QIAGEN® Ingenuity® Pathway Analysis

Project:	Sample Project XXXXX
Customer:	Dr. Sample Owner
Company/Institute:	QIAGEN
Date:	Thursday, December 5, 2019

Performed by:

QIAGEN Genomic Services

Genomic.Services@qiagen.com

QIAGEN.com/GenomicServices

Analysis reference: XXXXX

Sample to Insight

Contents

QIAGEN Ingenuity Pathway Analysis

QIAGEN Ingenuity Pathway Analysis (QIAGEN IPA) software utilizes extensive records maintained in the Ingenuity Knowledge Base to identify the changes in pathways, upstream regulators, and diseases and functions using the differentially expressed genes between samples.

Input dataset

QIAGEN IPA core analyses were performed with significantly differentially expressed genes. The input table describing the result of a differential gene expression analysis was imported with the following information:

- Gene ID
- Mean expression among all samples
- Log_2 fold change ratio of the gene expression between 2 groups
- *P*-value/FDR of the differential-expression test between the 2 groups describing the significance of the effect size

Three gene filtering criteria were applied:

- Mean expression ≥10 TPM
- FDR ≤0.01
- Absolute log fold change ratio ≥2

After filtering with the above criteria, 1259 genes (789 down regulated and 470 up regulated) were used for further analysis.

QIAGEN IPA core analyses

Ingenuity maintains databases on interactions between genes and other molecules, categorized by pathways and functions. In QIAGEN IPA core analyses, the enrichment or depletion of a pathway or function is identified by the number of significantly differentially expressed genes within a pathway or function:

- **Canonical pathways analysis**: Ingenuity maintains curated metabolic (302) and signaling pathways (390). Canonical pathway analysis predicts the changes in pathways based on the observed expression changes of genes within each pathway.

- **Upstream regulators analysis**: Causal effects of upstream regulator proteins (including miRNA, chemicals, drugs, etc.) and their downstream genes deriving from scientific literature are compiled in the Ingenuity Knowledge Base. Upstream regulator analysis uses the observed expression changes of the downstream genes in the dataset and predicts the changes of the upstream regulators.
- **Diseases-and-functions analysis**: Causal effects between genes and functions (i.e., downstream effects, such as cellular processes and biological functions) deriving from literature are compiled in the Ingenuity Knowledge Base. Diseases-and-functions analysis predicts the downstream functions that are expected to change given the observed gene expression changes in the dataset.

QIAGEN IPA comparison analyses

The comparison analysis ranks experimental core analyses results to determine the similarities and differences in canonical pathways (CP), upstream regulators (UR), and diseases and biological functions (DE) among other experimental observations analyzed previously with QIAGEN IPA.

Statistical analyses

All statistical tests were performed using Fisher's exact test to assess which biological attributes, such as a pathway or biological function, were significantly associated with the genes in the dataset after filtering (analysis-ready genes).

The Fisher's exact test essentially tests whether the overlap between 2 gene sets – the analysis-ready genes and the genes within a biological attribute – is statistically significant. The significance of the overlap is evaluated by *p*-values, which indicate the probabilities of getting the observed overlap if the pathway is neither enriched nor depleted, i.e., under the null hypothesis.

Here, we set a significance level of $p<0.05$. The probability of falsely calling an overlap significant increases with the number of tests. Since multiple pathways are tested for each analysis, the Benjamini-Hochberg (B-H) method is used to control the false discovery rate (FDR).

Neither *p*-value nor FDR take into account whether the genes are up regulated or down regulated in the dataset. Instead, the z-score activation predictions are provided to indicate the match between the expected directional relationship and the observed gene expression in the dataset. Z-scores greater than 2 or less than –2 should be considered as significantly activated or inhibited, respectively.

Glossary of Terms

Activation z-score	A summary value that predicts the activation (positive value) or inhibition (negative value) of an canonical pathway, upstream regulator or downstream function based on the gene expression changes within the network. Null value is shown when there is no clear directionality in the causal effect between the genes and the pathways annotated in the Ingenuity database.
Analysis-ready genes	Genes in the dataset after filtering
B-H *p*-value/FDR	Benjamini-Hochberg (B-H) method corrected *p*-value
Number of molecules	The number of analysis-ready genes that belongs to a disease, function, or tox function category
Predicted activation status	An activation z-score greater than 2 and less than –2 should be considered as significantly activated or inhibited
***P*-value of overlap**	*P*-value from Fisher's exact test that measures the significance of overlap between analysis-ready genes in the dataset and genes within a pathway or function
Ratio/Overlap	The ratio of analysis-ready genes in the dataset associated with a pathway or function

Canonical Pathways

The canonical pathway analysis identifies QIAGEN IPA pathways that were significantly enriched in the dataset. The significance of the association between the dataset and a canonical pathway was determined from a *p*-value of overlap calculated using a right-tailed Fisher's exact test. The outputs from canonical pathways analysis include:

- Bar charts of significantly-enriched QIAGEN IPA canonical pathways
- Tables listing all tested QIAGEN IPA canonical pathways

Bar charts of significantly-enriched QIAGEN IPA canonical pathways ($p \leq 0.05$ or B-H $p \leq 0.05$)

Each bar represents a canonical pathway's p-value on a negative logarithmic scale, such that the taller bars are more significant than the shorter bars. Bars are shaded according to their z-score activity predictions.

- **Orange**: pathways with positive z-scores.
- **Blue**: pathways with negative z-scores.

- **White**: pathways that have z-score of 0, indicating that the differential gene expression data did not allow for a clear determination of the activity prediction; in other words, the weight of the evidence for a prediction of activation is equal to that of inhibition.
- **Grey**: pathways for which no activity predictions can currently be made due to a lack of information in the Ingenuity Knowledge Base.
- **Orange dotted line**: the number of analysis-ready genes that are members of the pathway, divided by the total number of genes in the reference dataset that make up that pathway. The x-axis at the bottom represents the range of the ratio values. The *p*-value significance threshold of 0.05 is represented by an orange threshold line (–log=1.3).

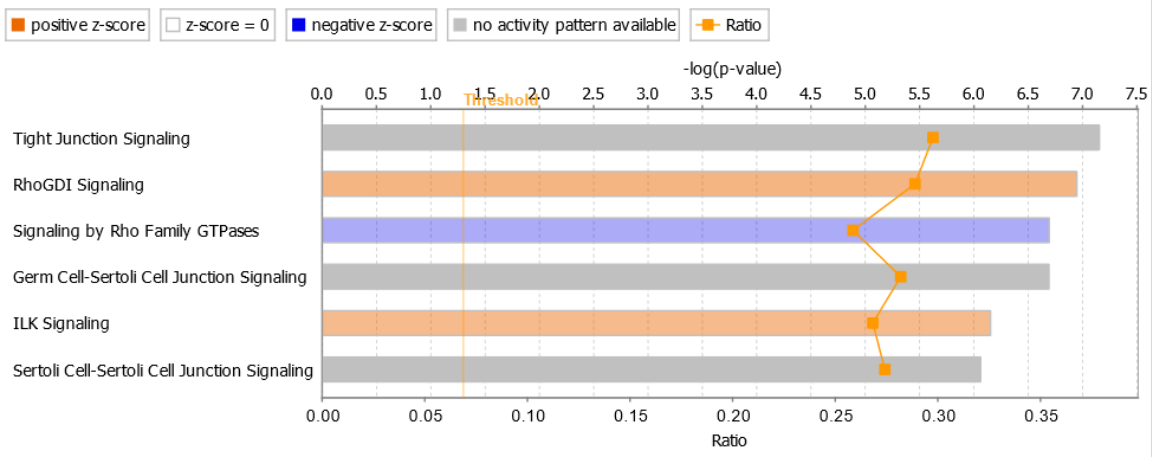

Figure 1. Bar chart of significantly-enriched QIAGEN IPA canonical pathways ($p \leq 0.05$)

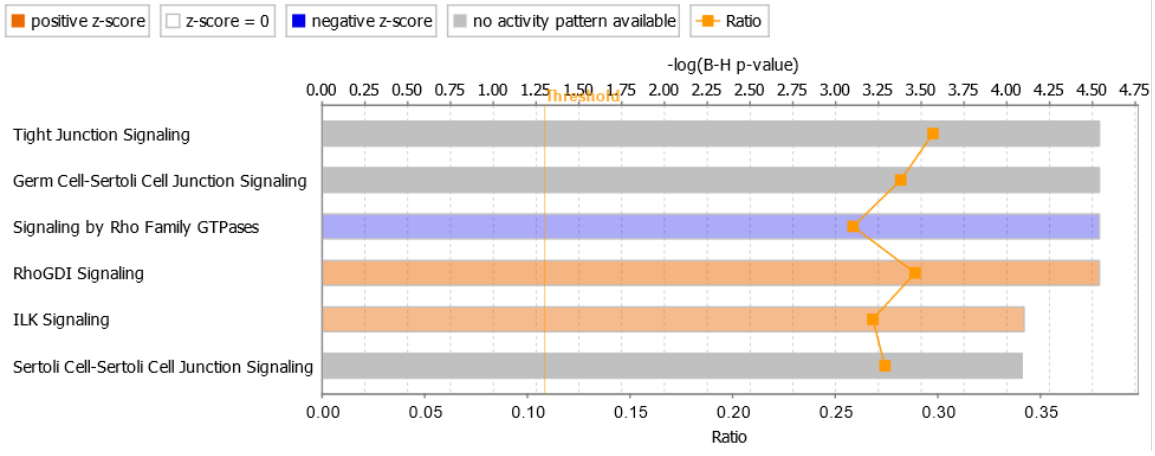

Figure 2. Bar chart of significantly enriched QIAGEN IPA canonical pathways (B-H $p \leq 0.05$).

Tables listing all tested QIAGEN IPA canonical pathways (including the ones that did not pass the *p*-value significance threshold)

The tables contain the following columns:

- **Ingenuity canonical pathways**: The pathways annotated in QIAGEN IPA.
- **–log(*p*-value)**: The *p*-value or Benjamini-Hochberg (B-H) method corrected *p*-value indicates the statistical significance of the overlap of analysis-ready genes that are within the pathway.
- **Ratio**: The ratio is calculated as the number of analysis-ready genes in a given pathway, divided by the total number of genes in the reference dataset that make up that pathway.
- **Z-score**: The activation z-score predicts the activation state of the QIAGEN IPA canonical pathway, using the gene expression patterns of the genes within the pathway. An absolute z-score ≥2 is considered significant.
- **Molecules**: Genes in the dataset that are within the given pathway.

Table 1. All tested QIAGEN IPA canonical pathways showing *p*-values (including ones that did not pass the *p*-value threshold)

Ingenuity canonical pathways	–log(*p*-value)	Ratio	Z-score	Molecules
Tight junction signaling	7.15	0.298	N/A	ACTA1, ACTB, ACTC1,…
RhoGDI signaling	6.95	0.289	1.206	CD44, CDC42, CDH1,…
Signaling by Rho family GTPases	6.69	0.259	–1.05	CDH26, CDH7, CFL2,…
Germ cell–Sertoli cell junction signaling	6.69	0.282	N/A	ACTN4, AKT1, CDC42,…
ILK signaling	6.15	0.268	1.109	RSU1, SNAI2, TGFB1I1,…

N/A: not applicable.

Table 2. All tested QIAGEN IPA canonical pathways showing B-H corrected *p*-values (including ones that did not pass the B-H corrected *p*-value threshold)

Ingenuity canonical pathways	–log (B-H *p*-value)	Ratio	z-score	Molecules
Tight junction signaling	4.54	0.298	N/A	ACTA1, ACTB, ACTC1,…
Rho GDI signaling	4.54	0.289	1.206	CD44, CDC42, CDH1,…
Germ cell–Sertoli cell junction signaling	4.54	0.282	N/A	ACTN4, AKT1, CDC42,…
Signaling by Rho Family GTPases	4.54	0.259	–1.05	CDH26, CDH7, CFL2,…
ILK Signaling	4.1	0.268	1.109	RSU1, SNAI2, TGFB1I1,…

N/A: not applicable.

Upstream Regulator Analysis

Upstream regulator analysis identifies upstream regulators that may be responsible for the gene expression changes observed in the dataset based on information from the Ingenuity Knowledge Base. These predictions are based on:

- ***P*-values of overlaps**: The *p*-value indicates the statistical significance of the overlap between the analysis-ready genes and the downstream genes targeted by the upstream regulators (as calculated by the right-tailed Fisher's exact test)
- **Activation z-score**: The activation z-score predicts the activation state of the upstream regulator using the gene expression patterns of the target genes. Absolute z-score values ≥2 are considered significant, with an upstream regulator being considered as activated if the z-score is greater than or equal to 2 and inhibited if the z-score is less than or equal to –2.

Output for the upstream regulator analysis include a **table listing of predicted upstream regulators** and the respective target molecules in the analyzed dataset.

The table contains the following columns:

- **Upstream regulator**: The name of the upstream regulator.
- **Expr fold change**: The fold change value of the upstream regulator extracted directly from the input dataset of the upstream regulator.
- **Molecular type**: The category of the upstream regulator.
- **Predicted activation state**: The predicted activation state of the upstream regulator based on the activation z-score.
- **Activation z-score**: The activation z-score predicts the activation state of the upstream regulator, using the gene expression patterns of the genes downstream of an upstream regulator.
- ***P*-value of overlap**: The *p*-value of overlap indicates the statistical significance of the overlap of the analysis-ready genes and genes that are downstream of the upstream regulator.
- **B-H *p*-value**: Benjamini-Hochberg–method–corrected *p*-value to account for multiple hypothesis testing.
- **Target molecules in dataset**: The genes in the dataset with changed expression that are downstream targets of the upstream regulator.

Table 3. Upstream regulator analysis

Upstream regulator	Expr fold change	Molecule type	Predicted activation state	Activation z-score	*P*-value of overlap	B-H corrected *p*-value	Target molecules in dataset
CST5	1.399	Other	Inhibited	–2.655	6.47E-09	2.36E-08	ACAT2, ANK3, ANXA2,…
Estrogen receptor	–2.14	Group	Inhibited	–3.823	9.61E-07	3.10E-07	ANXA1, AXL, CAPG,…
IL2	–2.407	Cytokine	NA	1.407	1.93E-06	4.84E-05	ABCB4, ACVR1B, AHR,…

Diseases-and-Functions (Downstream Effect) Analysis

Diseases-and-functions analysis identifies downstream effects that are expected to increase or decrease, given the observed gene expression changes in the dataset. Diseases-and-functions analysis is based on the expected causal effects, as derived from the literature compiled in the Ingenuity Knowledge Base, between genes and functions.

The analysis examines genes in the dataset that are known to affect functions and uses the expected causal effects of the genes, derived from the literature, to issue a prediction for each function based on the direction of change in gene expression. The z-score captures the direction of change as follows:

- **Increase in function**: Direction of change is consistent with the literature
- **Decrease in function**: Direction of change is inconsistent with the literature
- **No prediction**: No clear pattern related to the literature

Outputs from the diseases-and-functions analysis include:

- **Bar charts of significantly enriched diseases and biological functions (*p*-value ≤0.05 and B-H *p*-value ≤0.05)**. Each bar represents a high-level functional category as calculated from the Fisher's exact test. The *p*-values for diseases and biological functions are shown in the negative logarithmic scale, such that the longer bars are more significant than the shorter bars.

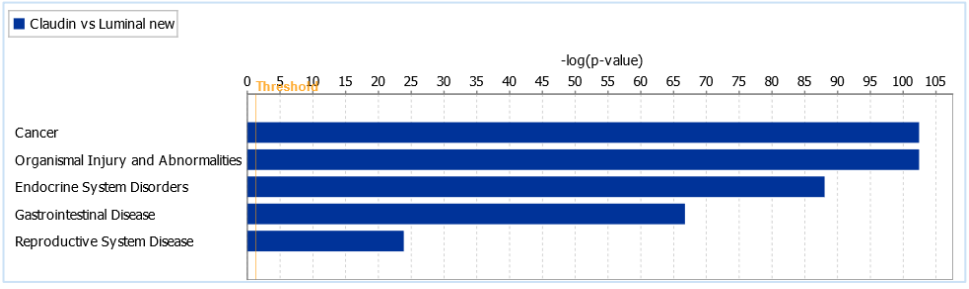

Figure 3. Bar chart of significantly-enriched QIAGEN IPA diseases and biological functions (*p*≤0.05).

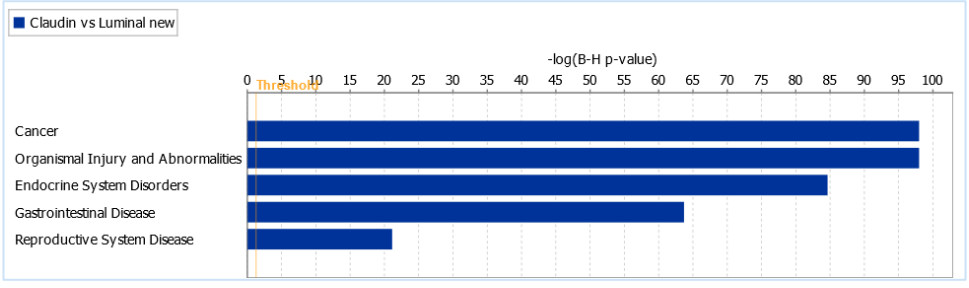

Figure 4. Bar chart of significantly-enriched QIAGEN IPA diseases and biological functions (B-H *p*≤0.05).

- Table listing of significantly enriched downstream functions from the diseases-and-functions analysis. The table contains the following columns:
  - **Category**: A high-level functional category (for example, cancer).
  - **Disease or functions annotation**: A specific function that is significantly calculated for the data (for example, growth of tumor).
  - ***P*-value**: Indicates the statistical significance of the overlap of analysis-ready genes and genes that influence the downstream function.
  - **B-H *p*-value**: Benjamini-Hochberg method corrected *p*-value to account for multiple hypothesis testing.
  - **Predicted activation state**: The predicted direction of change for the function, based on the z-score.
  - **Activation z-score**: The z-score predicts the direction of change for the function. An absolute z-score ≥2 is considered significant.
  - **Molecules**: The genes associate with each function.
  - **# molecules**: The number of genes associate with each function.

Table 4. Enriched downstream functions and annotations identified in the Diseases and Functions Analysis

Diseases or functions annotation	Categories	Function	*p*-value	B-H *p*-value	Predicted activation state	Activation z-score	Molecules	# molecules
Cancer, organismal injury, and abnormalities	Nonhematological solid tumor	Nonhematological solid tumor	3.74E-103	1E-98	Increase	2.325	AACS, ABAT, ABCA12, …	2872
Cancer, organismal injury, and abnormalities	Nonhematologic malignant neoplasm	Nonhematologic malignant neoplasm	1.7E-102	2E-98	N/A	1.085	ABCE1, ABCF1, ABCF2,…	2864
Cancer, organismal injury, and abnormalities	Neoplasia	Epithelial neoplasm	8.2E-102	8E-98	N/A	0.772	AKAP9, AKIRIN2, AKR1D1, …	2837
Cancer, endocrine system disorders, organismal injury, and abnormalities	Thyroid gland tumor	Thyroid gland tumor	9E-89	2E-85	N/A	–0.665	BPTF, BRD4, BRICD5, …	2300
Cancer, organismal injury, and abnormalities	Neck neoplasm	Neck neoplasm	2E-86	3E-83	Decrease	–3.338	DCAF8, DCBLD1, DCLK2,…	2311

N/A: not applicable.

Comparison Analyses

The comparison analysis ranks experimental core-analysis results to determine the similarities and differences in the top canonical pathways (20), upstream regulators (100), causal network (100), and diseases and functions (100) among other experimental observations analyzed previously with QIAGEN IPA.

Core-analysis data and the experimental observations are ranked and clustered with a hierarchical agglomerative clustering algorithm using average linkage and a Euclidean distance metric. The outputs for comparison analysis include:

- **Heatmaps for each comparison**. Each row is labeled by its type: canonical pathways (CP), upstream regulators (UR), diseases and functions (downstream effect; DE), and causal network (CN) with dendrograms displaying the hierarchical clustering.
    - **Heatmap of activation as calculated by z-scores**. Core-analysis data where at least one experimental observation had a significant *p*-value ≤0.05 and an absolute z-score value ≥2, are retained:
        - **Orange**: positive z-score results (greater than or equal to 2), indicating activation of CP, UR, DE, and CN.
        - **Blue**: negative z-score results (less than or equal to –2), indicating inhibition of CP, UR, DE, and CN.
        - **Gray**: insignificant z-score (absolute value ≤2).

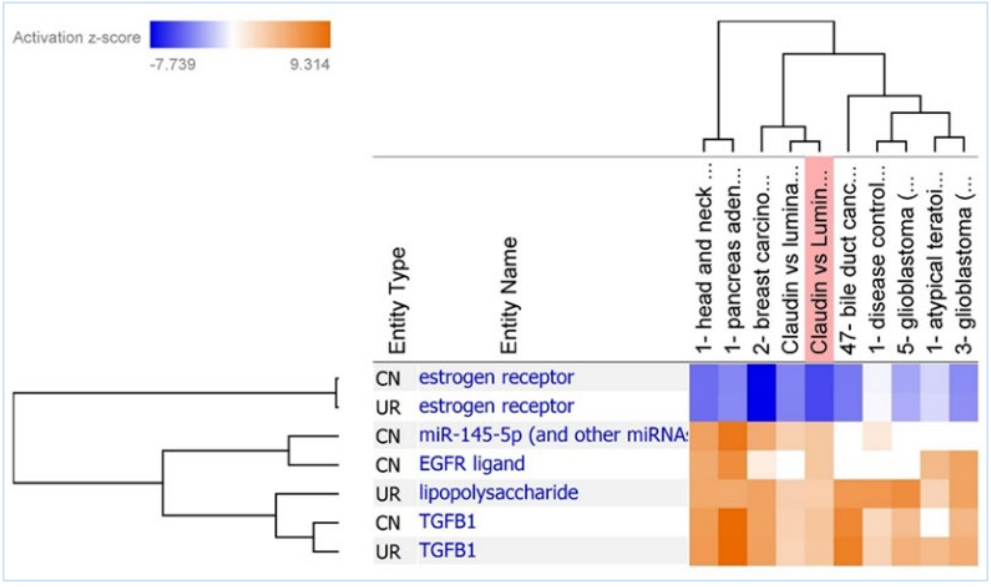

Figure 5. Heatmap of activations as calculated in z-scores.

- **Heatmap of *p*-value of overlap rankings**. Core analysis data where at least one experimental observation had a significant *p*-value ≤0.05 are retained. The *p*-value is on the negative logarithmic scale:
  - **Purple**: The most significant *p*-values (largest –log *p*-value).
  - **White**: The least significant *p*-values (smallest –log *p*-value).
  - **Gray dot**: Insignificant *p*-value threshold (–log *p*-value ≤1.3).

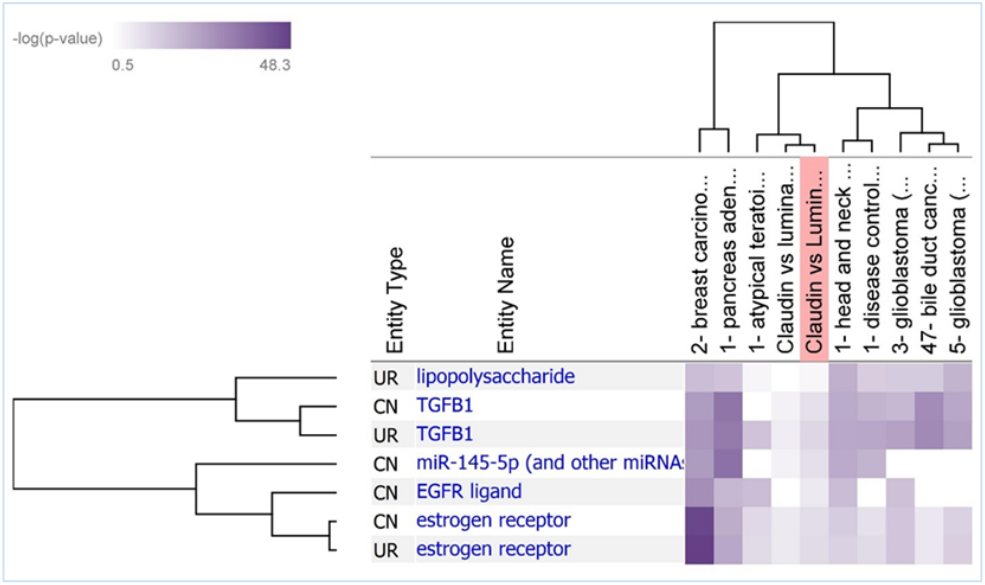

Figure 6. Heatmap of *p*-value of overlap rankings.

- **Table listing of comparison analysis z-scores and *p*-values for each matching analysis.** The table listed the comparison results between the experimental observations and other experimental observations analyzed previously with QIAGEN IPA on each row. The table contains the following columns:
  - **Analysis name**: The name of the experimental observation.
  - **Project**: The database source of the experimental observation.
  - **Case/disease state**: The disease state of the experimental observation.
  - **Case/Tissue**: The disease of the subject the sample is collected from (for example, esophagus adenocarcinoma, normal control).
  - **Comparison category**: General category for the comparison made in the model (for example, disease vs. normal, tissue 1 vs. tissue 2).
  - **Comparison contrast**: Statistical contrasts used to generate the comparisons.

- **Weblink**: A weblink to the detail of the experimental observation.
- **Z-score**: Matching z-scores are calculated for each of CP, UR, DE, and CN. For each core-analysis type, a raw z-score is calculated for each experimental comparison to capture the similarity and differences in the activation and inhibition states of the top pathways or functions. The matching z-score is then calculated by dividing the raw z-score by the maximum possible z-score, such that a matching z-score of 100 represents 100% identical effect, and a z-score of –100 represents the complete opposite effect.
- ***P*-value**: Matching *p*-value are calculated for each of CP, UR, DE, and CN. For each core-analysis type, the matching *p*-value is calculated by Fisher's exact test to determine the significance of overlapping top pathways or functions between experimental comparison.
- **Overall z-score**: The average z-score for all of CP, UR, DE, and CN.
- **Overall *p*-value**: For each of CP, UR, DE and CN, a *p*-value percentage is calculated for each experimental comparison by dividing the matching *p*-value by the maximum possible *p*-value in a negative logarithmic scale. The overall *p*-value is the average of these *p*-value percentages.

Table 5. Matching z-scores and p-values for comparison analysis

	Project 1a	Project 1b	Project 1c
Project	Example 1a	Example 1b	Example 1c
Case disease state	–	Breast carcinoma	Breast carcinoma
Case tissue	–	Breast	Breast
Comparison category	–	Treatment vs. control	Disease 1 vs. disease 2
Comparison contrast	–	Transfection → ER alpha siRNA vs control siRNA	Genetic subtype → basal B vs. luminal
Web link	Example link	Example link	Example link
CP (z-score)	–	–	–
UR (z-score)	74.83	57.74	57.74
CN (z-score)	57.45	52.92	47.96
DE (z-score)	63.25	42.43	46.9
Z-score overall score	48.88	39.27	38.15
CP (*p*-value)	0.000708946	–	–
UR (*p*-value)	1.66115E-59	1.13859E-27	1.13859E-27
CN (*p*-value)	3.81224E-45	7.61455E-36	3.15759E-27
DE (*p*-value)	7.71578E-35	2.1795E-12	2.0222E-15
P-value overall score	65.84	36.86	34.07

Notes

Notes