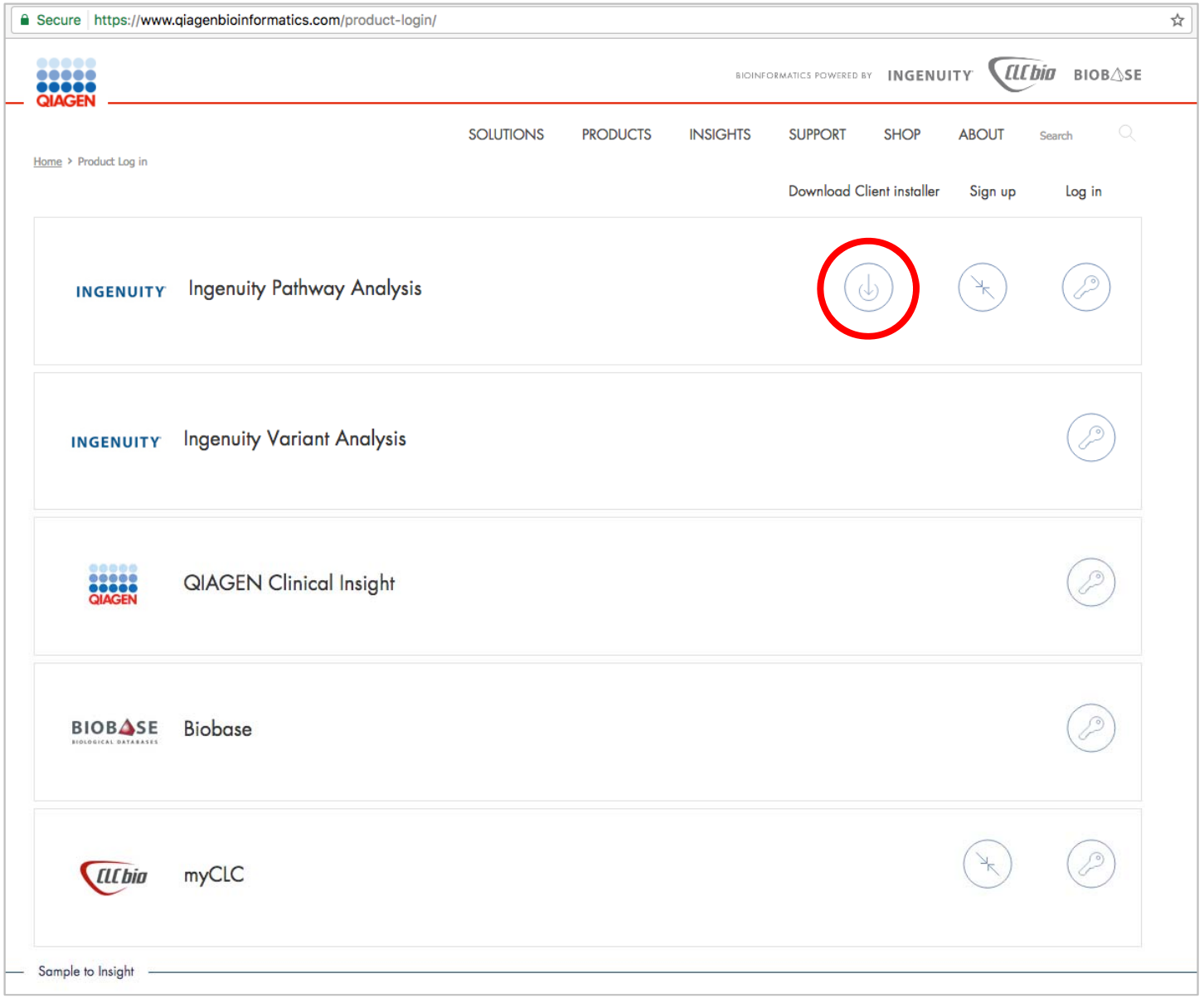
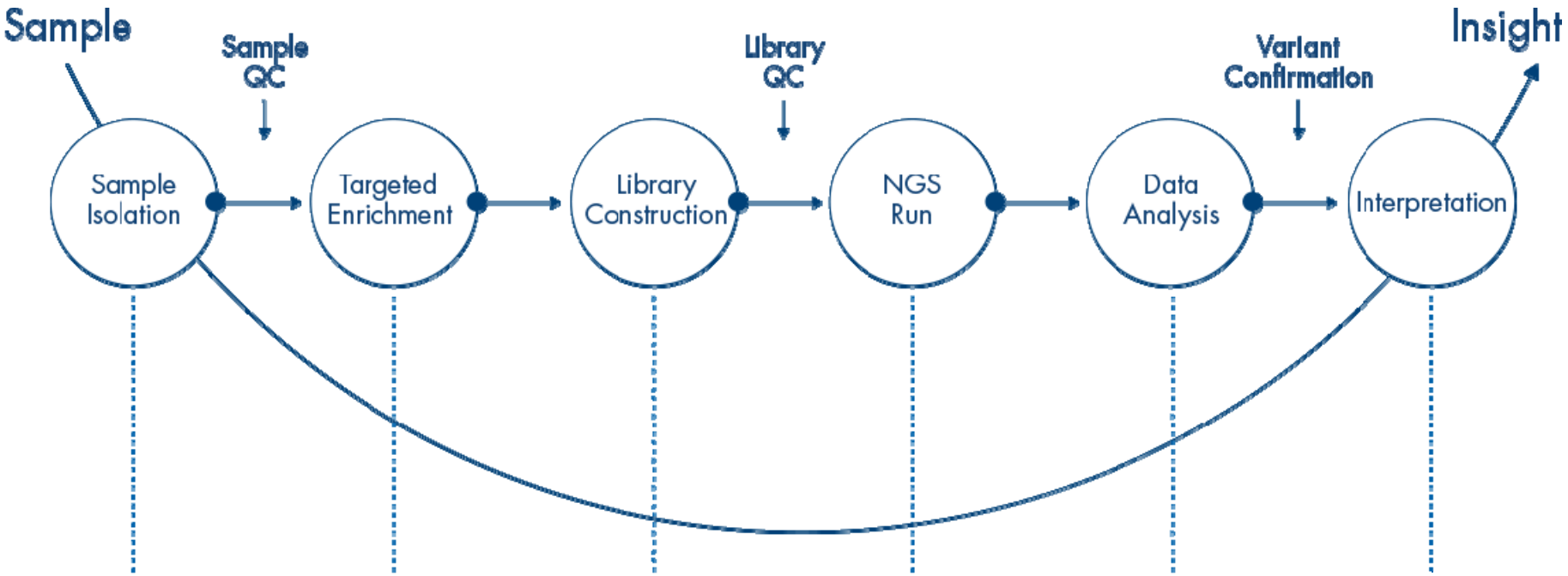
Install IPA: www.qiagenbioinformatics.com/product-login



Ingenuity Pathway Analysis (IPA) Training: Maximizing the Biological Interpretation of Gene, Transcript & Protein Expression Data with IPA

Jeff Knight, Ph.D.
Field Application Scientist, Bioinformatics
jeffrey.knight@qiagen.com

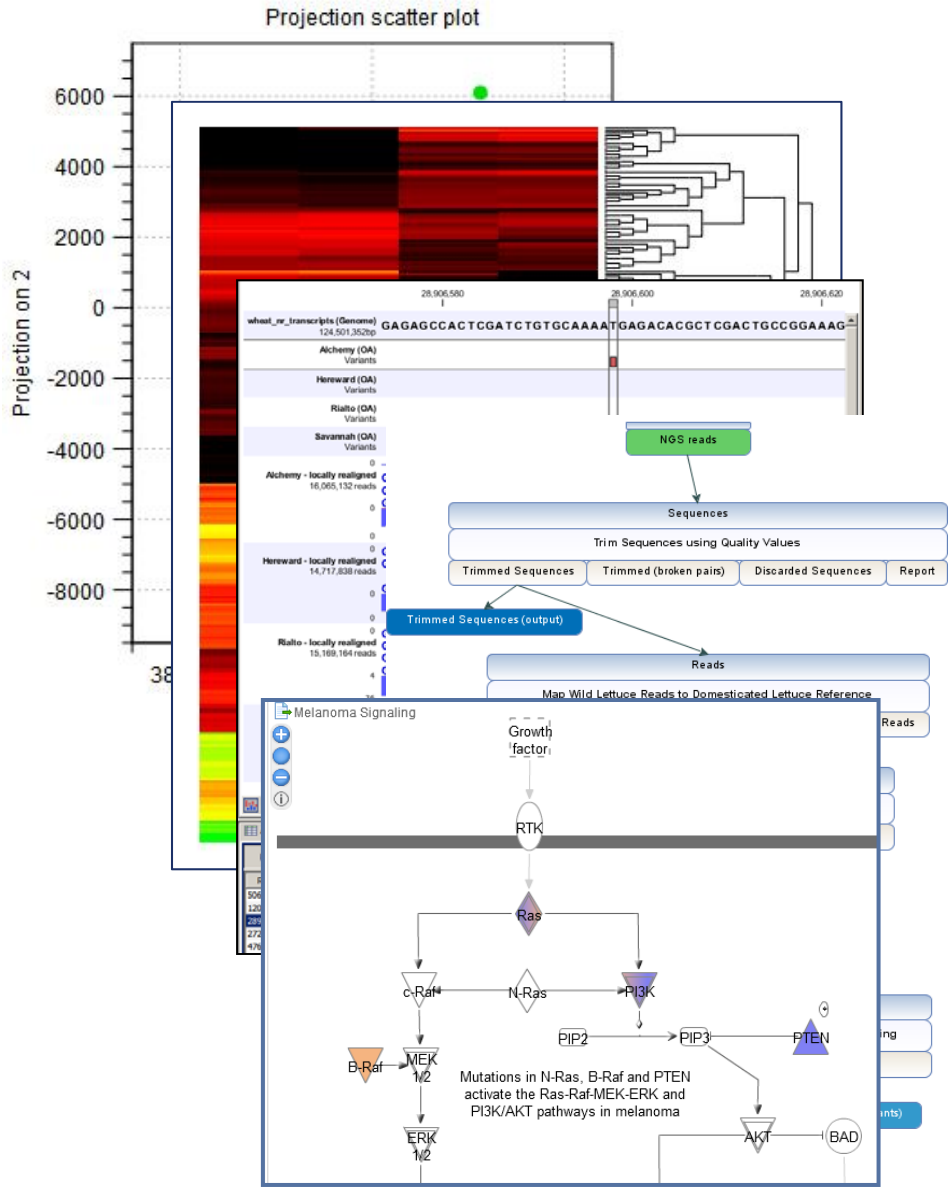
Sample to Insight



Sample Isolation	Targeted Enrichment	Library Construction	NGS Run	Data Analysis	Interpretation
GeneRead DNA FFPE Kit	GeneRead DNAseq Panel	GeneRead Library Prep Kits	Any Sequencer	Biomedical Genomics Workbench	Ingenuity Variant Analysis
QIAamp Circulating NA Kit		GeneRead Size Selection Kit		&	QIAGEN® Clinical Insight
exoRNeasy Kits		GeneRead DNAseq Library Quant Array		Biomedical Genomics Server Solution	Ingenuity Pathway Analysis
REPLI-g® Single Cell Kits					HGMD®
GeneRead DNA QuantiMIZE Kit					Inova Genomes
RNeasy Kits					

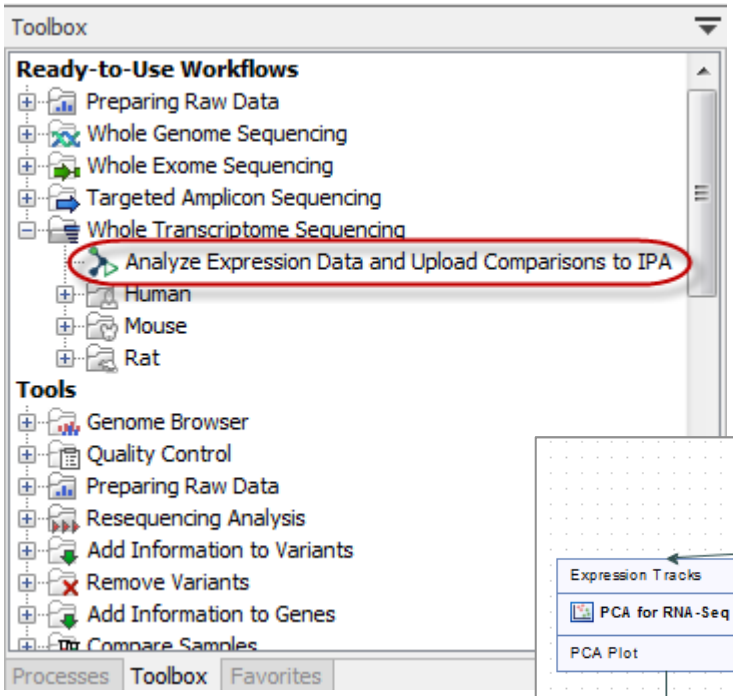
Genomics Workbench and Biomedical Genomics Workbench

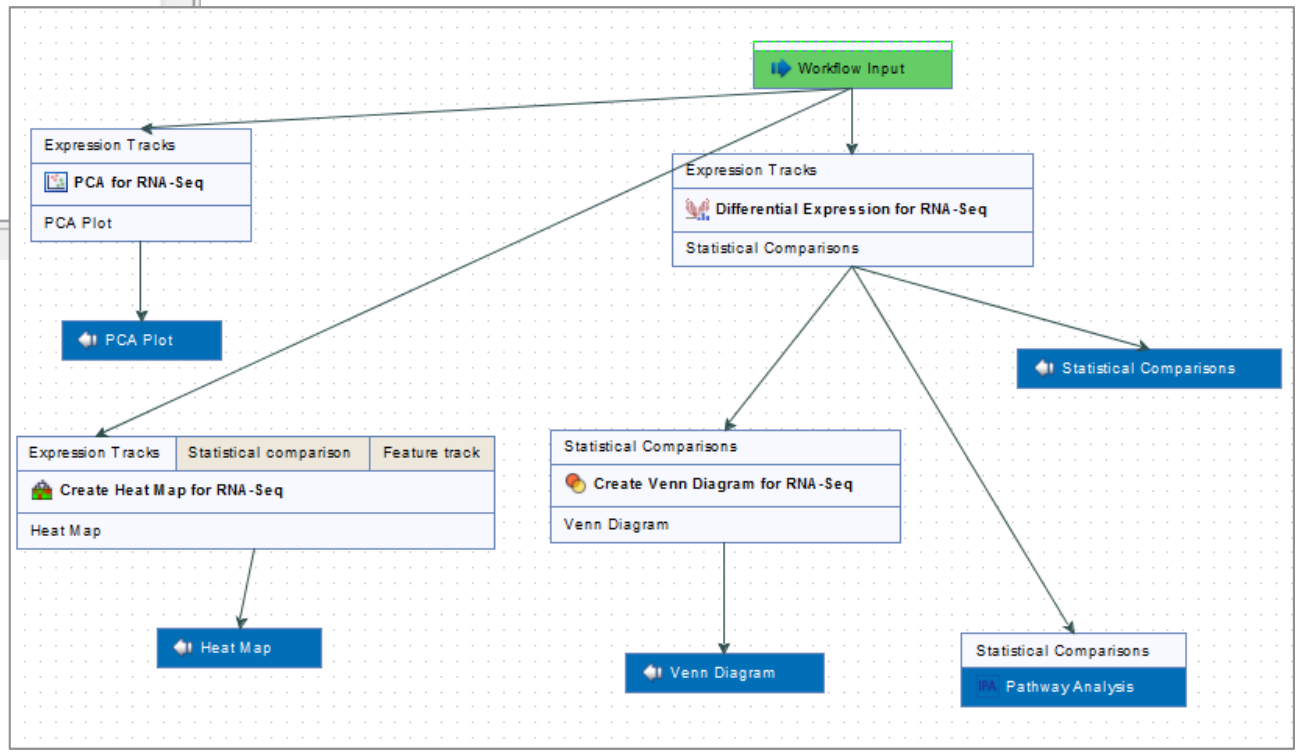
- QC and preprocess NGS data (RNA-Seq, smRNA, and DNAseq reads)
- Differential expression and statistical analysis for RNA-Seq and smRNA
- Generate, annotate, and compare high-confidence variant calls
- CNV detection
- ChIP-Seq, Bisulfite sequencing
- Genome assembly and finishing
- Microbial Metagenomics, typing
- Facilitate analysis with interactive visualization
- Construct automated workflows in user friendly interface
- Can scale to organization's needs



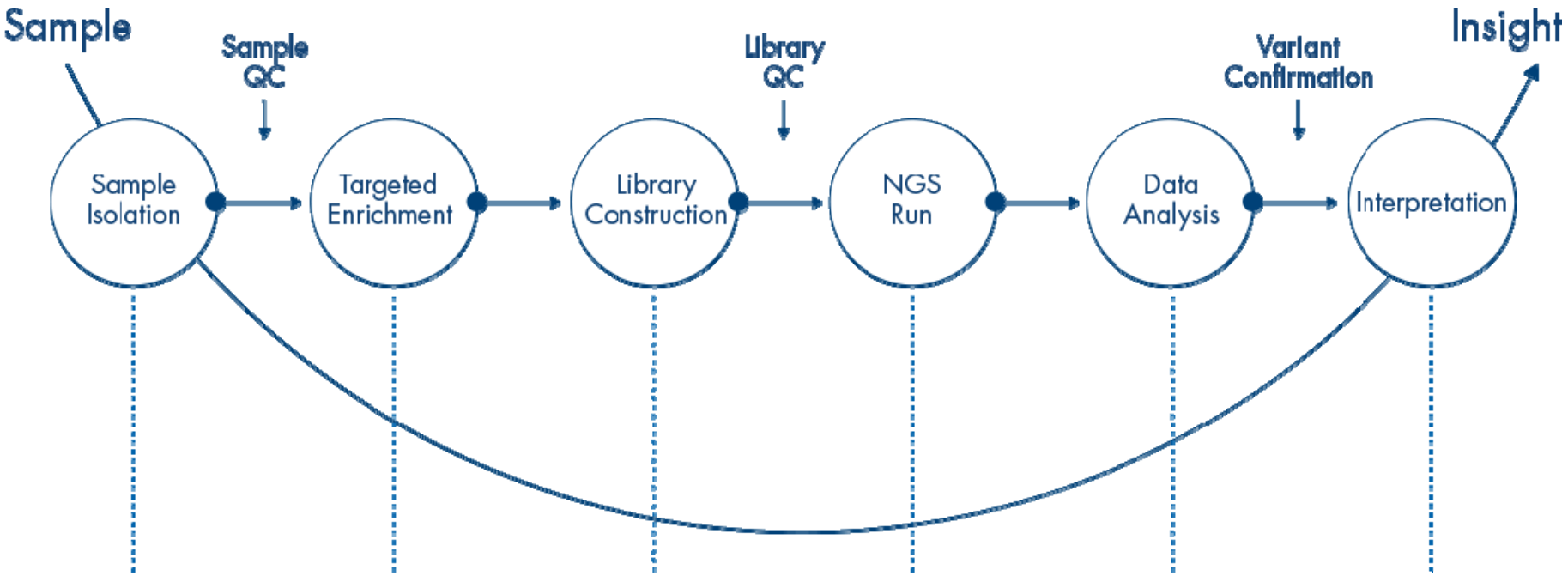
RNAseq Analysis: Identify Differentially Expressed Genes

Analyze Expression Data and Upload to IPA Workflow





Sample to Insight



Sample Isolation	Targeted Enrichment	Library Construction	NGS Run	Data Analysis	Interpretation
GeneRead DNA FFPE Kit	GeneRead DNAseq Panel	GeneRead Library Prep Kits	Any Sequencer	Biomedical Genomics Workbench	Ingenuity Variant Analysis
QIAamp Circulating NA Kit		GeneRead Size Selection Kit		&	QIAGEN® Clinical Insight
exoRNeasy Kits		GeneRead DNAseq Library Quant Array		Biomedical Genomics Server Solution	Ingenuity Pathway Analysis
REPLI-g® Single Cell Kits					HGMD®
GeneRead DNA QuantiMIZE Kit					Inova Genomes
RNeasy Kits					

How can IPA help you?

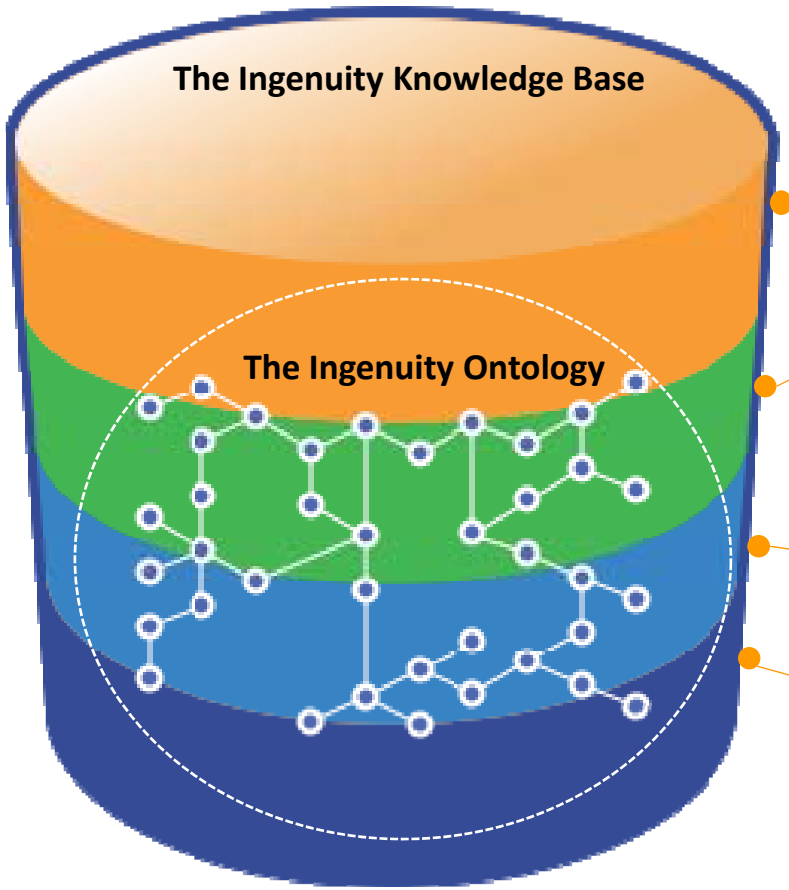
IPA

- Deep pathway understanding of a single gene/protein
  - Drug/therapeutic target discovery

- Biological understanding of large data sets, including
  - Transcriptomics – differential gene expression (array and RNAseq)
    - IsoProfiler: filter for transcript expression and annotation of interest
  - Proteomics – differential protein expression
  - Phosphoproteomics – differential protein phosphorylation
  - Genes with loss/gain-of-function variants
  - Metabolomics
  - miRNA expression
  - Methylation
  - Gene Lists
    - ChIP-Seq
    - siRNA screening

Ingenuity Content

Ingenuity Knowledge Base



Ingenuity Findings

Ingenuity® Expert Findings – Manually curated Findings that are reviewed, from the full-text, rich with contextual details, and are derived from top journals.

Ingenuity® ExpertAssist Findings – Automated text Findings that are reviewed, from abstracts, timely, and cover a broad range of publications.

Ingenuity Modeled Knowledge

Ingenuity® Expert Knowledge – Content we model such as pathways, toxicity lists, etc.

Ingenuity® Supported Third Party Information – Content areas include Protein-Protein, miRNA, biomarker, clinical trial information, and others

Gene View Pages in IPA

How can IPA help you?

IPA

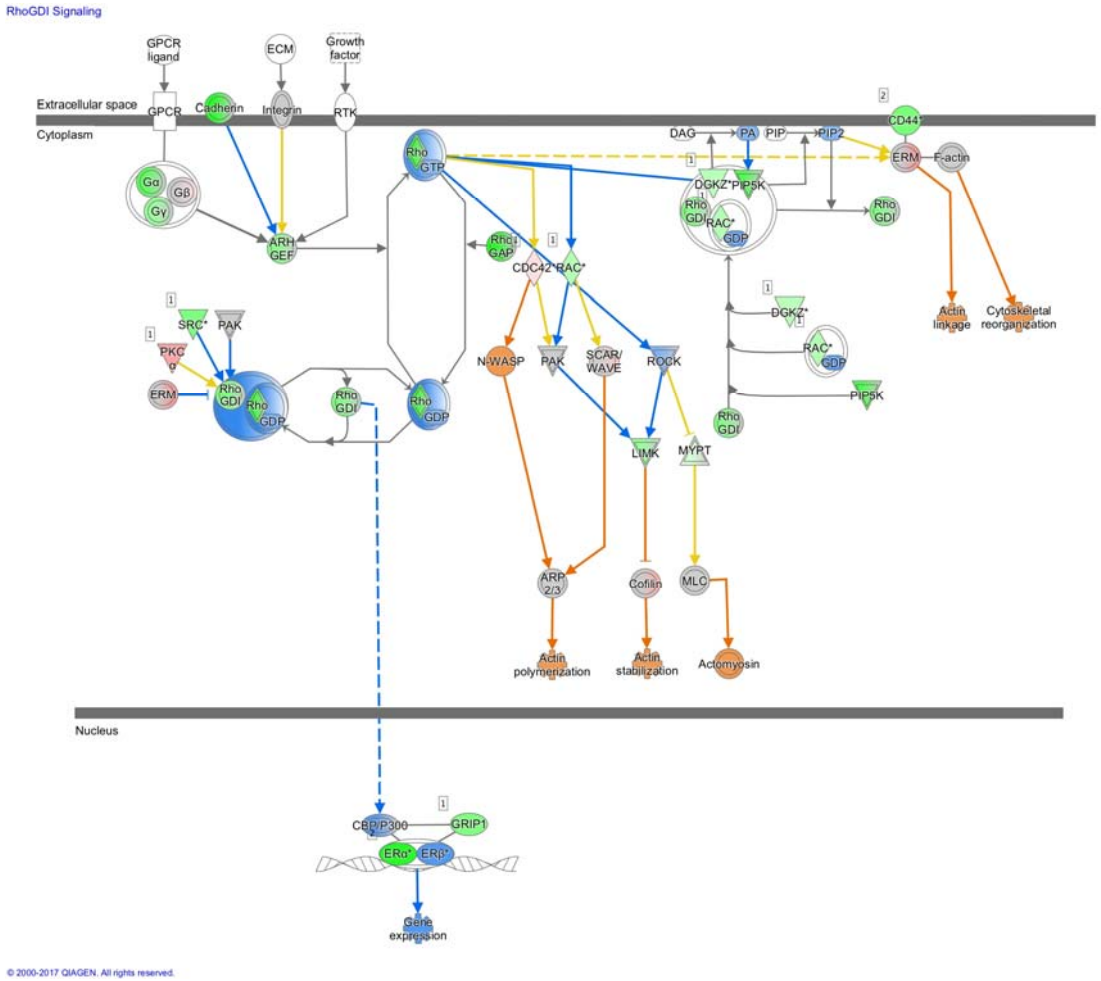
- Deep pathway understanding of a single gene/protein
  - Drug/therapeutic target discovery

- Biological understanding of large data sets, including
  - Transcriptomics – differential gene expression (array and RNAseq)
    - IsoProfiler: filter for transcript expression and annotation of interest
  - Proteomics – differential protein expression
  - Phosphoproteomics – differential protein phosphorylation
  - Genes with loss/gain-of-function variants
  - Metabolomics
  - miRNA expression
  - Methylation
  - Gene Lists
    - ChIP-Seq
    - siRNA screening

Identify key cellular pathways most likely to be affected

Canonical Pathways Analysis

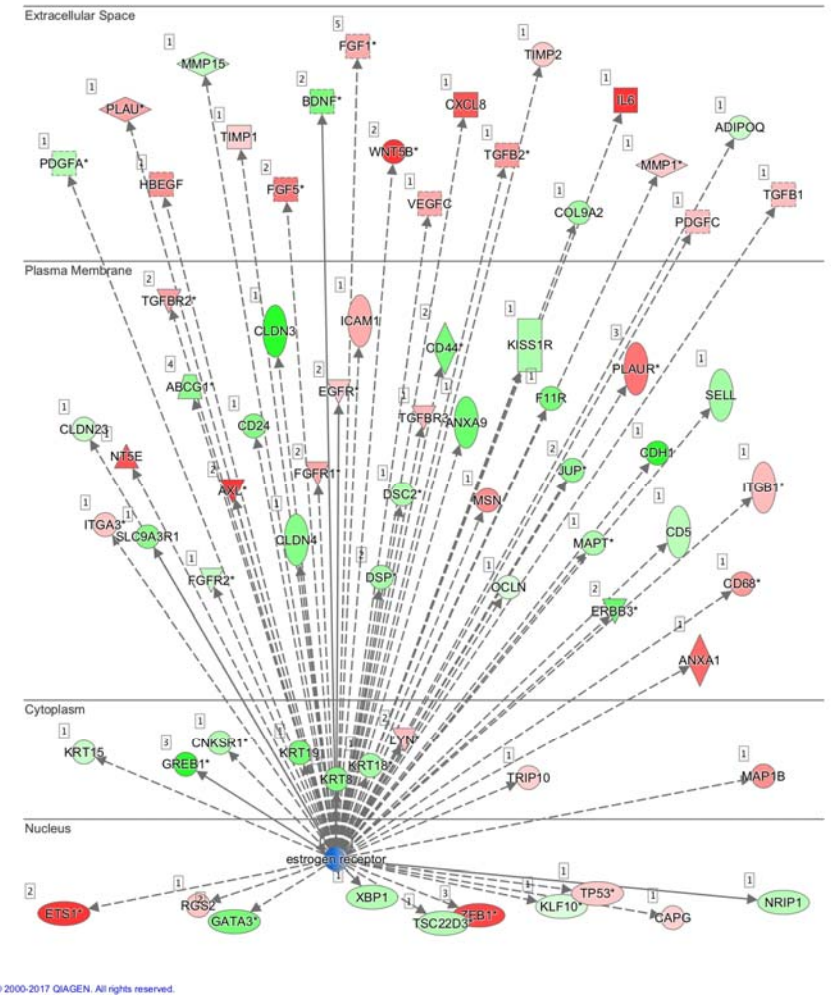
- Which metabolic and cell signaling pathways show a significance of enrichment for a group of genes?
- What are the predicted upstream and/or downstream effects of activation or inhibition of molecules in a pathway given molecules with “known” activity? (Molecule Activity Predictor)



Identify likely upstream regulators and their activity state

Upstream Analysis

- Use published experimental molecular interactions to identify upstream regulators
- Identify upstream regulators by determining gene enrichment in downstream genes
- Predict the activity state of regulators by correlating literature reported effects with observed gene expression



Upstream Analysis – Mechanistic Networks

Claudin low RNAseq LogRatio - L2R_1 P_.05 RPKM_15

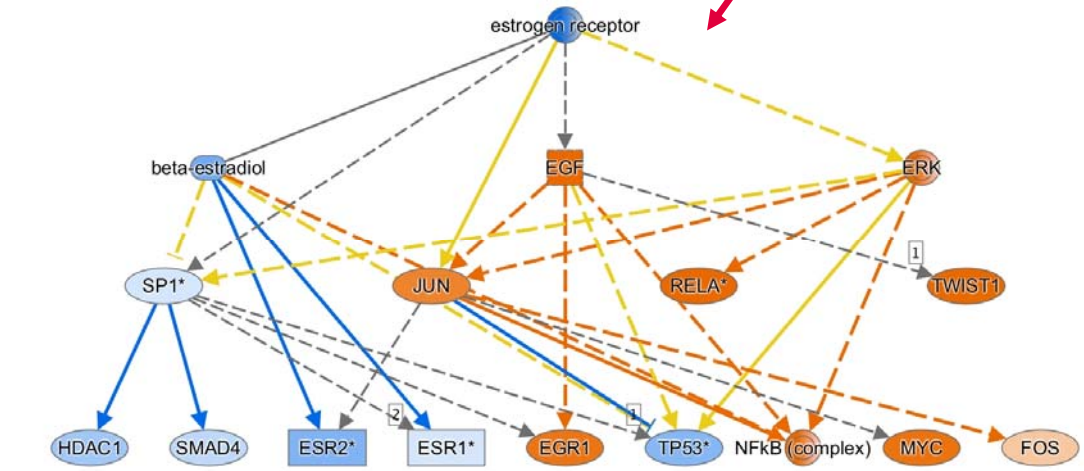
Summary \ Canonical Pathways \ Upstream Analysis \ Diseases & Functions \ Regulator Effects \ Networks \ Lists \ My Pathways \ Molecules \

Upstream Regulators \ Causal Networks \

ADD TO MY PATHWAY | ADD TO MY LIST | DISPLAY AS NETWORK | CUSTOMIZE TABLE | MECHANISTIC NETWORKS | p-value of over... 2.01E-16 - 3.95E-06 (p1 of 26) | More Info

Upstream Regulator	Expr Log Ratio	Molecule Type	Predicted Activation State	Activation z-score	p-value of overlap	Target molecules in d...	Mechanistic Network
TGFB1	↑3.000	growth factor	Activated	3.592	2.01E-16	↑ABCE1, ↑ABCF1, ...all 330	1066 (19)
ESR1	↓-12.090	ligand-dependent nuclear r...		-0.204	2.53E-15	↓ABCA3, ↓ABLIM1, ...all 268	936 (18)
beta-estradiol		chemical - endogenous ma...		-0.775	3.96E-15	↓ABCA3, ↓ABLIM1, ...all 345	1111 (19)
ERBB2	↓-4.440	kinase		0.360	2.69E-12	↑ABL1, ↑ACAA2, ...all 148	890 (19)
estrogen receptor		group	Inhibited	-5.346	3.00E-12	↓ABCG1, ↓ADIPOQ, ...all 71	925 (17)
MYC	↑0.890	transcription regulator		1.849	2.79E-11	↓ABCA2, ↓ABCA7, ...all 213	933 (16)
TP53	↑2.390	transcription regulator		-0.634	6.45E-10	↓ABAT, ↓ABCB4, ...all 263	1003 (22)
CST5	↓-1.030	other		-1.053	1.18E-09	↓ABLIM1, ↑ACAT2, ...all 74	
OSM	↑2.440	cytokine	Activated	2.416	1.98E-09	↑ABCC4, ↓ABCC8, ...all 119	971 (19)
NR3C1	↑4.400	ligand-dependent nuclear r...		-1.210	2.31E-09	↑ABL1, ↑ACTB, ...all 142	824 (14)

- Identify potential upstream regulator signal transduction
- Using shared downstream gene effects and gene-gene interactions, pathways (mechanistic networks) are created.



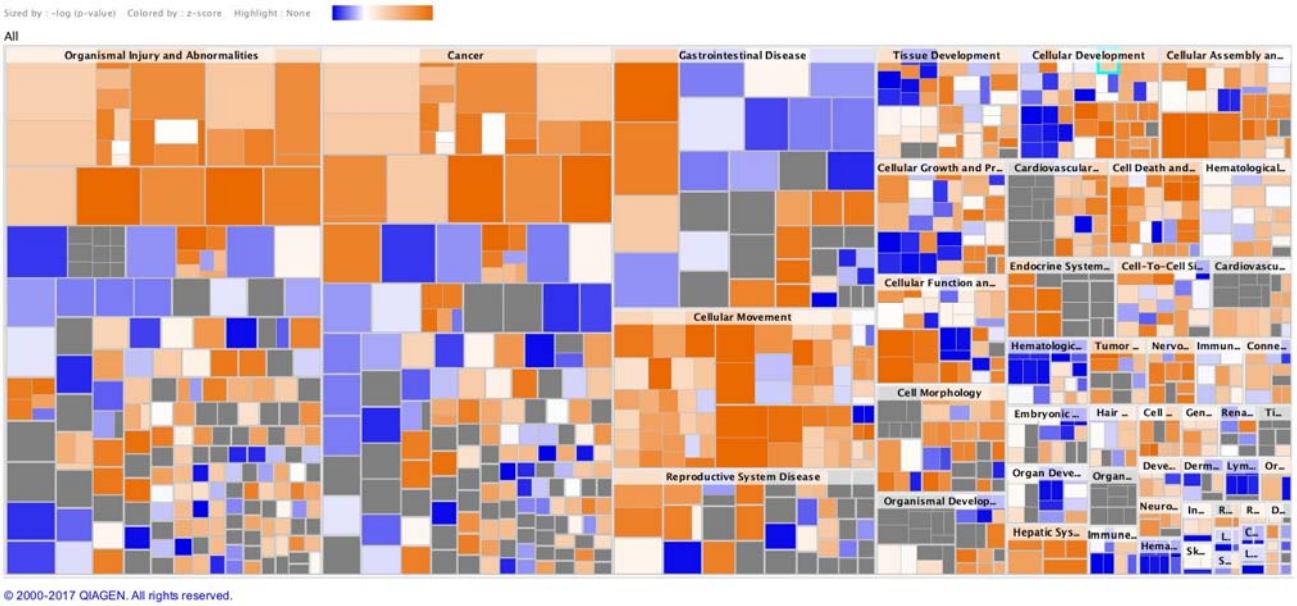
© 2000-2017 QIAGEN. All rights reserved.

Visualize and predict the biological impact of gene expression changes

Diseases & Functions

- Identify key biological processes influenced by differentially expressed genes

- Understand whether cellular processes are being driven up or down by correlating observed expression with reported experimental gene effects



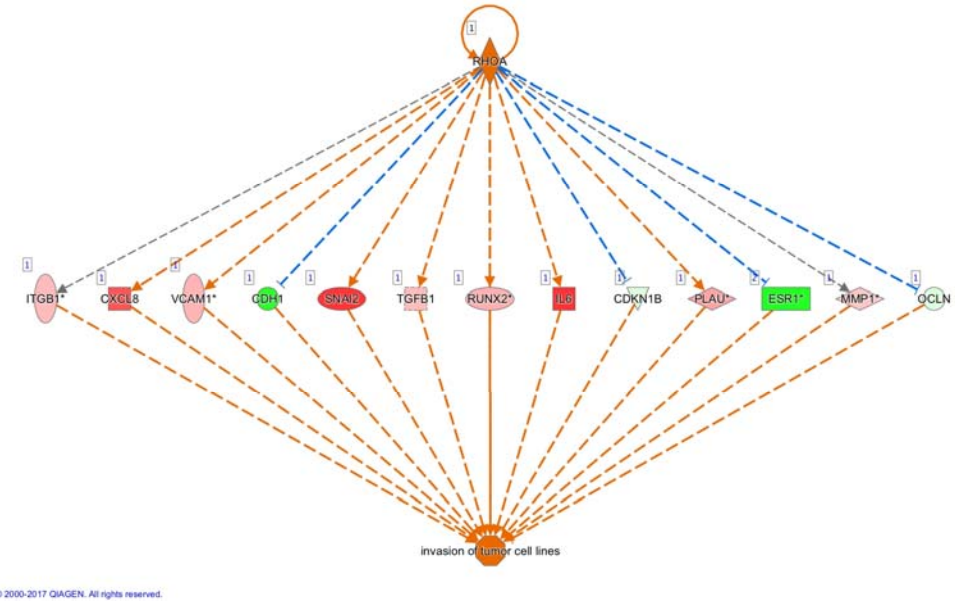
- Each box represents a biological process or disease
- The size of the box represents gene enrichment
- The color of the box indicates the predicted increase or decrease

Connect upstream, downstream analyses

Regulator Effects

- Hypothesis for how a phenotype, function or disease is regulated in the data set by activated or inhibited upstream regulators

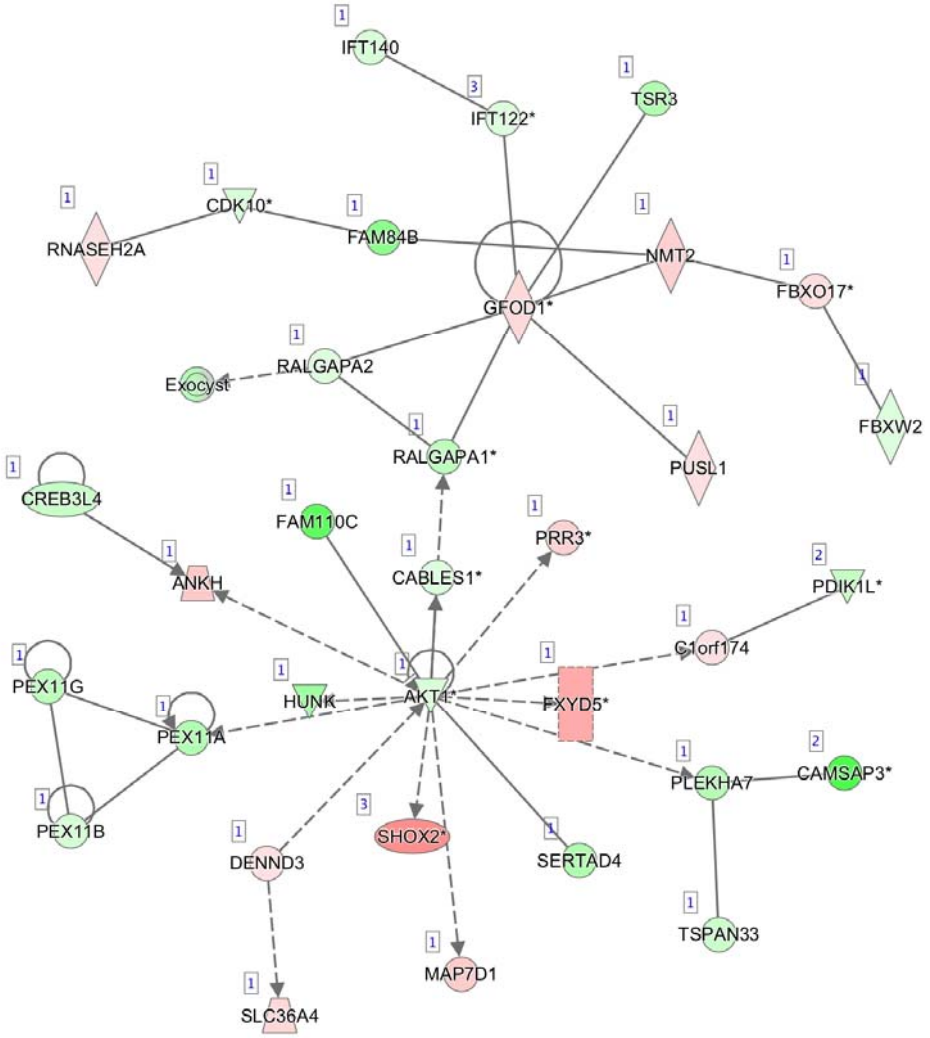
- Explain impact of upstream molecules on downstream biology
- Explain potential mechanism for a phenotype or drug
- Define drug targets
- Discover novel (or confirm known) regulator → disease/phenotype/function relationships



Discover additional interconnectivity within your data

Networks

- To show as many interactions between user-specified molecules in a given data set and how they might work together at the molecular level
- Highly-interconnected networks are likely to represent significant biological function



© 2000-2017 QIAGEN. All rights reserved.

Core Analysis Steps

1. Launch Core Analysis: File > New > Core Analysis
2. Upload Data (gene expression, protein expression, metabolomics, etc.)
3. Set Core Analysis Settings and Run Analysis
4. Interpret Results

What gets uploaded to IPA?

IPA Core Analysis input

- RNA-seq, microarray, miRNA, proteomic, genomic, SNP, or metabolic data
- Measurement calculations (e.g. differential expression and significance) are made outside of IPA prior to upload

- Observation – for a given experimental condition…
  - A list of molecule identifiers (gene, protein, etc.)
  - Corresponding measurement values (fold change, p-value, etc.)

- Single-observation data sets – one experimental comparison
  - Case vs. control
  - Mutant vs. wild-type
  - Treated vs. untreated
- Multiple-observation data sets – more than one experimental condition
  - A time course experiment with multiple time points
  - Dose response experiment with multiple doses
  - Measurement of multiple cell types or disease subtypes

Data upload format examples

ID (required)

Gene ID
NM_130786
NR_015380
NM_138932
NM_014576
NM_138933
NM_000014
NR_026971
NM_144670
NM_001080438
NM_017436
NM_016161
NM_015665
NM_023928
R 01 1035

Identifier Examples
Array IDs
dbSNP
Ensembl
Entrez Gene
GenBank
IPI
KEGG
PubChem
RefSeq
UniProt…

Data upload format examples

ID (required) Measurements (recommended)

Gene ID	Log2Ratio	p-value	Max RPKM
NM_130786	0.14	8.68E-01	2931.69
NR_015380	-0.99	2.24E-01	1649.26
NM_138932	-0.02	9.83E-01	1.67
NM_014576	-0.02	9.85E-01	1.77
NM_138933	0.02	9.79E-01	1.83
NM_000014	-4.79	1.02E-01	239.75
NR_026971	-0.67	6.17E-01	213.79
NM_144670	-5.96	1.30E-01	610.64
NM_001080438	-1.97	3.47E-01	3.91
NM_017436	-1.09	5.02E-01	6186.83
NM_016161	2.02	5.97E-02	149.85
NM_015665	-0.27	5.68E-01	13330.34
NM_023928	-1.42	1.03E-02	22828.45
[illegible]	[illegible]	[illegible]	[illegible]

Identifier Examples	Directional Comparisons	Other Measurements
Array IDs	Expr Ratio	Expr p-value
dbSNP	Expr Fold Change	Expr FDR (q-value)
Ensembl	Expr Log Ratio	Expr Intensity/RPKM/FPKM
Entrez Gene	Variant Loss/Gain	Variant ACMG Classification
GenBank	Phospho Ratio	Phospho p-value
IPI	Phospho Fold Change	Phospho FDR (q-value)
KEGG	Phospho Log Ratio	Phospho Intensity
PubChem		Phospho Site
RefSeq		
UniProt…		

Data upload format examples

ID (required) | Measurements (recommended) | Additional Observations (optional)

Gene ID	Log2Ratio	p-value	Max RPKM	Log2Ratio	p-value	Max RPKM
NM_130786	0.14	8.68E-01	2931.69	-0.01	9.82E-01	2117.73
NR_015380	-0.99	2.24E-01	1649.26	0.12	8.64E-01	14076.24
NM_138932	-0.02	9.83E-01	1.67	-1.62	1.46E-01	31.85
NM_014576	-0.02	9.85E-01	1.77	0.12	8.25E-01	10491.96
NM_138933	0.02	9.79E-01	1.83	2.02	4.44E-01	14788.5
NM_000014	-4.79	1.02E-01	239.75	-0.57	1.09E-01	273101
NR_026971	-0.67	6.17E-01	213.79	0.36	4.87E-01	11876
NM_144670	-5.96	1.30E-01	610.64	-0.17	7.48E-01	3339.36
NM_001080438	-1.97	3.47E-01	3.91	0.7	1.02E-01	37787.69
NM_017436	-1.09	5.02E-01	6186.83	4.09	1.74E-01	6988.43
NM_016161	2.02	5.97E-02	149.85	1.04	1.18E-01	27563.08
NM_015665	-0.27	5.68E-01	13330.34	-0.5	6.92E-01	760.71
NM_023928	-1.42	1.03E-02	22828.45	-3.85	6.92E-02	14.43
R 0[illegible]35	1.06	5.96E-01	6.13	5.41	3.18E-0[illegible]	103[illegible]3

Identifier Examples
- Array IDs
- dbSNP
- Ensembl
- Entrez Gene
- GenBank
- IPI
- KEGG
- PubChem
- RefSeq
- UniProt…

Directional Comparisons
- Expr Ratio
- Expr Fold Change
- Expr Log Ratio
- Variant Loss/Gain
- Phospho Ratio
- Phospho Fold Change
- Phospho Log Ratio

Other Measurements
- Expr p-value
- Expr FDR (q-value)
- Expr Intensity/RPKM/FPKM
- Variant ACMG Classification
- Phospho p-value
- Phospho FDR (q-value)
- Phospho Intensity
- Phospho Site

Experimental Comparison Examples
- Mutant vs. wild-type
- Treated vs. untreated
- Other case vs. control

- Additional time points
- Multiple dose responses
- Various cell lines

Data Upload

Best practices

- Calculate metrics outside of IPA (e.g. fold-change, p-value)
- Create an Excel spreadsheet or tab delimited file
  - Only 1 header row allowed
  - One column must have identifiers, preferably the left-most column
  - IPA will only look at the top worksheet in an Excel workbook
- Group related observations into a single spreadsheet if possible
  - Time course, drug concentration, cell lines, etc.
  - Can have up to 20 observations
- Specify array platform (chip) if possible
  - It is OK to use “Not specified/applicable”
- Pre-filter data at the lowest threshold that you have confidence in
  - For example, probe measurement p-value of .05 or other criteria
  - Further filter in Core Analysis

Expression Value Calculation

Verify the differential expression calculation

- Ratio differential expression

$$\frac{Experimental\ Condition\ Exp.}{Control\ Exp.}$$

- Log_2(ratio) differential expression (recommended)

$$Log_2\left(\frac{Experimental\ Condition\ Exp.}{Control\ Exp.}\right)$$

- Fold Change
  - If increased differential expression

$$\frac{Experimental\ Condition\ Exp.}{Control\ Exp.}$$

Fold change will never have values between -1 and 1

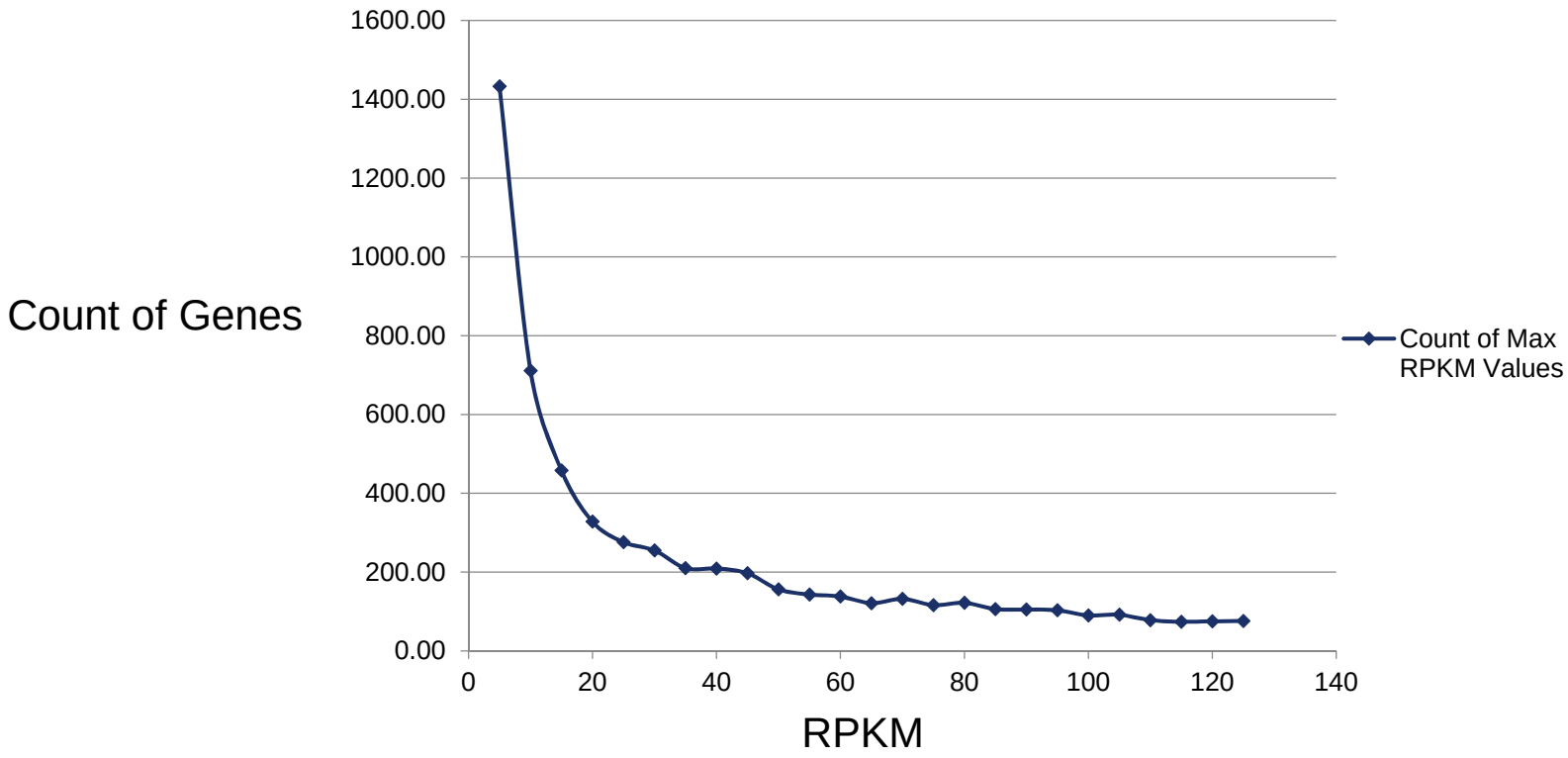
  - If decreased differential expression

$$-1\left(\frac{Control\ Exp.}{Experimental\ Condition\ Exp.}\right)$$

Why filter on max RPKM values in RNAseq data?

Filtering on absolute expression

- RNAseq measurements often result in many significant differential fold changes at low absolute transcript expression levels
- Including the maximum RPKM value of your experimental condition and control allows for later filtering on absolute expression value in addition to fold change and p-value



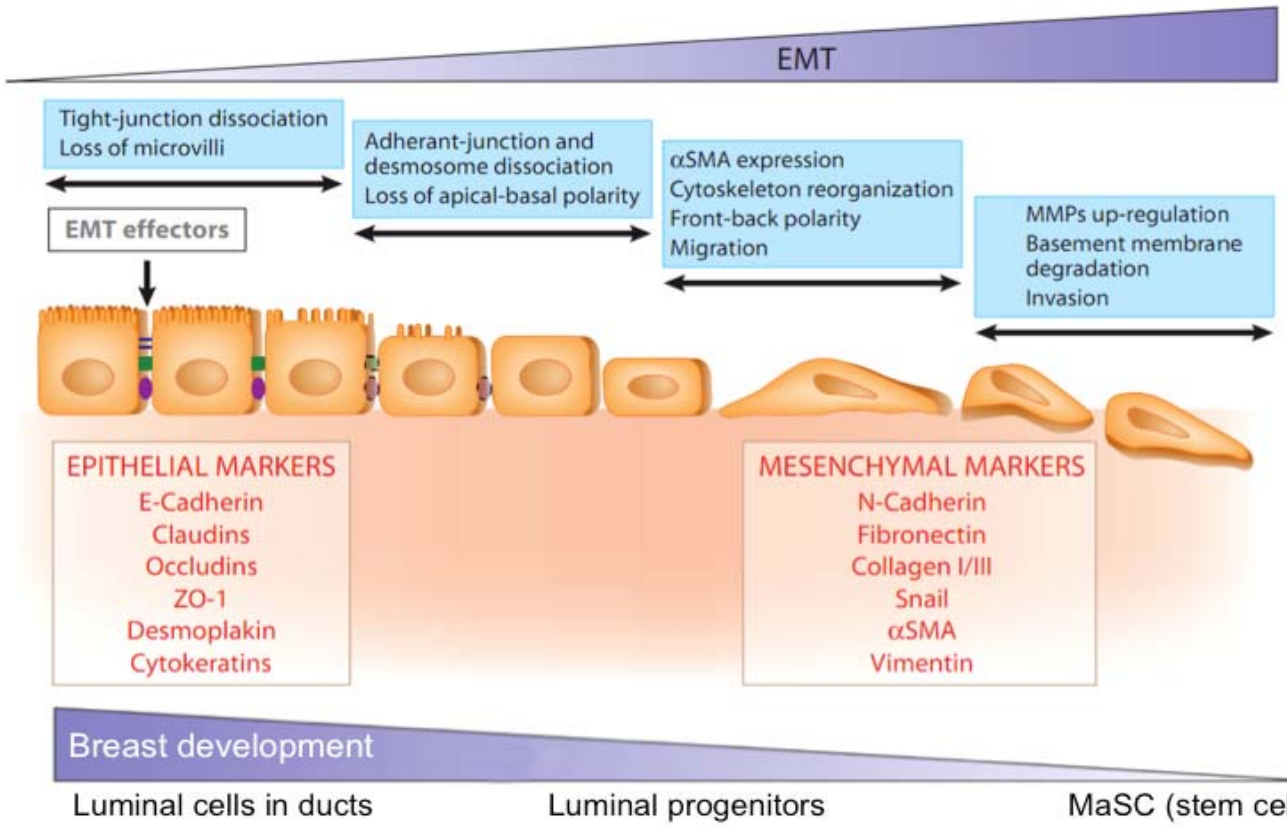
Core Analysis Steps

1. Launch Core Analysis: File > New > Core Analysis
2. Upload Data (gene expression, protein expression, metabolomics, etc.)
3. Set Core Analysis Settings and Run Analysis
4. Interpret Results

Data Upload and Core Analysis Set-up in IPA

Large Scale Data Analysis

Epithelial to Mesenchymal Transition



Adapted from Aroeira et al. J Am Soc Nephrol 18: 2004 –2013, 2007

Luminal Breast cancer | HER2-enriched | Basal | **Mesenchymal / stem cell-like breast cancer**

Luminal cell lines

Ratio Claudin-low to Luminal
5 vs 5 cell lines, RNA-Seq data

Claudin-low cell lines

Gene/Protein Expression Analysis, Overview

IPA Core Analysis

- Pathway Analysis
  - Identifies enriched canonical pathways and scores directional changes based on gene expression
- Upstream Regulator Analysis
  - Predicts what regulators caused changes in gene expression and the directional state of regulator
- Diseases and Functions Analysis
  - Predicts effected biology (cellular processes, biological functions) based on gene expression and predicts directional change on that effect
- Regulator Effects
  - Models pathway interactions from predicted upstream regulators, through differentially expressed genes, to biological processes
- Networks
  - Predicts non-directional gene interaction map

Analyzing and Interpreting Results

IPA calculates two distinct statistics as part of a core analysis

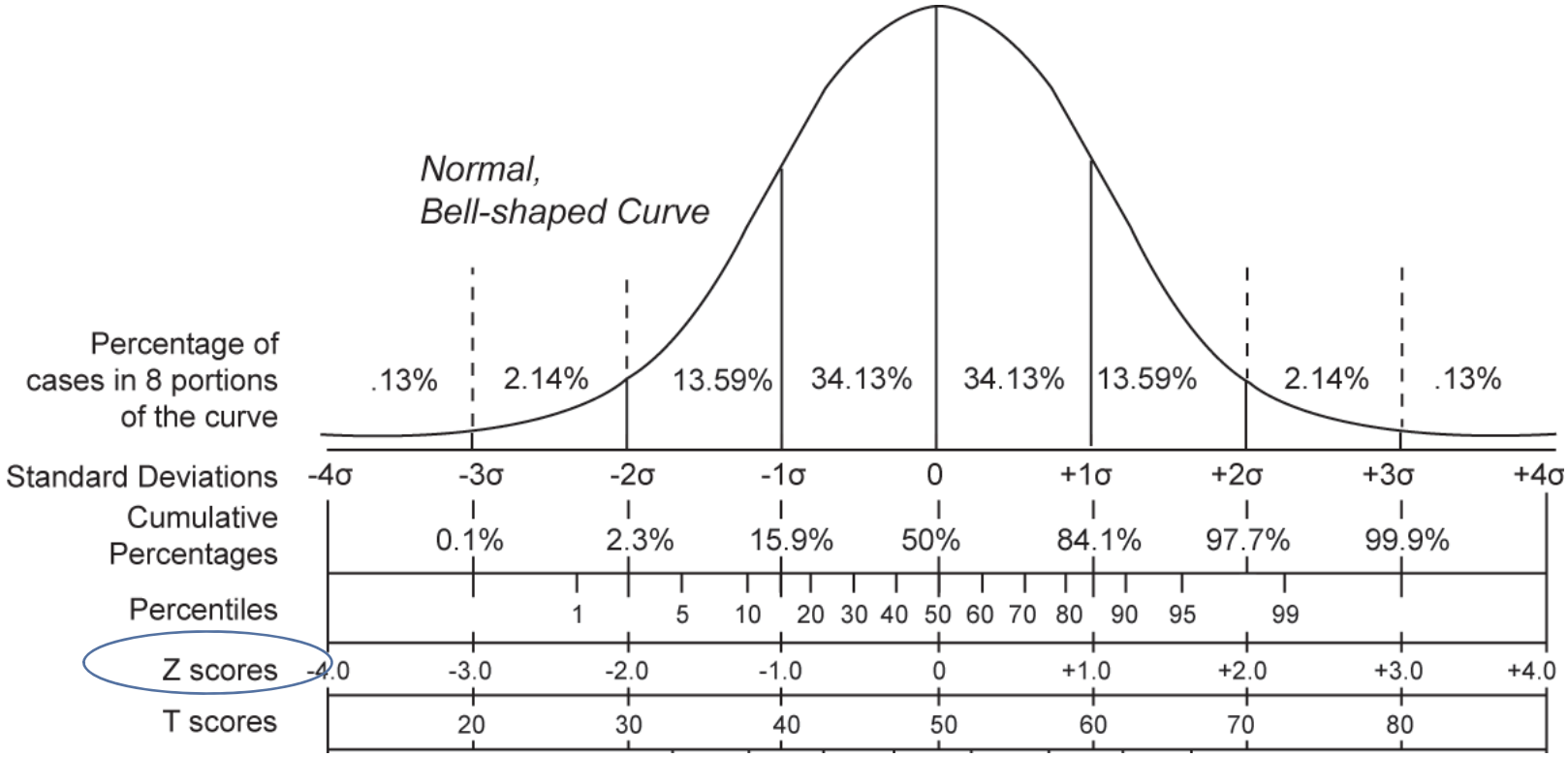
- P-value:
  - Calculated using a Right-Tailed Fisher's Exact Test
  - Reflects the likelihood that the association or overlap between a set of significant molecules from your experiment and a given process/pathway/transcription neighborhood is due to random chance. The smaller the p-value the less likely that the association is random.
  - The p-value does not consider the directional effect of one molecule on another, or the direction of change of molecules in the data set.

- Z-score:
  - Applied in some analysis types and provides predictions about upstream or downstream processes.
  - Takes into account the directional effect of one molecule on another molecule or on a process, and the direction of change of molecules in the data set.

z-scores and Normal Distribution

A set of genes chosen at random should be about equally likely to have an increasing or decreasing effect, thus, about 50% each direction, or a z=0.

A z-score represents the non-randomness of directionality within a gene set



Analyzing Results

Canonical Pathways

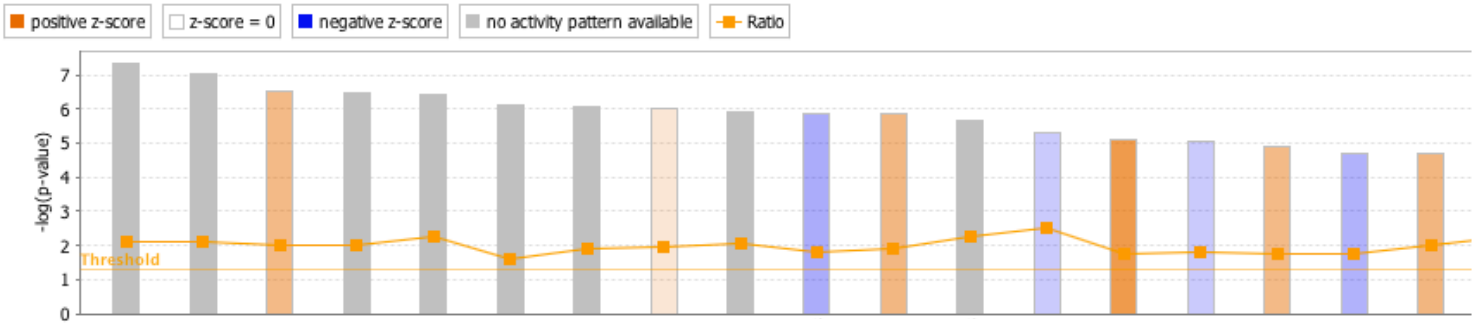
Canonical Pathway Z-scores

Pathway activity analysis

- Allows you to quickly determine if Canonical Pathways, including functional end-points, are increased or decreased based on differentially expressed genes or proteins in your data set

- Certain pathways within the knowledge base are directional (proceed from “A” to “Z”)
- As part of pathway curation, a subset of genes are selected to be active
  - Allows the directionality of other genes to be predicted
  - Result defines an “activated” state for a given pathway

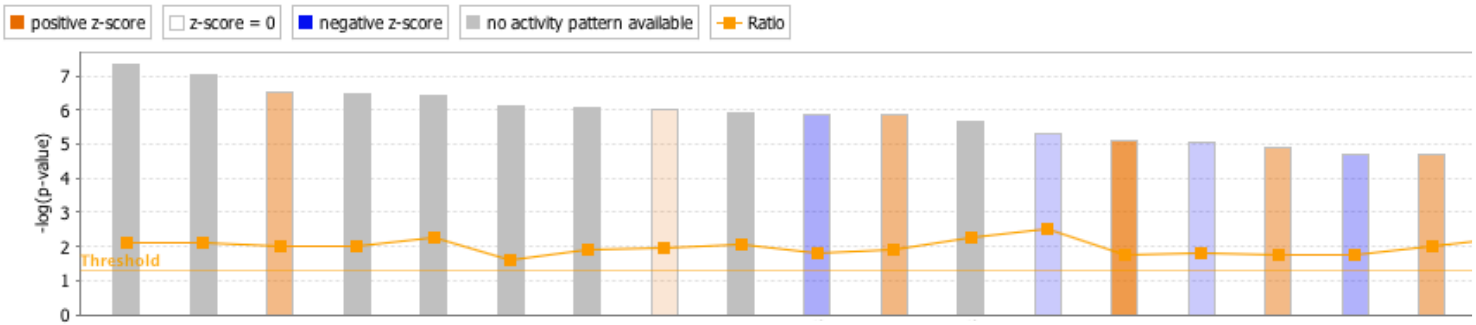
- Z-scores are calculated based on the data set’s correlation with the activated state



Canonical Pathway z-ccores

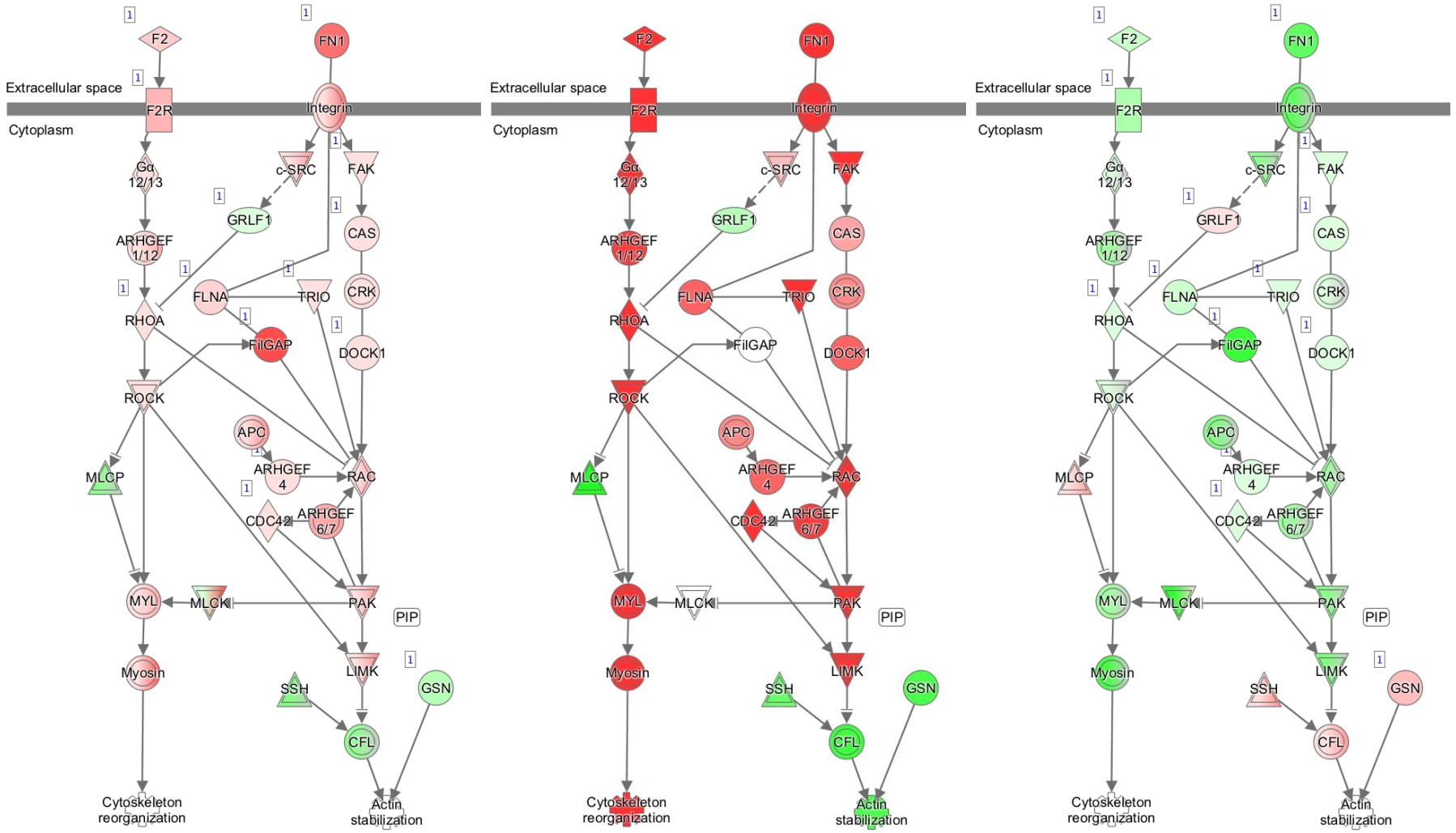
Pathway activity analysis

- Allows you to quickly determine if Canonical Pathways, including functional end-points, are increased or decreased based on differentially expressed genes or proteins in your data set



- Certain pathways within the knowledge base are directional (proceed from “A” to “Z”)
- Defining the “activated” state of a directional pathway
  - As part of the pathway curation, experts select a subset of genes within a pathway to be active, allowing the directionality of other genes to be predicted
- Z-scores are calculated based on the data set’s correlation with the activated state
  - Gray bar – no prediction can be made (pathway currently ineligible for a prediction)
- The “Expected” column (in the table that is displayed when you select of one of the bar charts) indicates the state that gene is predicted to have if the pathway were activated.
  - Press the “A” key when viewing a pathway to overlay expected activation state

Pathway Activity Analysis



Positive Z-score (Example Data)

Expected Activation State (Knowledge Base)

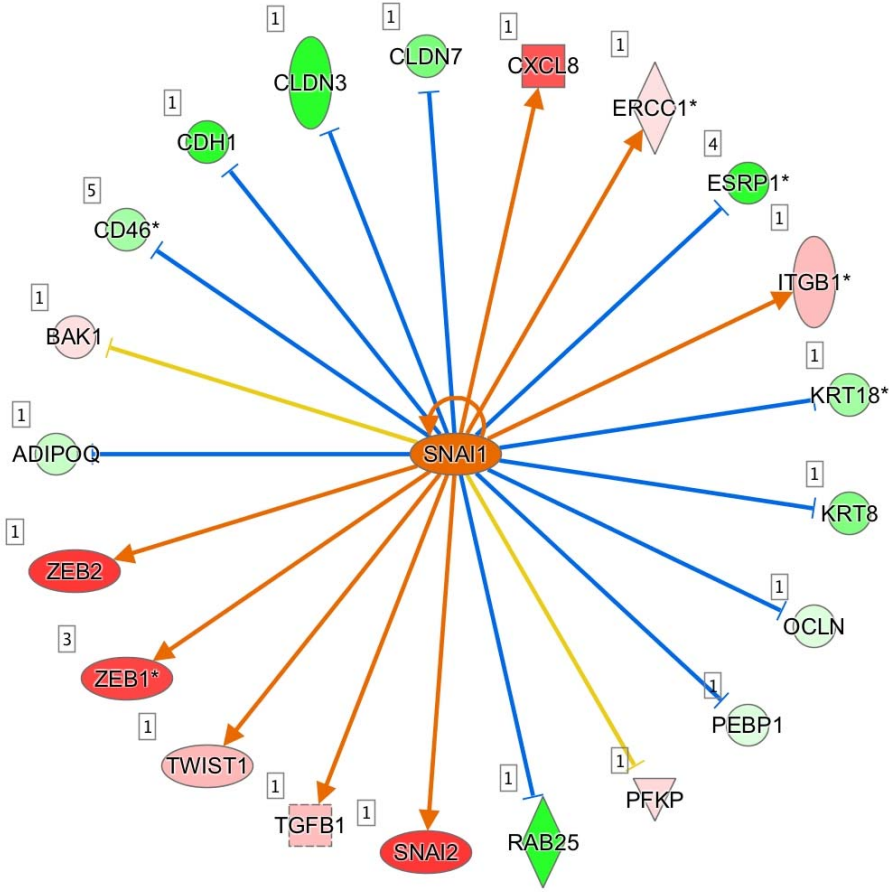
Negative Z-score (Example Data)

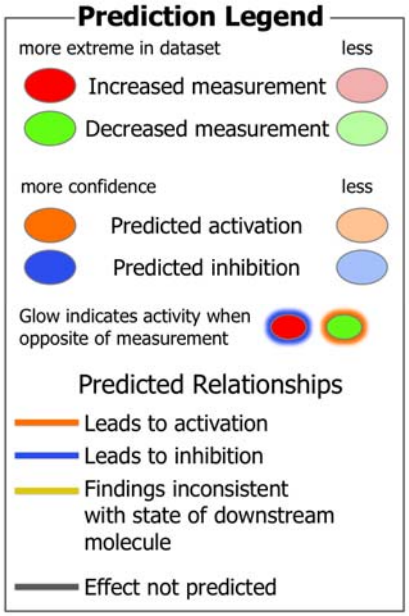
Analyzing Results

Upstream Regulators

IPA Upstream Regulator Analysis

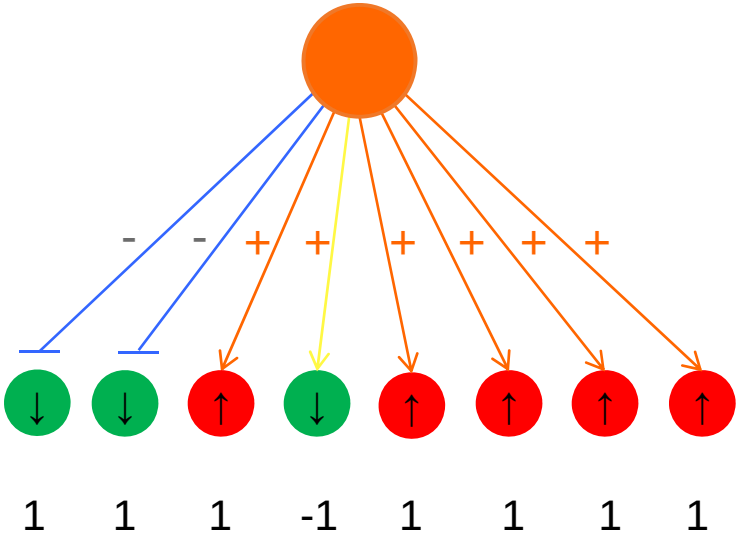
Directional Effects: Molecule Activity Predictor
Examine Expression Relationship Consistency





© 2000-2017 QIAGEN. All rights reserved.

IPA Upstream Regulator Analysis



← Every possible TF & Upstream Regulator in the Ingenuity Knowledge Base is analyzed

← Literature-based effect TF/UR has on downstream genes

← Differential Gene Expression (Uploaded Data)

← Predicted activation state of TF/UR:
1 = Consistent with activation of UR
-1 = Consistent with inhibition of UR

$$z = \frac{x}{\sigma_x} = \frac{\sum_i x_i}{\sqrt{N}} = \frac{N_+ - N_-}{\sqrt{N}} \quad = (7\text{-}1)/\sqrt{8} \quad = 2.12\ (= \text{predicted activation})$$

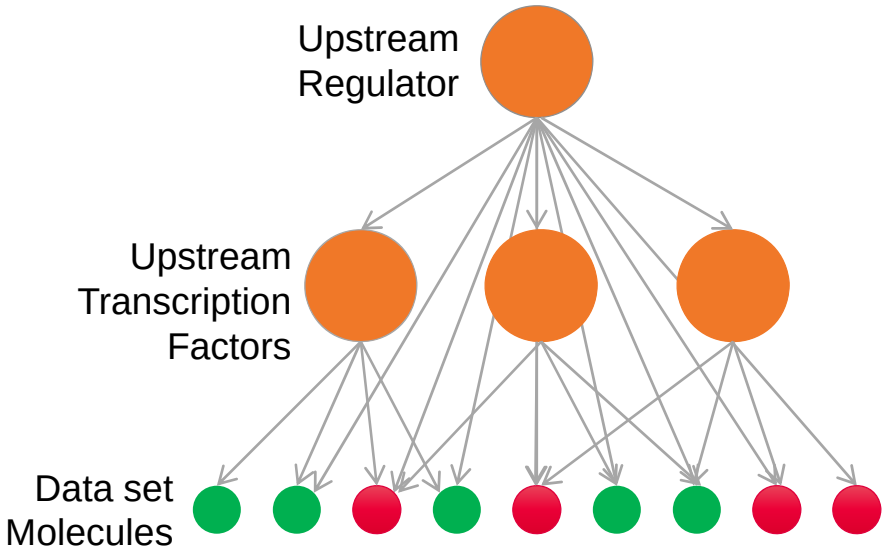
- z-score is a statistical measure of the match between expected relationship direction and observed gene expression
- z-score > 2 or < -2 is considered significant

Note that the actual z-score is weighted by the underlying findings, the relationship bias, and data set bias

IPA Mechanistic Networks

Goal: To discover plausible sets of connected upstream regulators that can work together to elicit the gene expression changes observed in a data set

How: Take IPA Upstream Regulator results and computationally seek pairs of regulators predicted to affect the expression of a similar set of genes. Repeat to build a network:



How might the upstream molecule drive the observed expression changes?

Claudin low RNAseq LogRatio - L2R_1 P_.05 RPKM_15

Summary | Canonical Pathways | Upstream Analysis | Diseases & Functions | Regulator Effects | Networks | Lists | My Pathways | Molecules

Upstream Regulators | Causal Networks

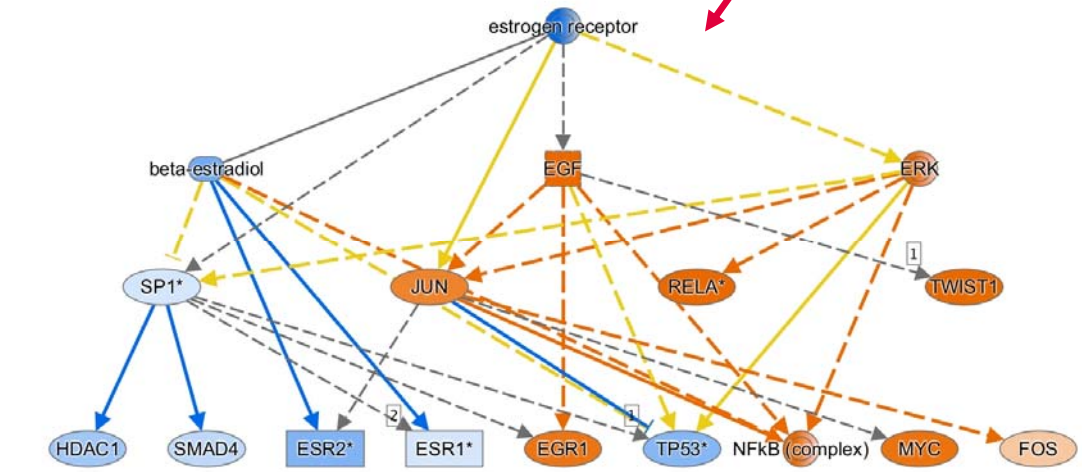
ADD TO MY PATHWAY | ADD TO MY LIST | DISPLAY AS NETWORK | CUSTOMIZE TABLE | MECHANISTIC NETWORKS

p-value of over... 2.01E-16 - 3.95E-06 (p1 of 26) More Info

Upstream Regulator	Expr Log Ratio	Molecule Type	Predicted Activation State	Activation z-score	p-value of overlap	Target molecules in d...	Mechanistic Network
TGFB1	↑3.000	growth factor	Activated	3.592	2.01E-16	↑ABCE1, ↑ABCF1, ...all 330	1066 (19)
ESR1	↓-12.090	ligand-dependent nuclear r...		-0.204	2.53E-15	↓ABCA3, ↓ABLIM1, ...all 268	936 (18)
beta-estradiol		chemical - endogenous ma...		-0.775	3.96E-15	↓ABCA3, ↓ABLIM1, ...all 345	1111 (19)
ERBB2	↓-4.440	kinase		0.360	2.69E-12	↑ABL1, ↑ACAA2, ...all 148	890 (19)
estrogen receptor		group	Inhibited	-5.346	3.00E-12	↓ABCG1, ↓ADIPOQ, ...all 71	925 (17)
MYC	↑0.890	transcription regulator		1.849	2.79E-11	↓ABCA2, ↓ABCA7, ...all 213	933 (18)
TP53	↑2.390	transcription regulator		-0.634	6.45E-10	↓ABAT, ↓ABCB4, ...all 263	1003 (22)
CST5	↓-1.030	other		-1.053	1.18E-09	↓ABLIM1, ↑ACAT2, ...all 74	
OSM	↑2.440	cytokine	Activated	2.416	1.98E-09	↑ABCC4, ↓ABCC8, ...all 119	971 (19)
NR3C1	↑4.400	ligand-dependent nuclear r...		-1.210	2.31E-09	↑ABL1, ↑ACTB, ...all 142	824 (14)

Hypothesis generation and visualization

Each hypothesis generated indicates the molecules predicted to be in the signaling cascade



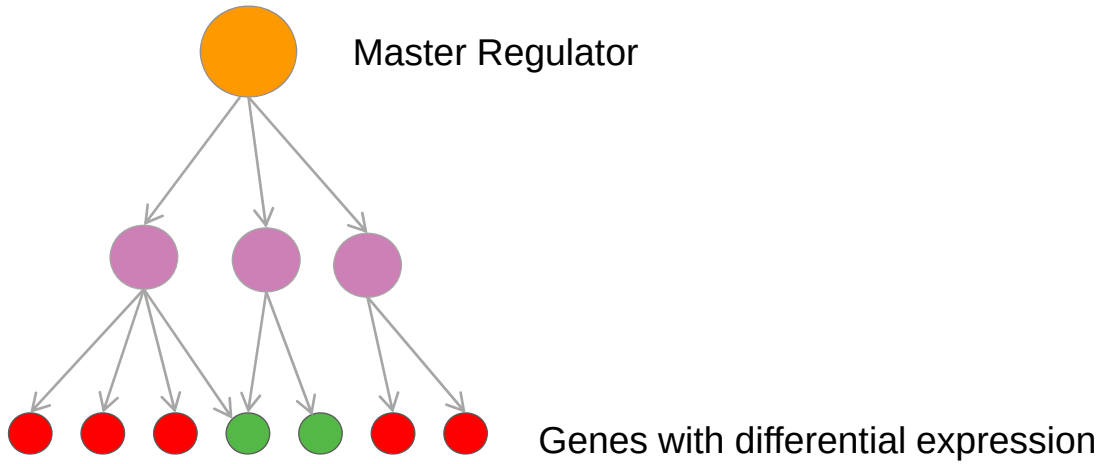
© 2000-2017 QIAGEN. All rights reserved.

Causal Network Analysis

Advanced Analytics

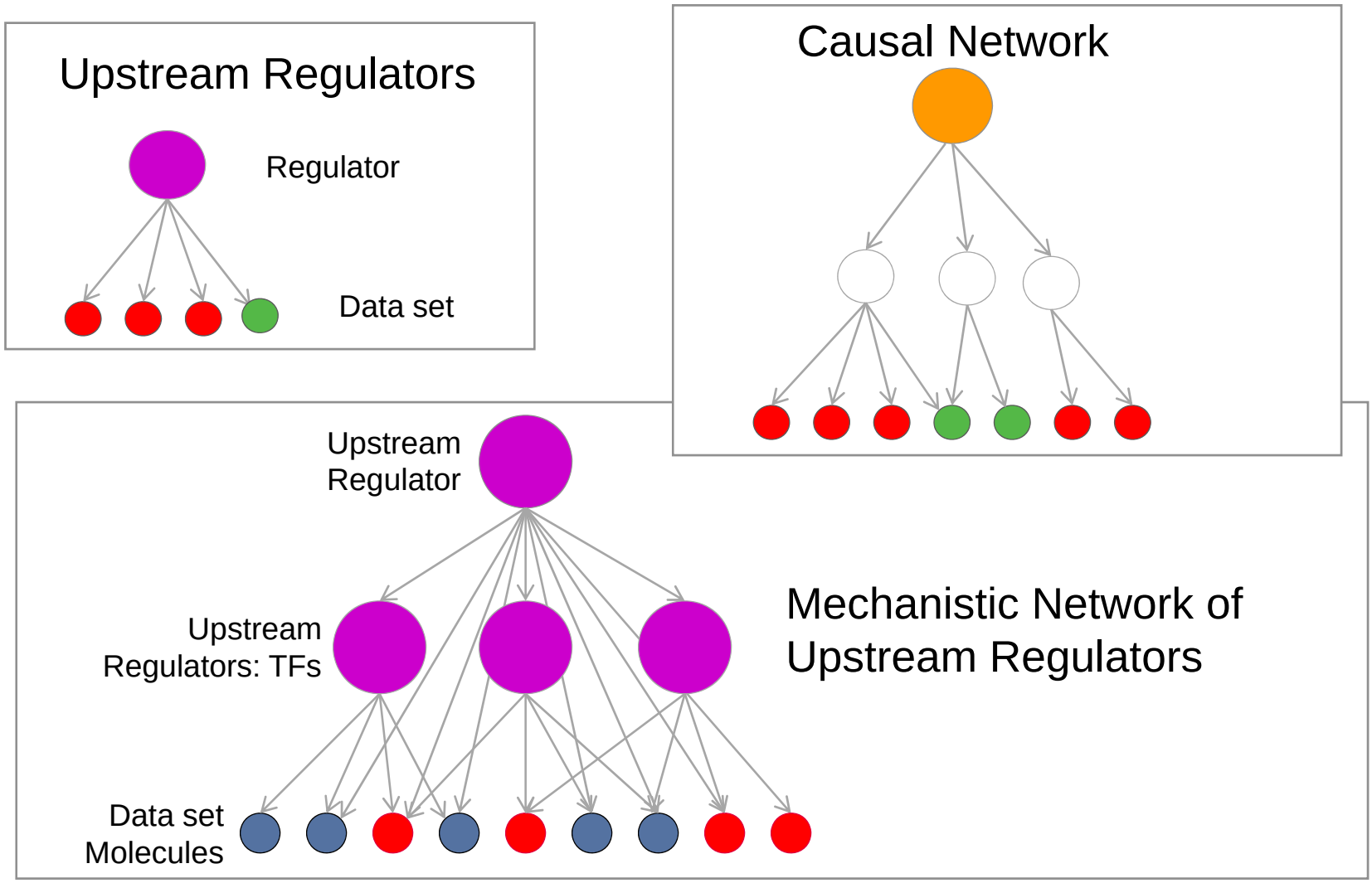
- Alternate method of predicting upstream regulators based on causal relationships and allowing multiple interaction steps to gene expression changes
- Identify potential novel master-regulators of your gene expression by creating pathways of literature-based relationships
- Expands predictions to include indirect upstream regulators not in mechanistic networks

Causal Networks



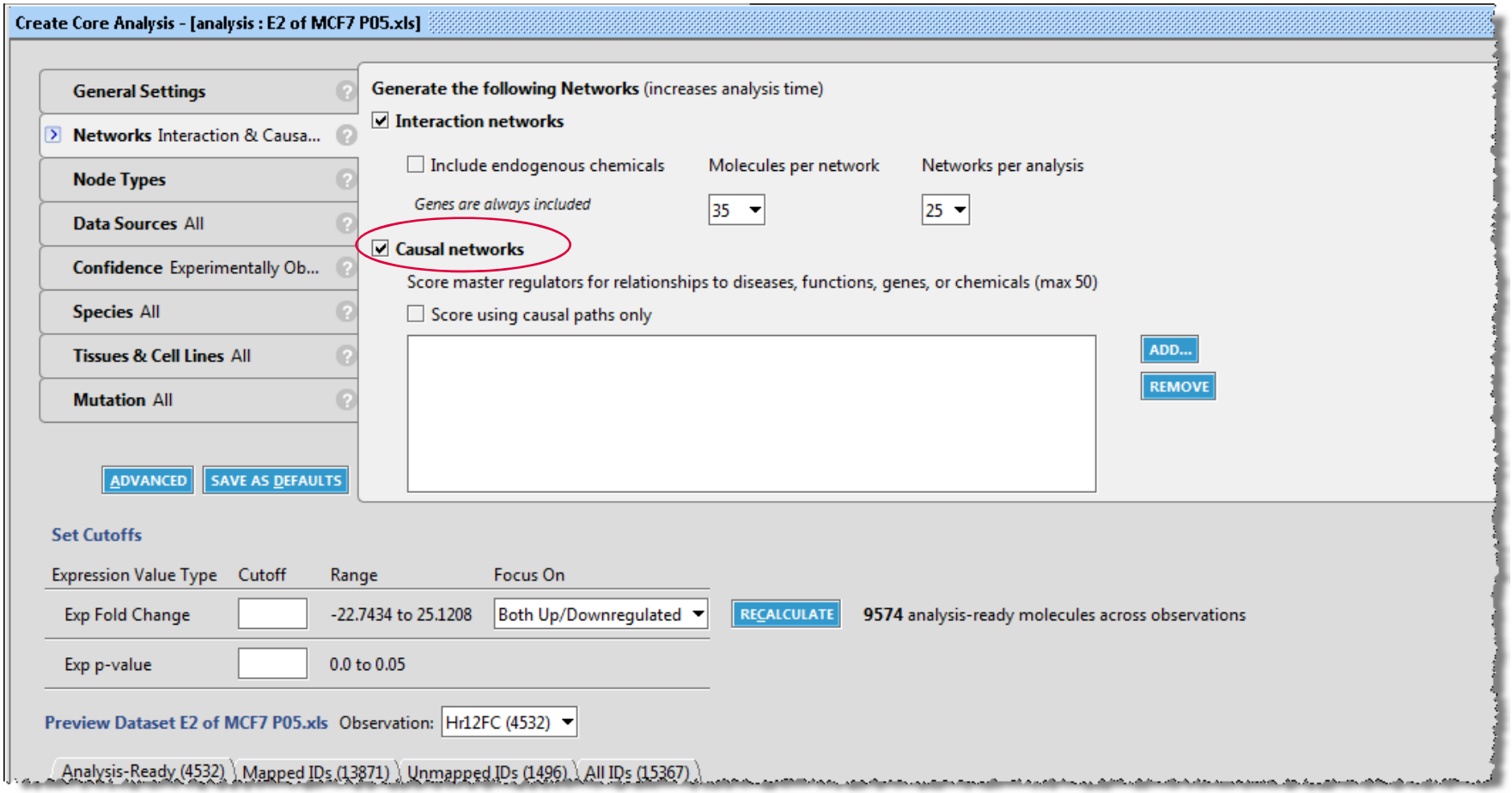
Single- vs. Mechanistic- vs. Causal Networks
Leveraging the network to create more upstream regulators

Advanced Analytics: Causal Network Analysis



Turning on Causal Networks (with Advanced Analytics)

DG1



Slide 45

DG1 Add slide and move to training slide deck

Darryl Gietzen, 3/25/2016

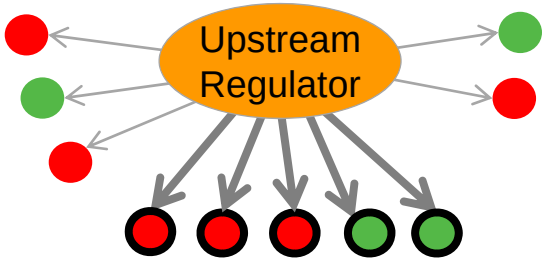
Analyzing Results

Regulator Effects

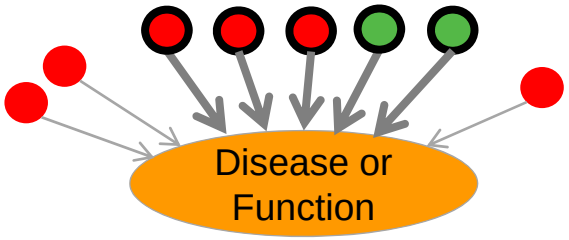
Regulator Effects

Hypotheses for how activated or inhibited upstream regulators cause downstream effects on biology

Upstream Regulator Analysis



Targets in the data set

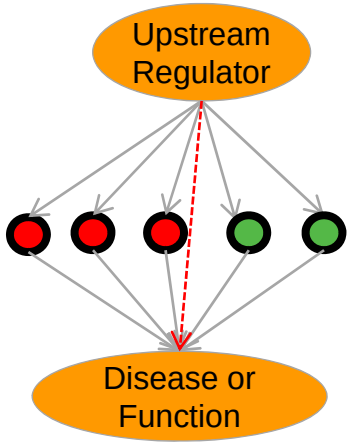


Downstream Effects Analysis

Algorithm

First iteration

Simplest Regulator Effects result

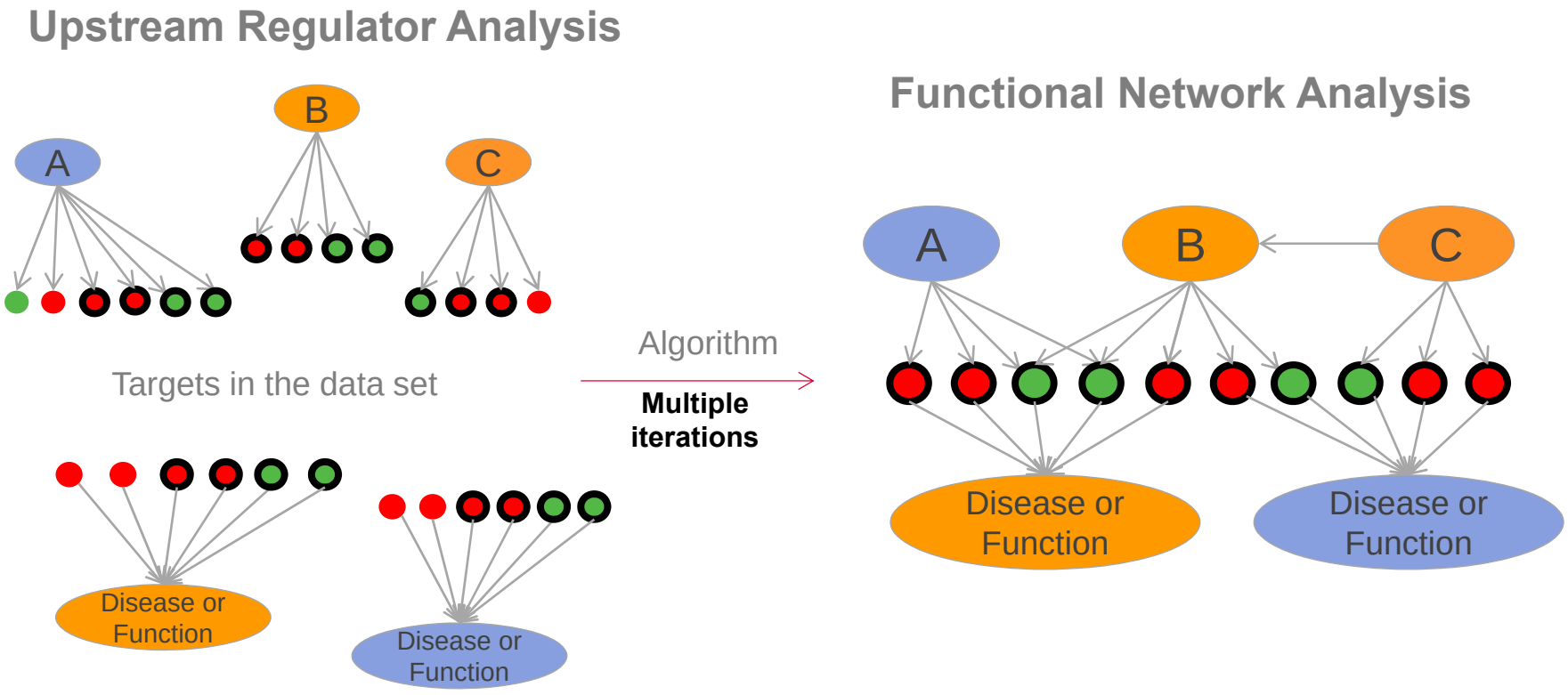


Displays a relationship between the regulator and disease/function if it exists

Causally consistent networks score higher

The algorithm runs iteratively to merge additional regulators with diseases and functions

Concept of Functional Network Analysis



Analyzing Results

Networks

Networks in IPA

- To show as many interactions between user-specified molecules in a given data set and how they might work together at the molecular level
- Highly-interconnected networks are likely to represent significant biological function

Networks

- Networks are assembled based on gene/molecule connectivity with other gene/molecules.
  - Assumption: the more connected a gene/molecule, the more influence it has and the more “important” it is.
- Networks are assembled using decreasingly connected molecules from your data set.
- Genes/molecules from the Knowledge Base may be added to the network to fill or join areas lacking connectivity.
- A maximum of 35, 70, or 140 genes/molecules can comprise a network based on parameter settings.
- Networks are annotated with high-level functional categories.

How Networks Are Generated

Focus molecules are “seeds”

Focus molecules with the most interactions to other focus molecules are then connected together to form a network

Non-focus molecules from the data set are then added

Molecules from the Ingenuitys KB are added

Resulting Networks are scored and then sorted based on the score