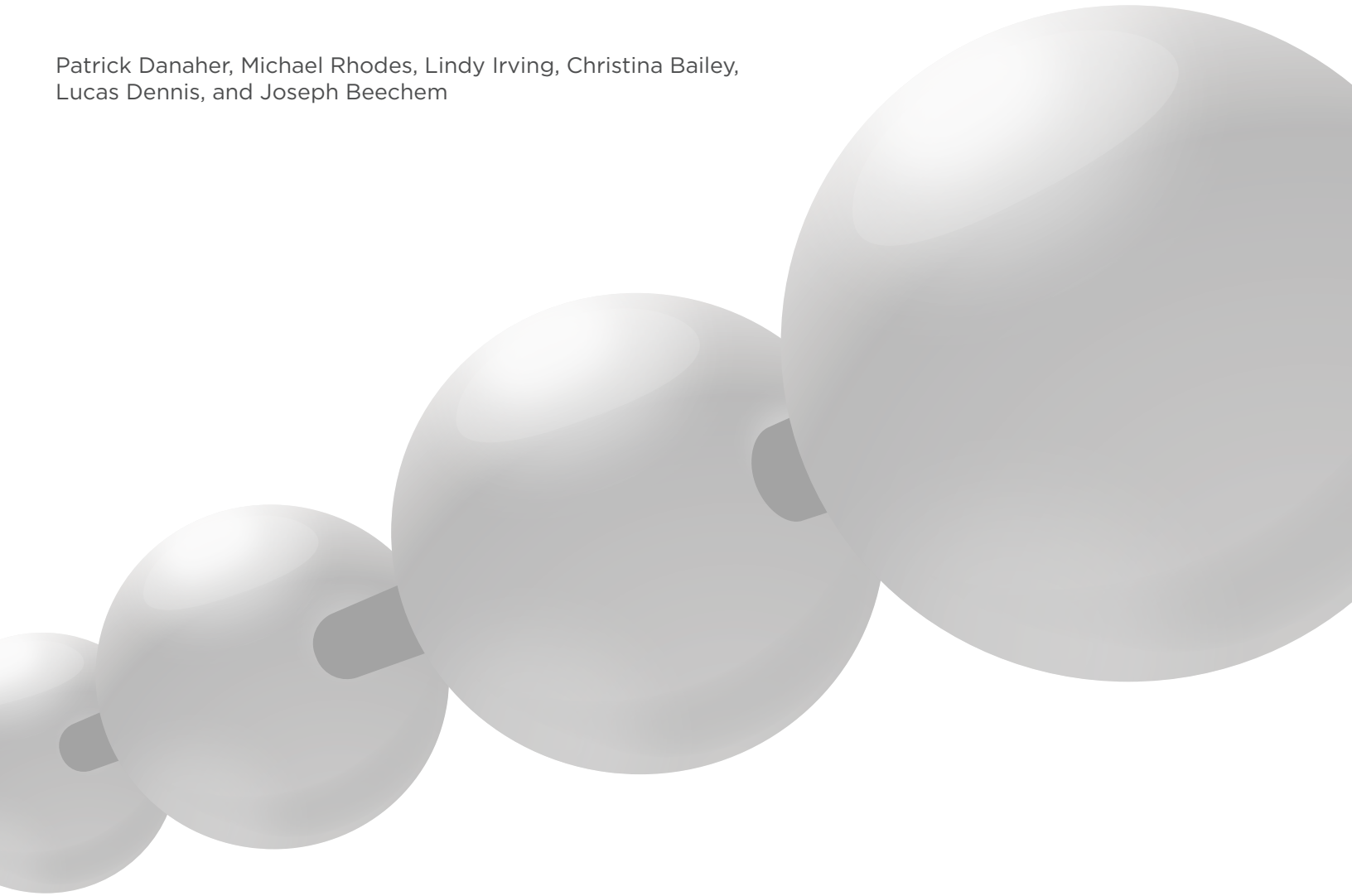
nanoString TECHNOLOGIES®

Using the PanCancer Pathways Analysis Module

for Analysis of nCounter® PanCancer Pathways Data

Patrick Danaher, Michael Rhodes, Lindy Irving, Christina Bailey,
Lucas Dennis, and Joseph Beechem



NanoString Technologies®, Inc.
www.nanostring.com

AUGUST 2015 LBL-10210-01

Using the PanCancer Pathways Analysis Module
for Analysis of nCounter® PanCancer Pathways Data

Introduction

The PanCancer Pathways Analysis Module was created to help scientists perform pathway-centered and statistically principled analyses of their nCounter PanCancer Pathway Panel data. It brings together powerful academic open-source analysis tools, provides a simple interface to guide the user through the analysis, and displays the results in an interactive HTML document. The collection of advanced analysis capabilities that define the PanCancer Pathways Analysis Module includes six modules for QC, Normalization, Pathway Scoring, Differential Expression, Gene Set Analysis, and Pathview Plots. These advanced analyses are performed using R, a powerful statistical software program. However, familiarity with R is not required as users only need to interact with a simple wizard within nSolver™ 2.5 or higher.

Results of an advanced analysis are displayed in two formats. A results directory contains every plot and table created by the analysis. The second format is an interactive HTML analysis report. This white paper describes an example analysis detailing the choices available to the user and then explaining the potential outcomes of these decisions in the results. It is presented in the style of a vignette that shows the complete analysis of an actual PanCancer Pathways Panel dataset.

Performing the nCounter® PanCancer Pathway Panel Advanced Analysis

The ***nSolver Analysis Software User Manual*** explains the basics of how to install and operate nSolver 2.5 or higher; this white paper will begin with the process of setting up an advanced analysis using the PanCancer Pathways Analysis Module. The analysis described below uses the example breast cancer data provided with the nSolver 2.5 or higher installation *.ZIP file (labeled "PanCancer annotated demo data") and can be used as a training tool if you wish to follow along.

Advanced analyses in nSolver 2.5 or higher can only be applied to one of two levels of data—raw data or normalized data—and an experiment must be created within nSolver to run the advanced analysis. If raw data are used, then the PanCancer Pathways Analysis Module can automatically choose optimal normalization genes and use them to perform normalization. Performing the advanced analysis using normalized data allows you to preserve the normalization and/or background subtraction already performed in nSolver.

Analysis modules can be imported into nSolver using the Analysis menu at the top of the application window. Once the advanced analysis wizard opens, choose a name for the analysis and select an analysis module. The PanCancer Pathways Analysis Module will only work with files generated using the PanCancer Pathways Panel and its accompanying Reporter Library File (RLF). A matching RLF is required if additional genes are spiked into the PanCancer Pathways Panel. Spiked-in genes will be excluded from pathway-specific analyses but included in differential expression analyses.

You will also need to determine where the analysis results should be stored. Advanced analysis results are not stored in the nSolver database, so this directory will be used to store the data that nSolver later displays in your HTML viewer, i.e., a web browser. The analysis directory will contain results plots and *.CSV files. Depending on the analysis options selected and structure of the data, these files will require 15-30+ MB of free disk space for 200 samples. Click

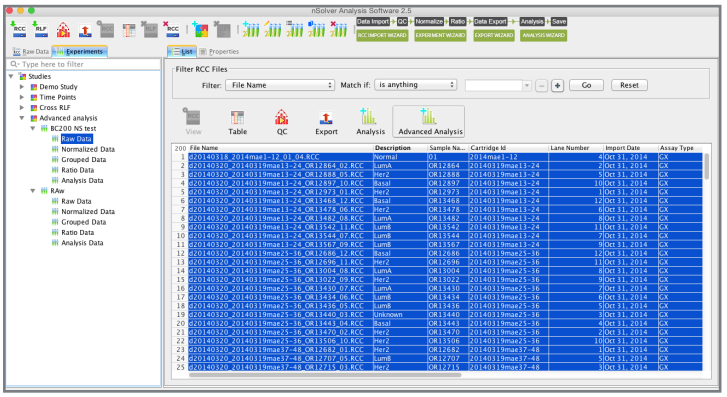


FIGURE 1 Start an Advanced Analysis. First select the desired samples from the Experiments view, then select the **Advanced Analysis** icon.

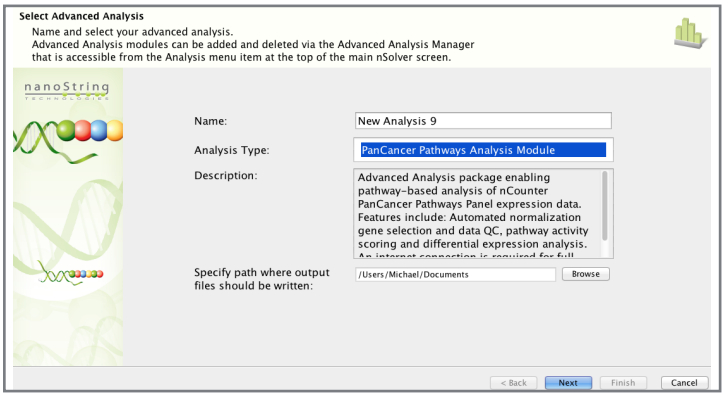


FIGURE 2 Select an Advanced Analysis. Enter an analysis name, confirm the analysis type, and enter the directory where analysis files should be saved.

nanoString TECHNOLOGIES®

Select Sample Annotations

The annotations screen is the first of four screens in which analysis parameters are entered. Select a variable to serve as a unique identifier for every lane. The *.RCC file name will always be a valid identifier, and in this example, File Name has been selected as the identifier. However, these file names tend to be lengthy; it may be preferable to select a more meaningful sample identifier if available.

Next, select the annotations (covariates) to be used in the analysis. Only the covariates selected here will be available for analysis. Standard annotations from nSolver (e.g., Cartridge ID, Scanned Date, and Binding Density) as well as any annotations you previously added to the data in nSolver will be available. A *.CSV file containing additional annotations can be imported and merged with existing annotations. (These custom annotations will not be saved in the nSolver database for future use.)

In most experiments, it will be appropriate to include one or more biological annotations in the analysis. It can also be useful to include technical annotations, either to confirm that they are not influencing the results or to account for their effects in your analyses. For example, CodeSet lot and hybridization time may be technical annotations that deserve consideration.

Three types of annotations—categorical, continuous, and true/false—can be included in the advanced analysis. For nSolver fields, the software attempts to provide logical default annotation types. However, all imported annotations must be specified as categorical, continuous or true/false.

Options Chosen for Example Analysis

Sample Annotations

- Subtype: categorical variable with reference level “normal”
- Binding Density: continuous variable
- Scanned Date: categorical variable, choice of reference level unimportant

Normalization

- Dynamically Choose Housekeepers
- Threshold low count data: under 20 counts in 50% of samples

Pathway Scoring

- Method: PC1
- Baseline: Subtype, using “Normal”
- Adjust for: Binding Density
- Plot vs.: Subtype

Differential Expression

- Predictors: Subtype
- Confounders: Binding Density
- p-Value Adjustment: Benjamini-Yekutieli
- Run GSA
- Run Pathview
- Additional KEGG IDs: 05200

Categorical – These are annotations for which the samples exist in a number of distinct categories. In this example, Subtype and Scanned Date are categorical. A categorical covariate may contain text or numbers and must always have a defined “categorical reference” or baseline. The choice of a reference shapes differential expression analysis, which will compare all the levels of the categorical annotation to the chosen reference level.

Continuous – Continuous annotations have values that can be interpreted meaningfully as numbers. Binding Density is a good example of a continuous variable: if two samples have binding densities of 1.0 and 1.2, this can be interpreted to mean the second sample has binding density 0.2 units greater than the first. However, some numeric variables, such as Disease Grade, describe more arbitrary measures. Classifying this annotation as “continuous” would be dubious because it would imply that the difference between Grade I and Grade II disease is the same as the difference between Grade II and Grade III, i.e., one “unit” of disease. Numeric variables like Disease Grade are thus better modeled as categorical annotations.

True/False – These annotations must take only the values TRUE or FALSE. For the purposes of the PanCancer Pathways Analysis Module, such annotations are equivalent to categorical annotations with FALSE as a reference level.

This example dataset contains results from 215 breast cancer and healthy breast tissue samples assayed with the PanCancer Pathways Panel. For each cancer sample, the subtype is known and was annotated in nSolver as A, B, C, or D. Thus the biological annotation **Subtype** was selected for the analysis. Additionally, the technical annotations **Binding Density** and **Scanned Date** were selected for analysis so that their effects can be adjusted for in the analysis. The annotations screen is used to specify that Binding Density is continuous, and Scanned Date and Subtype are categorical. Normal was specified as the categorical reference for Subtype and will be the baseline to which all other subtypes will be compared. An arbitrary reference can be used for Scanned Date because there is no natural choice to designate as a baseline.

Click **Next** to continue to the normalization screen.

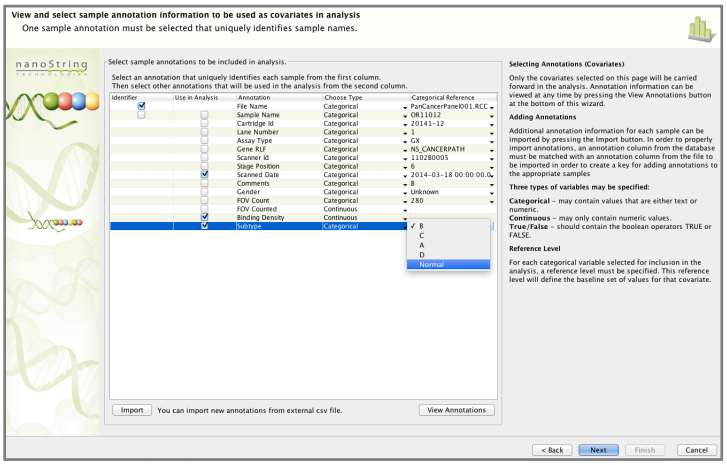


FIGURE 3 Select Annotations. In this example, values of the **Subtype** annotation have been loaded from a *.CSV file and merged with the standard annotations. Select **File Name** as the Identifier. Select **Scanned Date**, **Binding Density**, and **Subtype** as covariates to use in the analysis. Set Subtype to **Categorical** and choose **Normal** as its categorical reference level. (All categorical annotations require a reference level.)

Normalization Options

The PanCancer Pathways Panel has 40 candidate normalization genes ("housekeeping genes") that were selected based on stability in TCGA gene expression data from multiple cancer types. The stability of any of these 40 genes will vary between datasets because not all potential housekeeping genes are expressed stably in all cancer types or when exposed to a given treatment. Therefore, optimal analyses will perform normalization using only the most stable subset of these genes.

The PanCancer Pathways Analysis Module's normalization module uses the popular geNorm algorithm (Vandescompele et al., 2002) to identify an optimal subset of housekeeping genes. 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. geNorm relies on this theory, iteratively removing candidate housekeepers with the least stable expression relative to other candidates. The user may also specify the desired number of housekeeping genes. Note that the PanCancer Pathways Analysis Module cannot automatically detect whether normalized or raw data are used, so be sure to select appropriate normalization options during the advanced analysis. Normalization performed using the PanCancer Pathways Analysis Module will override any previously performed normalization.

It is possible that some genes may not be expressed in some or all samples because the PanCancer Pathways Panel is designed to work with a wide variety of cancer types. Setting the threshold for low count data helps to avoid spurious conclusions based on analysis of background rather than signal by removing genes that fall below a given low count level more than a set percentage of the time. Take care when setting this threshold. For example, if there are three treatments and the threshold is set to 25% of samples, genes that were silenced by one treatment—i.e., genes that were expressed in two groups but not in the third—could be eliminated despite their biological significance. If the effect of this filter is a concern, you can run the analysis with and without filtering. Conclusions that are robust to the choice of data cleaning method are more likely to be reproducible.

The PanCancer Pathways Analysis Module creates *.PNG images of all plots and inserts them into the final interactive report. If another plot type is chosen, duplicates of all *.PNG images will be made in the desired format. These images can be found in the analysis results directory specified on the first page of the Advanced Analysis Wizard.

Click **Next** to continue to the pathway scoring options page.

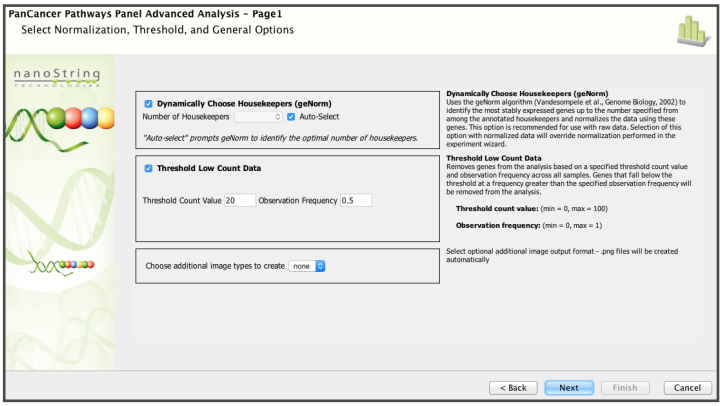


FIGURE 4 Select Normalization, Threshold, and Graphics Options. Use the default analysis settings as shown above.

Pathway Scoring Options

The PanCancer Pathway Panel was designed around thirteen canonical cancer pathways (Vogelstein et al., 2013; Dennis et al., 2014). It is therefore useful to explore panel data at a pathway level before exploring single-gene results. The PanCancer Pathways Analysis Module provides three methods for summarizing the data from a pathway's genes into a single score.

1st PC (First Principal Component)

A common approach to extracting pathway-level information from a group of genes uses the first principal component (PC) of their expression data (Tomfohr et al. 2005). For a given pathway, PC analysis scores each sample using a linear combination (i.e., a weighted average) of its gene expression values, weighting specific genes to capture the greatest possible variability in the data. Thus, the first PC will reflect whatever factor is the main driving force of variability in gene expression for that dataset.

Data for each pathway are scaled before taking the first PC by dividing each gene's $\log_2$ expression values by the greater of either their standard deviation or 0.05, thereby avoiding the assignment of excess importance to genes whose variance may be entirely attributable to technical variation.

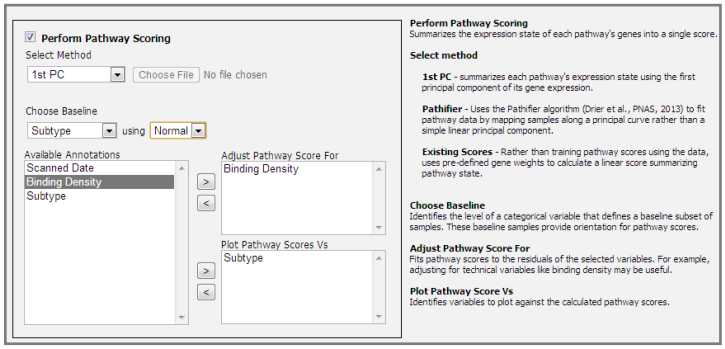


FIGURE 5 Select Pathway Scoring Options. Select **1st PC** as the pathway scoring method, select **Subtype** as the baseline, and use **Normal** as the baseline sample. Adjust pathway score for **Binding Density** and plot the pathway scores **vs. Subtype**.

Pathifier

The Pathifier algorithm (Drier et al., 2013) is an increasingly popular method for summarizing pathway state (Taherian-Fard et al., 2014; Tian et al., 2014). Pathifier is similar in spirit to principal component-based pathway scores, but it uses more advanced methodology in its attempt to better fit the data. Rather than assuming a purely linear effect of pathway status on gene expression, Pathifier scores samples by their locations along a "principal curve" (Hastie et al., 1989). This approach sometimes results in a very different pathway score; whether a Pathifier score is a better representation of pathway status than the first principal component depends on the dataset.

A drawback of the Pathifier method relative to the simpler first PC method is that Pathifier scores are non-linear and consequently difficult to apply to future datasets. Pathifier may run very slowly on large datasets, potentially taking many hours. It will often be practical to initially analyze this dataset using the first PC method and later run Pathifier after deciding on a final form for the analysis.

nanoString TECHNOLOGIES®

Existing Scores

The PanCancer Pathways Analysis Module also allows the user to score pathway status using pre-defined weights. The user may provide the path to a *.CSV file containing these weights, including the option of entries for pathways or other gene signatures beyond the original 13 pathways included in the panel. Genes in these files that do not appear in the data will be ignored.

Pre-defined scores can be used to implement a previously published gene expression signature with genes that overlap those in the panel. This option can also be useful in datasets that are too small to train stable pathway scores, e.g., pathway scores could instead be trained on very large publicly available databases like TCGA and then applied to a small cancer panel dataset. Although a pathway score trained in one dataset may be higher or lower than in another dataset due to differences in normalization or platform, its interpretation may persist.

In an example of how this option can be used, gene weights for pathway scores could be trained on a dataset of patients receiving varying doses of a drug. Pathway scores trained on this dataset would arguably capture the effects of dose response. Using the option of pre-existing scores, these "dose response" pathway scores could then be applied to a new study in which all patients received the same dose; in this new dataset, the scores might measure varying responses of patients to a common treatment.

TABLE 1 An example of a properly formatted *.CSV file that provides pre-defined scores.

	Notch	Wnt	Tamoxifen Response	Histones
PIK3R2	0	0	0.5	0
IL11RA	0	0	0.8	0
BAMBI	0	-0.045	0	0
ETV1	0	0	0	0
...	...	...	...	...

Select a Baseline for Pathway Scores

For methods like principal component analysis and Pathifier, the sign of a pathway score is arbitrary: a pathway score will fit the data equally well if it is multiplied by -1. This makes it difficult to compare scores from different pathways; for example, it is possible for the MAPK score to increase with pathway dysregulation and for the PI3K score to decrease with dysregulation. To avoid this confusion, a set of samples must be designated as a baseline. These samples are identified by choosing one level of a categorical variable. Once baseline samples have been identified, the signs/directions of the pathway scores (for principal component analysis and Pathifier only) will be changed so that for every pathway the baseline samples have below average scores.

Adjust for Covariates

For both the principal component analysis and Pathifier options, the PanCancer Pathways Analysis Module allows the user to adjust for covariates before performing pathway scoring. 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.

It is usually advisable to adjust for technical variables that are suspected to widely influence gene expression. Adjusting for a biological variable is a more difficult decision. In some cases, you may want to score pathway status independent of one biological variable in order 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.

In this example the pathway scores are to be adjusted for Binding Density. We have opted against adjusting for Scanned Date with the assumption that it has no effect on gene expression because it is correlated with Subtype in our data. Adjusting for variables correlated with a variable of interest will remove some of that variable's signal from the data. Finally, as we wish to use this experiment to compare subtypes, we have refrained from adjusting the pathway scores for subtype; doing so would strip its signal from the data.

Click **Next** to continue to the differential expression analysis screen.

Differential Expression Options

The PanCancer Pathways Analysis Module uses linear regression to investigate differential gene expression in response to multiple covariates simultaneously. This approach isolates the independent effect of each covariate on gene expression and avoids confounding due to technical variables. For example, when variables are confounded, this approach supports statements such as, "case vs. control status is associated with a 2-fold increase in BCL2 expression, holding age and sex constant."

To perform differential expression analysis, select at least one variable as a predictor. Additional variables may be selected as confounders. The linear regressions treat predictors and confounders identically, but results are only reported for predictors.

Two covariates are included in this example analysis: Subtype and Binding Density. A third annotation, Scanned Date, was omitted from the analysis because it is not expected to influence gene expression, and so there is no benefit in using the data to fit meaningless terms for each date. Recall that we have specified on the Annotations page that Subtype is a categorical variable with five levels and "Normal" designated as the reference level. The linear regression will fit a separate term modeling the difference of each of the four remaining subtypes from Normal samples. Finally, Binding Density is designated as a potential confounder rather than as a covariate of interest.

A linear regression will be run for each gene using the following model:

$$\mathrm{E}\ \log_2(expression) = \beta_0 + \beta_1(Subtype\ A) + \beta_2(Subtype\ B) + \beta_3(Subtype\ C) + \beta_4(Subtype\ D) + \beta_5(Binding\ Density)$$

where "SubtypeA", "SubtypeB", "SubtypeC", and "SubtypeD" are variables taking the values 0 or 1 depending on each sample's subtype, and each β_n is a constant to be estimated by the linear regression.

The large number of genes in the CodeSet makes the use of raw p-values problematic: when 730 genes are tested for association with a covariate, 36.5 genes are expected to have $p < 0.05$ by chance alone. The differential expression module provides two methods for adjusting p-values: the Benjamini-Yekutieli false discovery rate (FDR) and the Bonferroni correction. FDR is the proportion of genes with equal or greater evidence for differential expression 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. The Benjamini-Yekutieli method returns conservative estimates of FDR. The Bonferroni correction is a more 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.

Once a differential expression analysis has been set up, the PanCancer Pathways Analysis Module provides methods for examining its results from a pathway perspective rather than the level of an individual gene. Select the **Run GSA** button to calculate global significance scores summarizing the overall level of statistical significance of each covariate in each pathway. This analysis is entirely distinct from the pathway scores calculated in the pathway scoring module, and it provides a statistically independent approach for evaluating pathways' biological interest.

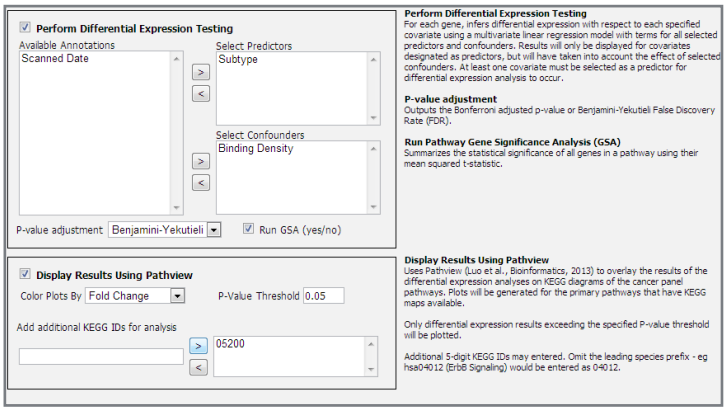


FIGURE 6 Select differential expression options. Perform differential expression testing using **Subtype** as predictor and **Binding Density** as confounder. Use the other default settings as shown above. Enter KEGG pathway ID **05200** in the bottom box on the left and use the arrow to move it right.

Finally, the option to **Display Results Using Pathview** will overlay the differential expression results on KEGG pathway graphs using the Pathview R package (Luo et al., 2013). **FIGURE 23** shows an example Pathview plot. For each variable in the differential expression analysis and for each pathway with a KEGG graph (three of the pathways included in the PanCancer Pathways Panel are not represented in KEGG), the PanCancer Pathways Analysis Module will produce a Pathview plot. Each node in a KEGG pathway represents a protein family and may correspond to multiple genes. Pathifier colors nodes according to the differential expression of their genes, measured either by fold change, ignoring statistical significance, or by t-statistics, which reflect statistical significance and correspond imperfectly to fold change. For both coloring schemes, a p-value threshold can be selected so that genes above this threshold will have their log fold change and t-statistics set to zero before Pathview is run. Additional KEGG pathway IDs can be entered as 5-digit numbers. This example uses the ID 05200 for the very high-level KEGG graph "Pathways in Cancer." Note that Pathview requires an Internet connection to run.

Click **Finish** to begin the analysis. If the Pathifier option was selected then analysis may take several hours; otherwise it will likely require between 2 and 15 minutes depending on the number of samples and the number of covariates. While the analysis is running, its progress will appear in place of the Analysis Results.

View the Analysis Results

When completed, results of the analysis can be viewed by selecting the appropriate data from the Experiments view and then selecting the Analysis Data icon above (**FIGURE 7**). This will open an HTML document. On most computers, HTML files will open in the default web browser.

The analysis is a navigable document with multiple layers of information, highlighted in **FIGURE 8**.

1. The first menu selects the analysis module to display. Available results depend on the analysis that was run.
2. The second menu links to different results within an analysis module. These choices will often have submenus for selecting individual covariates.
3. The third menu selects a pathway to focus on within a module.
4. At the bottom of each page are additional details on the interpretation of active plots.

Select the Summary tab in the first menu to list every choice made in the analysis, providing a record of the methods used to generate the results in the analysis and a means to replicate the analysis on a new dataset. This page also gives the location of the analysis directory where plots and tables are stored.

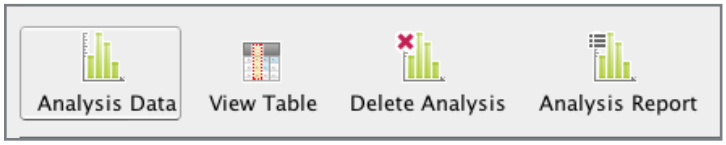


FIGURE 7 Click the **Analysis Data** button in the Navigation menu used to access the analysis results.

Data Exploration and QC Module

The PanCancer Pathways Analysis Module creates numerous plots that explore the structure of the data. NanoString recommends examining these plots before viewing the main analysis results because they give other results context and may even suggest changes to the analysis.

Before looking at any gene expression data, it is useful to examine the basic details of the study design. The PanCancer Pathways Analysis Module draws plots examining the relationships between all covariates included in the analysis. **FIGURE 9** shows examples of such plots, which can be seen in the HTML analysis report by clicking on the **Pair Plots** link. In **FIGURE 9A**, **Subtype** is compared to **Binding Density**. The two variables have no notable association, meaning Binding Density (a surrogate for RNA input) will not dramatically confound analyses of Subtype. **FIGURE 9B** shows the number of each subtype at each scanned date.

A perfectly balanced experiment would have equal proportions of the subtypes at each date, and a poorly-designed experiment would have each subtype scanned at a different date. The case here is somewhere in between. In particular, the final scanned date has a disproportionate share of Normal samples, and any anomaly in that date's scan will influence comparisons of Subtype vs. Normal samples. If there is a reasonable concern that Scanned Date might influence the data, this imbalance could be accounted for by adjusting for Scanned Date and including it as a confounder in the analyses.

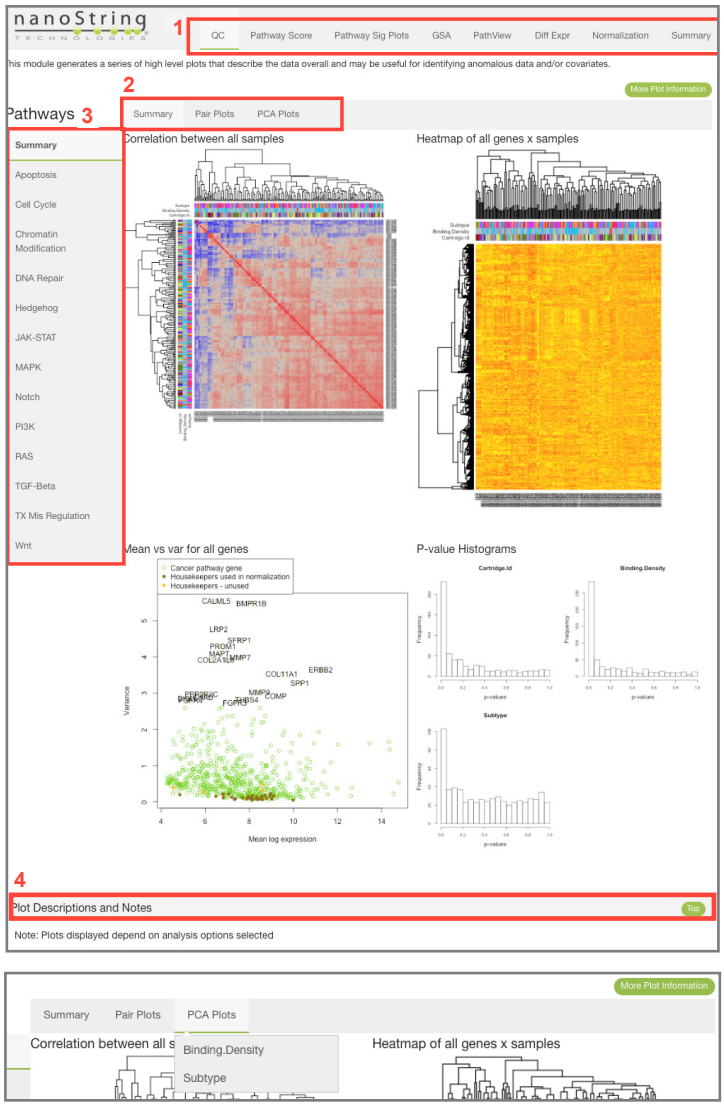


FIGURE 8 An overview of four key areas used to navigate the analysis results and an example of submenus within the secondary navigation menu when viewing PCA plots (highlighted as area 2).

The PanCancer Pathways Analysis Module creates numerous plots that explore the structure of the data. NanoString recommends examining these plots before viewing the main analysis results because they give other results context and may even suggest changes to the analysis.

Before looking at any gene expression data, it is useful to examine the basic details of the study design. The PanCancer Pathways Analysis Module draws plots examining the relationships between all covariates included in the analysis. **FIGURE 9** shows examples of such plots, which can be seen in the HTML analysis report by clicking on the **Pair Plots** link. In **FIGURE 9A**, **Subtype** is compared to **Binding Density**. The two variables have no notable association, meaning Binding Density (a surrogate for RNA input) will not dramatically confound analyses of Subtype. **FIGURE 9B** shows the number of each subtype at each scanned date.

A perfectly balanced experiment would have equal proportions of the subtypes at each date, and a poorly-designed experiment would have each subtype scanned at a different date. The case here is somewhere in between.

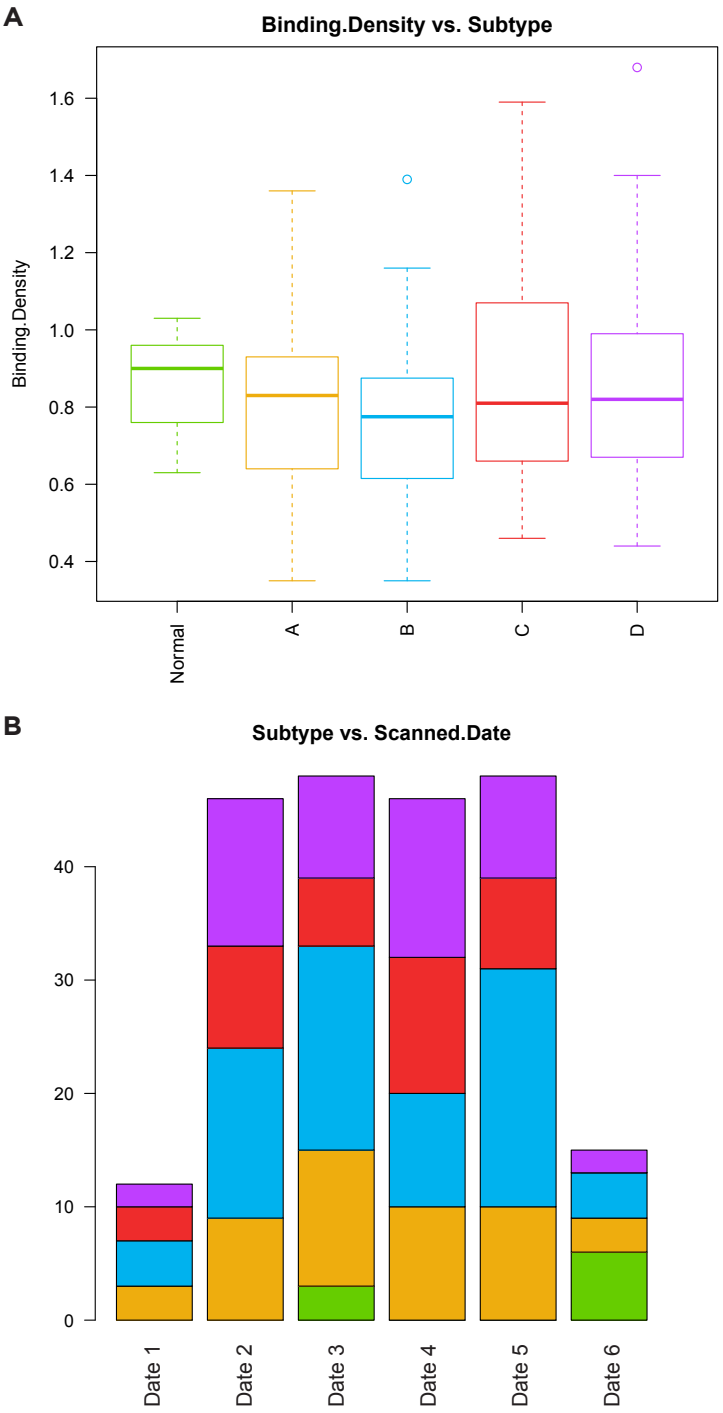


FIGURE 9 Example plots describing study design through comparison of covariates.

In particular, the final scanned date has a disproportionate share of Normal samples, and any anomaly in that date's scan will influence comparisons of Subtype vs. Normal samples. If there is a reasonable concern that Scanned Date might influence the data, this imbalance could be accounted for by adjusting for Scanned Date and including it as a confounder in the analyses.

It is good practice to understand the effect of every available covariate on gene expression. **FIGURE 10** draws histograms of p-values for the associations

between all genes and each covariate. Covariates with no association with gene expression appear mostly flat, and covariates with widespread effects on gene expression have peaks near zero. Technical covariates with such left-weighted histograms should be adjusted for in downstream analyses so as 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. This phenomenon is apparent in the histogram for Scanned Date, whose association with so many genes is most likely a result of its separate association with the covariate Subtype and not a real effect present in the data. In larger datasets there is little harm in adjusting differential expression analyses for likely unimportant technical variables like Scanned Date. In contrast, pathway scoring analyses should only be adjusted for technical variables of serious concern. If we were to adjust the pathway scores for Scanned Date, the association between Scanned Date and Subtype would result in some of the signal of Subtype being stripped from the data.

Binding Density is more likely to be associated with expression of some low-count genes, serving as a surrogate for total RNA input. Samples with low RNA input will be given larger normalization factors. A sample's normalization factor should have no effect on the normalized expression level of genes with moderate to high signal, but normalized expression of genes in or near the background of the system will vary with a sample's normalization factor.

Once the covariates of a study are understood, examine the structure of the gene expression data. **FIGURE 11** shows the mean and variance on the $\log_2$ scale of each gene in the normalized data. It confirms that the selected housekeeping genes are stable, and it shows the genes with the greatest variability, which will often be the most interesting genes for further study.

A heat map of the data is provided in **FIGURE 12**. Genes and samples are organized by hierarchical clustering, and colored bars indicate each the value of each sample for each covariate. Each row is a single gene, and each column is a single sample. Sample names will be illegible in large datasets, in which

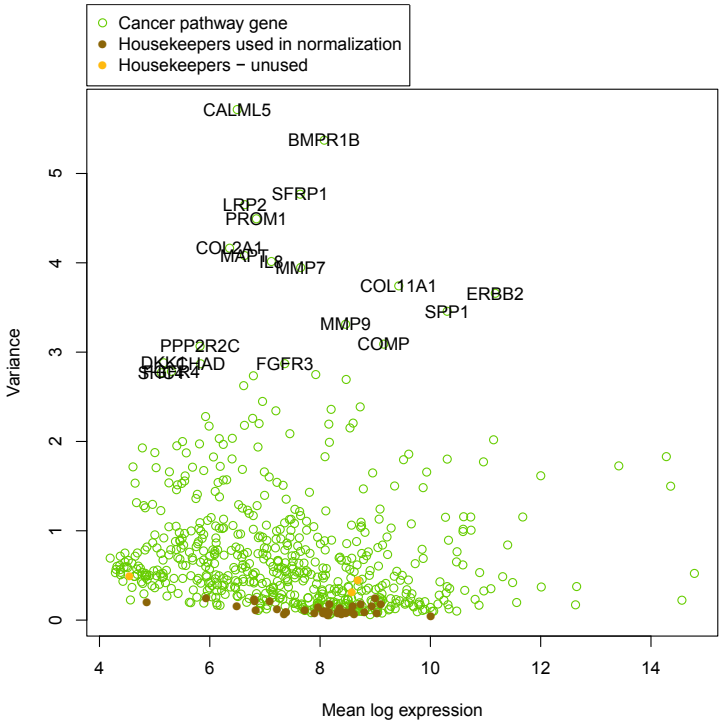


FIGURE 11 Variance and mean on the log scale of each gene in the data set.

case nSolver's interactive heat map functionality can provide a more "zoomed-in" view. **FIGURE 13** shows a heat map of the samples' correlation matrix. These heat maps can be useful to understand the similarity of samples to each other in more detail than a dendrogram can provide.

A heat map is also drawn for each pathway; **FIGURE 14** illustrates the DNA Repair pathway for the example data set. Interestingly, the Normal and Subtype A subtypes (colored green and orange) are clustered distinctly apart from the other subtypes in this pathway.

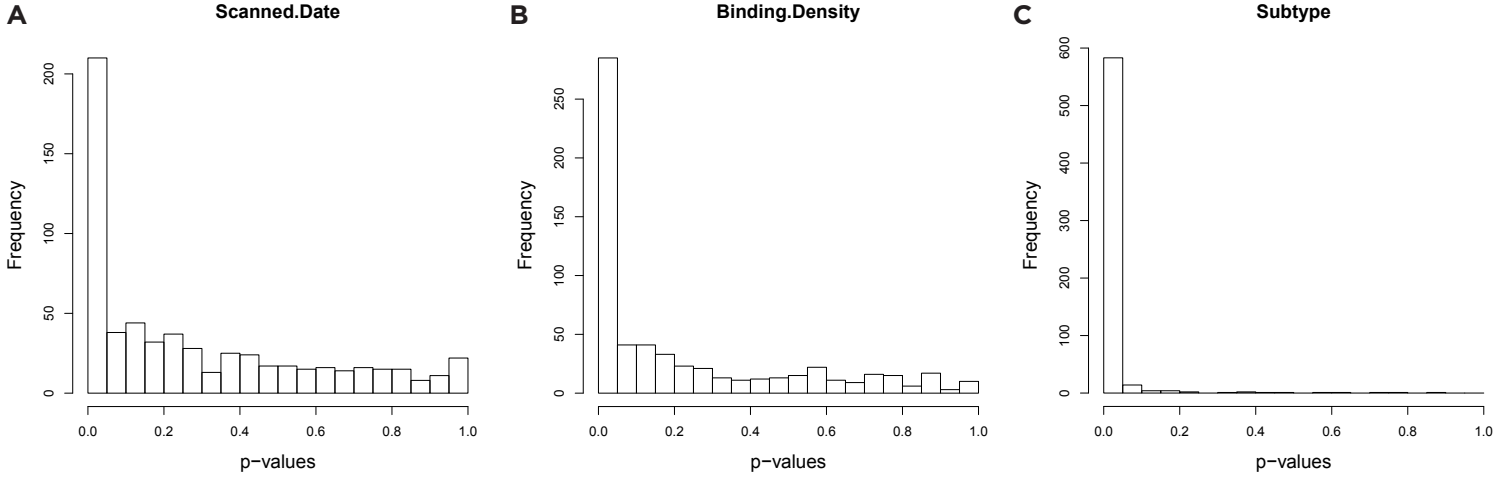


FIGURE 10 Histograms of p-values for each gene's association with each variable.

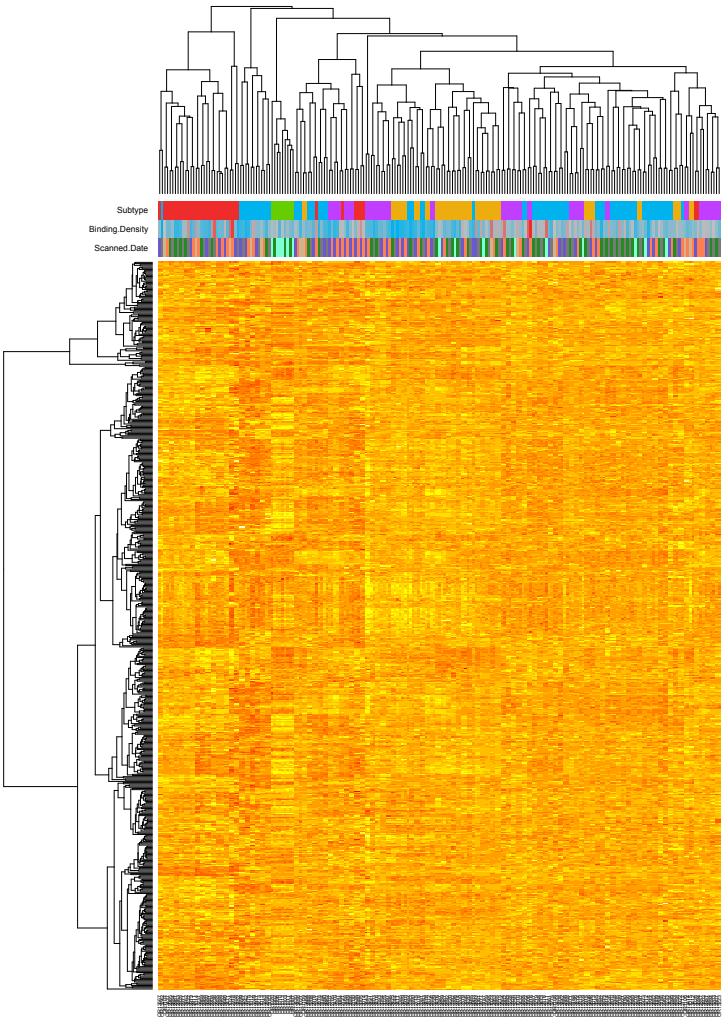


FIGURE 12 Heat map of all gene expression data. Yellow indicates high expression, and red indicates low expression. Expression values are scaled within genes.

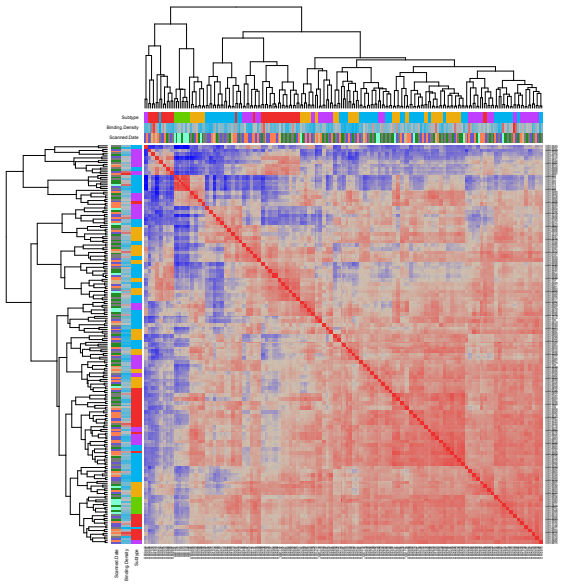


FIGURE 13 Heat map of the sample's correlation matrix. Blue indicates low correlation, and red indicates high correlation.

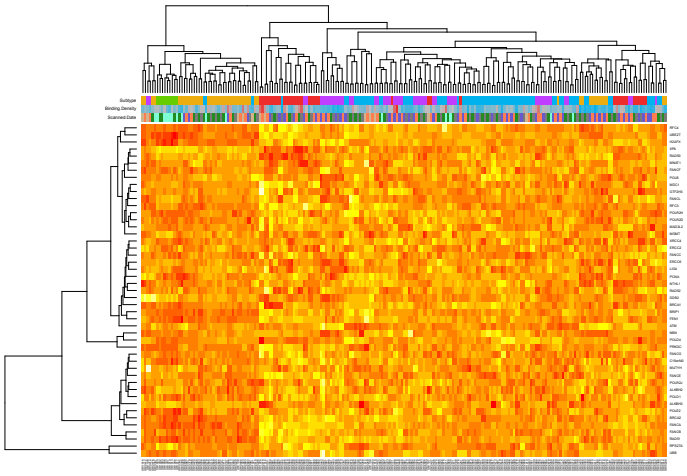


FIGURE 14 Heat map of gene expression data only from the DNA Repair pathway. Yellow indicates high expression, and red indicates low expression. Expression values are scaled within genes.

FIGURE 15 plots the first four principal components against each other. These plots can be viewed using colors determined by the values of each covariate. **FIGURE 15A** is color-coded with respect to the covariate Subtype. The powerful effect of Subtype is evident in the first two principal components of the data, which together capture 27% of the variability in the data. The third principal component appears to separate the few normal samples from the tumor samples. **FIGURE 15B** is color-coded with respect to Binding Density, and under this schema the fourth principal component, which reflects 7% of the total variability in the data, is seen to respond to Binding Density. Binding density in this example is an important variable for which adjustment may be necessary.

Samples that are outliers in any of the first four principal components of the data are indicated to the user in a file named "outliers in first 4 principal components.csv" and saved in the QC folder of the analysis results directory. Outliers may be biologically interesting or caused by technical artifacts like failed reactions. Samples that were defined as outliers by the PanCancer Pathways Analysis Module and initially flagged by nSolver for any reason should be treated with caution. Confirm that any important analysis results hold even when these samples are removed.

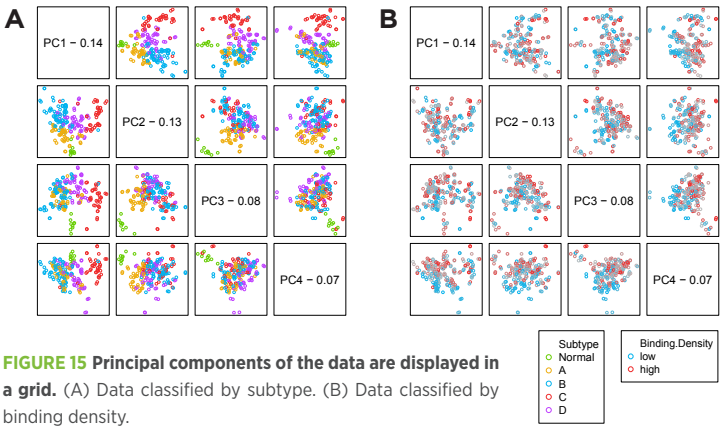


FIGURE 15 Principal components of the data are displayed in a grid. (A) Data classified by subtype. (B) Data classified by binding density.

Normalization Module

The PanCancer Pathways Analysis Module displays two plots detailing the performance of the selection of normalization genes. **FIGURE 16** shows the results of the geNorm algorithm applied to the example dataset. The horizontal axis shows the order in which candidate genes were removed from consideration, and the vertical axis shows a measure of pairwise stability amongst the remaining candidate genes. Black points indicate the selected subset of housekeeper genes. The algorithm only removed three genes before attaining optimal pairwise agreement. Looking back to **FIGURE 11**, three of the candidate housekeepers had significantly higher variance than the others. The list of selected housekeepers can be seen by selecting the link "view selected HK genes."

The effects on the data of normalizing to the chosen housekeepers are displayed in **FIGURE 17**. Histograms of average log gene expression of each sample are drawn from the pre- and post-normalization data. The lower graph displays a tighter histogram of the normalized data, indicating that normalization has successfully reduced variability in total gene expression.

If the desired subset of housekeeper genes has already been identified, the normalization should be carried out in nSolver using the desired housekeepers before running the PanCancer Pathways Analysis Module. Running the analysis on nSolver's normalized data and selecting the "no normalization" option will preserve the normalization performed using these genes.

Pathway Scoring Module

The pathway scoring module is a powerful tool for understanding gene expression data on the level of pathways. **However, pathway scoring results should be interpreted with extreme caution.** Both principal component analysis and Pathifier are "unsupervised" methods that fit a signal to the data without input from informative covariates like Subtype or Treatment Group. Therefore, interpretation of their results can be difficult.

Some authors argue that first principal component and Pathifier-based pathway scores capture the extent of pathway dysregulation in cancer samples (Drier et al., 2013). In biologically homogeneous datasets, pathway scores may capture pathway activity level. In datasets with a powerful clinical variable like Treatment Group or Subtype, pathway scores will often reflect the influence of the clinical variable on gene expression. It is also conceivable that pathway scores will reflect technical effects such as lot-to-lot differences. Thus there is no guarantee that pathway scores represent easily interpretable quantities like pathway activity, pathway dysregulation, or pathway response to treatment. Similarly, there is no guarantee that pathway scores have the same "direction." For one pathway, a high score may mean high dysregulation whereas in another a high school would mean low dysregulation. In light of these uncertainties, pathway scores should be considered as a descriptive tool for data exploration and hypothesis generation.

FIGURE 18 shows selected results from the pathway scoring analysis. **FIGURE 18A** shows a heat map of the 13 pathway scores across the samples in this example. (Colors indicating the subtypes can be seen in **FIGURE 18C.**) These pathways break into 3 distinct clusters. The 8 pathways at the bottom of **FIGURE 18A** have very high scores in Subtype B, while Chromatin Modification, DNA Repair, and Cell Cycle pathways have their highest scores in Subtype C and D samples. The STAT and Apoptosis pathways cluster tightly together and far from the others.

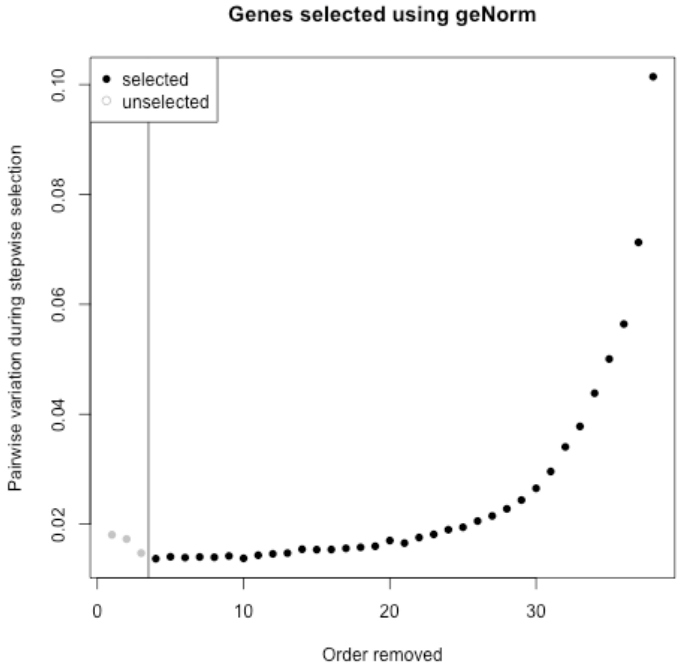


FIGURE 16 Results of the geNorm algorithm applied to the data.

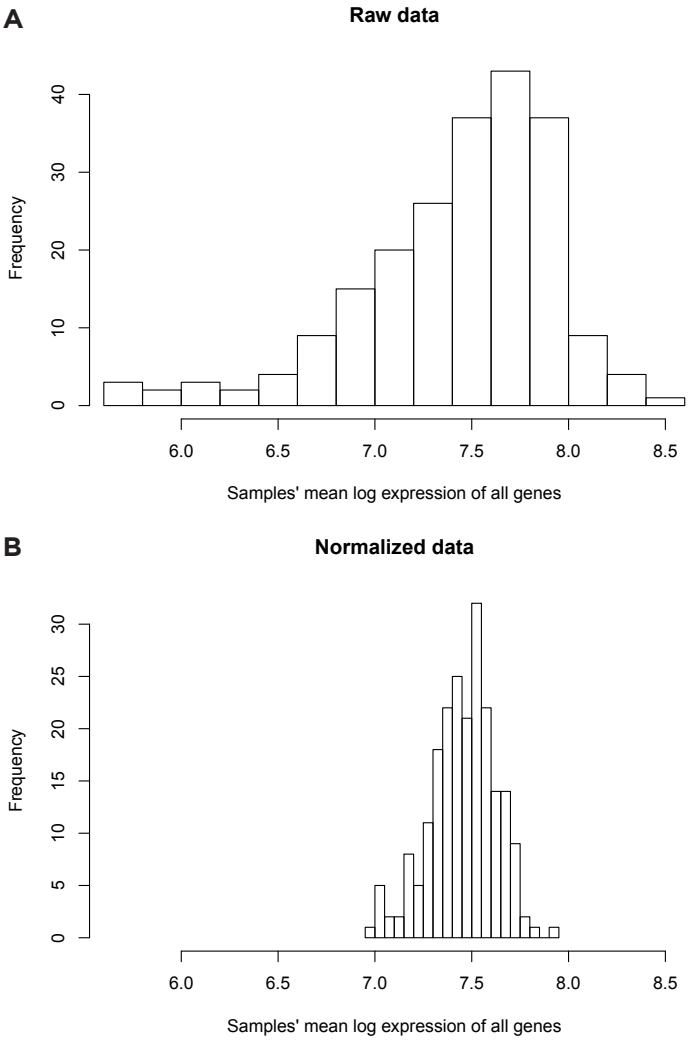


FIGURE 17 Distribution of samples' mean log gene expression before and after normalization.

nanoString TECHNOLOGIES®

FIGURE 18B shows the scores of selected pathways plotted against each other. The DNA Repair and Cell Cycle scores are highly correlated in the left panel, leading to the conclusion that the same underlying factor drives variability in these pathways. The middle panel shows that MAPK and Chromatin Modification pathway scores that are correlated but far from redundant. Subtype C samples have high Chromatin Modification scores, while Subtype B samples have higher MAPK scores. This observation suggests that tumors of these subtypes rely on dysregulation of different pathways to maintain growth. Finally, the Apoptosis and Cell Cycle scores in the right panel are almost completely uncorrelated, indicating that they reflect very different biological events.

STAT and Apoptosis pathway scores appear to have an opposite polarity from the other 11 pathways in FIGURE 19A. However, the direction of these scores is probably due to chance. **FIGURE 18C** shows that Normal samples, which are used as a baseline to orient pathway scores, fall near the middle of the Apoptosis score range; the plot for STAT scores has a similar pattern. Because STAT and Apoptosis scores do not appear to reflect a continuum from normal to highly deregulated samples, they likely have a different interpretation than other pathway scores.

FIGURE 18C also shows Cell Cycle and Apoptosis pathway scores stratified by subtype. The Cell Cycle scores exhibit clear subtype differences, with Subtype C samples getting by far the highest scores. However, Subtype C samples can be further divided into two clusters; a sizeable subset has extremely high Cell Cycle scores. These samples may be a particularly proliferative subset of tumors, or they may have particularly deregulated cell cycle checkpoints. Analysis of Apoptosis pathway scores appears quite different: all subtypes exhibited a range of scores, with no clear tendencies for one subtype to have

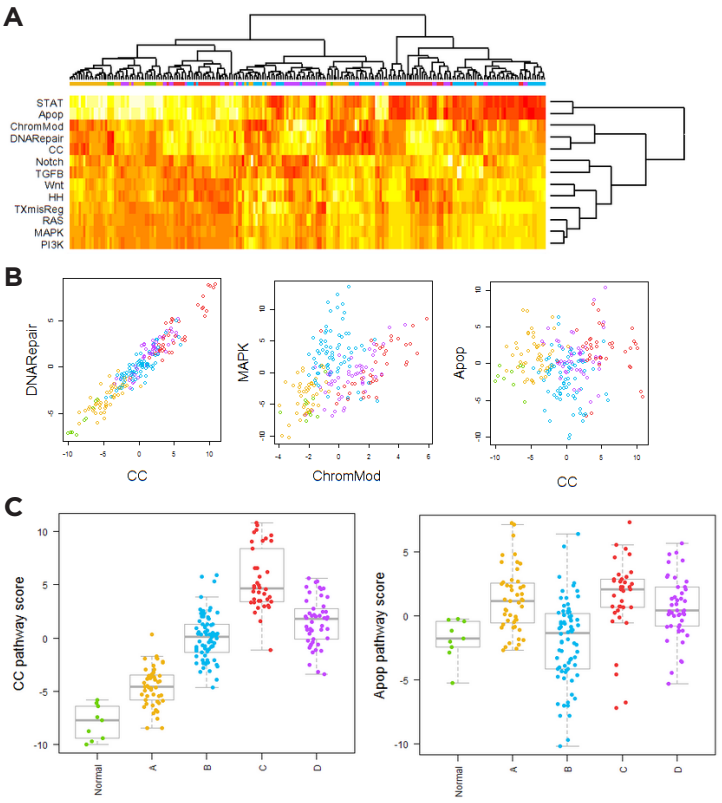


FIGURE 18 (A) Heat map of the 13 pathway scores across all samples illustrates clear differences between three groups of pathways. (B) Selected pathway score comparisons illustrate different levels of correlation as well as clusters of each subtype. (C) Box plots of pathway scores within each subtype represent clear differences in scores for Cell Cycle but less distinct variation for Apoptosis.

Differential Expression Module

Results from the differential expression analysis are presented separately for each predictor as a table providing:

- The estimated log fold-change in expression of each gene associated with that predictor
- A 95% confidence interval for that estimate
- The p-value associated with the fold-change
- An adjusted p-value derived using either the Bonferroni correction or the Benjamin-Yekutieli FDR
- A list of the pathways to which the gene belongs

The analysis report will show results for the 20 genes with the lowest p-values, and a table of full results is written as a *.CSV file in the results directory. **TABLE 2** shows the results from two genes in the comparison of Subtype D vs. Normal. The log fold change column gives the estimated differences in gene expression (measured on the $\log_2$ scale) between Subtype D samples and samples in the reference category, Normal. To convert these numbers into a fold change in linear space, raise 2 to the power of the log fold-change (e.g., $2^{4.31} = 19.83$, so ERBB2 is estimated to be 20-fold higher in Subtype D samples than in Normal samples. Similarly, $2^{-0.29} = 0.82$, so JAK1 is 18% lower in Subtype D samples.

Log fold change values have a slightly different interpretation for continuous variables. If **TABLE 2** gave the results for Binding Density, one could conclude, "A unit increase in Binding Density is associated with a 0.29 decrease in $\log_2$ expression of JAK1, holding Subtype constant."

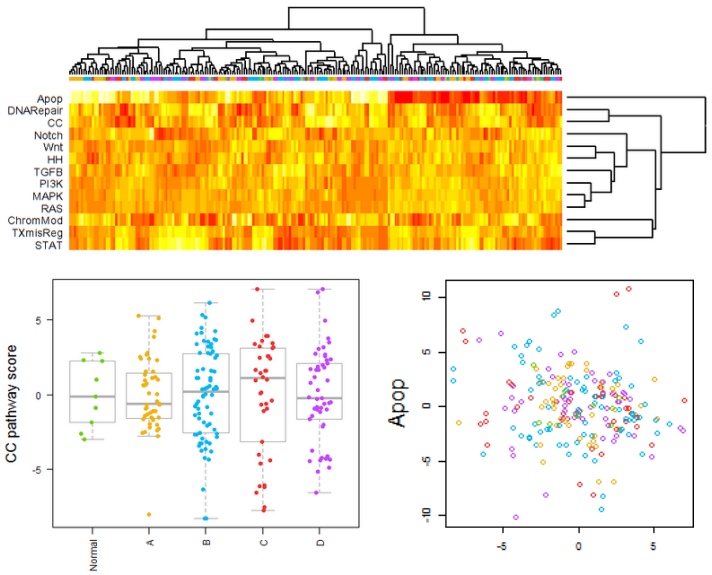


FIGURE 19 Pathway scoring results after adjusting for Subtype show fewer obvious patterns.

TABLE 2 Results for two genes' differential expression in subtype D vs. Normal samples. The results for JAK1 are correctly interpreted as follows: "Subtype D is associated with a 0.29 decrease in $\log_2$ expression of JAK1 relative to normal samples, holding the value of binding density constant. The data are consistent with a true decrease between 0.55 and 0.036. This association is statistically significant, with p = 0.026, although 32% of genes with similarly strong evidence will be false discoveries."

	Log Fold Change	Lower Confidence Limit	Upper Confidence Limit	p-Value	FDR	Pathways
ERBB2	4.31	3.56	5.06	2.84E-23	1.36E-20	0
JAK1	-0.29	-0.55	-0.036	0.026	0.32	STAT
...	...	...	...	...	...	...

Thus for continuous variables, the fold change must be read in the context of the range of the variable. Binding Density has a small range, so 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 1 mg of the drug has a small effect.

Differential expression analyses are often summarized using "volcano plots" in which the $-\log_{10}$ p-value of each gene is plotted against its log fold change. 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).

The PanCancer Pathways Analysis Module draws a volcano plot for each variable in the regression analysis. **FIGURE 20** shows example results for the comparison of Subtype D vs. Normal. Highly statistically significant genes are denoted by color, and the 40 most significant genes are named. One of the most impressive genes determined by both p-value and differential expression is ERBB2, which encodes HER2. ERBB2 has a p-value of roughly 10^{-20}, and it is up-regulated by roughly 2^4 (16-fold) relative to normal samples. Similarly, KIT is highly statistically significantly down-regulated in subtype D samples. Because ERBB2 is upregulated, but KIT is downregulated, they are located on opposite sides of the volcano plot.

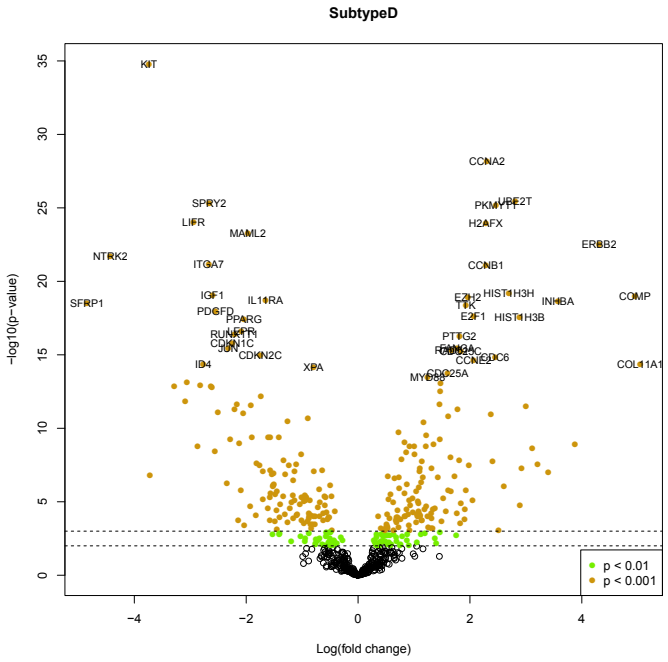


FIGURE 20 Volcano plot of the comparison between Subtype D and Normal samples.

Gene Set Analysis Module

Differential expression results at the individual gene level are important, but interpreting results from 730 genes is difficult. It is useful to first examine differential expression at the pathway level to gain a sense of which biological processes have the most profound and pervasive differential expression.

The PanCancer Pathways Analysis Module summarizes differential expression at the pathway level using two statistics: the "global significance statistic" and the "directed global significance statistic." Similar to how pathway scores condense the expression data from 730 genes into 13 pathway scores, global significance scores condense the differential expression results from 730 genes into 13 pathway-level measurements of differential expression. Note that pathway scores and global significance statistics rely on unrelated mathematics and have entirely distinct interpretations. These simple statistics are well-suited to PanCancer Pathways Panel data and serve as alternatives to gene set analysis methods designed for microarray data such as GSEA (Subramanian et al., 2005). They are calculated from the t-statistics of pathway genes, which are calculated from linear regressions run in the differential expression analysis. Global Significance Statistics are calculated separately for each variable in the regression.

The global significance statistic measures the cumulative evidence for the differential expression of genes in a pathway. For each covariate, it is calculated as the square root of the pathway's average squared t-statistic:

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

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

The directed global significance statistic is similar in spirit 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. For each covariate, it is calculated as the square root of the average signed squared t-statistic:

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

$$\text{where } U = \frac{1}{p}\sum_{i=1}^{p} sign(t)_i \cdot t_i^2$$

and where $sign(U)$ is -1 if U is negative or 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.

FIGURE 21 shows heat maps of the global and directed global significance statistics. The heat map of global scores on the left shows that the Cell Cycle pathway in Subtype C samples is by far the most enriched with differentially expressed genes, and it shows that Subtypes B and C have the most extensive differential expression; DNA repair and Cell Cycle genes are particularly differentially expressed. The heat map of directed global significance scores on the right shows that Cell Cycle genes are highly overexpressed in tumors compared to the Normal samples, while STAT, RAS and Notch genes are under-expressed in cancer, particularly in Subtype B.

For each pathway, the volcano plot from the differential expression analysis is redrawn with the genes from that pathway highlighted (**FIGURE 22**). The example suggests an overwhelming tendency of Cell Cycle genes to be up-regulated in Subtype C samples.

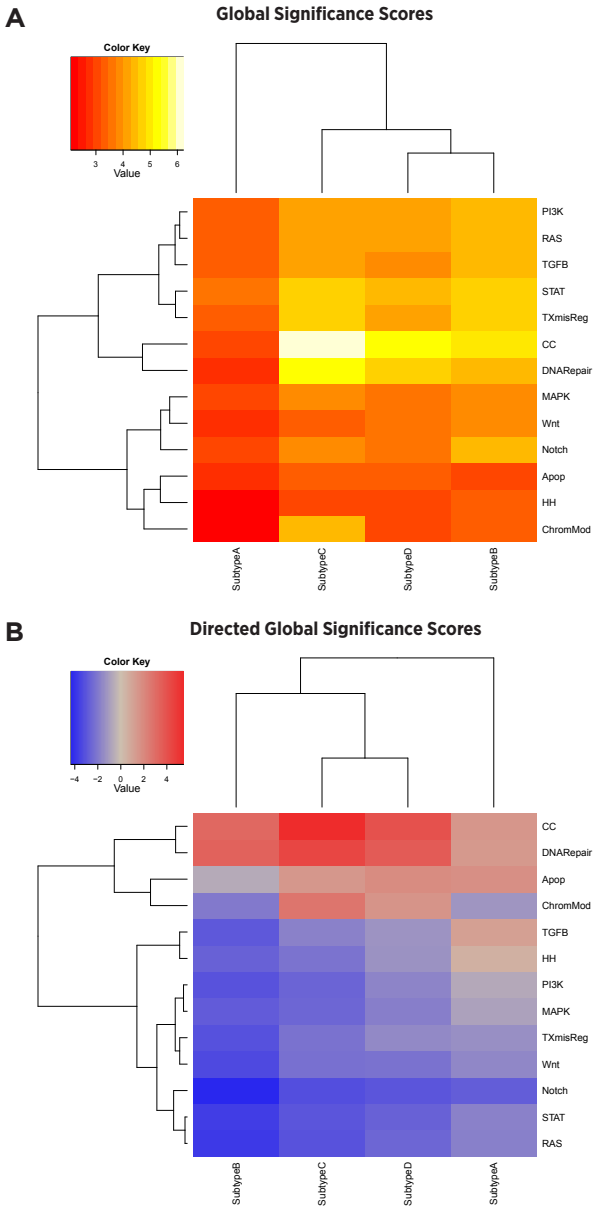


FIGURE 21 Global significance statistics and directed global significance statistics plotted for each subtype in each pathway. High global significance statistics indicate extensive differential expression. Very high or low directed global significance statistics indicate extensive up- or down-regulation, respectively.

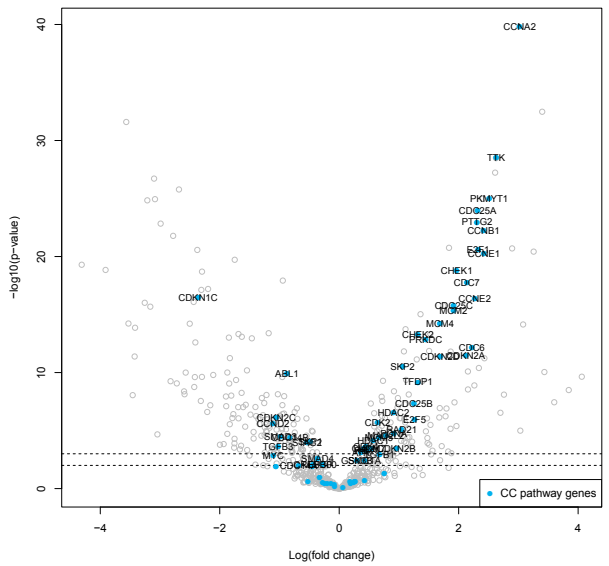


FIGURE 22 This volcano plot for the Cell Cycle pathway in the comparison of Subtype C to Normal samples shows the same results as the Subtype C volcano plot in the Differential Expression module (see **FIGURE 20**), but here the genes in the Cell Cycle pathway have been highlighted.

Pathview Plots Module

FIGURE 23 illustrates a Pathview plot of differential expression between Subtype D and Normal samples in the TGF-Beta pathway for this example dataset. Each node represents a protein family and may correspond to multiple genes, in which case the node is colored by the average fold-changes or t-statistics of its genes. Some biological results were expected, such as up-regulated p15 expression accompanying lower expression of its inhibitor, c-Myc. Other results were unexpected, such as the low expression of ActivinRII despite very high expression of its activator, Activin. Biologically unexpected results like these may indicate breakdowns in signaling pathways. However, careful interpretation is required: a relationship between proteins displayed in a KEGG graph may not apply at the level of their mRNA transcripts.

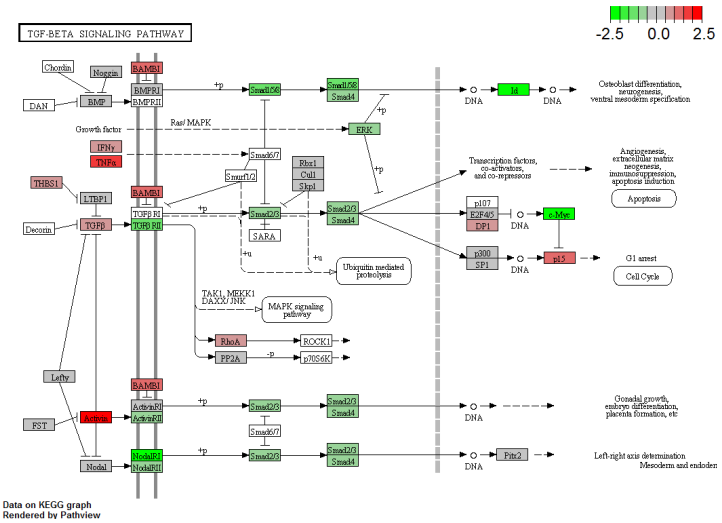


FIGURE 23 Pathview plot of differential expression between subtype D and Normal samples in the TGF-Beta pathway. Green nodes indicate down-regulated genes, red nodes indicate up-regulated genes, and gray nodes do not meet the p-value threshold for coloring. Nodes in white are not represented in the PanCancer Pathways Panel.

Pathway Significance Plots

The PanCancer Pathways Analysis Module employs an additional technique for summarizing pathway-level behavior. For each of the pathway scores created in the Pathway Scoring analysis, it performs a differential expression analysis using the same regression model as in the gene-level differential expression analysis. These regressions are used to calculate a p-value for the association of each pathway with each variable. These pathway score p-values represent an entirely different approach to summarizing pathway behavior than global significance statistics.

FIGURE 24 shows the highest-level view of cancer panel data. For the comparisons of Subtype B vs. Normal samples and Subtype C vs. Normal samples, the global significance score of each pathway is plotted against the $-\log_{10}$(pathway score p-value) for its association with subtype difference. The Cell Cycle pathway is the most powerfully deregulated in both subtypes under both metrics.

In Subtype C samples, the global significance score and pathway score p-value agree on the identities of the pathways with the greatest expression changes, whereas in Subtype B samples the extent of differential expression at the gene level does not agree with the extent of association between pathway score and Subtype B. In general, agreements between the two pathway metrics can strongly suggest the most biologically relevant pathways. Disagreements may hold other biologically interesting interpretations. For example, the STAT pathway is one of the most differentially expressed between Subtype B and Normal samples, but its pathway score is very weakly associated with the Subtype B vs. Normal contrast. These results could suggest that the Cell Cycle pathway score captures a continuum from normal to deregulated expression (and that Subtype B has a highly deregulated Cell Cycle pathway), while the STAT pathway score reflects an entirely different biological phenomenon.

Discussion

The tools in the PanCancer Pathways Analysis Module enable in-depth, end-to-end, pathway-centered exploration of panel data that enables researchers to analyze, visualize, and hypothesize follow-on studies. These modules include plots to encourage best-practice data QC, convenient implementation of a complex normalization technique, an easy approach to statistically principled differential expression analysis, and advanced academic methods for pathway-level analyses that thus far have only been accessible to R programmers. However, these automated tools do not replace expert analysts. Every dataset and scientific question will have its own optimal set of analyses that the tools provided here can only approximate. In particular, interpretation of the pathway scoring methods remains incredibly nuanced.

The analysis report is intentionally non-linear. Users may explore their results in whatever order they choose. Though many will want to first examine high-level results to confirm that their data is interesting, others will want to start with the data QC to confirm the results are not spurious.

Analysis techniques described in this tech note will be useful for understanding your data and for planning follow-on experiments. They will point to the most interesting genes and pathways, and they will detail the relationship between biological variables and the behavior of genes and pathways. Additionally, many of the analyses were built to return results suitable for publication. The differential expression analysis module uses standard methods that should be

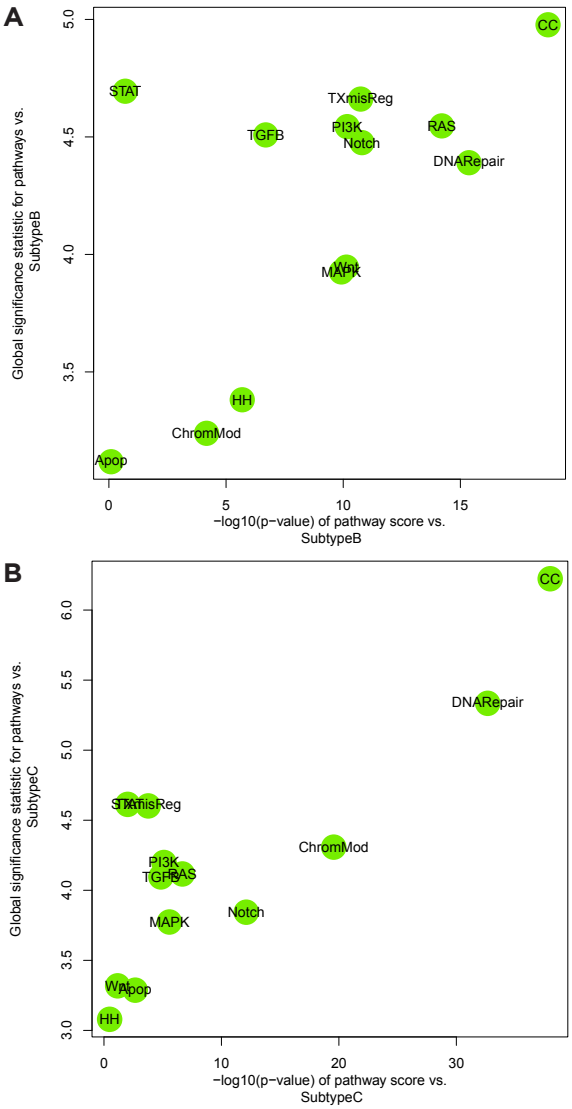


FIGURE 24 Global significance scores, and linear associations of pathway scores are plotted for each pathway. (A) Subtype B vs. Normal. (B) Subtype C vs. Normal.

familiar to reviewers. Global significance statistics used by nSolver are not a standard method, but they are simple and statistically principled enough that they could be included in a publication with a short methodological description. Pathway scores require more careful interpretation, ideally supported by ample evidence in the data, but numerous publications also make use of principal components or Pathifier-derived pathway scores (Taherian-Fard et al., 2014).

The opportunity for error with any statistical method tends to increase with its power and complexity, and the analyses provided by the PanCancer Pathways Analysis Module all have potential for misuse. A list of potential pitfalls follows:

- Study design: Failing to balance or randomize the biological variables over the technical variables (e.g. running all the tumor samples on one cartridge with one hybridization time and running all the normal samples on another cartridge with a different hybridization time).
- Normalization: Including housekeeping genes that vary with a covariate of interest.

- Normalization: Performing the advanced analysis on raw data without selecting the geNorm option
- Low signal genes: Filtering out too many genes, or filtering too few and having signal dominated by RNA input.
- Confounding variables: Failing to annotate important covariates or failing to adjust for them in differential expression analyses and pathway scoring.
- Pathway scores: Misinterpretation of or excessive confidence in pathway scores.

For assistance when installing and running nSolver advanced analyses, please contact nSolver support (**nsolversupport@nanostring.com**). For questions on data analysis options and interpretation, consult an expert at your institution.

References

1. Dennis L et al. (2013) Multiplexed cancer pathway analysis: nCounter PanCancer Pathways Panel for gene expression. *NanoString.com*
2. Drier Y, Sheffer M, Domany E (2013) Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci USA* 110(16):6388-6393.
3. Hastie T, Stuetzle W (1989) Principal curves. *J Am Stat Assoc* 84(406):502-516.
4. Luo W, Brouwer C (2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29(14):1830-1831.
5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102(43):15545-15550.
6. Taherian-Fard A, Srihari S, Ragan MA (2014) Breast cancer classification: linking molecular mechanisms to disease prognosis. *Brief Bioinform* 16(3):451-474.
7. Tian F, Wang Y, Seiler M, Hu Z (2014) Functional characterization of breast cancer using pathway profiles. *BMC Med Genomics* 7(45).
8. Tomfohr J, Lu J, Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6:225.
9. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3(7):research0034.
10. Vogelstein B, Papdopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW (2013) Cancer genome landscapes. *Science* 339(6127):1546-1558.

NanoString Technologies, Inc.
530 Fairview Ave N
Suite 2000
Seattle, Washington 98109 USA

LEARN MORE

Visit **www.nanostring.com/nsolver** to learn more about the nSolver 2.5 Analysis Software.

Toll-free: +1 888 358 6266 | Fax: +1 206 378 6288
www.nanostring.com | **info@nanostring.com**

SALES CONTACTS

United States: **us.sales@nanostring.com**
EMEA: **europe.sales@nanostring.com**
Asia Pacific & Japan: **apac.sales@nanostring.com**
Other Regions: **info@nanostring.com**

AUGUST 2015 LBL-10210-01