QIAGEN

White Paper

Regulator Effects in IPA®

Sample to Insight

Introduction

The goal of the Regulator Effects analytic in IPA is to provide insight into the causes and effects of differentially expressed genes or proteins in a dataset. Regulator Effects explains how predicted activated or inhibited upstream regulators might cause increases or decreases in phenotypic or functional outcomes downstream. These causal hypotheses take the form of directionally coherent networks formed from the merger of Upstream Regulator networks with Downstream Effects networks, which are described in a recent publication (Krämer et al. 2014) as well as in white papers available on www.qiagenbioinformatics.com. The resulting Regulator Effects networks can provide possible drug targets, mechanisms of toxicity, mechanism of efficacy and more.

The networks are derived dynamically based on the user's input dataset and findings in the Ingenuity Knowledge Base and are displayed in three tiers. The top tier is comprised of one or more upstream regulators, while the bottom tier is comprised of one or more diseases, functions, or phenotypes. The middle tier is made up of the dataset molecules that connect to the regulators above and to the diseases and functions below, and are predicted to be the intermediaries that carry the signal from the upstream regulators to downstream outcomes. If there are known relationships between upstream regulators and downstream diseases or functions in the bottom tier they are called out in the results and displayed in the networks. A Consistency Score is calculated for each Regulator Effect network, where higher scores are awarded to networks that are directionally consistent, meaning that most of the paths from regulator to target to disease/function are consistent with the predicted state of the regulator, the observed direction of expression of the target in the dataset and the expected impact on the disease/function downstream, based on findings from the literature.

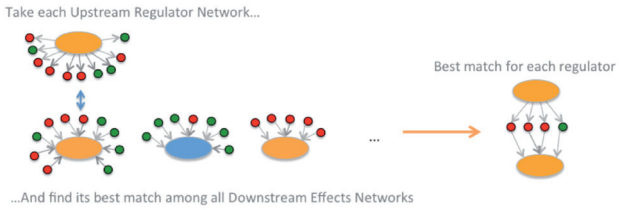
The Algorithm

The Regulator Effects algorithm generates hypotheses that explain how the activation or inactivation of regulators leads to an increase or decrease of function and disease-related outcomes based on the evidence provided by a dataset. These hypotheses are visualized as networks. The building blocks of these networks are the Upstream Regulator and Downstream Effect results from the analysis of the experimental dataset. The algorithm takes these building blocks as inputs and merges them to generate hypotheses through a process typically involving three phases.

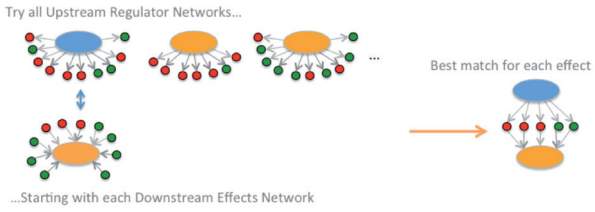
Phase 1: Initialization

In this first phase, networks from Upstream Regulator Analysis are paired with networks from Downstream Effects Analysis to create many simple networks consisting of a single regulator, a single function/disease, and the dataset molecules involved with both. This is described in the diagram below. Each Upstream Regulator is paired with the Downstream Effect that is its best match, and likewise, each Downstream Effect is paired with its best Upstream Regulator match.

Phase 1: Find the single best downstream effect network for each upstream regulator



Likewise, find the single best matching regulator for each downstream effect



In order to be considered a 'best match' pair, the combination of regulator and function/disease must meet the following criteria:

- The set of dataset molecules downstream of the regulator (i.e. regulator targets) and the set of dataset molecules upstream of the function/disease (i.e. molecules driving the effect on function/disease) must overlap by a minimum of three dataset molecules.
- The Fisher's Exact Test p-value of the overlap between the regulator and function/disease dataset molecules must be <= 0.05. The four inputs for this statistical test are

  1) The intersection of the regulator target set and the function/disease molecule set,

  2) The molecules in the regulator target set not in the intersection,

  3) The molecules in the function/disease set not in the intersection, and

  4) The molecules in the reference set (i.e. all molecules neighboring all regulators and all functions/diseases from the Upstream Regulator and Downstream Effect results that met the filter criteria) minus the molecules in the regulator and function/disease sets.

To find the best match for an Upstream Regulator, the algorithm merges the regulator and its targets with each Downstream Effect (if the pair meets the above criteria). For each best match pair, a Consistency Score (described in its own section below) is calculated on the resulting network and the network with the highest score is retained. If best match pairs have the same Consistency Score and involve the same dataset molecules, the networks are merged into a larger network containing the union of the regulators and functions/diseases.

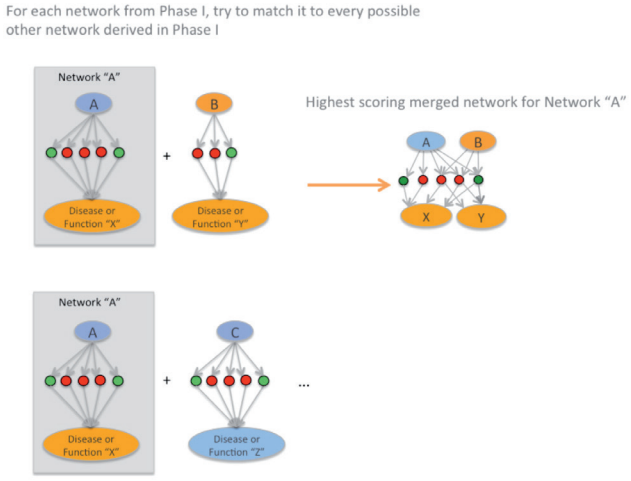
At the end of the Phase 1, the algorithm will have produced a collection of small networks generally consisting of a single regulator, a single function/disease, and their shared dataset molecules. If the user chose the "Minimal regulator to function networks" option when initiating Regulator Effects, the algorithm will stop at this point.

Phase 2: Merge by Score

The second phase of the algorithm starts with the set of simple networks produced by the first phase. For each network in the set, the algorithm pairs it with every other network in the set, looking for the merger that results in the highest Consistency Score. As in the first phase, the pair must meet

the minimum overlap and maximum p-value threshold in order to be considered. If the resulting merged network has a Consistency Score better than the individual Consistency Scores of the two original networks, the newly created network is retained; otherwise the two original networks are retained. This is depicted in the diagram below. An iteration consists of comparing every network in the set to all other networks, evaluating if the merged networks have a better Consistency Score. At the end of the iteration, if new networks have been created, then the process repeats again, this time operating only on the newly created networks. The iterations continue until no networks can be combined to create new larger networks with better Consistency Scores.

Phase 2: Merge pairs of networks if the merger creates a higher scoring network



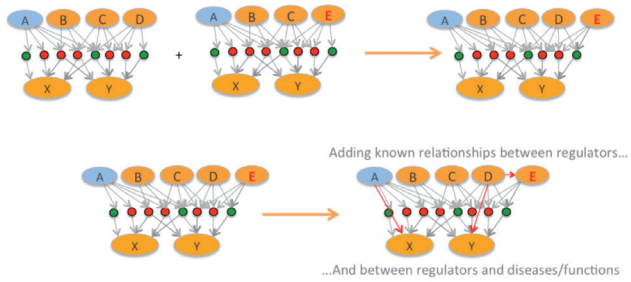
Phase 3: Merge by Similarity

By the end of the second phase, the Consistency Scores of the networks cannot be improved by merging. However, it is possible that some of the networks may be very similar in terms of the regulators, functions, and dataset molecules they contain, and they may appear almost redundant. Therefore, the third phase of the algorithm will merge networks that are so similar they essentially describe the same hypothesis. Similar networks will be merged in this step if:

- They have positive Consistency Scores
- The networks differ by three or fewer regulators and functions/diseases.
- The Consistency Score of the resulting merged network is at least 75% the original scores, i.e. reducing redundancy is prioritized over increasing the Consistency Score.

After networks are merged by similarity, any known relationships between regulators and between regulators and diseases/functions from the Ingenuity Knowledge Base are added to the networks for informational purposes (as shown in the schematic below as red lines). The final merger and annotation is shown in the diagram below. At this point, the Regulator Effect networks are complete.

Phase 3: Reduce redundancy by merging networks that consist of almost the same regulators and disease/functions



The results are displayed in a table in the IPA user interface, where each row represents a distinct Regulator Effects network ranked by Consistency Score. Additional columns are included to enumerate various characteristics of each network, and a link is provided to visualize each network.

The Consistency Score

The Consistency Score is a measurement used to help rank or prioritize the most useful networks. Its intent is to reward smaller networks with nodes highly connected by consistent relationships. A consistent relationship is one in which the direction of node activity/expression that is observed or predicted is consistent with the direction one would expect based on the findings from the Ingenuity Knowledge Base. The Consistency Score is not itself a measure of statistical significance, but a heuristic to rank a set of already statistically significant networks within the same analysis and settings. The Consistency Score formula is given as:

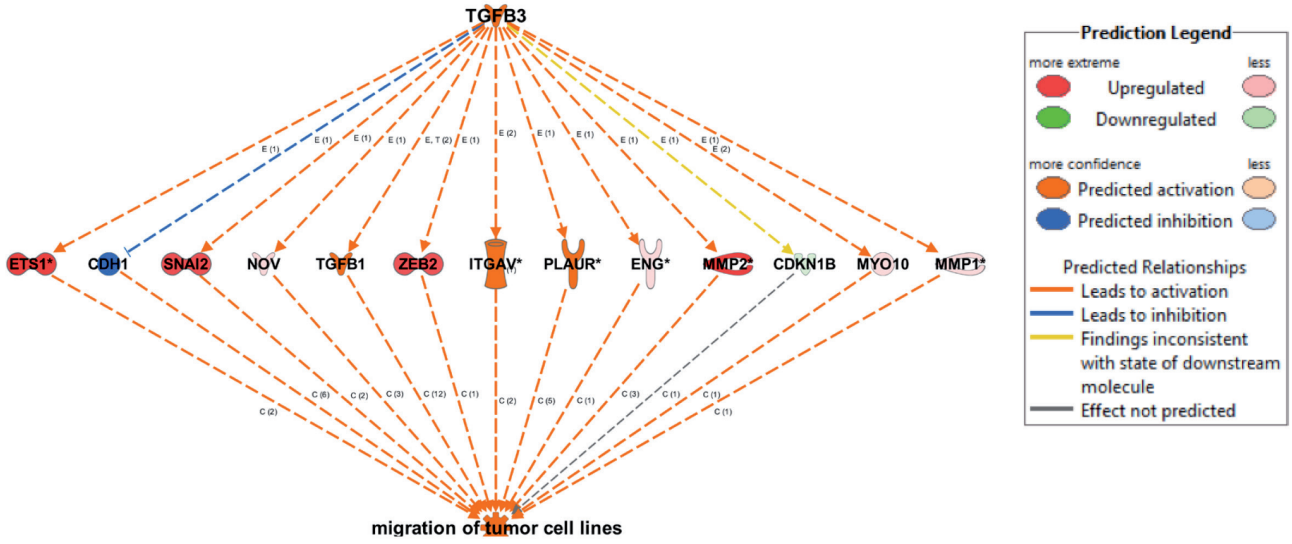
$$\text{Consistency Score} = \frac{Pc \cdot Wc + Pi \cdot Wi + Pn \cdot Wn}{(S)^{Ws}}$$

Where:

- Pc is the total number of consistent paths from regulator to function (through dataset targets).
- Wc is the weight that rewards for consistent paths and is set to 1.
- Pi is the total number of inconsistent paths.
- Wi is the weight that penalizes inconsistent paths and is set to -15.
- Pn is the total number of non-causal paths.
- Wn is the weight for non-causal paths. Set to 0 (e.g. non-causal paths don't affect the score).
- S is the size (the total number of dataset targets).
- Ws is a penalty weight for the size of the network and is set to 0.5.

Each path consists of two parts, from regulator to dataset molecule and from dataset molecule to disease/function. Note that in cases where one segment of the path is not causal, then the entire path is considered non-causal. Additionally, if one segment of a path is inconsistent, all paths containing that segment will be considered inconsistent.

To clarify how the Consistency Score is calculated, take the following real network with Consistency Score = 3.328. (See below)



In this network, there are 13 paths from TGFB3 to the function "migration of tumor cell lines." On 12 of those paths, the direction of activation/expression for the nodes is consistent with the direction expected based on the findings from the literature between the nodes.

One path, TGFB3->CDKN1B->migration of tumor cell lines is different. The path segment TGFB3->CDKN1B is inconsistent. The findings from the literature suggest that TGFB3 activates CDKN1B (as depicted by the arrowhead of the edge between them); therefore, one would expect that if TGFB3 were increasing in activity, CDKN1B would also increase. However in the dataset, the expression of TGFB3 was observed to be decreasing. Therefore the direction expected based on findings differs from the direction observed/predicted, and the edge is inconsistent and colored yellow. Additionally, the findings that support the path segment CDKN1B->migration of tumor cell lines have no causal effect; therefore it is colored gray. If any segment along a path has no known directional causal effect, then the entire path has no effect, so for the Consistency Score, the path TGFB3->CDKN1B->migration of tumor cell lines has no causal effect.

As a result, the individual values used in the calculation are as follows:

Pc = 12	(There are 12 consistent paths)
Wc = 1	(The weight for consistent paths is 1)
Pi = 0	(There are no inconsistent paths, only a non-causal path)
Wi =-15	(The weight for inconsistent paths is -15)
Pn = 1	(There is 1 non-causal path)
Wn=0	(The weight of non-causal paths is 0, these do not affect the score)
S = 13	(There are 13 dataset targets, shown in the middle layer of the network)
Ws = 0.5	(This constant is set to 0.5)

$$\text{Consistency Score} = \frac{Pc \cdot Wc + Pi \cdot Wi + Pn \cdot Wn}{(S)^{Ws}}$$

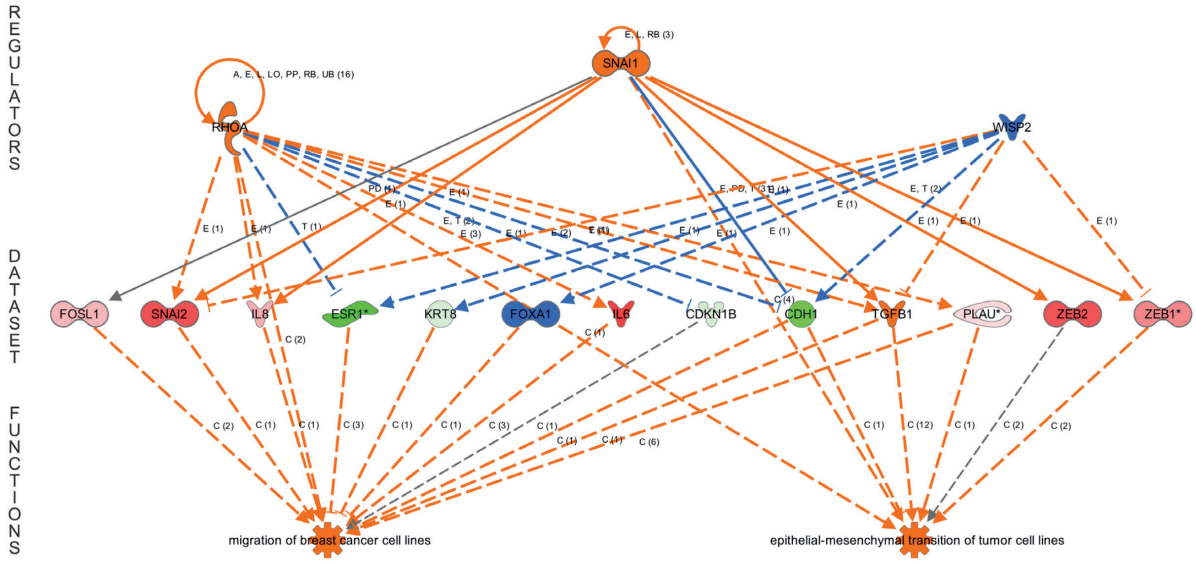
= (12*1) + ((0*(-15)) + (1*0) / (13^0.5) = 3.328

Regulator effect example from a real dataset

On the next page is an example of a top scoring network obtained from the Regulator Effects analysis of RNA-Seq gene expression data from "claudin-low" type breast cancer cell lines that have been ratio'ed to breast cancer lines with a more luminal-like gene expression pattern. These claudin-low cell lines have been characterized as a relatively more aggressive type of breast cancer having potentially undergone an epithelial to mesenchymal transition (Prat and Perou, 2011).

In the network shown, orange and blue lines represent relationships with causal consistency (though in this particular network, there are no inconsistent relationships). For example, an orange line connects SNAI1 (which is itself filled with orange color because the Upstream Regulator Analysis predicted it to be activated) to ZEB2 (which is overexpressed in the claudin-low lines and is therefore filled with red color) because SNAI1 is known from the literature to increase the expression of ZEB2 (See for example Taube et al 2010).

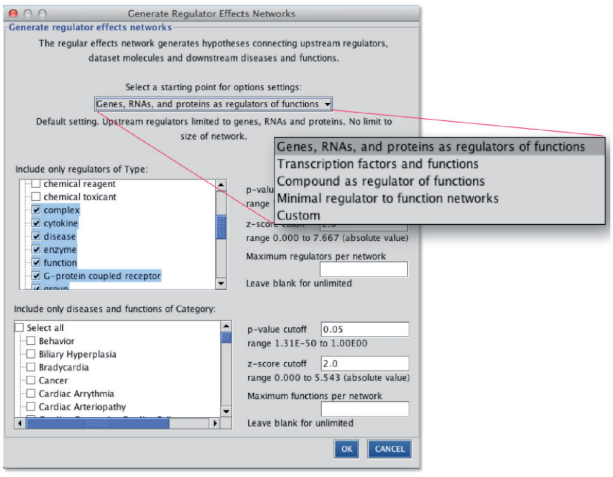
Note that some of the dataset targets displayed in the middle tier may themselves have been predicted to be activated or inhibited upstream regulators in the analysis, and in such cases will be colored orange or blue respectively (for example TGFB1 in the network above). In such instances, it is possible that the dataset target's differential expression



value may be in conflict with its upstream regulator z-score. For example, the gene may be downregulated in the dataset (and would have been colored green), but the upstream regulator analysis predicted it to be activated. Regulator Effects will prioritize z-score over differential expression, so such a gene would be considered "activated" when creating the Regulator Effects network and will be rendered orange rather than green. In this particular instance, TGFB1 is in agreement with the z-score – it is upregulated in the dataset.

As mentioned above, relationships between regulators and between regulators and diseases/functions are displayed in the network if they are known in the Ingenuity Knowledge Base. For example, relationships are shown between the regulator SNAI1 and the function "epithelial-mesenchymal transition of tumor cell lines". This feature enables the user to discover potentially novel relationships if there is no line shown between a regulator and a function for example.

There are a number of options provided by IPA so that the user can tailor the type of Regulator Effects networks that are generated to answer specific research questions. By default, Regulator Effects only includes upstream regulators that are genes, RNAs, or proteins (e.g. excludes chemicals/drugs), and demands that the regulators and diseases/functions that are fed into the algorithm have an absolute z-score >2 and p-value <0.05. However, the user can change these settings to include for example only single chemicals, microRNAs, or growth factors as upstream regulators or for example to exclusively consider cardiovascular diseases downstream. Or the user can use more lenient or more stringent z-scores and p-values as input when generating the networks. These settings are made in IPA as shown below:



Conclusion

Regulator Effects in IPA helps you make insights about your data by integrating the Upstream Regulator results with Downstream Effects results to create causal hypotheses that explain how upstream regulators may cause particular phenotypic or functional outcomes downstream. It provides simplifying and actionable hypotheses that increase the value of gene and protein expression experiments.

References

1. Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics (Oxford, England), (iv), 1–8. doi:10.1093/bioinformatics/btt703

2. Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. Molecular Oncology. doi:10.1016/j.molonc.2010.11.003

3. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, Hartwell K, Onder TT, Gupta PB, Evans KW, Hollier BG, Ram PT, Lander ES, Rosen JM, Weinberg RA, Mani SA. (2010). Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proc Natl Acad Sci USA. 2010 Aug 31;107(35):15449-54. doi: 10.1073/pnas.1004900107.

You can learn more about IPA, or sign up for a free trial, at www.qiagenbioinformatics.com.

QIAGEN Bioinformatics

EMEA
Silkeborgvej 2 · Prismet
8000 Aarhus C
Denmark
Phone: +45 7022 5509

Americas
1001 Marshall Street, Suite 200
Redwood City, CA 94063
USA
Phone: +1 (617) 945 0178