

Help file

Key Pathway Advisor (KPA)

Table of contents

Chapter 1: Functional overview	3
1.1 Transcriptome and proteome analysis in Key Pathway Advisor– technique and method	3
1.2 Systems analysis categories.....	4
Chapter 2: KPA browser settings and functionality	5
2.1 Data sets and file types.....	5
2.2 Settings	6
2.3 Data analysis process	7
2.4 Report panel – example.....	7
Chapter 3: Algorithm workflow.....	8
Chapter 4: Analysis report.....	10
4.1 General analysis guidelines and limitations.....	11
4.2 Molecular components analysis	11
4.2.1 Your data.....	12
4.2.2. Key hubs.....	12
4.2.3 Key hubs networks	14
4.2.4 Analysis guidelines and limitations.....	19
4.3 Key processes analysis	20
4.3.1 Key processes analysis.....	21
4.3.2 Pathway maps.....	22
4.3.3 Process networks	27
4.4 Drug targets (prior knowledge).....	30
4.5 Putative biomarkers.....	31
4.6 Content browsing	32
5.1 Causal reasoning analysis	34
5.2 Overconnectivity analysis.....	34
5.3 Enrichment analysis.....	35
5.4 Signaling Pathway Impact Analysis (SPIA)	35
Chapter 6: Glossary	37
Bibliography	40
Appendix A: How to get NCBI GEO data for KPA easily	41

Appendix B: Differential gene expression calculations	47
Appendix C: P-value	48
Appendix D: Upstream analysis.....	51
Appendix E: Molecular interactions mechanisms	54

Chapter 1: Functional overview

1.1 Transcriptome and proteome analysis in Key Pathway Advisor– technique and method

The *Key Pathway Advisor* (KPA) is a streamlined workflow primarily focused on analysis of transcriptome/gene expression data. These techniques focus on the identification of genes with either change in expression associated with the different conditions under investigation by the researcher. These genes can be studied in two directions: downstream and upstream.

Downstream analysis tries to interpret the consequences of expression changes defining which cellular processes can be affected by differentially expressed genes (DEG). Downstream analysis, therefore, looks for molecules that the genes shown to be changing under the experimental conditions (differentially expressed genes) have been shown to have an effect upon.

Upstream analysis tries to understand the molecular reasons that could cause expression changes between two phenotypes (including transcriptional factors and signaling pathways). It searches for molecules that have been shown to have an effect upon the gene expression changes observed in the data.

Using a one-click comprehensive workflow function, KPA identifies the following crucial information for DEG list interpretation:

- Predicted Key Hubs (Protein Activity) - molecules whose activity changes could explain the gene expression changes in your uploaded data. These hubs have the potential to change the expression level of experimentally defined DEGs but may not necessarily be identified by gene expression experiments as changing in expression themselves. Their activity changes may, therefore, be observed on other biological levels, such as activity changes.
- Key Pathways – a list of pathways and cellular processes maps enriched with DEGs and Key Hubs gives an interpretation of potentially affected biology.
- Enriched Gene Ontologies – sets of various *MetaCore* and GO ontologies enriched with DEGs and Key Hubs.
- Putative Biomarkers - identifies which of your DEGs or Key Hubs have previously been associated with your disease of interest (or similar diseases) and whether the direction of change observed/predicted is the same as previously seen in the literature.
- Drug Targets (Prior Knowledge) - DEGs or Key Hubs associated with drugs (from *Clarivate Analytics Integrity*) in preclinical to launched phases for your disease of interest (or similar diseases).

Note: KPA does not currently allow you to define fold-change or p-value cut-off levels upon upload. Please pre-filter your data before uploading.

Note: KPA is limited to analysis of 2,000 genes or less in a file no larger than 10 megabytes (MB);

- If your input file contains fold change (FC) values, KPA will use the 2,000 records with the highest FC

- If your input file does not contain FC values, the first 2,000 records in your file will be used.
- If you are uploading a gene variant file in addition to your gene list, the upload limit is 2,000 genes plus 1,000 gene variants, and a combined file size of no larger than 10 MB

1.2 Systems analysis categories

“Systems” or functional analysis is believed to help simplify complex biological responses involving hundreds or thousands of DEGs, gene variants and other molecular components. Functional analysis is described by several “knowledge-based” methods (such as ontology enrichment and biological network generation) which use well-annotated databases of protein interactions, multi-step pathways and cellular processes. These analytical methods can be divided into two categories:

The first category deals with molecular objects, such as DEGs, and consists of sub-dividing the sets of genes according to common functionality (ontology enrichment analysis). The genes / proteins of interest have different functions, and this functionality is multi-dimensional. The encoded proteins can belong to the pathways, groups and complexes, normal and pathological cellular processes. At the same time, alterations in gene expression can be associated with a phenotypic “endpoint” such as a disease. This complex response can be assessed by sub-categorizing the gene / protein content into different folders (terms) of ontologies, and the terms are further ranked based on a relative representation of genes of interest within the term. Ontology enrichment is an intuitive but “low resolution” descriptive method which alone usually does not provide a researcher with an executable hypothesis for subsequent follow-up experiments. Another enrichment method can be used for identifying enrichment synergy between multiple datasets, especially when comparing functionally relevant but “incomparable” datasets such as mutations and amplifications¹.

The second category of functional methods deals with protein interactions connecting genes / proteins within the dataset or with the whole proteome and consists of identifying and ranking genes based on their “biological relevance” for the phenotype. The “relevance” is assumed to be dependent on the number of protein-protein (or other biological molecules) interactions and multi-step pathways for each protein. This follows from the fact that proteins work in groups and are connected by interactions. Therefore, the relative “importance” of a protein for a certain dataset can be defined by the relative number of interactions with proteins / genes from this dataset. This “local topology” is normalized vs. the whole human interactome and can be ranked by p-values or a “connectivity ratio”. The interaction- focused methods include topologically significant “causal nodes”^{2, 3} and one-step overconnectivity calculations for different protein functions¹.

Note: topologically significant genes are often missed in expression profile experiments, as the key regulatory genes like transcription factors or kinases change expression only transiently and on a small scale. Topologically significant genes are complementary to differentially expressed genes and are important in the reconstruction of pathways and networks responsible for a given phenotype. The Causal Reasoning and Overconnectivity analyses help identify both the key direct regulators (i.e., one step away) of the dataset and the major “master” regulators of the global protein network. Concurrent enrichment analysis of both differentially expressed genes and their ranked direct and indirect regulators (Key Hubs), allows the researcher to reconstruct the mechanisms underlining differential gene expression on a more comprehensive level.

In order to understand the functional processes behind the DEG list, *the Key Pathway Advisor* (KPA) uses both approaches mentioned above and combines them in a comprehensive pathway analysis workflow to capture upstream and downstream processes

Chapter 2: KPA browser settings and functionality

KPA is fully supported in Google Chrome, Firefox, Safari and Microsoft Edge. Internet Explorer is no longer supported. The application functions as a ‘one-click’ analysis workflow, where you upload experimental data in an appropriate format and start the analysis using ‘drag and drop’ functionality. Important steps are described below.

Please refer to Appendix A for how to get gene expression datasets from NCBI Gene Expression Omnibus (NCBI GEO) without any data processing, bioinformatics skills or special tools.

2.1 Data sets and file types

- A human gene list with or without expression values (in “fold change” format) is required for KPA analysis. The list can also contain p-values of differential expression (not utilized by analysis). Valid file formats are tab or semi-colon delimited TXT or Excel XLS / XLSX. The following ID types are recognized:
 - EntrezGene (LocusLink) IDs
 - Gene symbol (e.g., TP53, etc.)
 - Affymetrix tag (expression)
 - Illumina tag (expression)
 - Agilent tag (expression)
 - Codelink tag (expression)
 - OMIM
 - RefSeq
 - Unigene
 - ENSEMBL
 - SwissProt
 - GeneBank
 - Panther
 - *MetaCore* gene

The current version of KPA interprets fold change values as differential expression measure and therefore if your file contains both positive and negative values KPA accounts positive as up-regulation (over-expression) and negative as down-regulation (under-expression) of gene comparing to control samples.

If your file contains only positive fold change values, then KPA confirms that you are studying only up-regulated (over-expressed genes). Otherwise, it is assumed that values bigger than zero but less than one indicates down-regulation (under-expression). In this case, such values will be logarithmically transformed (please see Appendix A for details).

- Gene variant data for no more than 150 gene variants are optional to submit to accompany gene list data set. Files should consist of the four columns listed below (in the same order).

- Chromosome
- Position
- Reference allele
- Alternative allele
- Gene variant data can be submitted in three file formats:
 - Gene variant list in TXT format (only four tab-delimited data columns)
 - Standard VCF
 - XLS file generated from data exported using a Genomic Analysis Tool filter (must contain the four columns listed above)

2.2 Settings

The user has the option to change the application's default analysis settings such as

- Ontology selection for Key Processes identification. The default ontologies list includes the following:
 - Pathway maps
 - Diseases
 - Processes networks
 - Pathway groups
- It is also possible to also select the following ontologies for the analysis:
 - GO processes
 - GO molecular functions
 - GO localizations
- If you wish to align your data with Prior knowledge on Putative Biomarkers and/or Drug Targets, then check the appropriate box(es). Then specify the relevant condition (disease) for your experiment to identify any known putative biomarkers and drug targets from your molecular component's gene list. Start typing the condition in the search field and the most relevant MESH terms will be shown.

- The Advanced Settings section allows the researcher to select the p-value threshold for statistically significant processes identification on the enrichment steps of the workflow (default is 0.05). Signaling Pathway Impact Analysis (SPIA) is automatically enabled for datasets with expression changes (fold changes).
- Overconnectivity (also known as interactome analysis in *MetaCore*) analysis can be selected instead of Causal Reasoning (default) analysis for data that associated with fold change values and select p-value threshold for Key Hubs identification (default is 0.01).

2.3 Data analysis process

After the data has been submitted by pressing “Run Analysis” the analysis process begins. An e-mail message containing a direct hyperlink to the online report or PDF / XLS will be sent to the email address associated with your KPA account when the report is complete. If the analysis should fail, you will also receive e-mail containing the error message. The report includes the selected ontology Key Processes list and a list of significant Molecular Entities associated including links to the appropriate *MetaCore* entity pages.

Note: if the default thresholds used for enrichment and Key Hubs analysis are overly strict for your dataset an empty results set may be returned. Entering more relaxed thresholds (like 0.1) manually in the advanced options of the second analysis step and re-submitting the same dataset should resolve this.

2.4 Report panel – example

An example entry from the KPA report panel is shown in Figure 1.

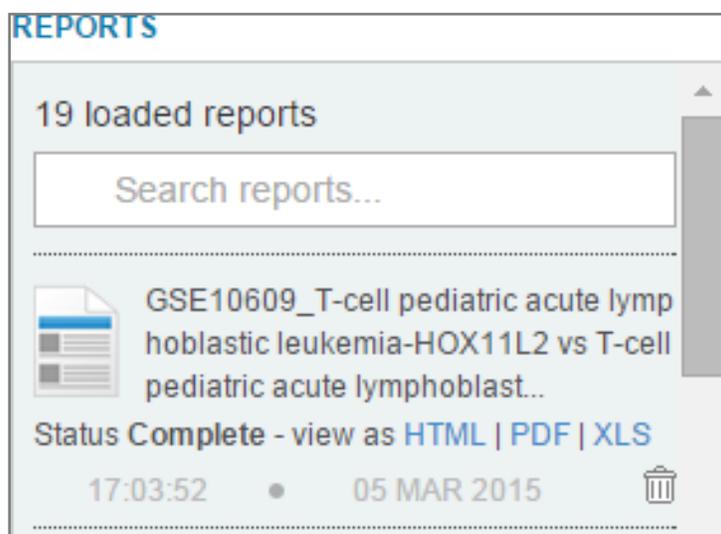


Figure 1: KPA report panel

- Clicking on the report name will open an online version of the report
- To download the report in PDF / XLS format, click on the PDF or XLS links under the report title.
- To delete a report, click on the bin icon in the bottom right-hand corner of the panel.

Chapter 3: Algorithm workflow

Key Processes are defined as ontology terms / entities (i.e., pathway maps) enriched with both input genes and corresponding topologically significant Key Hubs. They are identified by the following workflow (see Figure 2). Enrichment analysis is performed for the list of differentially expressed genes (DEGs) and gene variants (if submitted).

1. Statistically significant ontology entities (enrichment p -value < 0.05) for differentially expressed genes are calculated for your data set. Enrichment is calculated for several proprietary functional ontologies (see Section 2 for details).
2. Predicted Key Hubs (Protein Activity) (see section 4.3.2) are calculated using a Causal Reasoning approach (if DEGs associated with expression values) or Overconnectivity analysis (if DEGs uploaded without expression values). For further analysis statistically significant hubs with p -value < 0.01 are identified.
3. The same enrichment analysis is then performed for the list of predicted Key Hubs calculated in step 2. Statistically significant ontology entities (enrichment p -value < 0.05) for Key Hubs are identified.
4. The list of key hubs calculated in step 2 and your dataset are then combined into a single list and the enrichment analysis run for the union.
5. The final list resulting significant results from steps 1, 3 & 4 are then filtered to show only those where the result for step 4 (the union) is more significant than either of the individual lists (steps 1 & 3) alone.
6. For pathways in the final list a Signaling Pathway Impact Analysis (SPIA) is calculated to identify an impact of differentially expressed genes on the activity of downstream molecules. SPIA aims at the identification of perturbed pathways in a given condition by combining enrichment of perturbed genes in the pathway with the actual amount of perturbation, leading to the most promising candidate pathways and thus candidate genes.

Note: this is known as the synergy method of enrichment analysis. This method was originally designed to allow the comparison of lists which are functionally relevant but poorly overlapping at the gene level, for instance mutated and amplified genes in breast cancer¹. Genes derived from such lists may not overlap directly but populate the very same pathway or process, thus suggesting that they are functionally complimentary. To determine whether two distinct gene lists cooperatively alter a certain cellular pathway or process, we calculate the synergy between them by ontology enrichment. An ontology term (pathway or process) is considered synergistic if the enrichment p -value for the non-redundant union of compared gene lists is lower than p -values for individual lists. More significant enrichment for the union is reflected in the functional connectivity of two gene lists and their complementary effect on the pathway. Ontology Entities that display “synergistic” behavior for the list of differentially expressed genes and the list

of corresponding Key Hubs are defined. The final list of synergistic ontology entities includes all ontology terms with synergistic expression pattern for the union of DEGs and predicted Key Hubs and p-value < 0.05.

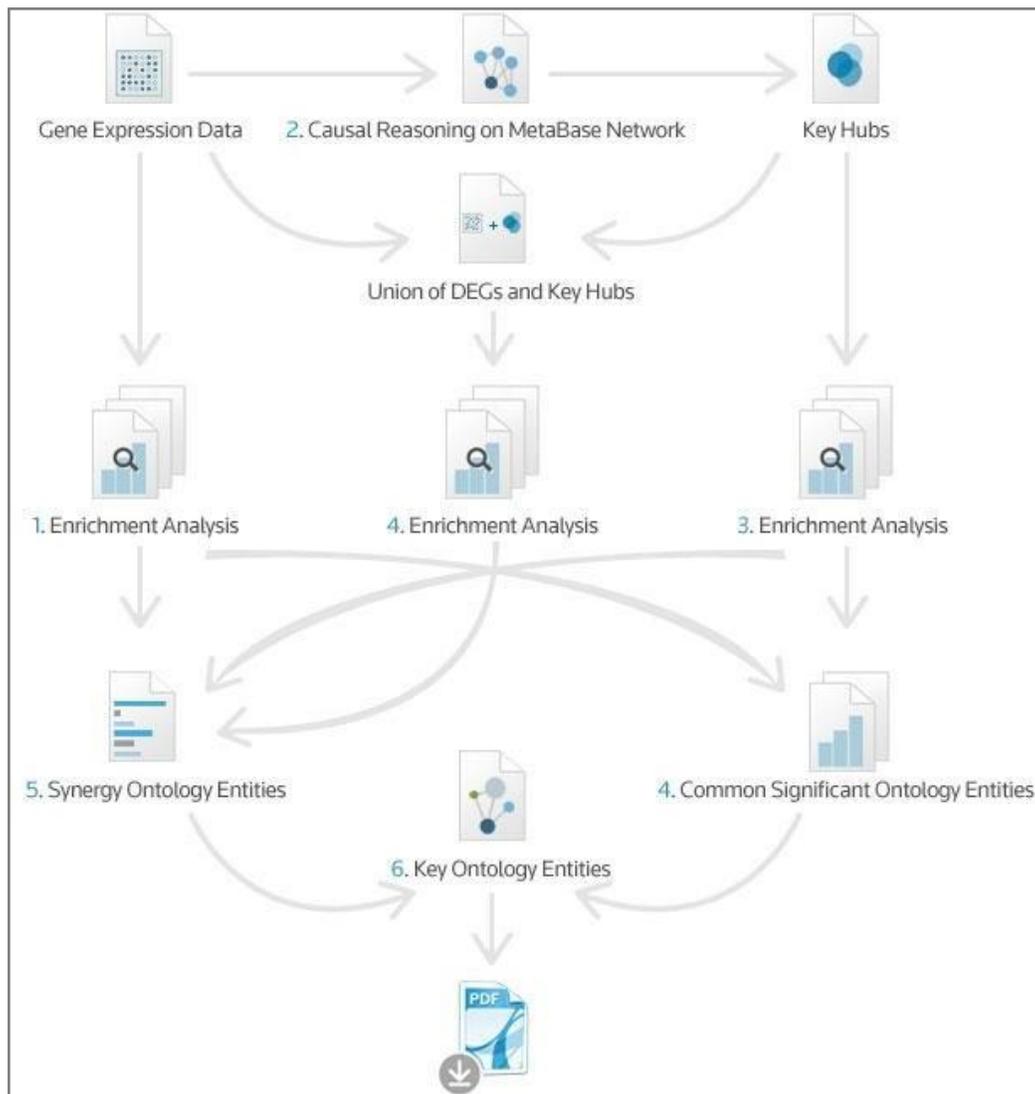


Figure 2: KPA workflow scheme. See description for each step in Section 3(above)

Chapter 4: Analysis report

This section describes how to study a KPA report to explore both up- and downstream analysis directions for your experimental data. Completed analysis reports are available in the Reports panel on the application home page or via the link in the e-mail confirmation.

There are three formats of the report:

1. View report online (HTML) – browse interactive enrichment distributions, pathway maps and significant molecular components. Good if you want to quickly check analysis results, study pathway maps and generate new results by changing analysis settings (e.g. re-submit the data with a broader range of default p-value thresholds).
2. Download a PDF report – shows static lists of enrichment distributions and top ten pathway map pictures overlaid with significant molecular components. Good if you want to show pathway snapshots and distributions to your colleagues.
3. Download an XLS report – contains more expanded lists of molecular components and cross-references between them. Allows you to sort / filter all molecules on the basis of various molecular characteristics narrowing down your data making them appropriate for input into the other system biology tools and workflows.

In general, the report contains four major information categories:

1. Your Data – a list of submitted gene IDs. KPA maps this to the Molecular Entities ontology. Expression values/foldchanges (if uploaded) are shown in the uploaded file.
2. Key Hubs – a list of Molecular Entities (proteins, protein complexes, miRNAs) that a) are predicted to regulate expression of your uploaded DEGs or b) molecules overconnected with the your uploaded genelist, depending on analysis type selected.
3. Key Processes – subsets of genes that are connected with each other performing specific cellular processes (e.g. pathway maps, diseases, interaction networks) that are significantly enriched with your input dataset (differentially expressed genes and optionally genes with genomic variants) and calculated with predicted Key Hubs (molecules that are highly connected with input genes). KPA gives you the ability to study up to seven process ontologies. The final list will depend on your application settings you choose before the analysis starts.
4. Drug Targets (Prior Knowledge) – a list of drugs, from the manually curated *Clarivate Analytics Integrity* database, associated with molecular entities in your dataset and key hubs results for your chosen disease.
5. Putative Biomarkers - indications that connect genes from your experiment and the selected disease taken from the complete set of indications collected in *Key Pathway Advisor* database. Putative biomarkers identify changes in gene characteristics associated with disease manifestation. As these indications could be found for different classes of gene and its products (DNA, RNA, protein isoforms) KPA compares gene changes from your experiment with our data to form a granular overview

4.1 General analysis guidelines and limitations

The following list of key points will give you tips how to manage the data generated in the KPA reports.

- Differentially Expressed Genes (DEGs) indicate gene products where the abundance is changed in examining condition comparing to control. Differential expression does not always indicate active / inhibited status of a protein, but most likely overexpression of ligands, receptors and transcriptional factors can magnify signal transduction that goes through a pathway. The opposite interpretation is valid for under-expressed genes.
- In general, over-expressed genes are usually more significant because they more likely change signal transduction: ligand abundance may start various signaling via various receptors; receptor abundance may act with increased number of ligands; mediator molecules can spread signal by different pathways via cross-talks between them. On the other hand, under-expressed genes, in general, may give less impact on a pathway because signal transduction can go around via pathway cross-talks.
- Key Hub molecules are associated with hypothetical activation/ inhibition status which is calculated to explain experimentally observed differential expression changes. The Key Hubs appearance on a map may fill gaps between differentially expressed genes and increasing interpretation value. Such predictions will always require separate laboratory validation to check whether predicted protein activity changes are indeed occurring.
- Maps are static drawings of a signal transduction 'story' from triggers to effectors (results of transcriptional regulation or entire processes) but sometimes they do not capture every detail. Sometimes they describe the general scheme of the process focusing on specific triggers and effectors but excluding signal transducers e.g., disease processes maps.

4.2 Molecular components analysis

System biology allows a researcher to analyze conditions on the cellular level as a system process. However, it is important to browse through a set of molecular components because not all of them will appear as a Key Process (e.g. Pathway Maps show only well studied canonical processes). KPA also allows you to browse molecular components such as Differentially Expressed Genes and Key Hubs as well as explore congruency with current biomarker and drug target knowledge.

4.2.1 Your data

KPA allows you to upload a gene list with or without associated expression changes between two conditions. For each uploaded Gene ID KPA shows the expression changes from your submitted file and associates them with the official gene name and "Network Molecular Entity".

Molecular Entities in this context refer to proteins, protein complexes, miRNAs, compounds, and drugs that interact with each other in cellular systems.

You can access tissue expression via the link to the Human Protein Atlas project ([proteinatlas.org](https://www.proteinatlas.org)) available from a gene's entity page.

4.2.2 Key hubs

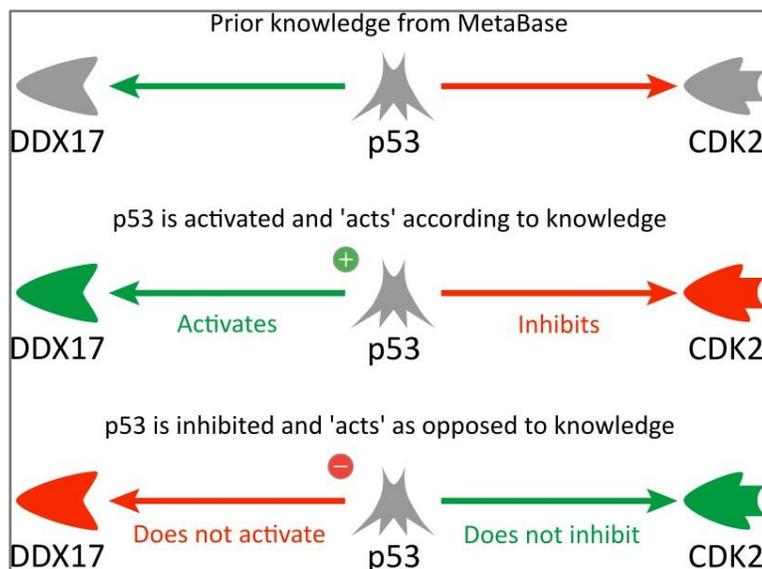
The main goal of Key Hubs (also known as Causal Reasoning) analysis is to identify molecular regulators that most likely directly affect/cause the changes observed in your uploaded expression data. By using this analysis, it is possible to predict potential transcriptional factors and other molecules with potentially altered activity which explains your observed differential expression patterns. Both algorithms for obtaining Key Hubs implemented in KPA can give your insight into transcription regulation causes; however, a Causal Reasoning approach shows the best performance here (see Appendix B) for differentially expressed genes.

The list of predicted Key Hubs is shown on the corresponding tab of your html analysis report, or in Appendix C in the XLS and PDF report versions. You can see the following details in the causal reasoning results:

- Molecular Entity - the name of predicted Key Hub. Each name has a link to an entity page should you require more information. An entity may represent protein, protein complex, regulatory RNA or peptide. Near each name you will see a view button to visualize regulatory network. You can access tissue expression via a link to the Human Protein Atlas project ([proteinatlas.org](https://www.proteinatlas.org)) available from a gene's entity page.
- Gene - name of a gene that corresponds to a Key Hub molecular entity.
- Molecular Function - the type of protein that describes the function of a Key Hub in a cell, e.g. transcriptional factors, kinase, etc.
- Predicted Activity - causal reasoning analysis allows the prediction of Key Hub potential activity changes that potentially explain the transcriptional changes observed in your uploaded data. This is possible because each Molecular Entity has outgoing activation and inhibition interactions to other objects (Figure 3) in the database. For example, a Key hub with predicted increase in activity shows increased expression for those objects it is known to activate and decreased for those it is known to inhibit.

An example of the Molecular Entities for KPA is shown in Figure 3.

Figure 3: Basic description of Causal Reasoning prediction based on interpretation of the prior knowledge about possible molecular interactions. Arrows indicate prior knowledge molecular interactions stored in the database. According to the prior knowledge p53 increases expression of DDX17 and decreases expression CDK2. Though, if in experimental gene expression data DDX17 and CDK2 have increased and decreased expression correspondingly we can assume that p53 might have increased activity. On the contrary, if DDX17 has decreased expression level and CDK2 is increased we can assume that p53 is inhibited. A statistical approach is applied to measure of significance of such prediction (see Methods section).



- Supportive/total data - all DEGs which are known to be regulated by a Key Hub according to Information from the database. Supportive data represent DEGs which fold change direction (+ or -) in your uploaded dataset aligns with the expected direction of change according to the interactions associated with the Key Hub and that molecule (see Figure 5 & Appendix B).
- P-value – This is calculated using a binomial test to assess the probability of making given number of supportive data out of all defined DEGs in your data file. The more consistent supportive data to actual data, the better. This test is adopted from Pollard et al's method³.
- Path length - indicates a maximal length of a shortest path (number of steps) between Key Hub and DEGs. InKPA Key Hubs may currently be located a maximum in three steps from DEGs.
 - A Key Hub located within one step regulates expression of correctly predicted DEGs directly (usually these are transcriptional factors).
 - A Key Hub located within two steps regulates transcriptional factors that influence DEG expression.
 - A Key Hub within three steps regulates regulators of transcriptional factors influencing DEG expression.

As an alternative analytic the overconnectivity test identifies direct regulators of the dataset that are one step remote and statistically overconnected with the objects from the dataset. The method is based on an assumption that proteins functionally important for a particular phenotype have relatively a lot of interactions with proteins encoded by genes altered (e.g., differentially expressed) in the phenotype. This allows for the assessment of the level of connectivity that individual objects in the database have with the genes from the analyzed gene list and whether this level of connectivity is more than would be expected purely by chance. This analysis does not use any of the fold change data uploaded in your data file and therefore is the default option for gene lists uploaded without data.

The following columns are shown for analysis results:

- Key Hub name - a name of a Molecular Entity predicted to have an activity change that may explain expression changes observed in your data. Each name has a link to a *MetaCore* entity page for more information.
- Molecular function - type of protein that describes a function of a Key Hub in a cell, e.g. transcriptional factors, kinase, etc.
- Correct predictions - predictions are DEGs that connected are regulated by Key Hub according to algorithm. Correct predictions are a number of all predicted connections minus all observed ones.
- Significance p-value – hypergeometric p-values (see Methods section).

4.2.3 Key hubs networks

Understanding a nature of gene expression pattern change under a condition (i.e. disease) is a crucial step in molecular biology research. *Key Pathway Advisor* utilizes previously published studies about transcriptional gene regulation and protein activity on a pathway level to identify Key Hubs – molecules which most probably caused over and under expression of genes in your data. Knowledge of what Key Hub is and what regulated genes do allows identifying the Hub as a disease driving gene and promising drug target. A collection of analyzed published studies that are focused on which transcriptional factor activates or suppresses transcription of specific genes as well as which protein activates transcriptional factors is a cornerstone piece required for the analysis. All molecular entities are connected with interactions that could be activating (highlighted by green), inhibitory/suppressive (red) or interactions for which it is hard to identify final effect (gray). All interactions are associated with corresponding reference(s) to sourcing scientific publications available from the Table view.

Key Hub is a molecule that is connected with a significant subset of genes from your data. Just like a puppeteer manipulates puppets by touching strings a Key Hub is predicted to manipulate gene expression by activating and inhibiting transcriptional factors or affecting gene expression directly.

Green (+) sign icon indicates that a Key Hub should exist in predominantly **activated** state in order to regulate related gene expression according to your data; red (-) icon indicates that a Key Hub should exist in predominantly **inhibited** state in order to regulate related gene expression pattern according to your data. The nature of predominant inhibition could be different depending upon the condition being studied. That might be a gene variant in this gene or a corresponding signal goes from upstream pathway (due to ligand binding, phosphorylation, etc). You can study such reasons by Key Pathway maps analysis on a corresponding tab of this report.

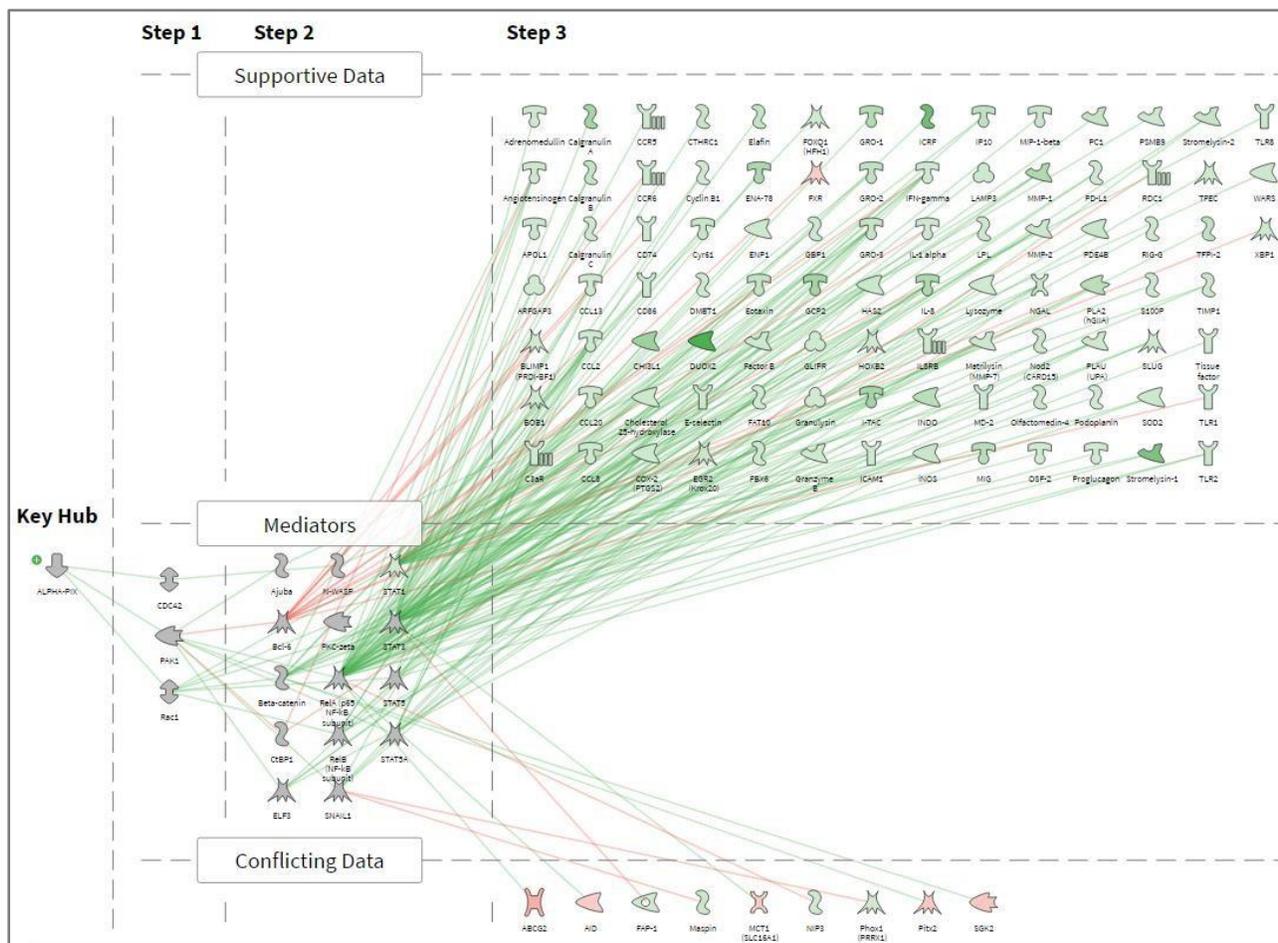


Figure 4: Key Hub regulatory network visualization. See explanation in the text below.

Supportive data panel contains over and under-expressed genes from your data set which support a hypothesis that a Key Hub is in predicted predominant state. Conflicting data panel contains over and under-expressed genes from your data set which are discordant with a hypothesis that a Key Hub is in predicted predominant state (Figure 4). The nature of that discordance may combine several reasons: epigenetic changes of different nature which are not identifiable by gene expression analysis methods or simply concurrent action of other transcriptional factors (which might be different Key Hubs). Different shades of green and red indicate over and down-expressed genes from your data respectively. Grey indicates proteins with unchanged expression. By analyzing which genes are over and down-expressed it is possible to identify a key hub as a disease driver and promising drug target.

Mediators panel contains other transcriptional regulators employed by a Key Hub to concordantly regulate expression of genes from your data. Key Hubs might activate or inhibit co-regulators by many mechanisms like binding, complex formation, post-translational modifications and more.

As this diagram visualizes a pathway or cascade all of the genes are placed several steps away from the Key Hub to make it more intuitive for interpretation. The word 'Step' here is directly related to a step on a pathway map diagram and indicates a chemical reaction, protein binding or factor binding to a DNA regulatory sequence. One step means that genes are directly related with the Key Hub. Key Hub directly affects expression of genes from Supportive and Conflicting panels or interacts with other Mediators. Genes located in two or three steps away from the Key Hub are not directly regulated by it but through mediators.

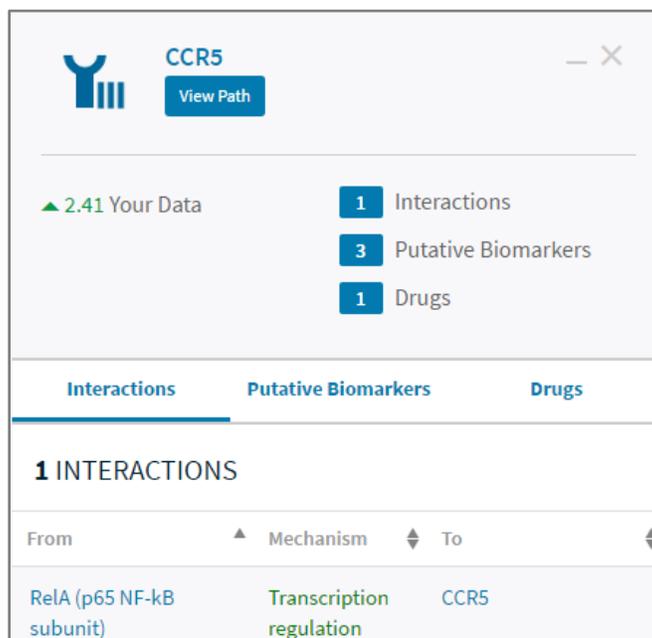


Figure 5: A pop-up summary information about gene selected on the network. See explanation in the text below.

By clicking on a molecule, you select it and a path from the Key Hub is highlighted (Figure 11). A pop-up with summary information appears to give more data about the molecule. If additional data are available, it shows expression change from your data, interactions on the network, associated drugs and putative biomarkers for the condition (details are available by the click). View Path button will take highlighted pathway and visualize in more details giving access to molecular interactions overview (Figure 6). This pathway shows exact molecules binding and modification that leads to expression changes. It is useful when browsing all transcriptional factor targets or studying how multiprotein complexes influence an expression of a single gene from your data. Click Back to Network to return to the initial diagram.

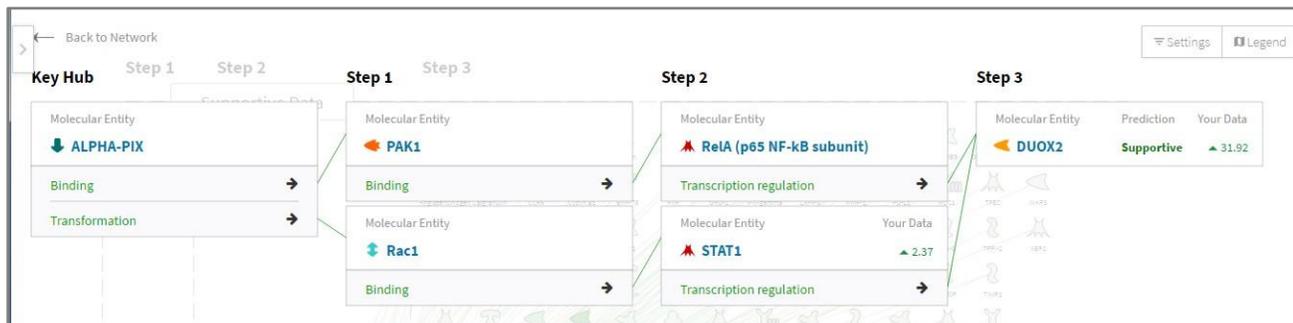


Figure 6: Diagram shows a path that connects the Key Hub and selected molecule. Each interaction is connected to corresponding entity page with more experimental details about it.

Click on the Settings to access a few more useful options (Figure 7).

1. Full Color Style checkbox will convert visualization into a style when an icon color also indicates a molecule function and gene expression changes are shown in green and red sectors around each molecule. Increased expression value corresponds to the green sector which size increases clockwise around the molecule icon. Decreased expression value corresponds to the red sector which size increases counterclockwise.



2. Connectivity Size will change size of molecules accordingly to the number of interactions they have on this diagram. Using a pathway analogy here we speak about avenues and interchanges that connect the Key Hub with differentially expressed genes. The bigger the molecule, the more important interchange it is.

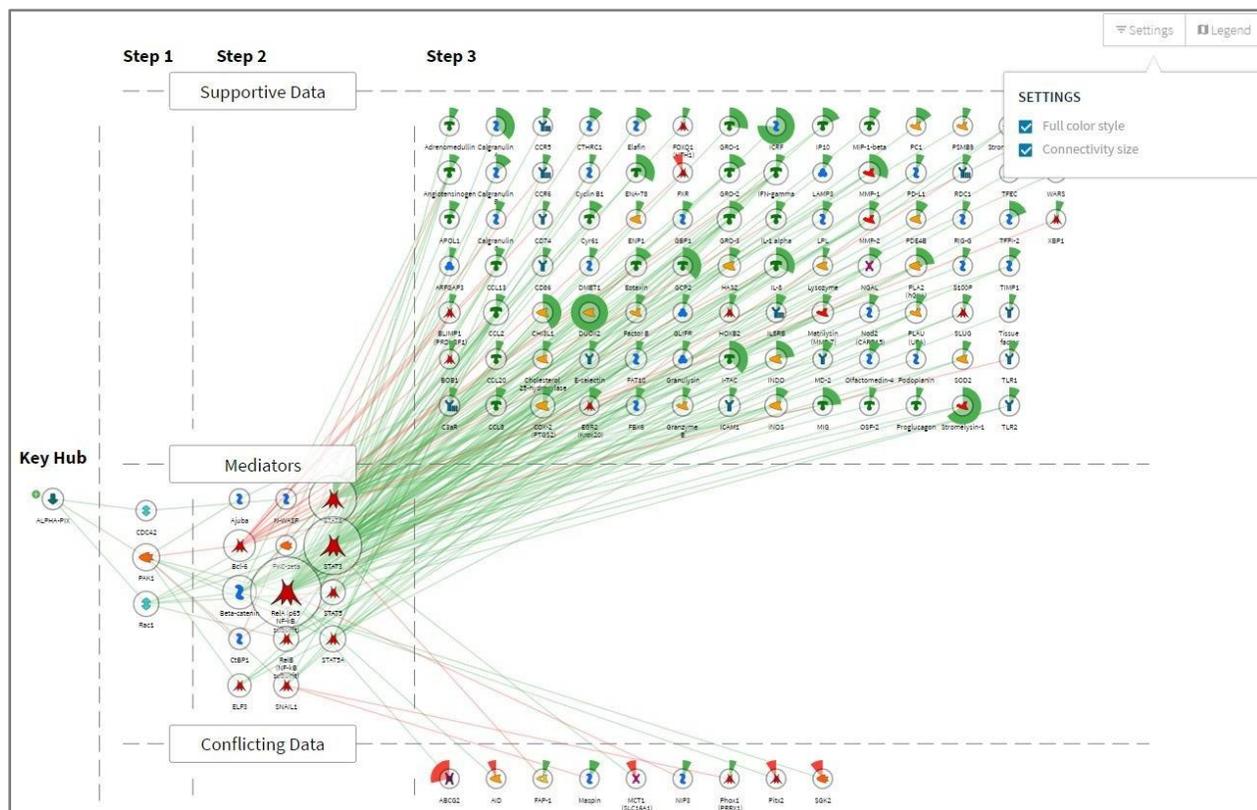


Figure 7: Key Hub regulatory network visualization. Full color style and Connectivity size options are applied.

You can browse pathway content in table view by clicking on the corresponding button in the top left corner of the page (Figure 8). Molecular entities table lists all molecules shown on the map associated with your experimental data, Key Hub information, disease causal associations and drug target data. Clicking on putative biomarkers or drug targets cells for each molecule will open a pop up with summary data and links to corresponding entity pages. You can click on 'View Interactions' arrow for a molecule to visualize its interactions from network on the left side of the page.

Molecular Entities							Interactions		
Name	Your Data	Hypothesis Basis	Steps	Putative Biomarkers	Drug Targets	View Interactions	FROM	Mechanism	TO
CHI3L1	▲ 13.54	Correct	3	●●	-	→	Ajuba	co-regulation of transcription	LPL
Cholesterol 25-hydroxylase	▲ 2.59	Correct	3	-	-	→	ALPHA-PIX	Transformation	CDC42
COX-2 (PTGS2)	▲ 2.95	Correct	3	●●●	✓	→	ALPHA-PIX	Binding	PAK1
CTHRC1	▲ 4.02	Correct	3	-	-	→	ALPHA-PIX	Transformation	Rac1
Cyclin B1	▲ 2.03	Correct	3	-	-	→	Bcl-6	Transcription regulation	BLIMP1 (PRDI-BF1)
Cyr61	▲ 4.1	Correct	3	-	-	→	Bcl-6	Transcription regulation	CCL3
DMBT1	▲ 2.46	Correct	3	●●●	-	→	Bcl-6	Transcription regulation	CCL2
DUOX2	▲ 31.92	Correct	3	-	-	→	Bcl-6	Transcription regulation	CCL8
E-selectin	▲ 2.51	Correct	3	●●●	✓	→	Bcl-6	co-regulation of transcription	CCR6
EGR2 (Krov20)	▲ 3.78	Correct	3	-	-	→	Bcl-6	Transcription regulation	CD74

Figure 8: Table that contains all molecules and connecting interactions represented on the network. The following data are shown for molecules if available: predicted activity state, expression change value from your data, putative biomarkers and drugs for disease selected in the analysis (each line is interactive, and details are available by click). Filters allow selecting molecular entities depending on drug target status and biomarker match, interactions on the basis of effect and mechanism.

4.2.4 Analysis guidelines and limitations

- Each list (DEGs, Key Hubs) contains a cross-reference with each other which may give you insight into cross-talk between biological layers.
- Differentially expressed genes are usually the direct output from an experiment and to investigate a particular set of conditions or treatments and the resulting aberrant signaling. When genes found to be changing in expression are also predicted to be a Key Hub – a molecule which drives aberrant gene expression- it may therefore indicate a particularly critical molecule or driver of your phenotype (further lab validation would potentially be required however to confirm this hypothesis).
- A Key Hub which is associated with activated / inhibited status is a valuable hypothesis that explains gene expression changes. If such Key Hub has a non-synonymous gene variant – it could be additional proof of a hypothesis about that Key Hub’s importance and gain / loss of function for a gene variant. e.g. if a Key Hub with a gene variant is active then it may be a gain of function mutation and if a Key Hub is inhibited than it is potential loss of function for associated gene variant.
- The Key Hubs calculation is a hypothesis generation process that is likely to require further laboratory validation. Gene expression may be caused by multiple factors (including changed expression via Copy Number Variations, promoters’ methylation, miRNA action, complex and dimerization ratios, etc.) that may need to be investigated as the potential source of aberrant expression or Key Hub activity. Since availability of a complete set of all these data layers is rare for the majority of experiments it is possible to use the Key Hub hypotheses as a starting point for further investigations.

The XLS report allows you to view more detail on the results, and on the data you uploaded, like expression change values, Key Hubs and gene variants intersections, Key Processes counts where each DEG appeared (in what number of processes a molecular component appears), in a single place. Accumulating sorting and filtering strategies for all these files you can specify a set of initial Gene IDs which are worthy to study further by more precise experimental methods in your study pipeline. This format also allows you to input the data into other software solutions/workflows for further analysis.

4.3 Key processes analysis

The Key Processes page is used to view results of the downstream analysis on cellular processes which have been enriched by experimental data and the Key Hubs. You can see a separate tab for each major process ontology specified when the data was submitted. The tab contains the most relevant processes list associated with p-value of statistical significance (see Appendix A to find the intuitive description of p-value concept).

Usually, enrichment analysis is made for an input gene set (e.g., differentially expressed genes - DEGs), but its value could be leveraged by adding Key Hubs to the analysis and using a synergy approach (see Section 3). This will define processes that are significant not only for DEGs (results of aberrant signaling) but also considering Key Hubs and gene variants information (both supposed to represent original aberrant signaling itself).

All processes are grouped into specific categories called functional ontologies for KPA analysis (Table 1). Default search categories are listed in bold; the remaining categories should be set manually when analysis starts.

Table 1: Functional ontology search categories

Pathway map	Pathway maps are graphic images representing complete biochemical pathways or signaling cascades in a commonly accepted sense.
Physiological pathway maps	A pathway maps subset that shows only pathways or processes as they are in normal condition. The subset is useful for analysis as it allows tracing normal pathways changed by genes identified from your experimental data.
Pathological pathway maps	A pathway map subset that contains only pathways and processes specifically changed under various disease conditions that might give you a clue to your experimental condition interpretation via similarity.
Pathway groups	This is a collection of manually created pathway maps, grouped hierarchically according to main biological processes.
Processes networks	A recognized series of events (interactions or biochemical reactions) accomplished by one or more ordered assemblies of molecular functions with a defined beginning and end.
Diseases	This ontology is created based on the classification in Medical Subject Headings (MeSH). Each disease in diseases ontology has its corresponding biomarker gene or set of genes
GO processes	A GO ontology for biological processes. The processes are structured as hierarchical tree with branches defined according to the Gene Ontology controlled vocabulary. GO process folders are nested, i.e., each folder references all the proteins participating in its sub-processes
GO localizations	A GO ontology for localization of the gene products inside or outside the cell. A given molecule in a given localization is represented by a Molecular Entity in <i>MetaCore</i> .
GO molecular Functions	A GO ontology of hierarchically structured molecular functions. A protein may be linked to several different molecular functions.

Key processes are associated with three types of p-value for enrichment:

1. Input data only (default in PDF report) – sorting by this p-value is useful if you prefer to rely mostly on your experimental gene list as statistical significance indicator.
2. Key Hubs – sorting by this p-value is useful if you want to study processes more significant for topologically significant genes.
3. Union of input data and Key Hubs – sorting by this p-value gives you the most biologically relevant ranking.

4.3.1 Key processes analysis

The Key Processes page is used to view results of the downstream analysis on cellular processes which have been enriched by experimental data and the Key Hubs. You can see a separate page for each major process ontology specified when the data was submitted. The page contains the most relevant processes list associated with p-value of statistical significance (see Appendix A to find the intuitive description of p-value concept).

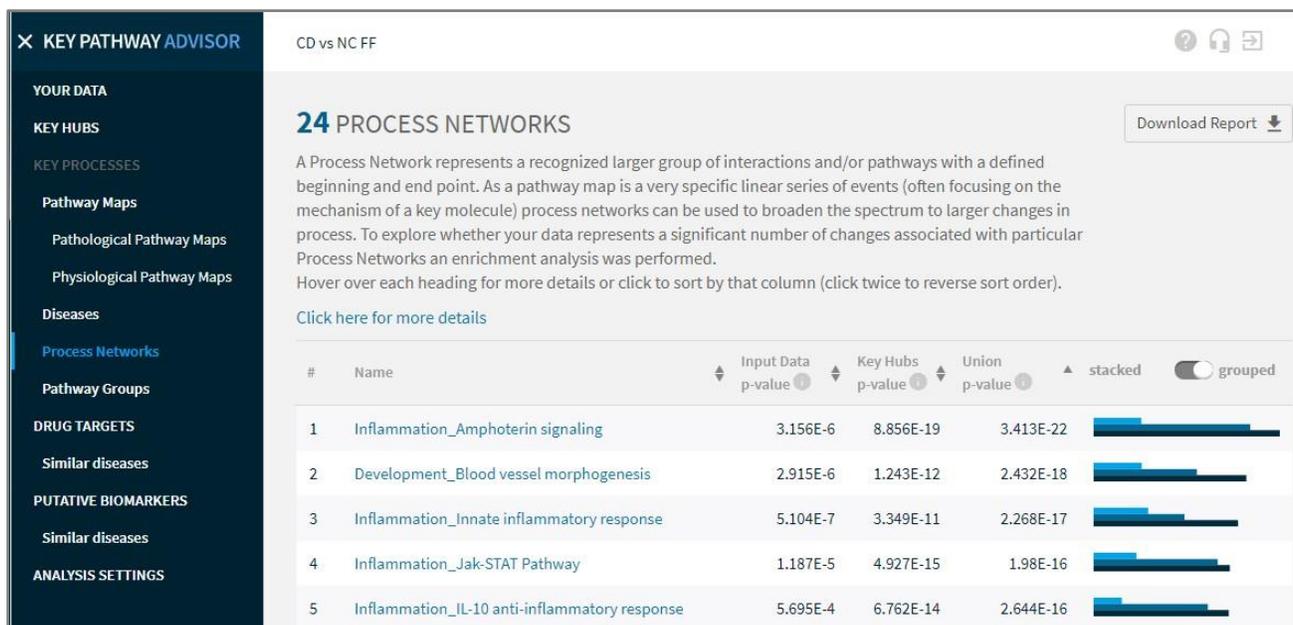


Figure 9: Key Process tab example

Each Key Process is associated with three enrichment p-value calculations.

- Input objects (your uploaded gene list with / without gene variants)
- Key Hubs (as per Predicted Key Hubs (Protein Activity) analysis results)
- The union list of input data and Key Hubs

Each set of p-values is also represented a set of colored bars in the far right hand column. The length of each bar displays the $-\log(p\text{-value})$, giving a more visual representation of the data. The bigger the bar is, the more significant the p-value.

4.3.2 Pathway maps

Pathway maps (Figure 10) are the most intuitive way of downstream analysis. You can study graphical visualization and text descriptions online and in a PDF report (only for top ten the most significant maps). KPA utilizes Signaling Pathway Impact Analysis (SPIA) to identify how differentially expressed genes from your dataset affect downstream molecular activity (see Methods section). Key Pathway Maps tab differs if SPIA calculation was activated (Figure 10).

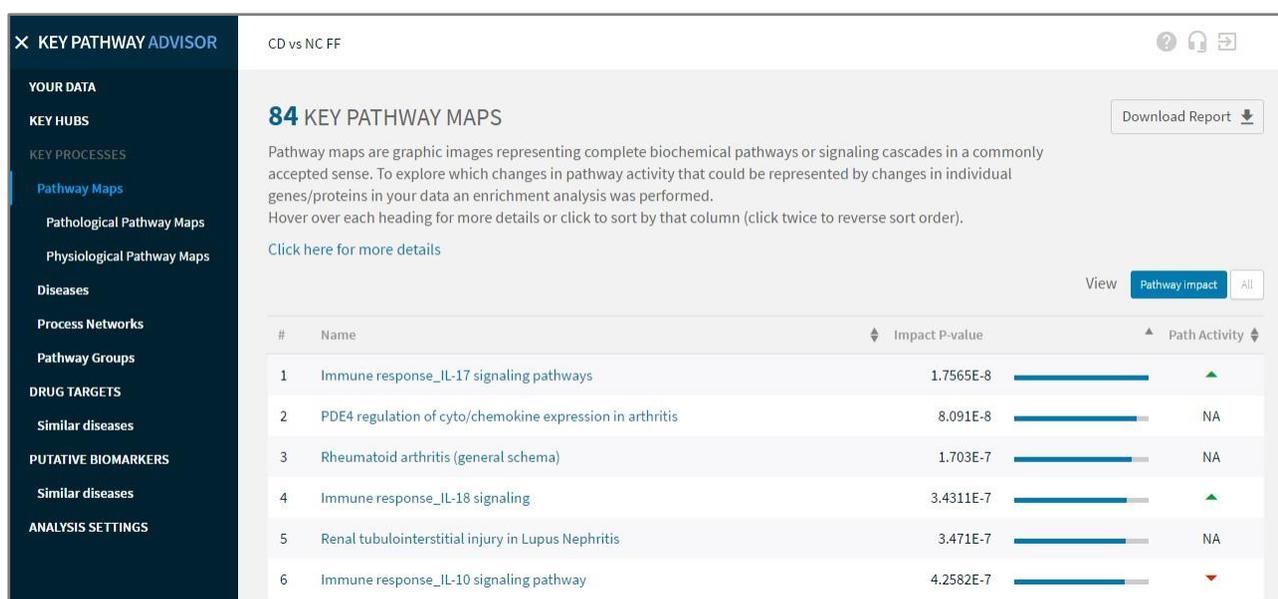


Figure 10: Key Pathway Maps tab example with SPIA analysis. Impact P-value indicates significance of pathway overlay with DEGs from input list as well as how differential expression impacts downstream molecular activity. Path activity shows if a pathway or a sub path was activated or inhibited on the basis of differential expression. NA highlights pathways for which it was impossible to identify activity impact (predominantly diagrams of pathological processes).

Maps represent canonical signal transduction pathways and overlaid with DEGs, Key Hubs (green (+) sign representative key hub from causal reasoning and hubs from over-connectivity; red (-) indicate inhibited hubs from causal reasoning). By default, different shades of green and red indicate over and down-expressed genes from your data respectively. Grey indicates proteins with unchanged expression.

Each pathway map is fully interactive. By clicking on a molecule, a pop-up with summary information appears to give more data about it (Figure 11). If additional data are available (only if a molecule is a DEG or a Key Hub) it shows expression change from your data, associated drugs for the similar condition and putative biomarkers for the condition (details are available from the table view). It also shows molecule perturbation factor calculated by SPIA algorithm

which consists of differential expression of a molecule combined with upstream activation and/or inhibition effects from other molecules (see Methods section). Molecule name is a link to dedicated entity page. All paths which go through this molecule from the map upstream and one step downstream will be highlighted.

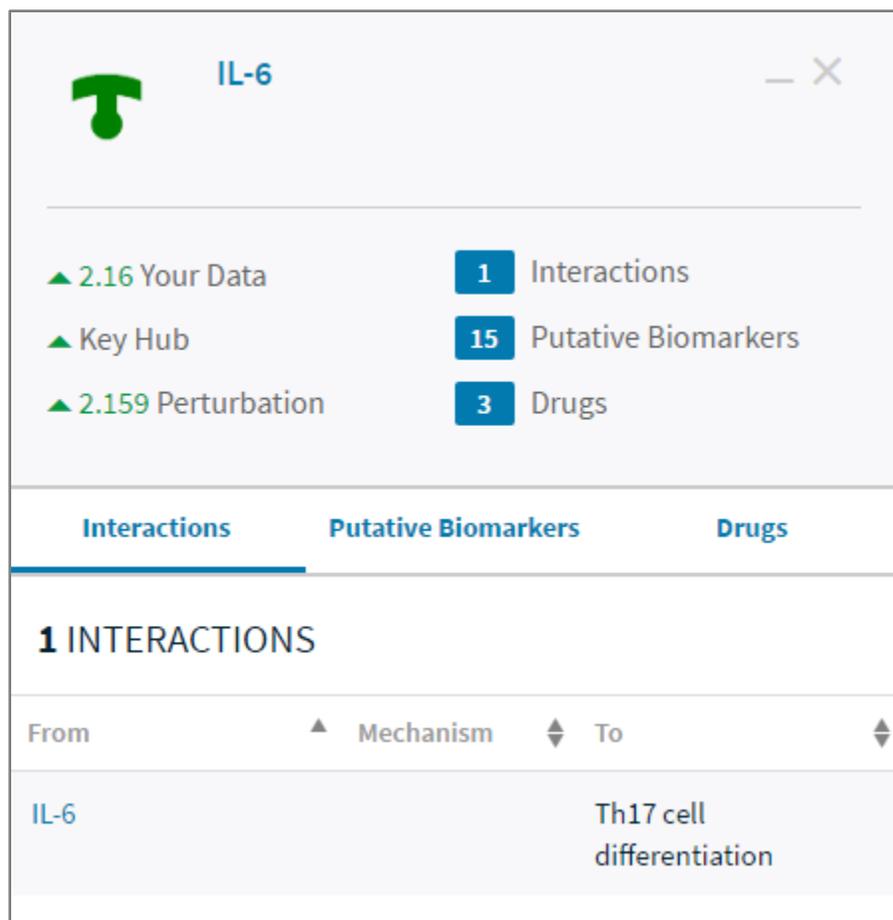
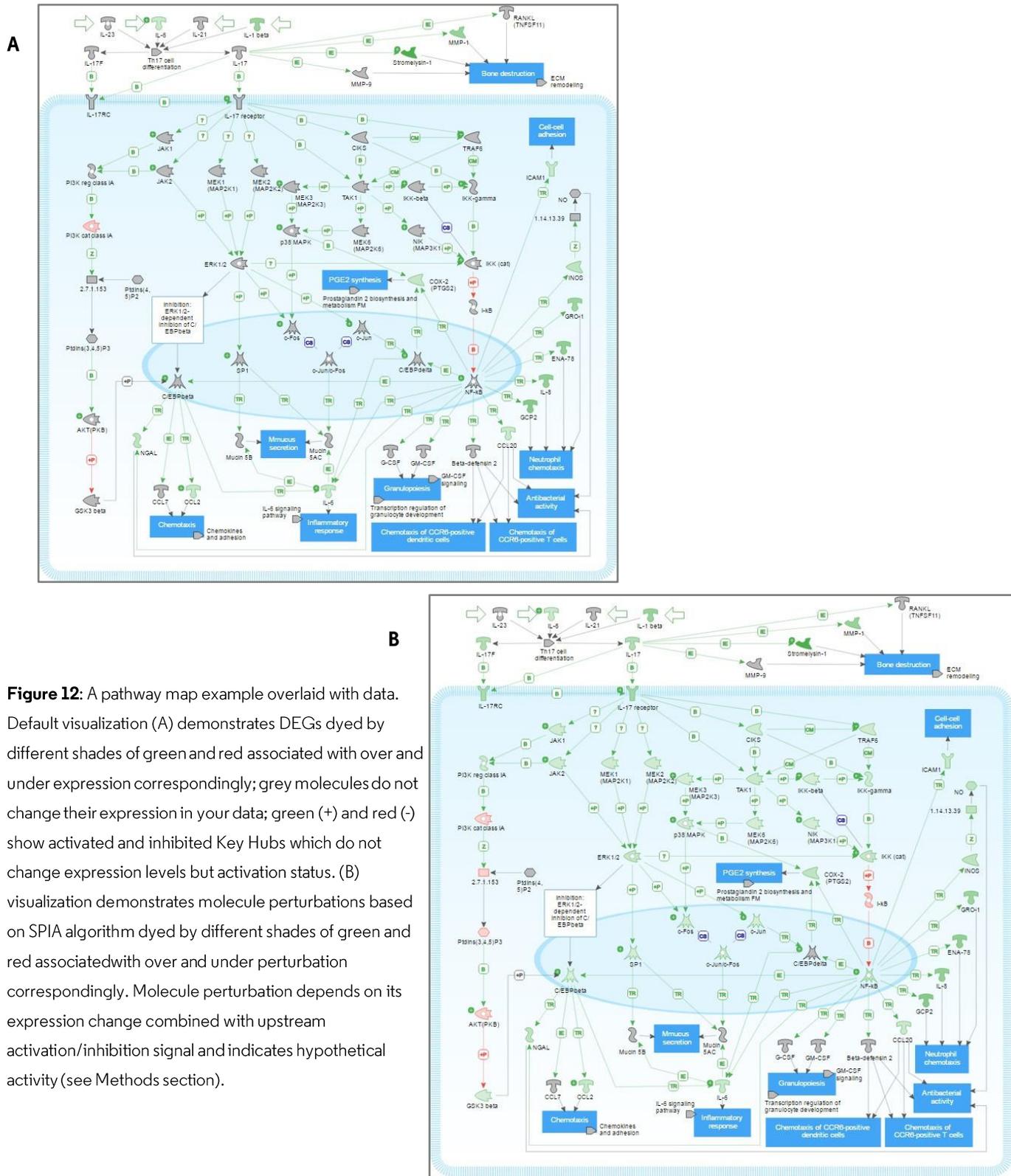
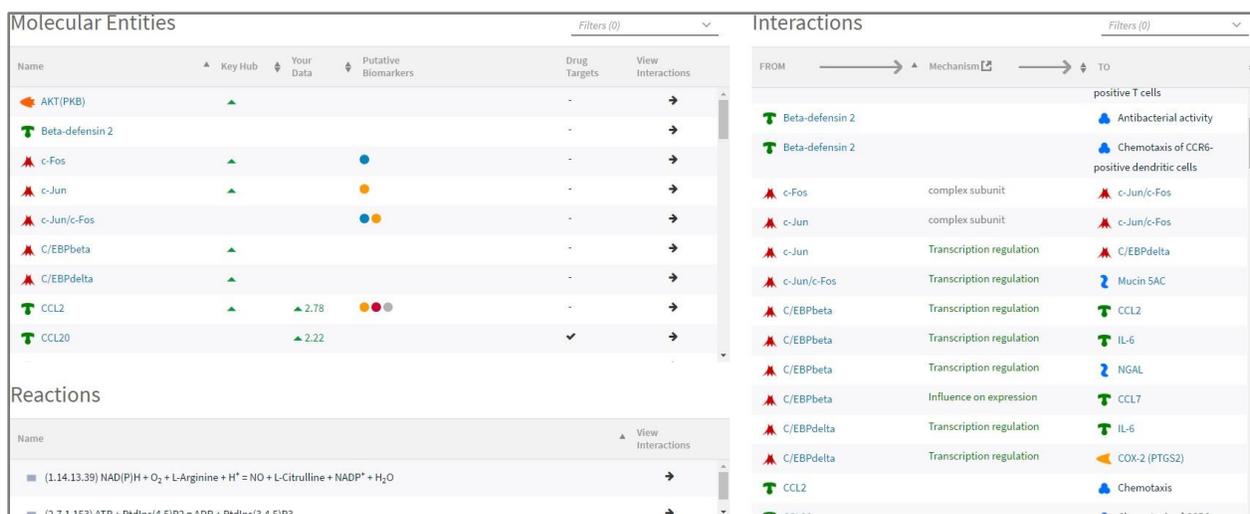


Figure 11: A pop-up summary information about gene selected on the network. See explanation in the text above.

Note: overconnectivity analysis output (Predicted Key Hubs (Protein Activity) tab in Molecular Analysis section) can also be used to carry out downstream analysis for Key Hubs where the molecular function is not miRNA or transcriptional factor. These Key Hubs may connect modules of interacting proteins and probably are associated with disease manifestation.



You can browse pathway content in table view by clicking on the corresponding button in top left corner of the page (Figure 13). Molecular entities table lists all molecules shown on the map associated with your experimental data, KeyHub information, putative biomarkers and drug target data. Clicking on putative biomarkers or drug targets fields for each molecule will open a pop up with summary data and links to corresponding entity pages. You can click on ‘View Interactions’ arrow for a molecule to visualize its interactions from the map on the left side of the page.

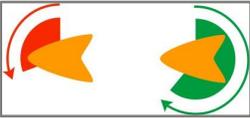


The screenshot shows two side-by-side tables. The left table is titled 'Molecular Entities' and has columns for Name, Key Hub, Your Data, Putative Biomarkers, Drug Targets, and View Interactions. It lists molecules like AKT(PKB), Beta-defensin 2, c-Fos, c-Jun, c-Jun/c-Fos, C/EBPbeta, C/EBPdelta, CCL2, and CCL20. The right table is titled 'Interactions' and has columns for FROM, Mechanism, and TO. It shows interactions between molecules like Beta-defensin 2, c-Fos, c-Jun, c-Jun/c-Fos, C/EBPbeta, and CCL2, with mechanisms such as 'Transcription regulation' and 'Influence on expression'.

Figure 13: Table that contains all molecules and connecting interactions represented on the network. The following data are shown for molecules if available: predicted activity state, expression change value from your data (might have several values in case if molecular entity is a complex or group of molecules), putative biomarkers and drugs for disease selected in the analysis (each line is interactive, and details are available by click). Filters allow selecting molecular entities depending on drug target status and biomarker match, interactions on the basis of effect and mechanism.

Click on the Settings to access a few more useful options.

- Your data (fold change) option highlights DEGs on the pathway map. Green molecules have increased expression, red – decreased.
- Molecule perturbation (SPIA) option highlights molecules with predicted activity changes caused by upstream DEGs (see Methods section). Green molecules have increased activity due to upstream signaling, red have decreased. Comparison of upstream activation predicted by SPIA and downstream activation prediction by Causal Reasoning gives an independent evaluation of molecular activity, especially if both predictions match.

- Full Color Style (Figure 14) checkbox will convert visualization into a style when an icon color also indicates a molecule function and gene expression changes are shown in green and red circles around each molecule. Increased expression and positive molecule perturbation value correspond to the green sector which size increases clockwise around the molecule icon. Decreased expression and negative molecule perturbation value correspond to the red sector which size increases counterclockwise.
 
- Connectivity Size will change the size of molecules accordingly to the number of interactions they have on this diagram. Using a pathway analogy here we speak about avenues and interchanges that connect the Key Hub with differentially expressed genes. The bigger the molecule, the more important interchange it is.

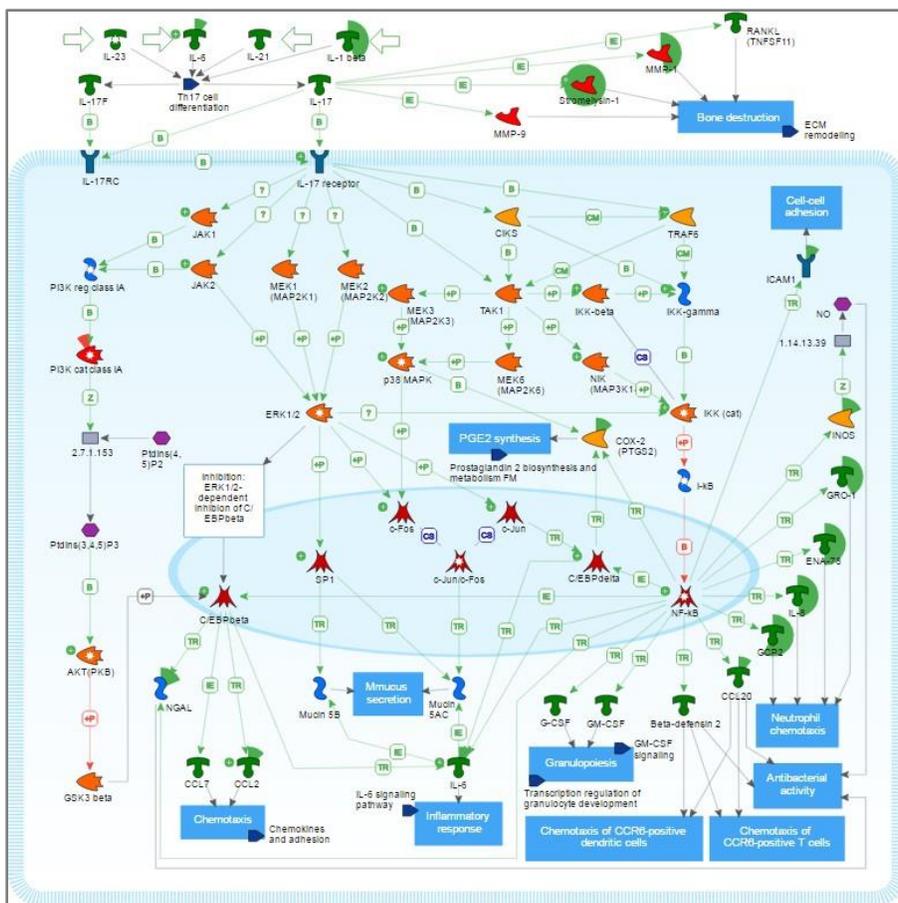
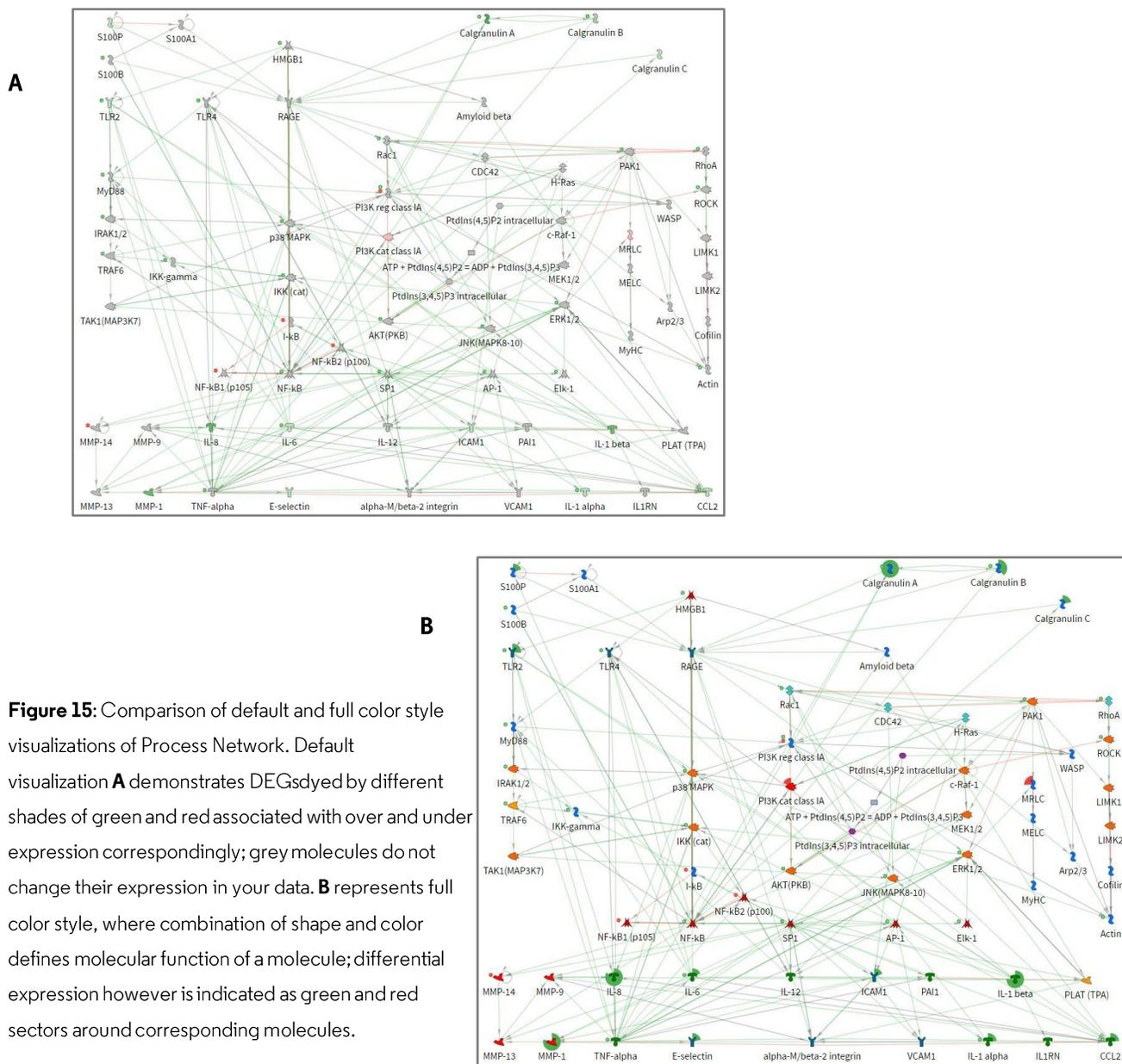


Figure 14: Pathway map visualization with full color style option activated.

4.3.3 Process networks

Process networks visualize cellular processes from more global perspective. Instead of focusing on specific pathway a process network visualizes the more complex manually curated relationship between several pathways involved in regulation of the same process.



All molecular entities are connected with interactions that could be activating (highlighted by green), inhibitory/suppressive (red) or interactions for which it is hard to identify final effect (gray). All interactions are associated with corresponding reference(s) to sourcing scientific publications available from the Table view.

Networks are overlaid with DEGs, Key Hubs (green (+) sign represents active key hub from causal reasoning and hubs from over-connectivity; red (-) sign indicates inhibited hubs from causal reasoning). By default, different shades of green and red indicate over and down-expressed genes from your data respectively. Grey indicates proteins with unchanged expression. This visualization differs from pathway maps as process networks usually contain significantly bigger number of objects which may decrease intuitive identification of genes from your data (Figure 15).

Each network is fully interactive and allows nodes rearrangement. By clicking on a molecule, a pop-up with summary information appears to give more data about it (Figure 16). If additional data are available (only if a molecule is a DEG or a Key Hub) it shows expression change from your data, associated drugs and discovery biomarkers/causal associations for the condition (details are available from the table view). Molecule name is a link to dedicated entity page.

Figure 16: A sample pop-up summary info which appears after you click on a molecule. In this example, Calgranulin A had an increased expression, FoldChange=12.11, in your uploaded dataset. It was predicted as a Key Hub in a predominantly active state, and is a putative biomarker for the selected disease. All details regarding interactions on the network and putative biomarkers are available from a table below. A protein name is a link that leads to corresponding molecule entity page.


Calgranulin A
— ✕

▲ 12.11 Your Data

▲ Key Hub

5 Interactions

4 Putative Biomarkers

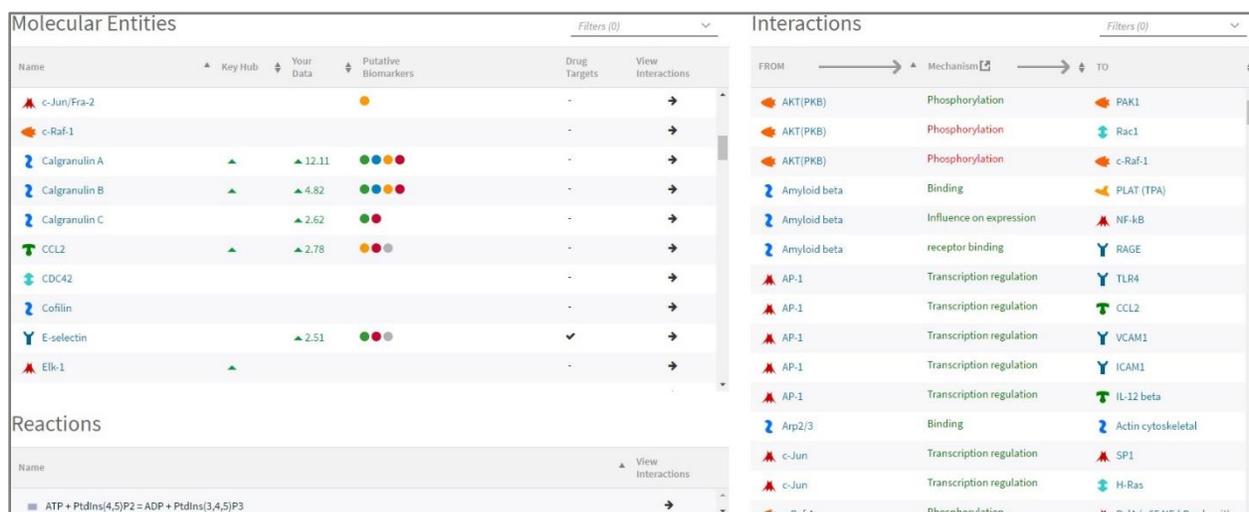
0 Drugs

Interactions
Putative Biomarkers
Drugs

5 INTERACTIONS

From	Mechanism	To
Calgranulin A	Unspecified	NF-kB
Calgranulin A	Binding	Calgranulin B
Calgranulin A	receptor binding	RAGE
Calgranulin B	receptor binding	Calgranulin A
TNF-alpha	Influence on expression	Calgranulin A

You can browse network content in table view by clicking on the corresponding button in top left corner of the page (Figure 17). Molecular entities table lists all molecules shown on the map associated with your experimental data, KeyHub information, putative biomarkers and drug target data. Clicking on causal associations or drug targets cells for each molecule will open a pop up with summary data and links to corresponding entity pages. You can click on ‘View Interactions’ arrow for a molecule to visualize its interactions from the network on the right side of the page.



The screenshot shows two side-by-side tables. The left table, titled 'Molecular Entities', has columns for Name, Key Hub, Your Data, Putative Biomarkers, Drug Targets, and View Interactions. It lists molecules like c-Jun/Fra-2, c-Raf-1, Calgranulin A, B, and C, CCL2, CDC42, Cofilin, E-selectin, and Elk-1. The right table, titled 'Interactions', has columns for FROM, Mechanism, and TO. It shows interactions such as AKT(PKB) leading to Phosphorylation of PAK1, Rac1, and c-Raf-1; Amyloid beta leading to Binding of PLAT (TPA) and Influence on expression of NF-κB; and AP-1 leading to Transcription regulation of TLR4, CCL2, VCAM1, ICAM1, and IL-12 beta.

Figure 17: The table that contains all molecules and connecting interactions represented on the network. The following data are shown for molecules if available: predicted activity state, expression change value from your data (might have several values in case if molecular entity is a complex or group of molecules), putative biomarkers and drugs for disease selected in the analysis (each line is interactive, and details are available by click). Filters allow selecting molecular entities depending on drug target status and biomarker match, interactions on the basis of effect and mechanism.

4.4 Drug targets (prior knowledge)

The Drug Targets tab gives an overview of known drug targets identified in your experiment/results. If a gene or key hub is a target for one or more of *Clarivate Analytics Integrity* drugs (between preclinical or launched phases) it will be shown with the corresponding phase.

There are two pages:

- Drugs associated with disease / condition specified in analysis start settings.
- Drugs that are associated with similar diseases to the condition previously specified according to MESH ontology. The list might indicate repurposing candidates.

The Under Active Development (UAD) label appears on records for drugs that are actively moving through the drug R&D pipeline from preclinical stages through registration (i.e., preclinical, IND filed, phase I, phase I/II, phase II, phase II/III, phase III, preregistered, recommended approval, and registered). The criteria for labeling a compound Under Active Development include the following, which must have occurred over the last 12 to 18 months:

- a) The company is actively informing the public on the development of the product via press releases, mention in annual reports, citation on the company's website (particularly, if the company publishes a 'pipeline' and the product is included in it).

and/or

- b) References to the compound that indicate its progress is being published in the biomedical literature (journals and congresses).

Compounds that are launched and are not being investigated for new conditions, in new regions or by new organizations are not considered Under Active Development.

4.5 Putative biomarkers

This analysis compares the direction of your gene expression (or Key Hub activity) with known directional changes manually curated from literature within the database. The database includes molecular alteration knowledge for DNA, RNA, Protein, and post-translational modifications levels. This table, therefore, allows you to understand how your experimental data aligns with current knowledge of these molecules in association with the chosen disease/indication. The following classification is used to categorize your results:

- **Perfect.** Known disease associated molecular changes are found within the database that are in the same direction as your experimental data (e.g. you have up-regulated gene expression in a dataset and up-regulation was shown to be associated with a disease). These results indicate that you have experimental results that are in line with current biomarker knowledge for this disease.
- **Conflicting.** Known disease associated molecular changes are found within the database that are in the opposite direction as your experimental data (e.g. up-regulated gene in your dataset and down-regulation of the same gene was reported to be associated with a disease). These results indicate that you have experimental results that conflict with current biomarker knowledge for this disease.
- **Similar.** Known disease associated molecular changes are found within the database in the same direction as found in your experimental data, but the aberrations occurred on different molecular level (e.g. you have increased expression in your dataset and we find increased protein activity in the database associated with the disease). These results indicate that you have experimental results that may be supported by current biomarker knowledge for this disease but validation on a different molecular level may be required.

- **Uncertain.** Known disease associated molecular changes are found within the database in the opposite direction as found in your experimental data and the aberrations occurred on different molecular level (e.g. you have increased expression in your dataset and we find decreased protein activity in the database associated with the disease). These results indicate that you have experimental results that conflict with current biomarker knowledge for this disease on another molecular level and therefore further experiments to clarify whether the changes observed are also found at the indicated molecular levels may be required.
- **Unknown.** Either the experimental dataset or association in the database lack directional information (e.g. you input a gene list without expression changes or in the database the evidence behind the disease association does not have a clear directional change). These results indicate that there is insufficient experimental data input or on file to draw conclusions.

Each line is associated with View link that leads to corresponding entity page where all details about gene-disease association study are available.

Statistically significant gene-disease associations are shown by default: molecular changes are significantly associated ($p < 0.05$) with pathology / its manifestation degree / prognosis. Unchecking this checkbox on the results page will add low trust risk, hypothetical and correlation less gene-disease associations found in scientific papers.

There are two pages:

- Putative Biomarkers associated with disease / condition specified in analysis start settings.
- Putative Biomarkers that are associated with similar diseases to the condition previously specified according to MESH ontology. This list gives an overview of associations that were not connected to target disease before and gives biomarker subset in case if no biomarker was found with the specified disease on the previous tab.

4.6 Content browsing

Key Pathway Advisor allows you to browse detailed information about molecular entities (genes, RNA, proteins, protein complexes and protein groups), molecular interactions and gene-disease causal associations.

By clicking on a molecule in KPA report you will be redirected to an entity page for the corresponding molecule. On the page there is a molecular function description, links to external resources (NCBI, Swiss-Prot, *MetaCore*) for more information and summary pie chart visualization of molecular interactions for the molecule. This entity page is separated by several levels, and you can study data related to original gene (all interaction for gene product will be summarized), RNA (RNA affecting interactions are listed, like transcriptional regulation or miRNA binding if applicable), protein (interaction of the protein) and, if applicable, mature forms of protein including peptides and post translational modifications.

Molecular interaction entity page is available from table views of pathway maps, process networks and Key Hub networks. Each scientific reference on the page is associated with PubMed link, short summary description of the study related to molecular interaction, its directionality, effect, molecular mechanism (see all interaction mechanism summary descriptions in Appendix D), tissue and cell type/line and detection method. If the interaction has several

supporting publications, then all references are sorted by publication date starting from the most recent one.

Putative Biomarker entity pages are available from Putative Biomarkers analysis page and from table views of pathway maps, process networks and Key Hub networks. Each scientific reference on the page is associated with PubMed link, short summary description of the study results related to the association. The following data are also extracted from publication for your convenience.

Strength of Association describes an association of GV with risk disease development, pathology progression, prognosis, manifestation degree. Reflects the power of statistical significance considering p-value meaning, number of patients involved into the analysis.

- **High.** Association with risk of pathology development and/or its manifestation degree and/or prognosis, in case-control studies, cohort studies, nested case-control studies, clinical studies and meta-analysis is statistically significant
- **Borderline.** Contradictory statistical results were given, i.e. cases when authors interpret effect size as a significant association but statistics does not allow to make the same conclusion. Authors do not report an association with risk of pathology development, but statistically significant correlation with certain symptoms, signs of pathology was found.
- **Low.** Correlation analysis was performed but significant association with disease development or disease signs was not found.
- **Unknown significance.** The reference doesn't provide information about the association of mRNA/protein expression (or activity) with disease development.
- **Insufficient evidence.** Applicable for gene variants. It is used when, in a clinical practice guideline, there is insufficient evidence to make a recommendation. Under this scenario, the guideline usually does not recommend the use of a variant because there is insufficient evidence to make a statement on this matter at this point (further evidence or studies are required).

Although statistical significance underlies gradation of these meanings there are some deviations from this rule because statistical analyses are not provided in some association studies (descriptive studies).

Effect of Association reflects an influence of molecular changes described in the annotation on disease development/progression/survival.

Biomarker Role allows indicating the function or utility of the biomarker in the context of the use.

Abundance and Activity change identify mRNA or protein expression change (activity for protein) that was studied to be associated with a condition.

Functional Consequences for a gene variant indicate if either abundance or activity of gene product is affected by variation (which was specifically outlined in the original study).

Study design, experimental patient cohorts, detection methods are briefly described for each reference. If the association has several supporting publications, then all references are sorted by publication date starting from the most recent one.

Chapter 5: Methods

The *Key Pathway Advisor* workflow uses combination of the following methods to define biologically relevant result.

5.1 Causal reasoning analysis

Causal Reasoning is a shortest path-based method aimed at the identification of upstream regulators that cause gene expression changes observed in transcriptomics data^{2,3}. Causal Reasoning relies on a directed network that is annotated with activation and inhibition edges as well as biological mechanisms (transcription regulation). Causal Reasoning identifies candidates ("hypotheses") in the network that can be reached via a pre-defined maximum shortest path length from the differentially expressed genes. Candidates are scored based on the number of differentially expressed genes that can be reached via the shortest paths and the correctness of the regulation. The correctness is assessed based on the activation and inhibition edges along the paths and the expected and observed direction of fold changes of the differentially expressed genes.

The significance of the predictions made by a hypothesis is assessed using a binomial test based on the following information:

- k - the sum of correct predictions
- n - the sum of correct and incorrect predictions
- The p-value is calculated as probability to get k successes in n predictions using binomial trials with $p=0.5$

$$p\text{-value} = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k}$$

- P-values are assigned in the score matrix and hypotheses above the p-value threshold are filtered out of the score matrix.

5.2 Overconnectivity analysis

An overconnectivity test identifies one step away direct regulators of the dataset that are statistically overconnected with the objects from the data set. The method assumes that proteins functionally important for a particular phenotype have a relatively high number of interactions with proteins encoded by genes altered (e.g., differentially expressed) in the phenotype. This method allows you to assess the level of connectivity that individual objects in the database have with the genes from the analyzed gene list and whether this level of connectivity is more than would be expected purely by chance. The significance of ‘over connection’ (when the number of interactions of tested objects with objects from the analyzed gene list is higher than expected by chance) is determined as a p-value of hypergeometric distribution.

$$p - value = \frac{R! n! (N - R)! (N - n)!}{N!} \sum_{i=\max(r, R+n-N)}^{\min(n, R)} \frac{1}{i! (R - i)! (n - i)! (N - R - n + i)!}$$

- N – number of Molecular Entities covered by the whole ontology
- R – number of Molecular Entities in a list under analysis
- n – number of Molecular Entities associated with a particular category from the ontology
- r – number of gene from input list intersecting with genes from a particular category

5.3 Enrichment analysis

Enrichment Analysis methods have become commonplace tools applied to the analysis and interpretation of biological data. This is the first “low resolution” step of functional analysis. The goal is to discover pathways or processes associated with the gene list of interest.

The significance of enrichment is defined by using the hypergeometric test⁴:

$$p - value = \frac{R! n! (N - R)! (N - n)!}{N!} \sum_{i=\max(r, R+n-N)}^{\min(n, R)} \frac{1}{i! (R - i)! (n - i)! (N - R - n + i)!}$$

- N – number of Molecular Entities covered by the whole ontology
- R – number of Molecular Entities in a list under analysis
- n – number of Molecular Entities associated with a particular category from the ontology
- r – number of gene from input list intersecting with genes from a particular category

As a result, all terms from the ontology are ranked according to calculated p-values. Ontology terms with p-values less than the p-value threshold 0.05 are defined as statistically significant and therefore relevant to the studied list of genes. In other words, the gene list is associated with a quantitatively ranked list of pathways and processes summarizing its effects at a systems-biology level.

5.4 Signaling Pathway Impact Analysis (SPIA)

SPIA aims at the identification of perturbed pathways in a given condition by combining enrichment of perturbed genes in the pathway with the actual amount of perturbation, leading to the most promising candidate pathways and thus candidate genes⁶.

SPIA captures two different probabilities for each pathway:

1. the enrichment of differentially expressed genes within the pathway (as described above)
2. the level of perturbation within the pathway as measured by propagating expression changes through the pathway (perturbation probability).

The enrichment can be calculated by applying a simple hypergeometric test. To estimate the level of perturbation within a pathway, a molecule perturbation (MP) is calculated for each gene as follows:

$$MP(G_i) = E(G_i) + \sum_{j=1}^n \beta_j \frac{MP(G_j)}{N(d,j)}$$

E represents the signed expression change of gene i (fold changes). The second part of the equation is the sum of molecule perturbations of all genes j that are directly upstream of gene i , normalized by the number of downstream genes of each such gene j . β reflects the type of interaction between genes i and j : in case of an activation edge, it is set to +1 and in case of inhibition to -1. Transcription regulation interactions are not accounted because that may cause discordance with input data leading to incorrect interpretation.

The net perturbation accumulation at the level of gene i is calculated as the molecule perturbation minus its observed fold change; $Acc_i = MP(G_i) - E(G_i)$. Thus, the more agreement with upstream events, the higher the molecule perturbation.

The overall pathway perturbation is computed as the sum of all perturbation accumulations; $t_A = \sum (Acc(G_i))$. If t_A is positive, then we conclude that the pathway is activated (or positively perturbed). If t_A is negative, then we assume that the pathway is inhibited (or negatively perturbed).

The computation of perturbation probability for a given pathway is based on a bootstrap procedure in which we test if the observed global activation or inhibition of the pathway computed with the real data, t_A is unusual compared to a multitude of random scenarios.

The step-by-step procedure we used is:

1. An iteration counter k is initialized ($k = 1$).
2. A set of $N_{de}(P_i)$ node IDs is selected at random from the input pathway P_i , where the $N_{de}(P_i)$ is the number of start nodes observed on the pathway with the real data. The log fold changes for these random gene IDs are assigned by drawing a random sample with replacement from the distribution of all DEGs to be analyzed. The permuted start nodes data frame is formed to compute the perturbation accumulations Acc , for each gene in P_i . The net total accumulation is computed as the sum of all perturbation accumulations across each pathway: $t_A(k) = \text{sum}(Acc(g_{ik}))$
3. Step 2 is repeated 2000 times
4. The median of $t_A(k)$ is computed and subtracted from $t_A(k)$ values centering their distribution around 0. The resulting corrected values are denoted with $t_{A;c}(k)$. The observed net total accumulation is also corrected for the shift in the null distribution median to give $t_{A;c}$.
5. The probability to observe such total net inhibition or activation just by chance, perturbation probability, is computed as:
$$2 * \text{sum}(I(t_{A;c}(k) > t_{A;c})) \text{ if } t_{A;c} > 0;$$
$$2 * \text{sum}(I(t_{A;c}(k) < t_{A;c})) \text{ otherwise,}$$

where the identity function $I(x)$ returns 1 if x is true and 0 otherwise. The multiplication by 2 accounts for a two-tailed test since we do not have a particular expectation regarding the pathway status (inhibited or activated).

Combined p-value is calculated as

$$\text{Impact P-value} = c_j - c_i * \ln(c_j)$$

Where **Impact P-value** represents the overall pathway perturbation probability and c_j is the multiplication of the two different probabilities described before (enrichment and perturbation probabilities).

Chapter 6: Glossary

Differentially Expressed Genes (DEGs)	Genes where expression differs between at least two phenotypical conditions (e.g., Disease / Control)
Diseases	This ontology is created based on the classification in Medical Subject Headings (MeSH). Each disease in diseases ontology has its corresponding molecular alterations in a gene or set of genes
Downstream Analysis	Analysis of cellular process disruption which molecular components were shown to be dysregulated (e.g., DEGs, gene variants, etc.)
Enrichment Analysis (EA) (also, Ontology Enrichment)	An analysis procedure that consists of mapping gene IDs of the dataset(s) of interest onto gene IDs in entities (terms) of built-in functional ontologies such as pathway maps, networks, diseases, etc. The terms in a given ontology are ranked based on "relevance" in the dataset. The statistical relevance procedure, a p-value of hypergeometric distribution, is calculated as the probability of a match to occur by chance, given the size of the ontology, the dataset, and the particular entity. The lower the p-value, the higher is the "non-randomness" of finding the intersection between the dataset and the particular ontology term. That, in turn, translates into a higher ranking for the entity matched. Everything equal, the more genes / proteins belong to a process / pathway, the lower the p-value. In EA multiple proprietary ontologies (canonical pathway maps, cellular processes, disease biomarkers etc., and public ontologies such as Gene Ontology (cellular processes, protein functions, localizations) are utilized.
Enrichment Synergy	The enrichment synergy method was offered for comparison of datasets that are functionally relevant but poorly overlapping at the gene level. For instance, mutated and amplified genes in breast cancer ⁵ . The genes derived from different datasets may not overlap directly but populate the very same pathway or process, which suggests that they are functionally complimentary. To determine whether two distinct gene lists cooperatively alter a certain cellular pathway or process, we calculate the synergy between them by ontology enrichment. An ontology term (pathway or process) is considered synergistic if the enrichment p-value for the non-redundant union of compared gene lists is lower than p-values for individual lists. More significant enrichment for the union reflects functional connectivity of two gene lists and their complementary effect on the pathway.

Entity	An element or a term in an ontology, e.g., a given disease, or a given process, etc.
Gene Variant	A DNA sequence variation
GO Localizations	A GO ontology for localization of the gene products inside or outside the cell. A given molecule in a given localization is represented by a Molecular Entity.
GO Molecular Functions	A GO ontology of hierarchically structured molecular functions. A protein may be linked to several different molecular functions.
GO Processes	A GO ontology for biological processes. The processes are structured as hierarchical tree with branches defined according to the Gene Ontology controlled vocabulary. GO process folders are nested, i.e., each folder references all the proteins participating in its sub-processes.
Key Entity	An ontology term (i.e., pathway maps) that enriched with both differentially expressed genes and corresponding key Hubs (see Introduction part for detailed workflow description).
Key Hub (KH)	A topologically significant Molecular Entity that that is known to regulate differential expression genes. KHs could be obtained by two approaches: causal reasoning network analysis and interactome analysis. Using causal reasoning the one could define one step KHs (transcriptional factors that statistically significant associated with experimental differential expressed genes regulation) and distant KHs (second step objects tht regulate one step transcriptional factors, etc., up to three steps). Interactome analysis gives Molecular Entities that are overconnected with experimental differentially expressed genes.
Pathway Groups	This is a collection of manually created pathway maps, grouped hierarchically into folders according to main biological processes.
Molecular Entity	Any type of molecule, (e.g., kinases, transcriptional factors, receptors, etc.) involved in interactions between molecules in the database.
Ontology	Functional controlled vocabularies developed for a key theme (e.g., Pathways or processes). Every ontology has a hierarchical tree structure and corresponding sets of pre-built networks and pathway maps, or, in case of the disease ontology, gene lists.
Pathway Map	Pathway maps are graphic images representing complete biochemical pathways or signaling cascades in a commonly accepted sense.
Processes Networks	A recognized series of events (interactions or biochemical reactions) accomplished by one or more ordered assemblies of molecular functions with a defined beginning and end.
p-value	Statistical measure of the likelihood that an event would happen purely by chance
Upstream Analysis	Search for molecules with aberrant activity that may be the root cause of observed differential gene expression

Bibliography

1. Nikolsky, Y., Sviridov, E., Yao, J., Dosymbekov, D., Ustyansky, V., Kaznacheev, V., Dezso, Z., Mulvey, L., Macconail, L.E., Winckler, W., et al. (2008). Genome-Wide Functional Synergy between Amplified and Mutated Genes in Human Breast Cancer. *Cancer research* 68, 9532–9540.
2. Chindelevitch L, Ziemek D, Enayetallah A, Randhawa R, Sidders B, et al. (2012) Causal Reasoning on Biological Networks: Interpreting Transcriptional Changes. *Bioinformatics* 28: 1114-1121.
3. Pollard J Jr, Butte AJ, Hoberman S, Joshi M, Levy J, Pappo J. (2005) A Computational Model to Define the Molecular Causes of Type 2 Diabetes Mellitus. *Diabetes Technol Ther.* 2005 Apr; 7(2):323-36.
4. Nikolsky, Y., Kirillov, E., Zuev, R., Rakhmatulin, E. & Nikolskaya, T. (2009). Functional Analysis of OMICs Data and Small Molecule Compounds in an Integrated “Knowledge- Based” Platform. In *Protein Networks and Pathway Analysis* (Nikolsky Y & Bryant J, eds), pp. 177–196. Humana Press, Totowa,
5. NJ. Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C. & Bugrim, A. (2009). Identifying Disease-Specific Genes Based on their Topological Significance in Protein Networks. *BMC Systems Biology* 3, 36+.
6. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. (2009) A novel signaling pathway impact analysis. *Bioinformatics* 25: 75-82

Appendix A: How to get NCBI GEO data for KPA easily

Public OMICs data repositories are very useful providers of gene expression datasets produced for other studies. Such datasets are good if you don't have your own datasets or to test a hypothesis made after an analysis of your in-house datasets. The major obstacle is however that public datasets contain raw data and though require a user to have dataprocessing and bioinformatics skills and tools. Only after that it will be possible to identify genes which expression levels change in experimental conditions when comparing to expression levels of a control condition. NCBI Gene Expression Omnibus (NCBI GEO) provides an easy tool to perform that by few clicks called GEO2R.

Go to ncbi.nlm.nih.gov/geo to get access to NCBI GEO repository. Search for a keyword e.g., 'prostate cancer' and you will be redirected to a page with list of datasets associated with the keyword. Follow the detailed description how to get a dataset shown at Figure 15. Please note that not all datasets might be processed by GEO2R function depending on their data structure.

The figure shows a sequence of three screenshots from the NCBI GEO website. The top screenshot shows the 'Gene Expression Omnibus' search page with 'prostate cancer' entered in the search bar. A red arrow points to a pop-up box that displays search results: 'There are 19439 results for "prostate cancer" in the GEO DataSets Database.' and 'There are 2227697 results for "prostate cancer" in the GEO Profiles Database.' A second red arrow points to the search results page, which shows a list of datasets. A third red arrow points to the 'Accession Display' page for dataset GSE2443, which includes details such as status, title, organism, and a table of platforms and samples. At the bottom of this page, the 'Analyze with GEO2R' button is highlighted with a red box.

Figure 15: Go to NCBI GEO web page to search for gene expression data. Type a keyword in a search area and press Search. Click on number of datasets GEO proposes you to look at (a number in the first line). You will be redirected to a search results page where you need to select Series in the top left menu called Entity type. By clicking on items in search results you will be redirected to a dataset entity page where you can get more details about experimental design and the study itself. Once you found a proper dataset you need to press Analyze with GEO2R button at the bottom of the entity page.

Once you are on a GEO2R page you first need to define sample groups for comparison and differential expression identification. Follow the instructions shown on Figures 16 and on.

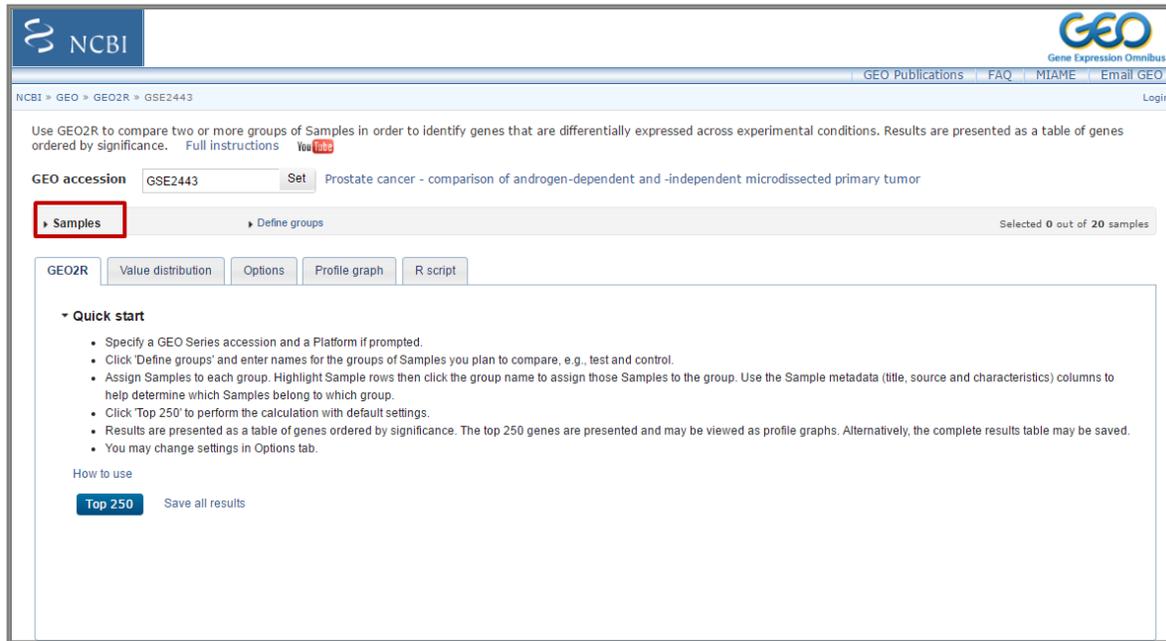


Figure 16: GEO2R page with activated dataset GSE2443 in this example. Proceed to *Samples*.

GEO accession Prostate cancer - comparison of androgen-dependent and -independent microdissected primary tumor

▼ Samples Selected 0 out of 20 samples

Columns

Group	Accession	Title	Source name
-	GSM45730	AD-01	untreated human prostate cancer
-	GSM45847	AD-02	untreated human prostate cancer
-	GSM45848	AD-03	untreated human prostate cancer
-	GSM45849	AD-04	untreated human prostate cancer
-	GSM45850	AD-05	untreated human prostate cancer
-	GSM45851	AD-06	untreated human prostate cancer
-	GSM45852	AD-07	untreated human prostate cancer
-	GSM45853	AD-08	untreated human prostate cancer
-	GSM45854	AD-09	untreated human prostate cancer
-	GSM45855	AD-10	untreated human prostate cancer
-	GSM45856	AI-01	androgen-independent primary human prostate cancer
-	GSM45857	AI-02	androgen-independent primary human prostate cancer
-	GSM45858	AI-03	androgen-independent primary human prostate cancer
-	GSM45859	AI-04	androgen-independent primary human prostate cancer
-	GSM45860	AI-05	androgen-independent primary human prostate cancer
-	GSM45861	AI-06	androgen-independent primary human prostate cancer
-	GSM45862	AI-07	androgen-independent primary human prostate cancer
-	GSM45863	AI-08	androgen-independent primary human prostate cancer
-	GSM45864	AI-09	androgen-independent primary human prostate cancer
-	GSM45865	AI-10	androgen-independent primary human prostate cancer

Figure 17: Once you are on *Samples* tab you see a list of samples in the dataset. There might be different parameters associated with each sample. In the case of GSE2443 there is only one important column called *Source name* which clearly shows that there are two groups of samples *untreated human prostate cancer* and *androgen-independent primary human prostate cancer*. These groups of samples will be used to identify differential expression between them. Click on *Define groups*.

Figure 18: Enter group names in a *Define groups* box. Two groups were created for this example. Select required set of samples and then click on a group name on *Define groups* panel. Ten *untreated prostate cancer* samples we assigned to corresponding group. Do the same for the other samples and groups. You need to create two groups to perform a comparison.

Figure 19: Once both groups are selected go to the bottom of the page and press *Save all results*. Resulting list of genes with expression comparison will appear (Figure 20).

"ID"	"adj.P.Val"	"P.Value"	"t"	"B"	"logFC"	"Gene.symbol"	"Gene.title"
"210784_x_at"	"0.124"	"0.0000187"	"-5.431091"	"2.5332"	"-1.639385"	"LILRB3"	"leukocyte immunoglobulin like receptor B3"
"201812_s_at"	"0.124"	"0.0000256"	"-5.299834"	"2.2829"	"-1.206232"	"C4orf46///TOMM7"	"chromosome 4 open reading frame 46//translocase of outer mitochondrial membrane 7"
"204121_at"	"0.124"	"0.0000261"	"-5.291437"	"2.2668"	"-2.133268"	"GADD45G"	"growth arrest and DNA damage inducible gamma"
"200909_s_at"	"0.124"	"0.0000277"	"-5.266255"	"2.2185"	"-1.439145"	"SNORA52///RPLP2"	"small nucleolar RNA, H/ACA box 52//ribosomal protein lateral stalk subunit P2"
"214885_at"	"0.124"	"0.0000346"	"5.174358"	"2.0415"	"1.750829"	"KAT8"	"lysine acetyltransferase 8"
"214021_x_at"	"0.124"	"0.0000346"	"-5.17427"	"2.0414"	"-1.822669"	"ITGB5"	"integrin subunit beta 5"
"203815_at"	"0.124"	"0.0000399"	"-5.124619"	"1.9453"	"-2.788959"	"GSTT1"	"glutathione S-transferase theta 1"
"214307_at"	"0.134"	"0.0000558"	"-4.976687"	"1.6575"	"-1.597092"	"HGD"	"homogentisate 1,2-dioxygenase"
"212254_s_at"	"0.134"	"0.0000599"	"4.953296"	"1.6118"	"1.473708"	"DST"	"dystonin"
"211347_at"	"0.134"	"0.0000604"	"4.943574"	"1.5927"	"1.676523"	"CDC14B"	"cell division cycle 14B"
"208982_at"	"0.134"	"0.0000683"	"4.892774"	"1.4932"	"1.393153"	"PECAM1"	"platelet and endothelial cell adhesion molecule 1"
"214061_at"	"0.134"	"0.0000723"	"4.869725"	"1.4479"	"2.307755"	"TBC1D31"	"TBC1 domain family member 31"
"209959_at"	"0.16"	"0.0000932"	"-4.765137"	"1.2419"	"-3.021774"	"NR4A3"	"nuclear receptor subfamily 4 group A member 3"
"203103_s_at"	"0.175"	"0.0001097"	"-4.698181"	"1.1095"	"-1.13884"	"PRPF19"	"pre-mRNA processing factor 19"
"221844_x_at"	"0.186"	"0.0001251"	"-4.644317"	"1.0027"	"-0.991627"	"SPCS3"	"signal peptidase complex subunit 3"
"215078_at"	"0.186"	"0.0001583"	"-4.547742"	"0.8107"	"-3.231901"	"LOC100129518//S002"	"uncharacterized LOC100129518//superoxide dismutase 2, mitochondrial"
"214815_s_at"	"0.186"	"0.0001711"	"-4.515977"	"0.7474"	"-2.441159"	"TRIM33"	"tripartite motif containing 33"
"203752_s_at"	"0.186"	"0.0001867"	"-4.48019"	"0.676"	"-1.348666"	"JUND"	"JunD proto-oncogene, AP-1 transcription factor subunit"
"203680_at"	"0.186"	"0.0001935"	"4.465525"	"0.6467"	"1.790872"	"PRKAR2B"	"protein kinase cAMP-dependent type II regulatory subunit beta"
"215016_x_at"	"0.186"	"0.0001991"	"4.45404"	"0.6238"	"1.342822"	"DST"	"dystonin"
"205666_at"	"0.186"	"0.0002085"	"4.435067"	"0.5859"	"1.888795"	"FMO1"	"flavin containing monooxygenase 1"
"207173_x_at"	"0.186"	"0.0002189"	"4.415184"	"0.5461"	"1.629408"	"CDH11"	"cadherin 11"
"217544_at"	"0.186"	"0.0002217"	"4.409899"	"0.5356"	"1.841755"	"MIR3916"	"microRNA 3916"
"201962_s_at"	"0.186"	"0.0002305"	"-4.393918"	"0.5036"	"-1.708343"	"RNF41"	"ring finger protein 41"
"213096_x_at"	"0.186"	"0.0002358"	"-4.384645"	"0.4851"	"-1.939099"	"FAM149B1"	"family with sequence similarity 149 member B1"
"217973_at"	"0.186"	"0.000236"	"-4.384341"	"0.4845"	"-1.94048"	"DCXR"	"dicarbonyl and L-xylulose reductase"
"AFFX-r2-Hs18S-rRNA-5_at"	"0.186"	"0.0002402"	"-4.377097"	"0.47"	"-1.419486"	""	""
"204556_s_at"	"0.186"	"0.0002489"	"4.36262"	"0.441"	"1.016636"	"DZIPI1"	"DAZ interacting zinc finger protein 1"
"AFFX-HUMRGE/M10098_5_at"	"0.186"	"0.0002594"	"-4.345611"	"0.4069"	"-1.45614"	""	""
"205645_at"	"0.186"	"0.000265"	"-4.336925"	"0.3895"	"-1.571905"	"REPS2"	"RALBP1 associated Eps domain containing 2"
"AFFX-r2-Hs28S-rRNA-3_at"	"0.186"	"0.0002654"	"-4.336327"	"0.3883"	"-1.417767"	""	""
"204293_at"	"0.186"	"0.0002708"	"-4.327995"	"0.3717"	"-1.285204"	"SGSH"	"N-sulfoglucosamine sulfohydrolase"
"211456_x_at"	"0.186"	"0.0002757"	"-4.320712"	"0.3571"	"-1.865821"	"MTIHL1"	"metallothionein 1H-like 1"
"204582_s_at"	"0.187"	"0.0002877"	"-4.303337"	"0.3222"	"-1.784828"	"KLK3"	"kallikrein related peptidase 3"
"216804_s_at"	"0.187"	"0.0002934"	"-4.295216"	"0.306"	"-2.581981"	"PDLIM5"	"PDZ and LIM domain 5"
"220415_at"	"0.198"	"0.0003228"	"4.256227"	"0.2278"	"1.557969"	"FPGT-TNMI3K///TNMI3K"	"FPGT-TNMI3K readthrough//TNMI3 interacting kinase"
"207510_at"	"0.198"	"0.0003295"	"4.247738"	"0.2108"	"1.3953"	"BDKRB1"	"bradykinin receptor B1"
"203477_at"	"0.216"	"0.0003735"	"4.196474"	"0.1079"	"1.41371"	"COL15A1"	"collagen type XV alpha 1 chain"
"212464_s_at"	"0.216"	"0.0003773"	"4.192259"	"0.0994"	"1.335921"	"FN1"	"fibronectin 1"
"205221_at"	"0.216"	"0.0003949"	"-4.173695"	"0.0622"	"-2.161355"	"HGD"	"homogentisate 1,2-dioxygenase"
"212158_at"	"0.216"	"0.0003972"	"4.17126"	"0.0573"	"1.643666"	"SDC2"	"syndecan 2"
"209978_s_at"	"0.216"	"0.0004086"	"-4.159692"	"0.034"	"-1.706408"	"PLG//LPA"	"plasminogen//lipoprotein(a)"
"211719_x_at"	"0.216"	"0.0004173"	"4.151059"	"0.0167"	"1.440912"	"FN1"	"fibronectin 1"
"218674_at"	"0.216"	"0.0004315"	"4.137364"	"-0.0108"	"1.037805"	"TRAPP13"	"trafficking protein particle complex 13"
"AFFX-HUMRGE/M10098_3_at"	"0.216"	"0.0004479"	"-4.122099"	"-0.0414"	"-1.048263"	""	""
"220044_x_at"	"0.216"	"0.0004497"	"4.120438"	"-0.0448"	"1.456752"	"LUC7L3"	"LUC7 like 3 pre-mRNA splicing factor"
"AFFX-r2-Hs18S-rRNA-3_s_at"	"0.216"	"0.0004556"	"-4.115068"	"-0.0556"	"-1.186905"	""	""
"AFFX-M27830_5_at"	"0.222"	"0.0005014"	"-4.075831"	"-0.1344"	"-1.243534"	""	""
"204619_s_at"	"0.222"	"0.0005063"	"-4.071862"	"-0.1424"	"1.629689"	"VCAN"	"versican"
"216388_s_at"	"0.222"	"0.0005362"	"4.048308"	"-0.1897"	"1.626603"	"LTBR4"	"leukotriene B4 receptor"

Figure 20: Example of a DEG file comparing gene expression between two samples. Save this file (in Windows: Ctrl+S or right click on the page) and then drag and drop into KPA, it will be automatically recognized and uploaded

Appendix B: Differential gene expression calculations

The major goal of gene expression measurement experiments is to identify genes which expression levels change in experimental conditions when comparing to expression levels of a control condition. For example, comparing disease samples with normal/healthy control samples, samples before and after drug administration or different subtypes or stages of disease.

Fold changes as a measure of difference

Let's say A is expression of a gene in a case group and B in a control. If $A > B$ then a gene is *over-expressed* (*up-regulated*), otherwise If $B > A$ then a gene is *under-expressed* (*down-regulated*). Fold Change (FC) is used as a measure of quantity of the change of gene expression between case and control groups and usually calculated as a ratio between A and B. FC is an intuitive way to represent such changes: a) if a FC value is positive, then gene is *over-expressed*, if FC is negative – *down-expressed*; b) the magnitude of FC value indicates how significantly a gene is over/down-expressed. These two features allow visualizing such values e.g. build heat maps, or highlight expression changes on pathway maps.

There are several methods to calculate FC or similar in sense values:

1. Ratio. A is always compared to B. B effectively becomes a fixed value for each gene to which A's expression levels are being compared.

$$FC = \frac{A}{B}$$

In result FC which value is between 0 and 1 means *under-expression* in A compared to B. If FC bigger than 1 it is *over-expression* in A compared to B. That representation is not very intuitive to work with. To make the results easier to visualize it is common to apply a logarithmic transformation (usually to base 2). KPA applies logarithmic transformation to base 2 for all data in such format by default (see Logarithmic Transformation section below for more details).

2. Fold Change as inverted ratio. If A is bigger than B (*over-expression* in A):

$$FC = \frac{A}{B}$$

If B is bigger than A (*under-expression* in A):

$$FC = -\frac{B}{A}$$

This results in an absolute level of expression change where the expression represents the size of the difference between the two values. In these results FC with a value is less than -1 means *under-expression*. If FC larger than 1 is *over-expression*. If you FC calculated by this method no further logarithmic transformation is required and KPA processes these data as is

Logarithmic transformation

In simple cases the logarithm counts repeated multiplication. Logarithm to base 2 of a value N is basically how many times should I multiply 2 on itself to get N. For example, the base 2 logarithm of 8 is 3, as 2 to the power 3 is 8 ($8 = 2 \times 2 \times 2 = 2^3$); the multiplication is repeated three times or

$$\log_2(8) = 3$$

Alternatively, the base 2 logarithm of 0.25 is -2 or

$$\log_2(0.25) = -2$$

That transforms results in the following: a) If the FC is less than 0 than a gene is *under-expressed*; b) If the FC is bigger than 0 than a gene is *over-expressed*.

So why log transformation?

Simply, it is easier to link log ratio to fold change and visualize more intuitively. Very small numbers become big numbers that can be easier to handle and visualize on graphical outputs.

If you submit a data file where each gene is associated with expression change calculated as A/B ratio the logarithmic transformation to base 2 will be applied by default. It is possible that your data are already log transformed (usually column with such values is called Log Ratio), in this case your values will be accepted without further transformation by KPA).

Appendix C: P-value

In most cases high-throughput experiments result in lists of genes or proteins of interest. These lists could, for example, be genes differentially expressed between two conditions or proteins identified in a sample. These datasets usually contain anywhere between few dozen and few thousand genes / proteins. Here we will explain the statistical concepts behind the enrichments KPA. No statistics knowledge is assumed and formulae for non-proprietary algorithms are shown in the previous sections of this help guide for more statistically knowledgeable users.

C.1 What is a p-value?

Before attempting to understand the statistics behind the analysis performed in KPA it is important to understand the basic concept of statistical significance. Statistical significance is a measure of the likelihood that an event would happen purely by chance. The number associated with this measurement is known as the p-value or probability-value. P-values relate the likelihood of an event occurring purely by chance as decimal percentage. For example, a p-value of 1 would mean that there is a 100% probability that this would occur purely by chance; a p-value of 0 would mean that there is a 0% probability and a p-value of 0.05 would mean there is a 5% (or 1 in 20) probability that this would occur by chance alone. Commonly a p-value of 0.05 is accepted as sufficient evidence for an event to be classed as unlikely to

have occurred purely by chance. Events that occur with a p-value of 0.05 or less are commonly referred to as statistically significant. However, it is increasingly debated whether this is the right cut-off for such biologically variable data. It is therefore recommended that although not all results are statistically significant they may still be biologically important and vice versa. See Table 2 for further examples of p values and % probabilities.

Table 2: Relationship between p-values, probabilities, and chance.

P-value	Probability (%)	Chance
1	100%	Definitely by chance
0.1	10%	1 in 10
0.05	5%	1 in 20
0.01	1%	1 in 100
0.005	0.5%	1 in 200
0	0%	Definitely not by chance

C. 2 Enrichment p-values and $-\log(p\text{-values})$

Enrichment analysis in KPA aims to understand the biology behind your dataset by examining the intersection between your dataset and the prebuilt pathway maps and networks in KPA (i.e., how enriched your dataset is in a particular map or network). However, if we were to look solely for the intersection between the dataset and the prebuilt content in KPA we would likely simply return the largest maps and networks that have the biggest intersection purely due to their size. We therefore utilize a statistic (known as the hypergeometric mean) to consider the number of objects in your dataset, the number of objects in the intersecting map or network and the number of objects in the entire database. This assessment therefore returns a p-value that tells us the likelihood that the intersection between your dataset and a particular map / network is obtained purely by chance (See Figure 21).

- A. Objects in your dataset
- B. Objects in Pathway Map / Process / Disease
- C. Overlap between your dataset and individual Map / Process / Disease
- D. All Objects in the database
- E. All objects covered by your experiment method that can be used as an alternative background list (e.g. all microarray probes)

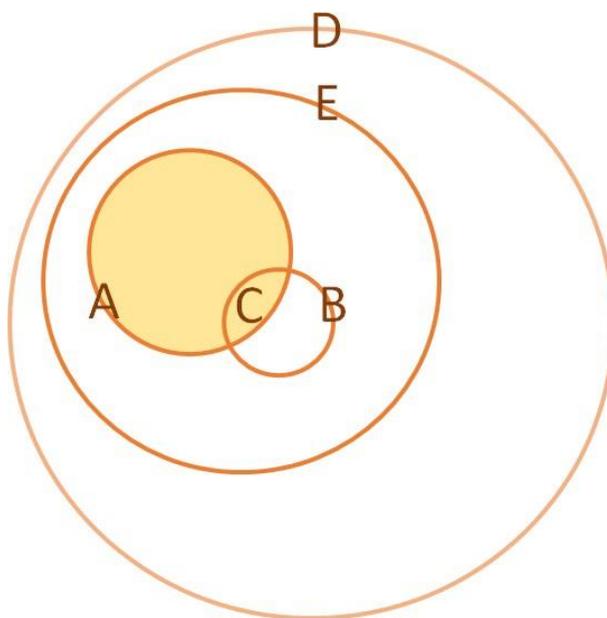


Figure 21: Calculation of enrichment p-values.

This p-value therefore gives us a method of statistically ranking the intersections between your data and the prebuilt content in KPA. The results of this analysis are displayed as a histogram. However, as a very significant map / network would have a very low numerical p-value (e.g., a p-value of 0.0001) they would have very small bars such bars would not be easy to see. We therefore multiply the p-values by $-\log_{10}$ to invert the bars. This transformation means that the smaller the p-value the larger the bar on the histogram that it has. For example, a very significant p-value of 0.0001 has a $-\log(p\text{-value})$ of 4 whereas a non-significant p-value of 0.1 has a $-\log(p\text{-value})$ of 1. For reference a p-value of 0.05 has a $-\log(p\text{-value})$ of 1.3 and therefore enrichments with a $-\log(p\text{-value})$ greater than this can be considered “statistically significant” in the commonly accepted use of the p-value cut-off. See Table 3 & Figure 22 for further examples.

Table 3: Relationship between p-values, probabilities and chance.	
P-value	$-\log(p\text{-value})$
1	0
0.1	1
0.05	1.3
0.01	2
0.005	2.3
0.0001	4

#	Name	Input Data p-value	Key Hubs p-value	Union p-value	stacked	grouped
1	Immune response_IL-3 signaling via JAK/STAT, p38, JNK and NF-kB	5.24E-4	1.031E-21	4.717E-22		<input type="checkbox"/>
2	Immune response_HMGB1/RAGE signaling pathway	2.178E-6	1.785E-11	4.049E-16		<input type="checkbox"/>
3	Immune response_IL-18 signaling	1.532E-7	2.874E-9	7.357E-15		<input type="checkbox"/>
4	Glomerular injury in Lupus Nephritis	6.005E-7	2.565E-9	1.742E-14		<input type="checkbox"/>
5	Immune response_IL-17 signaling pathways	1.526E-9	3.367E-7	3.21E-14		<input type="checkbox"/>

Figure 22: Example enrichment results showing p-values and $-\log(p\text{-value})$ colored bars.

Note: p-values in KPA should be used as a guide only. A result which does not fall below the p-value 0.05 threshold could still be biologically important.

Taking some time to review not only your most significant results but also the less significant results can help you bring together a more complete biological story.

Appendix D: Upstream analysis

The most likely regulator molecules (or Key Hubs) are overconnected by binding or functional interactions with each other and differentially expressed genes, so we need to inspect the whole interaction network within the cell to find

these molecules. There are two broad classes of global network analysis approaches widely used in the academic research and industry.

- Connectivity analysis (random walk; information flow; network propagation) – the identification of potential regulators using the network of molecular interactions and a list of ‘phenotype-associated’ nodes (for example, genes differentially expressed in disease or genes known to be associated with disease pathology). Network nodes which provide more connectivity between the ‘disease-associated’ nodes are thought to be good candidate for targeting. In KPA we use an Overconnectivity algorithm for this approach¹.
- Causal reasoning – identification of nodes which are most likely to cause observed expression changes associated with the phenotype. In this case, changes in expression of genes, direction, and effect of edges in the network are considered. For each node X, observed changes are matched with the expected changes inferred from network structure given hypothesis that X decreased or increased its activity.

Figure 23 contains an explanation of the difference between connectivity analysis and causal analysis.

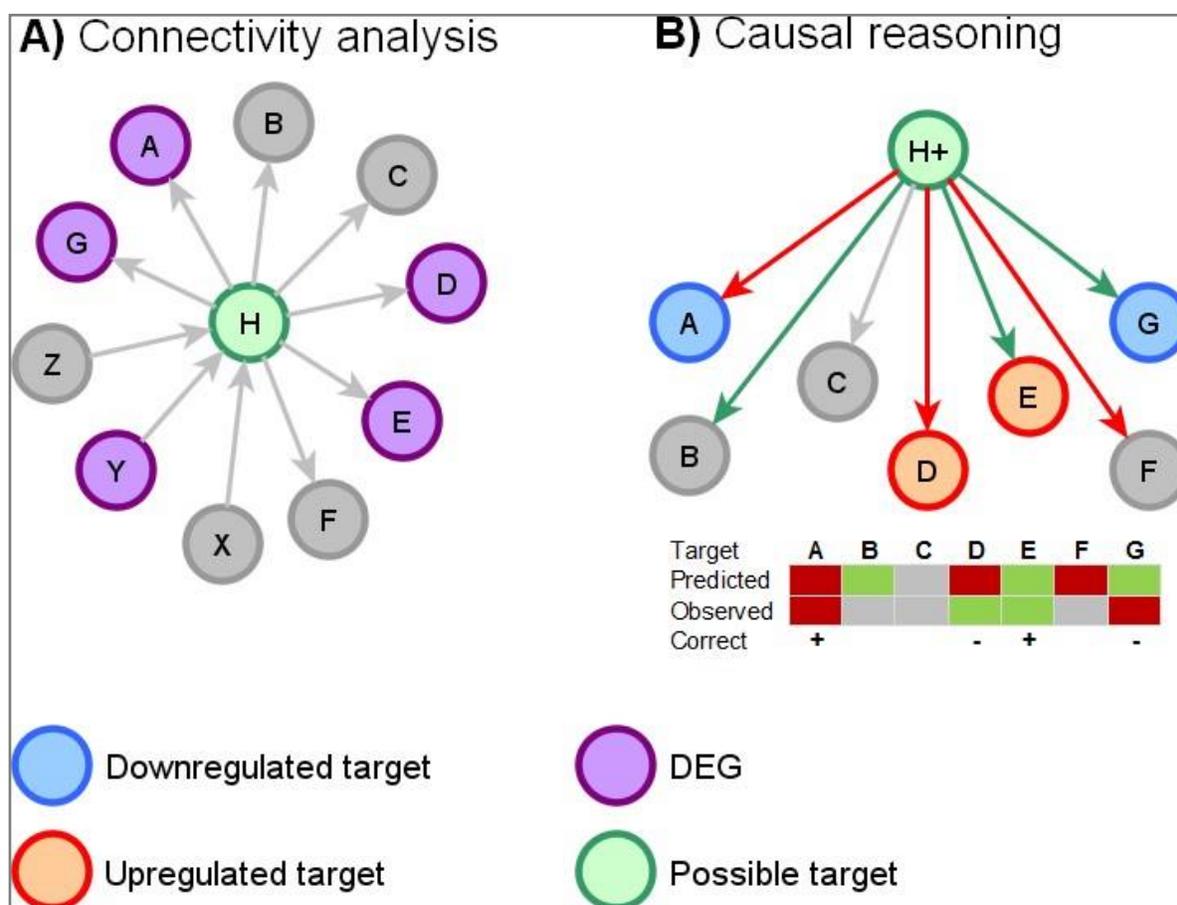


Figure 23: Connectivity analysis and casual reasoning.

In both cases (connectivity analysis and causal reasoning), we have the same network and the same set of differentially expressed genes and the same network, when the green object in Figure 8 is tested for ‘topological significance’ – i.e. we want to know if this object significantly influences system and is eligible for targeting in this phenotype.

Figure 23 A) Neighbors of this gene are only counted when they are DEGs and are then used to calculate enrichment p-value (given total number of DEGs and degree of a green node). In this case, 50% of green node’s neighbors are DEGs –which seems significant.

Figure 23 B) However, is the green node really influencing the expression? Without considering the actual fold change data it is not possible to fully hypothesis about the activity of the green node. KPA therefore uses Causal reasoning as it considers only downstream targets of green nodes and looks to see if they are consistent with observed differential expression genes. In this case, if we suggest that the green node is activated (hypothesis H^+), we see that only two of four downstream DEGs are consistent with this hypothesis, and two others contradict it.

In this case, the green node is looks like a less suitable candidate.

The example in Figure 24 uses causal reasoning to examine both the direct neighbors of the ‘nodes of interest’ (differentially expressed genes), but also more remote (several steps away) regulators.

The first step in identifying upstream regulators is to definite transcriptional factors with potential changed activity in your experimental data (which reflects the treatment or phenotype you are trying to understand). Using the causal reasoning algorithm, it is possible to identify the transcriptional factors that are statistically and logically (see Methods section) could work as direct regulators of expressional changes observed in your experiment. Applying a causal reasoning approach further we can therefore identify proteins that potentially regulate these transcriptional factors and reconstruct the most likely interpretation of altered regulation chain activity. Up to three steps up the regulatory chain are currently able to be traced using KPA.

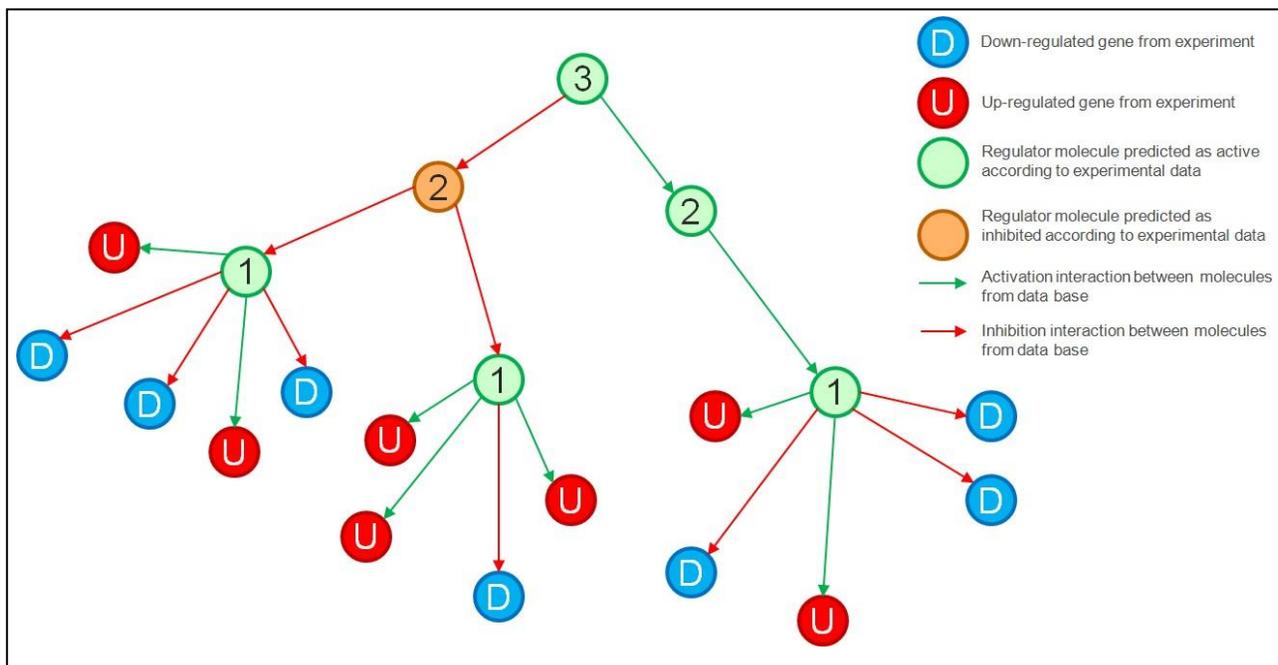


Figure 24: Casual reasoning example.

Appendix E: Molecular interactions mechanisms

Binding

Direct physical interaction between the molecules is proven by appropriate methods

Receptor binding

Interactions of natural ligands with membrane or nuclear receptors

Transport

Interaction between specific transporter protein and its macromolecular target resulting in a change of target localization

Complex formation

The Complex formation mechanism from the regulatory subunit of complex to catalytic subunit is used to show complex formation required for activation of a whole complex rather than a simple binding interaction

Transformation

Interactions between G-proteins and their activators/inhibitors – GEF and GAP proteins. The result of Transformation is GTP/GTP exchange on G α subunit of G-protein complex

miRNA binding

Interaction between microRNA and mRNA of corresponding target genes. Interactions with activation or inhibition effects are experimentally validated events. Interactions with this mechanism and unknown effect are predicted regulation events

Cleavage

Macromolecular cleavage is the process of breaking the bonds between amino acids in proteins (or between nucleotides in nucleic acids)

Covalent modification

Describes how the one macromolecule is modified by a specific enzyme with forming or breaking of a covalent bond. This mechanism type can also characterize protein function modification via covalent bond formation with a compound

Transcriptional regulation

- Interaction between Transcription factors and gene regulatory regions that results in expression change (increase or decrease of mRNA abundance)

Co-regulation of transcription

Interaction between different types of co-regulators of transcription and regulated genes, including:

- co-activators and co-repressors of transcription
- proteins providing epigenetic regulation of gene activity (DNA methylation-demethylation, histone acetylation-deacetylation)
- proteins (regulators of chromatin, etc) that interact with the regulatory region of a gene and play significant role in gene regulation

Influence on expression

Indirect influence of chemical compound (drug), ligand, receptor, or transcription factor on target gene expression. Influence on expression could be demonstrated both on RNA or Protein level

Unspecified

This mechanism type is used when there is reliable data about the effect (positive/negative), but no/conflicting evidence of the exact mechanism by which it occurs

Competition

Indicates competition between two molecules (proteins, xenobiotic compound, and natural ligand) for one binding site resulting in inhibition of expression or activity of the target molecule

Catalysis

Catalysis mechanisms indicate a reaction facilitated or accelerated by an enzyme.

Ubiquitination

An enzymatic post-translational modification in which a ubiquitin protein is attached to a substrate protein

Deubiquitination

An enzymatic post-translational modification in which ubiquitin is removed from substrate proteins

Sumoylation

An enzymatic post-translational modification in which a Small Ubiquitin-like Modifier (SUMO) protein is attached to a substrate protein

Desumoylation

An enzymatic post-translational modification in which a Small Ubiquitin-like Modifier (SUMO) is removed from substrate proteins

Neddylation

An enzymatic post-translational modification in which a ubiquitin-like protein NEDD8 is attached to a substrate protein

Deneddylation

An enzymatic post-translational modification in which a ubiquitin-like protein NEDD8 is removed from a substrate protein

Acetylation

An enzymatic post-translational modification in which an acetyl functional group is attached to a substrate protein

Deacetylation

An enzymatic post-translational modification in which an acetyl functional group is removed from a substrate protein

ADP-ribosylation

An enzymatic post-translational modification in which an ADP-ribose functional group is attached to a substrate protein

Glycosylation

An enzymatic post-translational modification in which a carbohydrate group is attached to a substrate protein

Methylation

An enzymatic post-translational modification in which a methyl group is attached to a substrate protein

Demethylation

An enzymatic post-translational modification in which a methyl group is removed from a substrate protein

S-nitrosylation

An enzymatic post-translational modification in which a NO is attached to a substrate protein

GPI-anchor

An enzymatic post-translational modification in which a Glycosylphosphatidylinositol (GPI anchor) is attached to a substrate protein



Sulfation

An enzymatic post-translational modification in which a sulfo group is attached to a substrate protein

Phosphorylation

An enzymatic post-translational modification in which a phosphate group is attached to a substrate protein

Dephosphorylation

An enzymatic post-translational modification in which a phosphate group is removed from a substrate protein

About Clarivate

Clarivate™ is a global leader in providing solutions to accelerate the pace of innovation. Our bold Mission is to help customers solve some of the world's most complex problems by providing actionable information and insights that reduce the time from new ideas to life-changing inventions in the areas of Academia & Government, Life Sciences & Healthcare, Professional Services and Consumer Goods, Manufacturing & Technology. We help customers discover, protect, and commercialize their inventions using our trusted subscription and technology-based solutions coupled with deep domain expertise. For more information, please visit clarivate.com

Contact our experts today:

+1 215 386 0100 (U.S.)

+44 (0) 20 7433 4000 (Europe)

clarivate.com