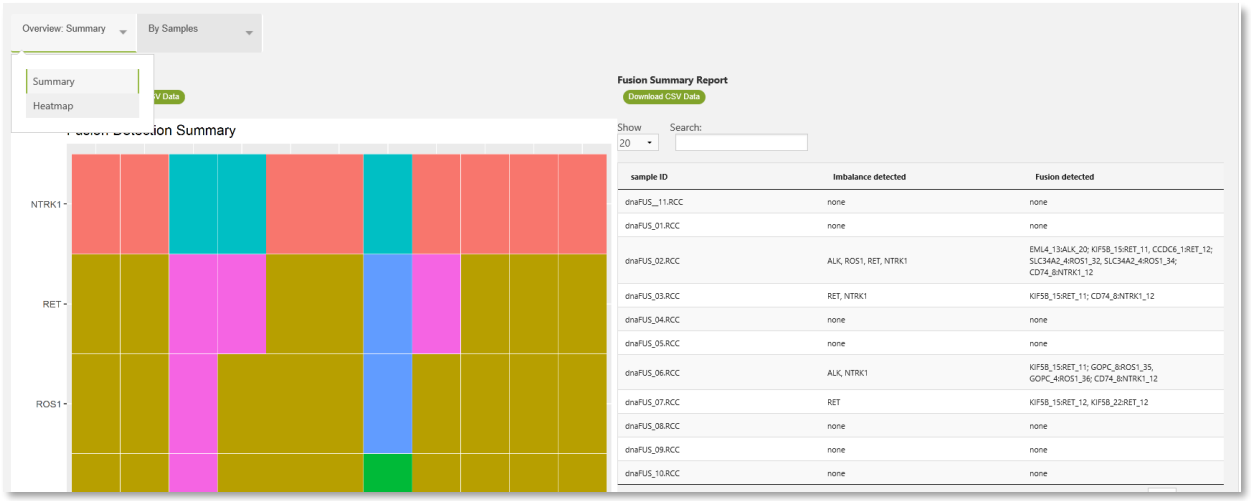


Figure 79: Fusion module view and options

Before You Start Fusion

Use caution when working with very few samples (fewer than 6) and/or when replicate samples are selected for analysis. The outlier test's power is sensitive to sample size and fusion frequency (see the *Fusion Algorithm Details* section).

Simplified sample names can streamline the plot labels and lists. Use the **Description** column in nSolver to assign these shortened sample names, then select **Description** as your **Identifier** when initiating Advanced Analysis (see the *Identifiers and Covariates* section).

Custom Options for Fusion

There is no Fusion custom options menu, however, the **General Options** menu will include a **Specify Fusion Parameters** button if it detects Fusion probes in the dataset. This button allows you to adjust the p-value threshold for Junction probe detection and End probe imbalance expression.

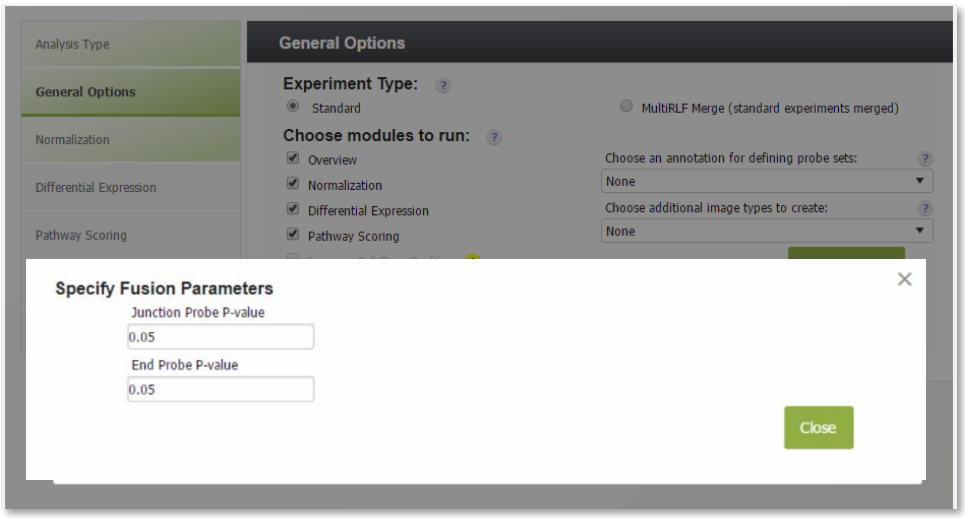


Figure 80: Windows associated with custom options for Fusion - General Options menu

Interpreting Results of Fusion Plots

Fusion Detection Summary

The **Fusion Detection Summary** is a sample-gene matrix, color-coded according to whether a gene tested positive or negative for evidence of a junction (relying on the Junction probe results) and/or positive or negative for evidence of an imbalance (using collaborated End probe data). This figure summarizes the fusion calls by plotting each gene tested on the vertical axis and each sample tested horizontally. Color (see plot key) indicates whether a fusion was detected and the type of evidence (Junction +/-, Imbalanced +/-, or both) used to make the call. The Junction and End probes provide different levels of evidence for fusion events (see Table 5). The fusion call is clearest when both types of probes exhibit positive results.

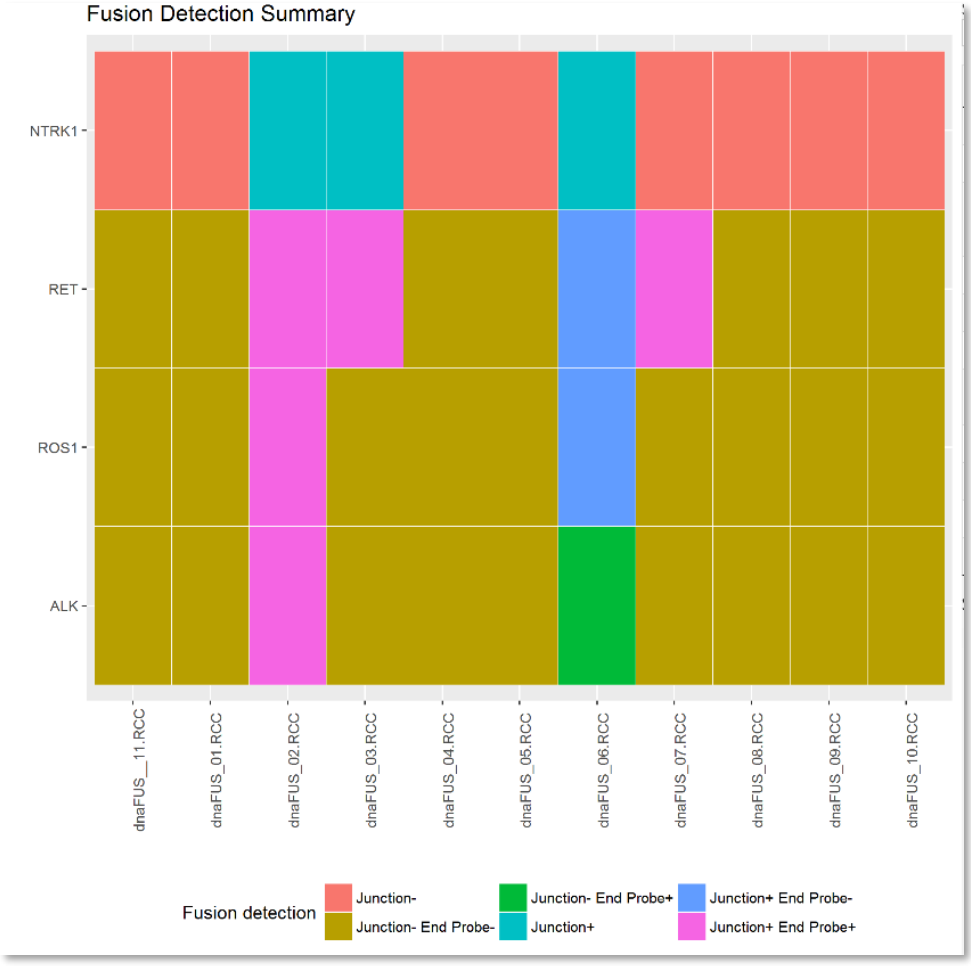
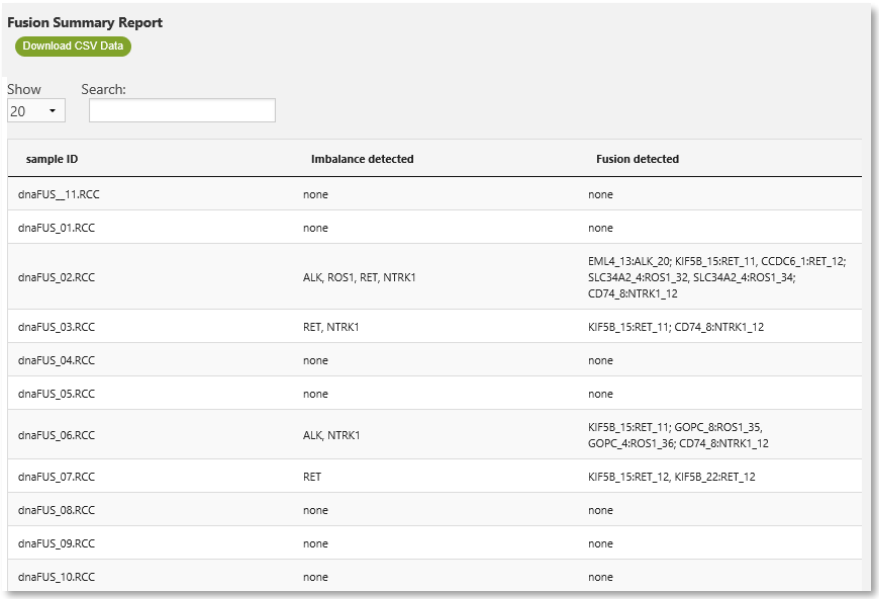


Figure 81: Fusion module - Fusion Detection Summary

These concordant fusion calls provide the strongest evidence for a fusion event. However, even when End probes and Junction probes provide discordant fusion calls, clear conclusions can often be made. A positive result for the End probes accompanied by a negative result for the Junction probe may result from fusions not targeted by the experiment's Junction probes. A positive result from a Junction probe accompanied by a negative result from an End probe may result from fusions that are expressed at much lower levels than the wild-type transcript; these fusion calls hold weaker evidence, and examination of the raw data for these samples is recommended.

Fusion Summary Report

This table summarizes the results from the call summary in table format and also gives specific details about the probe results used to make the conclusion(s) at each gene.



Fusion Summary Report

Download CSV Data

Show 20 Search:

sample ID	Imbalance detected	Fusion detected
dnaFUS__11.RCC	none	none
dnaFUS_01.RCC	none	none
dnaFUS_02.RCC	ALK, ROS1, RET, NTRK1	EML4_13:ALK_20; KIF5B_15:RET_11, CCDC6_1:RET_12; SLC34A2_4:ROS1_32, SLC34A2_4:ROS1_34; CD74_8:NTRK1_12
dnaFUS_03.RCC	RET, NTRK1	KIF5B_15:RET_11; CD74_8:NTRK1_12
dnaFUS_04.RCC	none	none
dnaFUS_05.RCC	none	none
dnaFUS_06.RCC	ALK, NTRK1	KIF5B_15:RET_11; GOPC_8:ROS1_35, GOPC_4:ROS1_36; CD74_8:NTRK1_12
dnaFUS_07.RCC	RET	KIF5B_15:RET_12, KIF5B_22:RET_12
dnaFUS_08.RCC	none	none
dnaFUS_09.RCC	none	none
dnaFUS_10.RCC	none	none

Figure 82: Fusion module - Fusion Summary Report

Table 5: Categories for Fusion Calls

Result	Category	Summary	Example conclusion
End probe detection call; Junction probe detection call	Detected Gene Fusion, Variant Conclusive	There is a high probability that the sample is positive for a specific gene fusion variant	Positive ALK gene fusion event at EML4_13:ALK_20
End probe detection call; no Junction probe detection call	Detected Gene Fusion, Variant Inconclusive	There is a high probability that the sample is positive for a fusion event but the variant is inconclusive. May indicate the variant is not currently included in the fusion-specific probes (potentially a new variant)	Positive ALK gene fusion event, location unknown
End probe undetected call; Junction probe undetected call	Non-Detected Gene Fusion	There is a high probability that the sample is negative for a fusion event	No gene fusion variants detected
End probe undetected call; Junction probe detection call	Inconclusively Detected Gene Fusion	It is possible that the junction probe hit is a false positive, or that a fusion is truly present but has insufficient expression to be detected with the End probe test.	Possible low-level expression of GOPC_4:ROS1_36 atop high wild type ROS1 expression.

Heatmap

The **Heatmap** displays the log_2 raw counts for the different fusion probes and allows you to view **All** probes, just the **End probes**, or just the **Junction probes**. Many fusions will be immediately obvious in these heatmaps, either through high counts of a Junction probe or through strongly imbalanced 5'/3' probes for a gene within a sample. These heatmaps can also reveal technical artifacts that may mislead the detection algorithms. Look for Junction probes with unexpectedly high background counts, and look for samples with unusually high or low signal across a wide range of probes. Although every experiment and every probe is different, counts below 20 (3.3 in log_2) are often background, and counts above 100 (6.6 in log_2) are very seldom background.

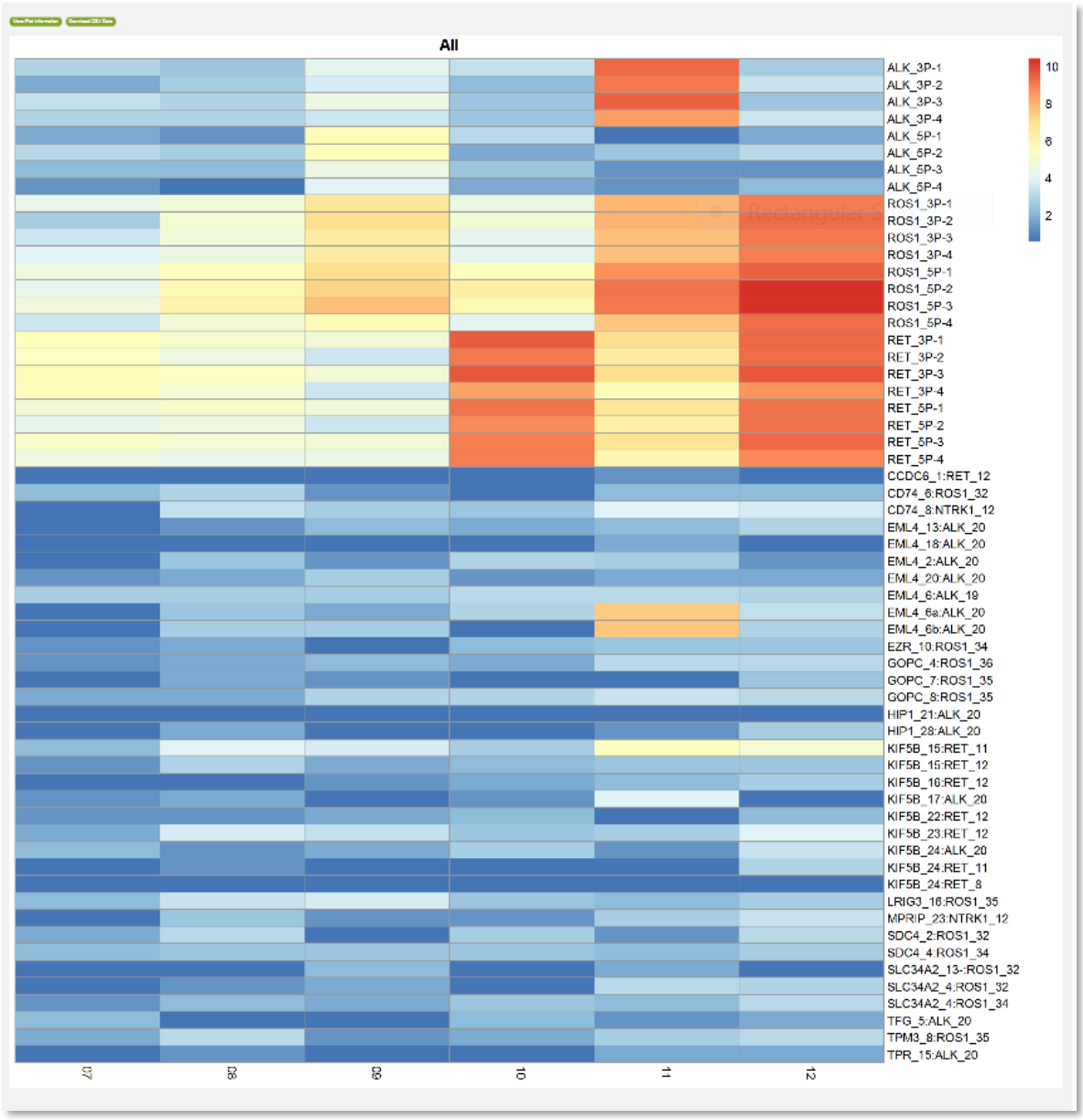


Figure 83: Fusion module - Heatmap

By Sample

The **By Samples** tab allows you to view bar plots of the log_2 raw counts from each probe of any sample in your dataset. The End probes plot is shown separately from the Junction probes plot. Review these plots for every detected fusion call.

The **End probes** plot shows the log_2 raw counts of the 5' and 3' probes. If an End probe detection call was made in the Fusion Detection Summary plot, confirm in this plot that the 3' probes' counts are visibly higher than the 5' probes' counts. Be aware, however, that in genes whose wild-type transcript has high expression, fusions may appear as slight but consistent increases in 3' probes relative to 5' probes. The null hypothesis here is that no fusion event occurred and therefore the mean 5' and 3' probes will be equal. A **p-value** for this is provided in the upper left of the plot.

The **Junction probes** plot shows the log_2 raw counts of the Junction probes. If a Junction probe detection call was made in the Fusion Detection Summary plot, this plot can provide a double check for the calling algorithm's results. Check that the probe detected in the Fusion Detection Summary is truly expressed above background and above the other Junction probes in this plot. The numbers above the bars show the ranking of detection where **1** conveys the highest confidence in detection and subsequent lower rankings convey decreasing confidence. A ranking of **0** means undetected. In the absence of a fusion, all probes will fall in the background of the system (with a 0 rank).

If a fusion is present in less than half the samples, the minimum sample size is > 6, and there is a Junction probe targeting it, that probe should have expression noticeably higher than the corresponding counts for samples lacking the fusion. In some infrequent cases, splice variants will affect a fused gene and two Junction probes will be elevated in the same sample.

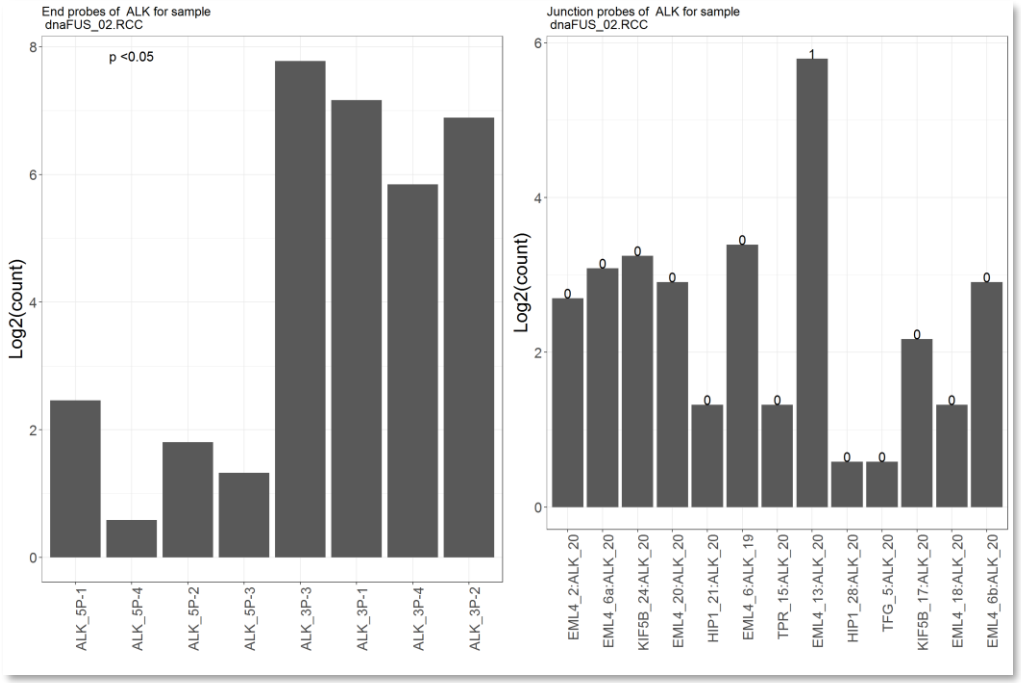


Figure 84: Fusion module - By Sample plots

Fusion Algorithm Details

End probe detection algorithm

Fusions result from a 5' promoter region fusing to the 3' region of a given gene. Highly-expressed fusions will therefore result in much higher expression of the driver gene's 3' end than its 5' end. To quantify the evidence for 3' overexpression, nSolver performs a t-test comparing the log-scale data from the 5' and 3' probes. Equal variance is assumed since, under the null hypothesis, there is no reason to suppose the 5' and 3' probes will behave any differently from each other. To prevent false detection at low counts, we set the mean of $\log_2$-transformed counts of 5' probes to 3 (8 on the raw count scale) if the real mean is less than 3. We also set the sample standard deviation to be 1.714, a fixed value derived from a vast database of background counts. Since we assume the mean count of 3' probes is higher than the mean count of 5' probes, no test will be performed if the mean count of 3' probes is lower than the mean count of 5' probes. Also, to avoid false fusion calling when the mean count of 3' probes is close to or within the background, no test will be performed if the mean count of 3' probes is less than 32 (5 in $\log_2$ scale).

Junction probe detection algorithm

If a patient has no known fusion for a given gene included in the panel, then all the Junction probes for that gene will lack target and return only background counts. Thus, fusion detection is simply an exercise in identifying Junction probes whose counts are sufficiently higher than background. To test whether each Junction probe count is above background, we estimate the distribution of background counts, and use a **sequential outlier test (robust Grubbs' test)** to call probes that fall above this background distribution. We apply the test separately for each Junction probe, looking for outliers among all the samples.
The sequential outlier test works as follows:

- We model the observed $\log_2$(counts) as coming from a normal distribution with mean equal to the sample median of the observed counts and standard deviation equal to either the sample standard deviation of the observed counts or an SD estimate predicted from a large NanoString database (details below).
- Given this estimated normal distribution, we derive a p-value for the maximal data point by calculating the probability that the maximal data point drawn from this distribution would be greater than or equal to its observed value.
- If the probability is less than our p-value threshold, we make a fusion call, remove the data point, and repeat the process until half samples are tested.
- The order in which data points are called as outliers/fusions is recorded; the fusion call with smaller order is generally more reliable than the fusion call with larger order.

The outlier test's power varies with sample size and fusion frequency: the more samples and the lower the frequency of fusion events, the more accurately the algorithm can make fusion calls. We have run simulations showing that a sample size of 6 is generally required for adequate power. In experiments with less than 6 samples, we recommend looking particularly carefully at the raw data bar plots to confirm each

call. Fusion frequency will seldom compromise the algorithm since it is usually lower than 10%. In extreme situations where the fusion frequency is greater than 50%, the test power will reduce dramatically.

A note on estimating standard deviation

Because many experiments will lack sufficient data to estimate the standard deviation (SD) of noise, we use a large historical dataset to model SD. We find most probes adhere to the following relationship:

$$SD(\log_2 counts) = 1.2172288 - 0.1250544 * mean(\log_2 counts).$$

To avoid excessively small SD estimates for probes with high means, we set the floor of SD to be 0.5095818. Once the mean of $\log_2$(count) is greater than 3.5, we will choose the greater between the real sample SD and 0.5095818. This use of the sample SD in place of the constant 0.5095818 is appropriate because when $\log_2$(count) is greater than 3.5, the counts are less likely to come from background noise and more likely to come from either nonspecific binding or other factors, making the SD estimate from our historical data less reliable.

Appendix A: 3D Bio Data Example for Advanced Analysis 2.0

The dataset, **3D Bio Data**, is included when you download the nSolver 4.0 Analysis Software. This data contains three biological replicates from two different melanoma cell lines, SKMEL28, which has a known mutation (c.1799T>A; p.V600E) in both copies of the BRAF gene, and SKMEL2, which has two normal copies of the BRAF gene (and a known mutation in the NRAS gene). Both cell lines were treated with either DMSO (vehicle) or vemurafenib (a specific inhibitor of the V600E mutant BRAF protein) dissolved in DMSO for 8 hours.

Throughout the Advanced Analysis 2.0 User Manual, you will find excerpts of this dataset's analysis.

nSolver Data Prep

To prepare your data for Advanced Analysis you must:

1. **Import** files and set **QC** parameters in nSolver 4.0.
    - Select the **Import RLF** button and follow the prompts of the Import Wizard to import the RLF for dataset, then repeat the process to import the RLF for the SNV references.
    - Select the **Import RCC** button to import RCC files for dataset and the SNV references. Accept the default **QC** parameters.
    - You may wish to assign shortened labels that uniquely identify each of the RCC files (including the SNV references) using the **Description** column. This will simplify downstream visualizations.

2. **Create an Experiment** using the **New Experiment** button.

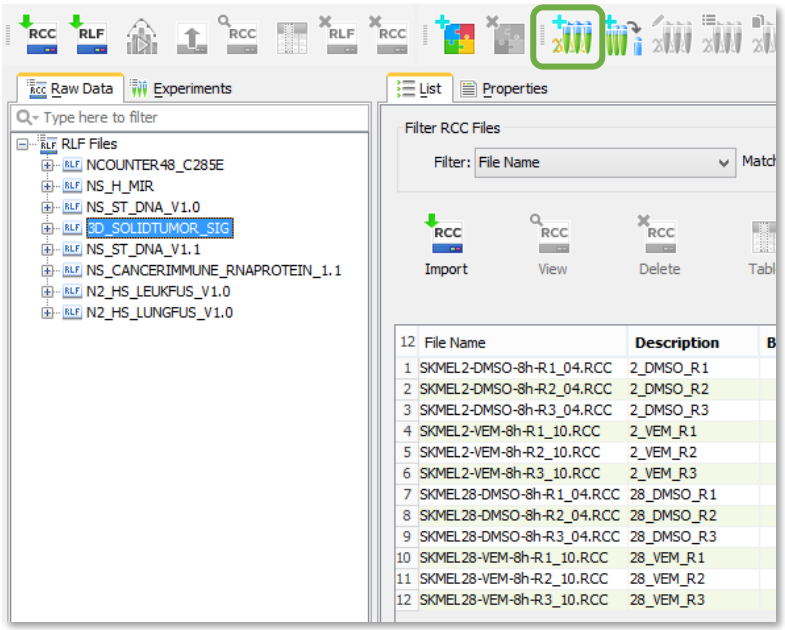


Figure 85: Creating an experiment in nSolver

As you follow the prompts in the Experiment Wizard, you can leave **background** correction off, create **annotations** that will be informative to you in your analysis (see below), accept defaults for **normalization,** and leave **ratio** creation off.

- Annotations: Create one annotation column titled **Treatment**, and assign **DMSO** or **VEM** according to what is documented in the sample names. Create a second annotation titled **BRAF Genotype** and assign **WT/WT** to the SKMEL2 samples and **Mut/Mut** to the SKMEL28 samples.

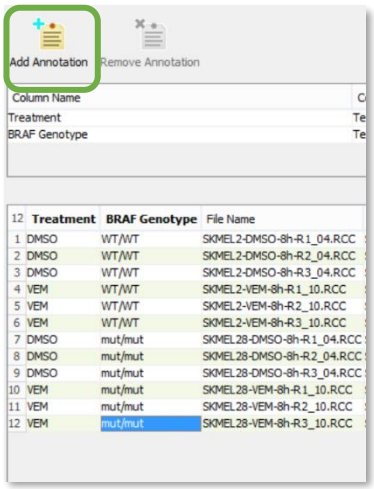


Figure 86: Creating annotations in nSolver

3. Once your experiment has been built, expand the navigation tree on the Experiments tab and highlight the **Raw Data** level. Select the **Advanced Analysis** button.

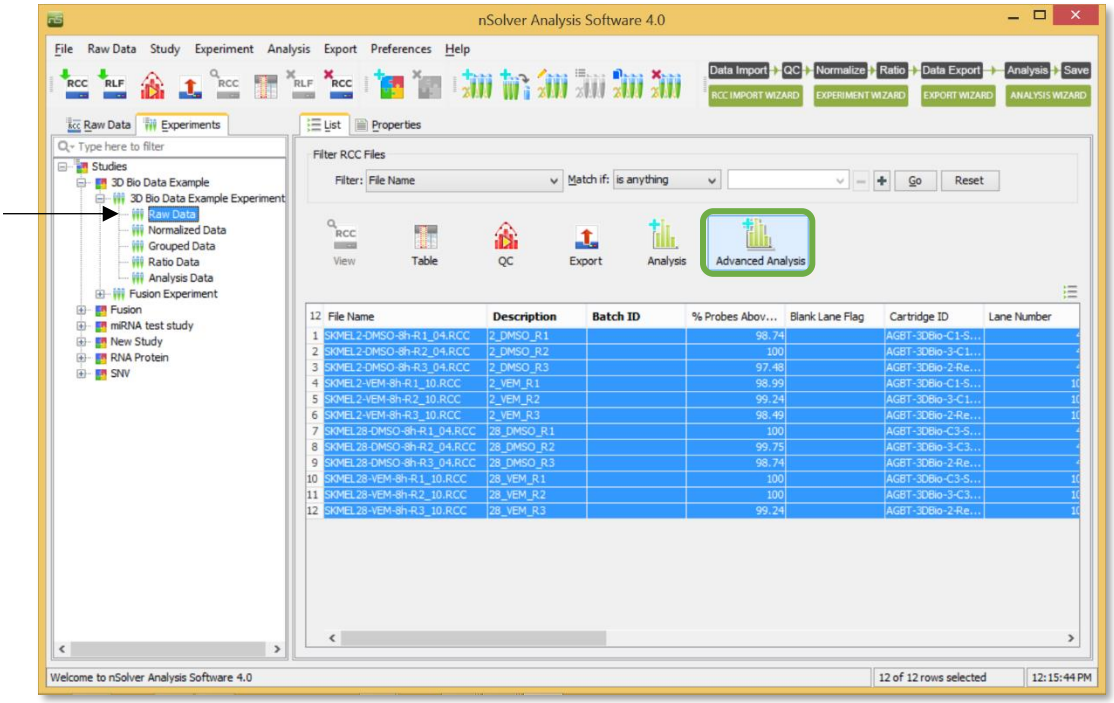


Figure 87: Creating an Advanced Analysis in nSolver

Setting up the Advanced Analysis

1. Choose a **Name** for your analysis and select nCounter Advanced Analysis 2.0 for **Analysis Type.** If you have not yet installed version 2.0, refer to the *Installation* section. You can **Browse** to choose where the output files should be saved. Select **Next.**

2. Select a unique **Identifier** – this field will be used to label samples in the resulting plots. If preferred, use the modified labels you entered in the Description column when preparing your data in nSolver. Select your annotations in the **Use for Analysis** column – this will select them as covariates for analysis. Select **Next.**

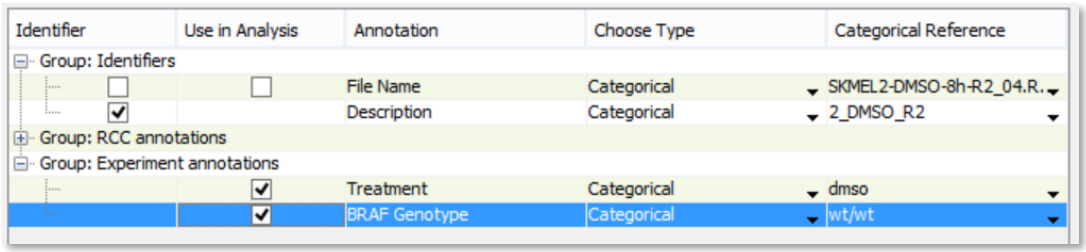


Figure 88: Selecting identifiers and covariates in Advanced Analysis

3. Select **Custom Analysis.**

4. Select the **General Options** tab. For the purposes of this example, de-select any modules other than **Overview**, **Normalization**, and **Differential Expression**. We will focus on these to get a general overview of our data and the samples and genes that are differentially expressed in it.

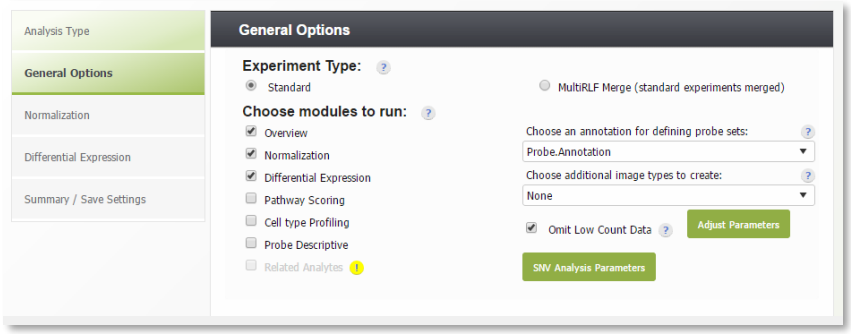


Figure 89: Advanced Analysis Custom Analysis menu - General Options

5. On the **Normalization** tab, you can customize the Normalization settings. For this example, leave all defaults.

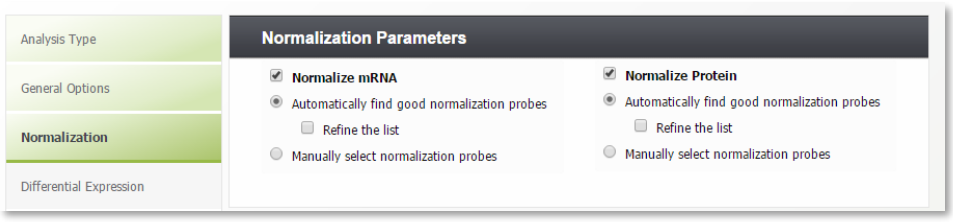


Figure 90: Advanced Analysis Custom Analysis menu - Normalization

6. On the **Differential Expression** tab:
    - Move **Treatment** and **BRAF Genotype** from the **Available Annotations** field to the **Selected Predictors** field using the green arrow button.
    - Leave the **Optimal** setting and **P-value Adjustment** method as defaults.
    - Leave the **Run GSA** and **Display Results Using PathView** boxes selected. These plots will provide more detail and context to the Differential Expression results. We can change the number **of top pathways to display** from 20 to **10** to speed up processing.
    - Leave the **Color Plots by** and **P-value Threshold** settings as defaults.

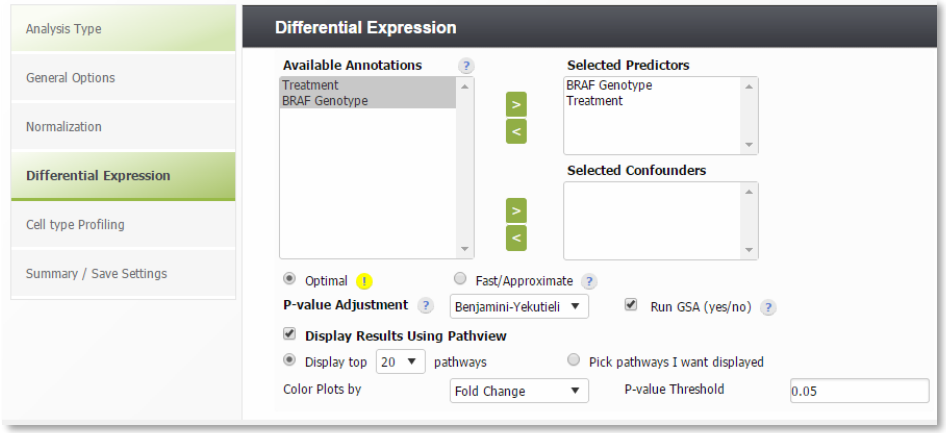


Figure 91: Advanced Analysis Custom Analysis menu - Differential Expression

7. Select **Finish.**

8. You will be returned to the nSolver dashboard. Expand the navigation tree of your experiment on the **Experiments** tab and highlight the **Analysis Data** level. Highlight the analysis you just ran in the central table and select the **Analysis Data** button.

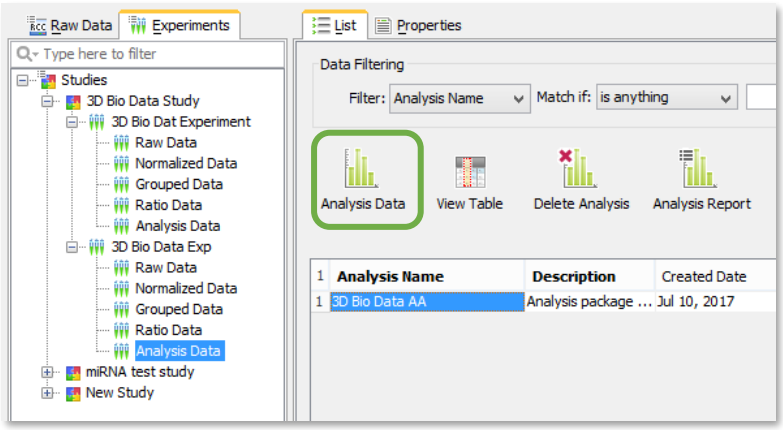


Figure 92: Selecting analysis to view in nSolver

9. Your analysis will start to load in an HTML window. You may need to **Allow blocked content**. Allow up to one hour (potentially) if running Advanced Analysis 2.0 for the very first time, as the initial downloading of R libraries is time-consuming. You will need an internet connection and permissive firewall settings for this step.

Advanced Analysis modules

1. Overview:

   The heatmaps cluster data with similar expression patterns. Colored bars along the top of each heatmap designate SNV and Fusion variant status, covariates, and QC flags. The colored bar along the left side provides information on probes that will be dropped from analysis due to low signal. In the **raw data heat map** (on the left) and the **normalized data heat map** (on the right) we can see roughly how the data is clustering and if that coincides with any of the variants or covariates (there are no QC flags in this dataset). We can see that one sample, 2_DMSO_R3, appears to be slightly different from the others; this sample does not deviate a great deal and so is no cause for concern, but serves as an example of the type of pattern you can look for to identify outliers in your data. We won't draw any additional conclusions from these plots, since this module is intended be used as a QC tool and way to get a general impression of your data.

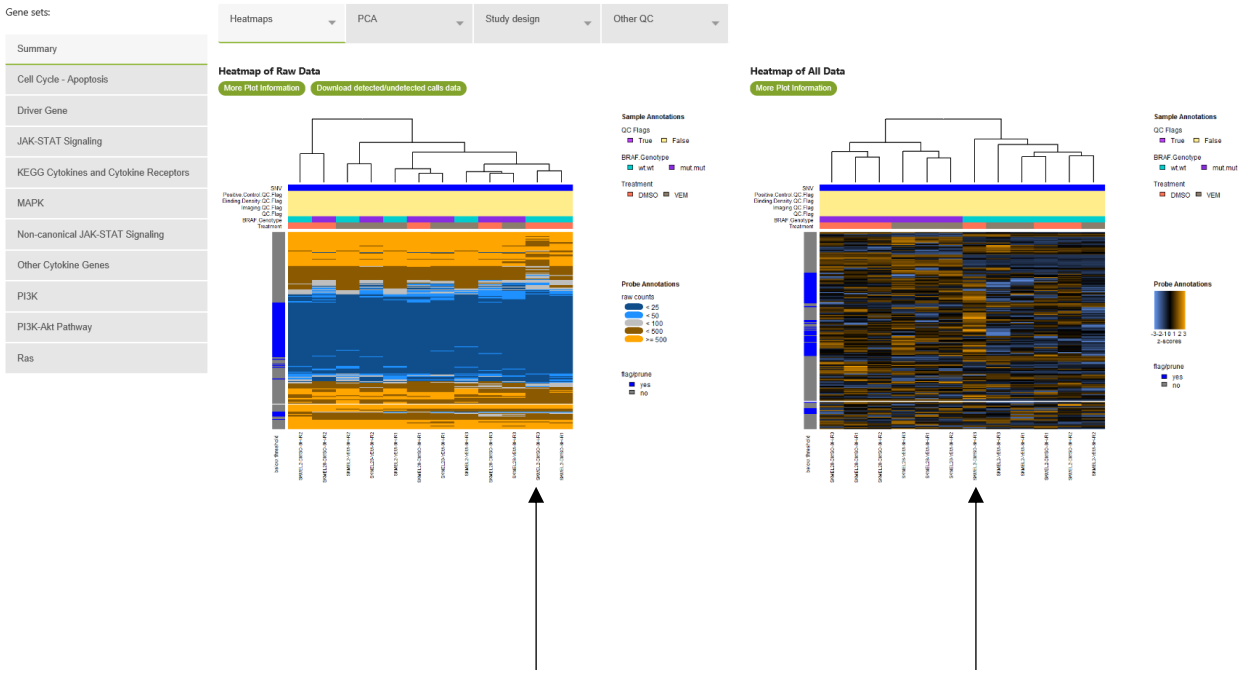
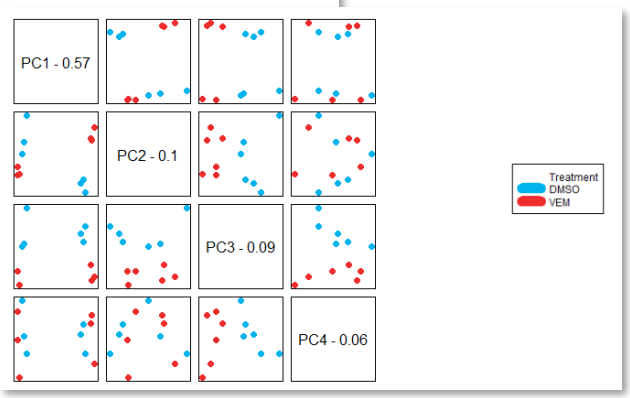


Figure 93: Advanced Analysis Overview module

Figure 94: Advanced Analysis Overview module - PCA plots

In the **Principal Component Analysis (PCA)**, a clear separation of **BRAF.Genotype** data points can be seen in PC 1 vs. PC 2 results, meaning that changes in this variable cause clear, consistent changes in the data. **Treatment** does not have the same effect.



This is reinforced by the p-value histograms under the **Other QC** tab, which shows a clear left- weighted plot for **BRAF.Genotype** samples, meaning there are a number of p-vaules in the significant range, close to zero. The **Treatment** p-values are more evenly distributed, indicating that relatively few genes appear to be differentially expressed between the treatment and control samples.

The scatter plot on this tab displays the housekeeping genes in color. Their placement at the bottom of the plot indicates that they are stable and require little further attention.

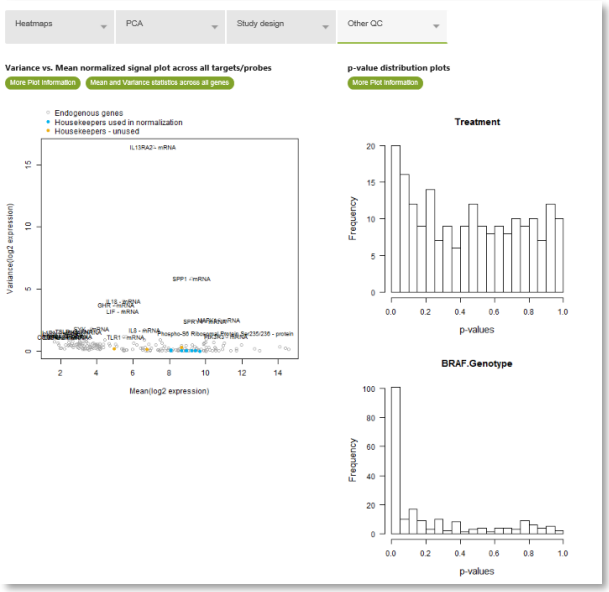


Figure 95: Advanced Analysis Overview module - QC plots

2. Normalization

Looking at **mRNA** data, we can see the effect of the geNorm algorithm's removal of each candidate reference gene on the variability of the normalized data; this information is then used to rank them. We can also see the Mean Square Error (MSE) of each sample plotted against its normalization factor in the **Normalization Summay** plot. Samples that are outliers are named on the plot; we can see that 2_DMSO_R3, again, appears as a bit of an outlier.

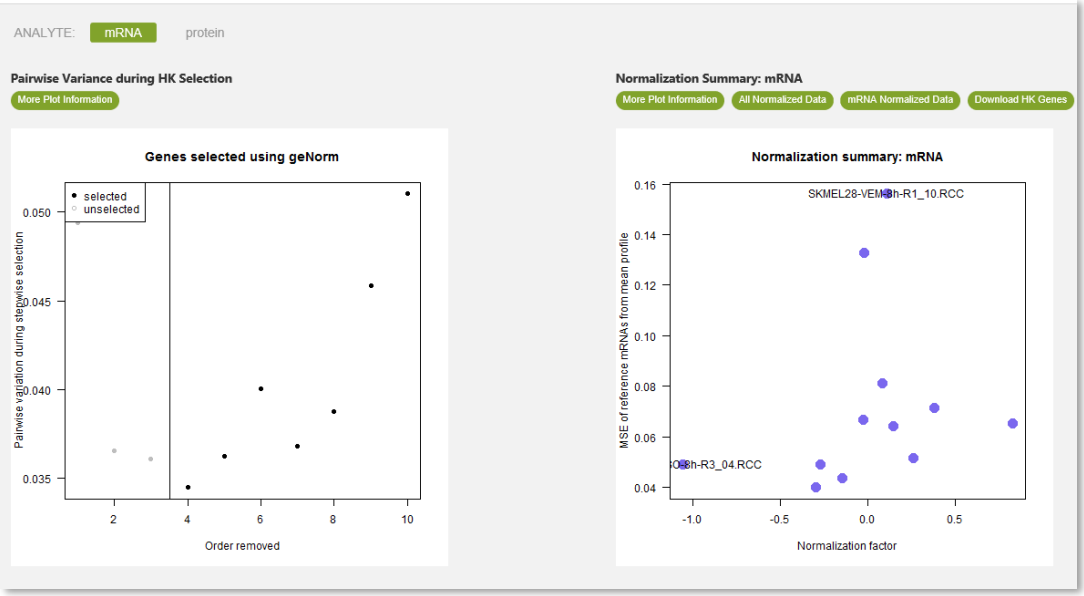


Figure 96: Advanced Analysis Normalization module - mRNA plots

Under the **Protein** tab, we can see similar plots referencing proteins that were unselected for Normalization and the MSE vs. normalization factor for each sample in the **Normalization Summary**. We can check protein controls on the **Protein Expression Background** plot.

Figure 97: Advanced Analysis Normalization module - Protein plots

3. Differential Expression

The **Volcano Plot** for the covariate **BRAF.Genotype** depicts the differential expression of genes in mut/mut samples relative to the wt/wt samples. It shows multiple p-value (significance level) thresholds. Only probes with p-values in the significant range are colored and named.

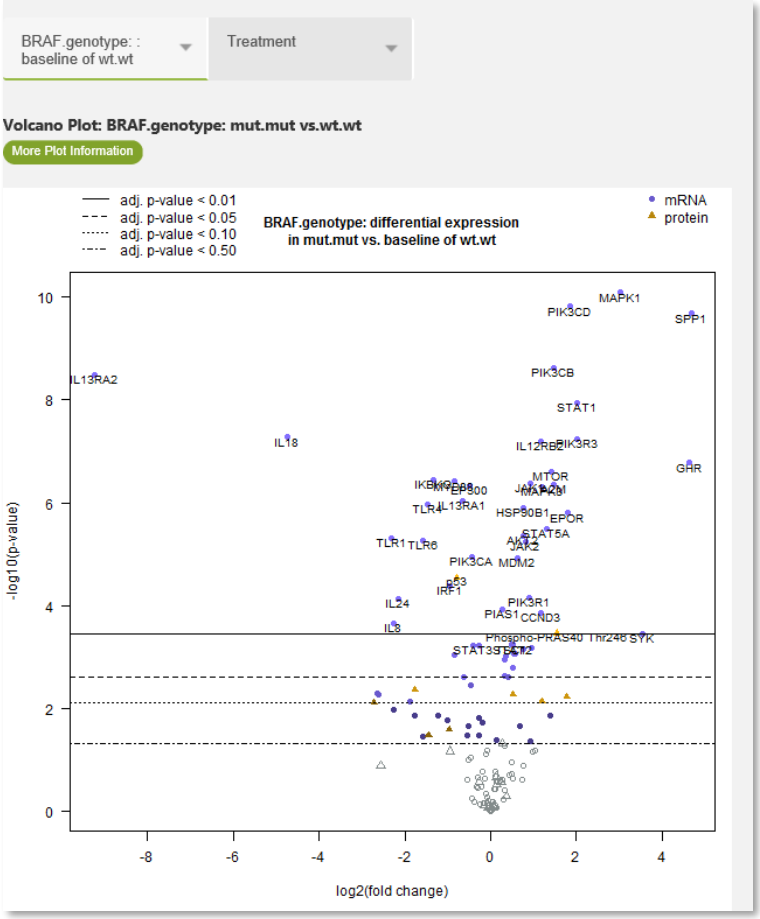


Figure 98: Advanced Analysis DE module - Volcano plot

Viewing this plot under the Treatment tab shows a colorless plot with no p-value thresholds, indicating Treatment did not result in signficant gene expression changes.

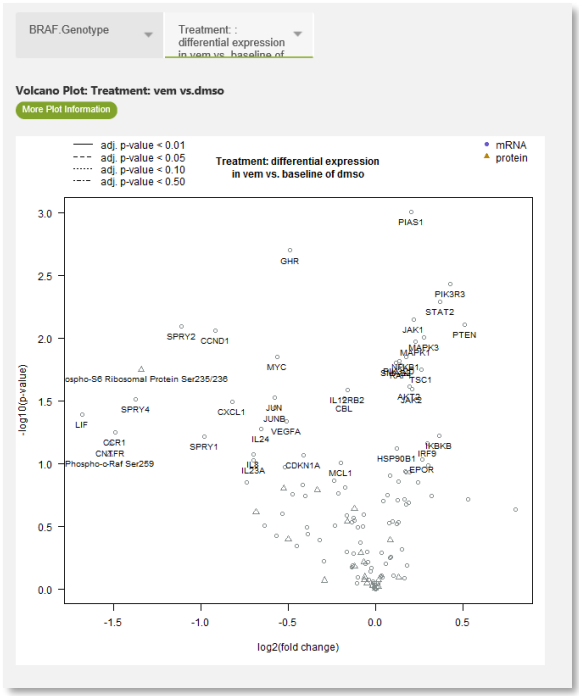


Figure 99: Advanced Analysis DE module - Volcano plot

4. GSA

Gene Set Analysis (GSA) shows us the variation in **Global Significance Scores** among the gene sets for each covariate. **BRAF.Genotype** is associated with more variable results among the gene sets than **Treatment**. We can see from the **Directed Global Significance Scores** plot that the **P13K-Alt Pathway** gene set has the highest score in the BRAF.Genotype category.

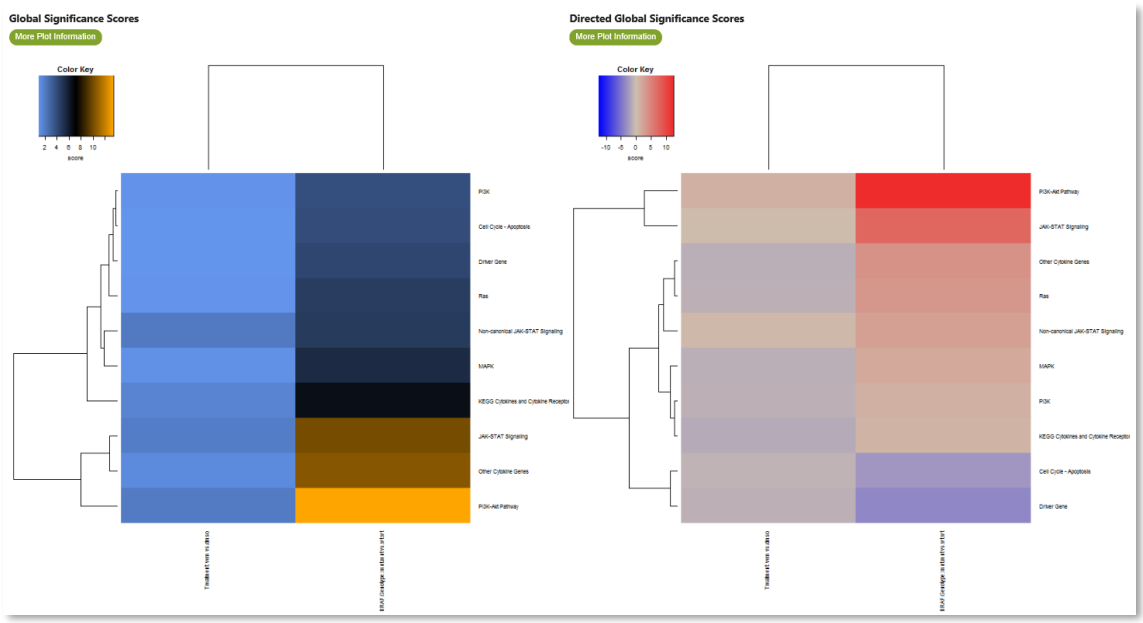


Figure 100: Advanced Analysis GSA module - heatmaps

Selecting the **P13K-Alt Pathway** gene set results in the Differential Expression volcano plot, overlaid with colored points which reflect the probes in that gene set. We can see that there are a number of probes from this gene set with significant results.

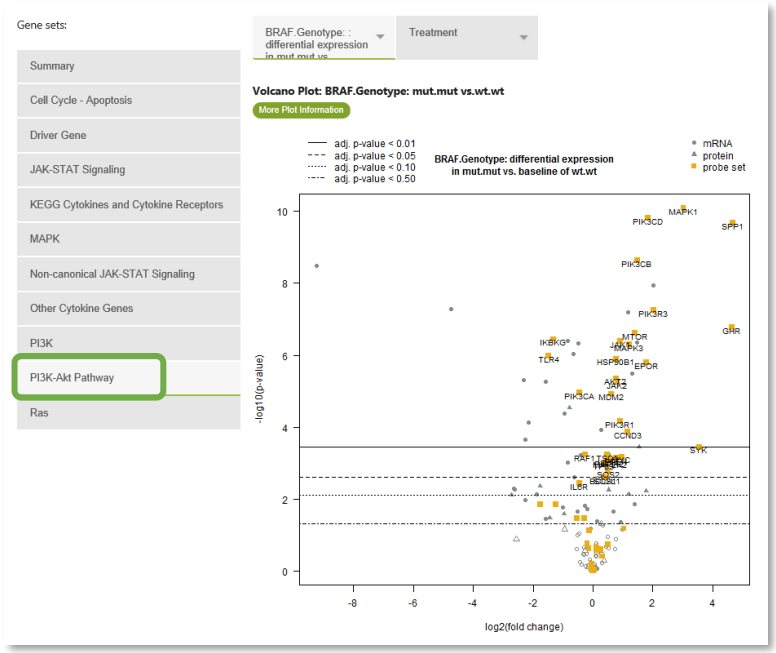


Figure 101: Advanced Analysis GSA module - Volcano plot

5. PathView

As a next step to the GSA analysis, we can view the pathways that include our gene set(s) of interest in the **PathView** module. Here, we select the **P13K-Alt Pathway** again to see where our genes of interest lie in this particular pathway. Colored boxes show the specific elements of the pathway that were differentially expressed and whether they are up- or down-regulated in our data. If we decided to later run the Probe Descriptive module, we would enter these genes for analysis.

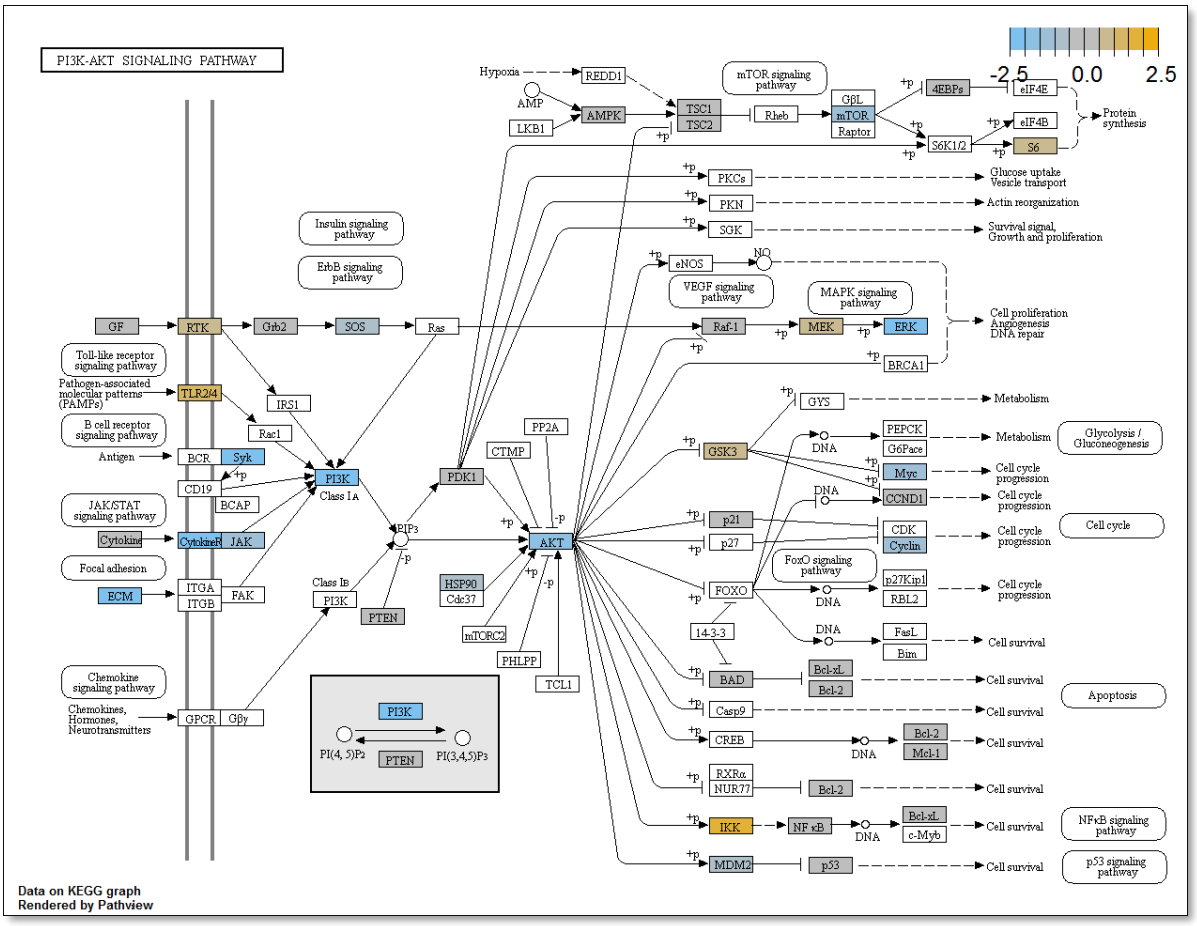


Figure 102: Advanced Analysis PathView module

6. SNV

The **SNV call summary** gives a clear depiction of the SNV calls made in this data. Results are as expected: SKMEL28 samples all exhibited variant calls in the BRAF gene, while SKMEL2 samples all exhibited variant calls in the NRAS gene.

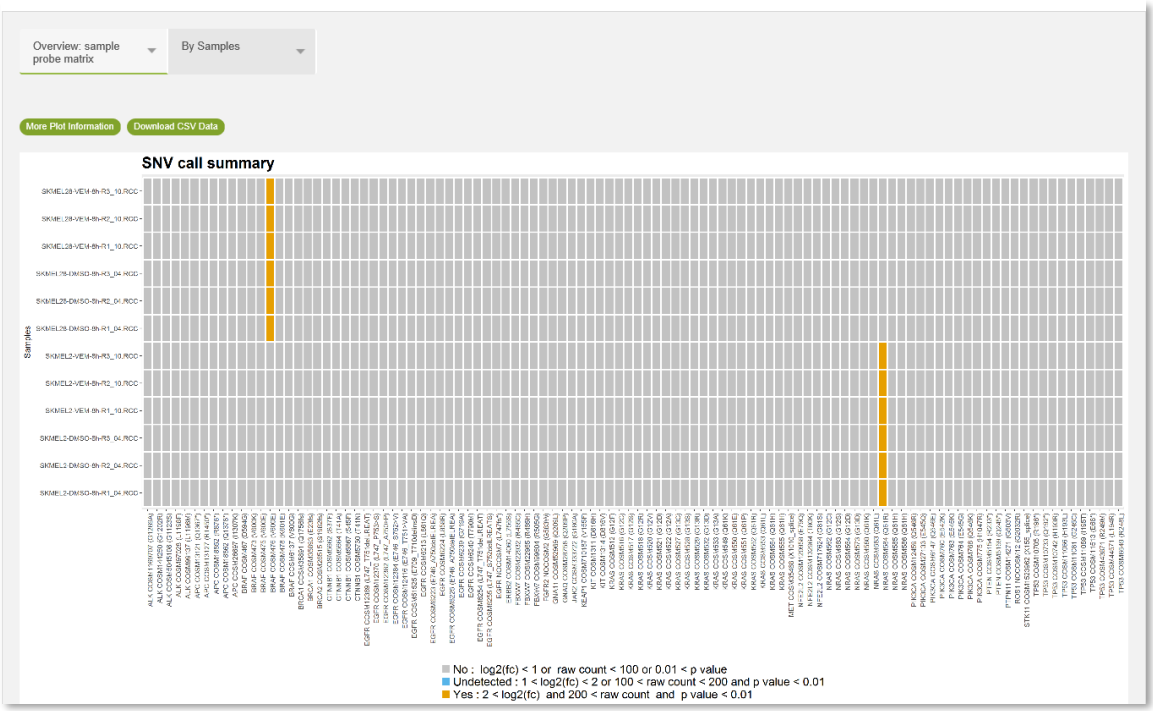


Figure 103: Advanced Analysis SNV module - call summary

The **By Samples** tab allows you to view a plot of each individual sample and the data for each individual gene tested for that sample.

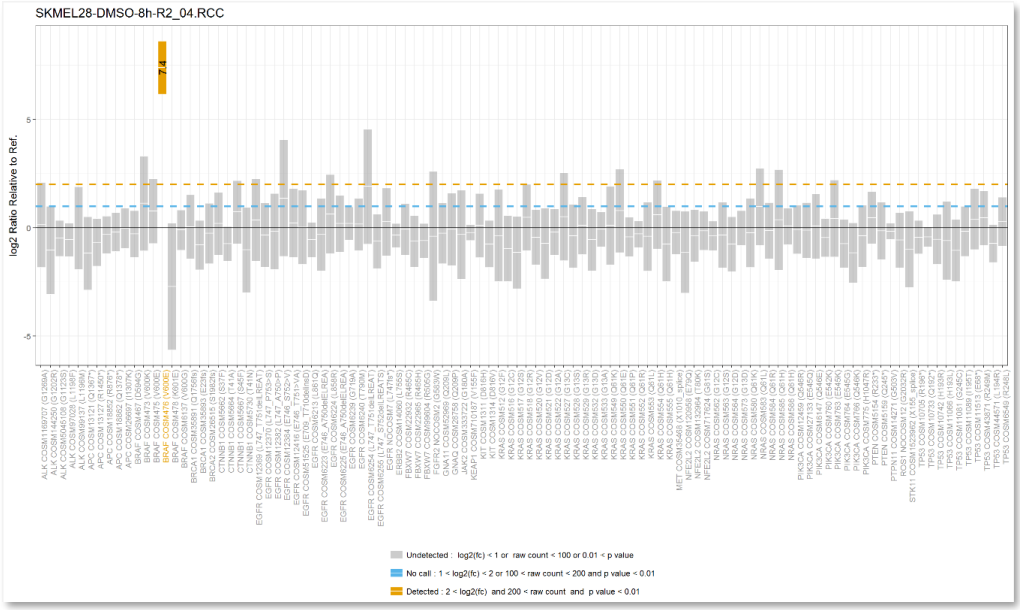


Figure 104: Advanced Analysis SNV module - sample plot

Appendix B: References

Vandesompele, J. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 2002; 3(7).

Wang H, Horbinski C, Wu H, Liu Y, Sheng S, Liu J, et al. NanoStringDiff: a novel statistical method for differential expression analysis based on NanoString nCounter data. *Nucleic Acids Research*. 2016; 44(20):151.

Tomfohr, J., Lu, J., & Kepler. T. B. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*. 2005; 6:225.

Danaher P, Warren S, Dennis L, D'Amico L, White A, Disis ML, et al. Gene expression markers of Tumor Infiltrating Leukocytes. *Journal for Immunotherapy of Cancer*. 2017; 5(1):18.

Glossary

This section defines terminology associated with the Advanced Analysis plug-in module.

Analysis Type: users can choose between two different levels of analysis (see Quick Analysis and Custom Analysis).

Analyte: a sample that can be identified as RNA, DNA, Protein, or a mixture of one or more types based on the composition for the purposes of using them in an nCounter assay.

Annotation: A type of notation that can be used to establish groups of samples or probes.

Boolean (true/false) variable: a variable with exactly two categories; yes or no.

Cartridge: the physical device that has 12 lanes which is put into the Digital Analyzer for counting.

Categorical variable: a discrete variable with two or more categories.

CodeSet: a collection of Capture and Reporter Probes designed against specific target sequences.

Confounder: a variable which affects your data but which is not scientifically relevant. Technical confounders are variables such as run date or cartridge lot. Experimental confounders are variables such as patient body mass index or age.

Covariates: variables which the Advanced Analysis tool can isolate and assess the effect of. At least one covariate must be selected for analysis.

Continuous variable: a variable with infinite possible values.

Custom Analysis: the user may select multiple covariates and customize settings in this analysis. In addition to the core modules, Overview, Normalization, Differential Expression, GSA, and PathView, the user has access to Related Analytes, Probe Descriptive, Cell Type Profiling, and Pathway Scoring.

Custom CodeSet: a CodeSet with probe content customized to meet a specific customer's needs. The probes and their respective target are designed in consultation with the NanoString Bioinformatics team and manufactured by the NanoString.

Directed global significance score: this value measures the extent to which a given gene set is up- or down-regulated relative to a given covariate. It is calculated similarly to the undirected global significance score, but it takes the sign of the t-statistics into account.

Cell Type Profiling module: this module quantifies cell populations using marker genes which are expressed stably and specifically in given cell types. These marker genes act as reference genes specific to individual cell types, as they are expressed only in their nominal cell type, at the same level in each cell.

Differential Expression module: this module is used to identify the specific genes which exhibit significantly increased or decreased expression in response to the chosen covariate. It provides the basis for the Gene Set Analysis (GSA) and PathView modules and should be viewed prior to both.

Gene set: a group of genes affiliated with a common cell type, disease, pathway, or function.

Gene Set Analysis (GSA) module: this module summarizes the change in regulation within each defined gene set (selected along the left side of the window) relative to the baseline (or in the case of continuous variable, per unit change in variable). The values calculated are the global significance score and the directed global significance score and are expressed in heatmaps and/or a data table.

General Options menu: the menu from which the user can begin to customize a Custom Analysis. Among other options, users may adjust parameters and choose modules to run.

Global significance scores: (also called undirected global significance scores) a measure of the overall differential expression of the selected gene set relative to selected covariates, ignoring whether each gene is up- or down-regulated.

Group: a category of samples, usually defined by an annotation.

KEGG pathways: KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways describe high-level functions of cell-signaling pathways.

Overview module: this module provides a general overview of the data through descriptive plots, organized into four categories: Heatmaps, PCA (principal component analysis), Study Design, and Other QC.

Identifiers: unique names that differentiate every sample from the others. The Sample File Name will always be unique, but can be long, so users may prefer to choose another type of identifier.

Normalization module: this module seeks to eliminate run-to-run and sample-to-sample technical variability in the raw counts, which arises from inconsistencies in effective sample input and fluctuations in the overall efficiency in capturing and counting target molecules. It normalizes each analyte-type separately, resulting in clickable analyte-type tabs which reveal respective plots.

Panel CodeSet: an off-the-shelf CodeSet with predesigned probe content manufactured by NanoString.

PathView module: this module overlays the Differential Expression analysis results with various KEGG pathways. Elements that are over-expressed in this pathway are colored gold, those that are under-expressed are colored blue, and those that are neutral are gray.

Pathway Scoring module: this module combines the expression from all genes in a gene set into a single "pathway score". Just as Differential Expression analysis of individual genes or gene sets is used to research the effect of covariates on a dataset, the Pathway Score can be used to summarize the data from a pathway's genes into a single score.

Predictor: a variable which affects the data and which is scientifically relevant. Examples include treatment type, treatment time, and cell line.

Principal Component Analysis (PCA): a way of analyzing data with multiple variables. Variables that naturally correlate with each other will be grouped as a principal component.

Probe annotation file: a .csv file containing annotations which document the biological significance of the probes and link them to the pathways with which they are associated. Users should check probe annotation files to ensure the fields they need are filled.

Probe Descriptive module: this module provides multiple plots which are focused just on the probes of interest, which the user designates on the Custom Analysis menu.

Quick Analysis: this type of analysis is performed with only a *single* covariate and default parameters set for the preselected core modules – Overview, Normalization, Differential Expression, GSA, and PathView.

Related Analytes: this module enables comparison of mRNA and protein expression levels when the gene and protein have been linked in the probe annotations file. It applies all the tools of the Probe Descriptive Module to each pair of related analytes. This module is especially powerful for describing the co-regulation of mRNA and Protein.

Sample annotations: these annotations are assigned to sample groups during experiment creation in nSolver and can be used to label both confounders and predictors.

SNV module: this module summarizes SNV variant events and QC information detected in the data.

Variable: a factor in or element of the experiment which is subject to change.

Use for Analysis: this column is available for covariate selection in setting up an Advanced Analysis.

Z-Score: a value that is used to indicate the distance of a certain number from the mean of a normally distributed dataset.