Manual

In Silico Experiment System for Testing Gene Function Hypothesis Using Three Networks in the Knockout Mouse Data

For functional study of gene, formulation a hypothesis and confirmation using biological experiments needs much time and effort. Also researchers generate NGS data (microarray or RNA-seq data) of mRNA and miRNA in knock-out (KO) mice to discover the functions of gene. Our in silico experiment system extracts candidate gene for hypothesis through BEST, which is context-aware based search tool, and verifies the hypothesis using the network analysis ,which searches genes related with regulator gene and target genes and integrates three different networks, TF, miRNA and PPI networks. Through this tool, researchers identify the correlation between the regulator gene and disease hypothesis with visual and numerical result and rapidly perform with various hypotheses to help formulating accurate hypothesis. This tool also allows us easily to discover previously unrecognized high association between genes and diseases and to design easily biological verification experiments using network result.

Methods

Our in silico experiment system is to search candidate target gene of hypothesis and verify the relationship between regulator gene and target gene in knockout mouse data. Our tool works with comparable two class mRNA, microRNA expression data, regulator gene name and hypothesis.

When user has two class comparison data of mRNA and miRNA, this tool show the correlation of regulator gene and candidate target genes of hypothesis with three-networks. Three-networks are miRNA, TF (transcription factor) and PPI (protein-protein interaction) networks.

User input

- Gene (mRNA) expression data
- microRNA expression data
- Regulator gene name (e.g. knock-out gene)
- Hypothesis (e.g. disease, pathway)

Our tool supports raw data of microarray (i.e. *.CEL files), processed data of microarray (i.e. matrix files of gene expression) and processed data of RNA-seq (i.e. matrix files of read counts). We do not support raw data of RNA-seq.

Output result

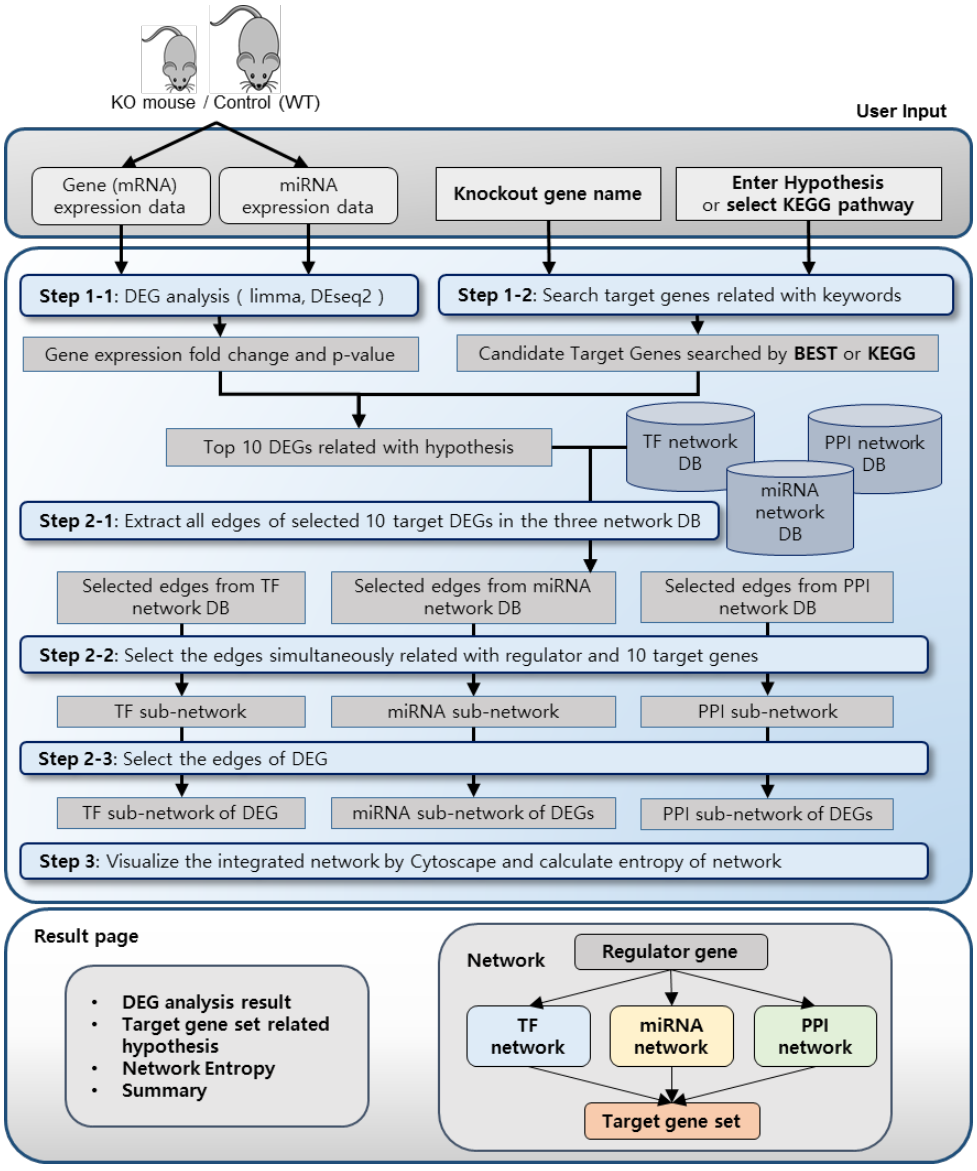
- Gene expression analysis result (download filtered_gene.txt)
- Candidate target gene list related with regulator gene and hypothesis
- Network of regulator gene and target gene within TF, microRNA and PPI network
- Entropy of the network and p-value

The in silico experiment system developed by our team provides services on the Internet. The user must input the expression data of mRNA and miRNA as microarray or RNA-sequence data in knockout mouse and wildtype for comparison. Our system was optimized to analyze within 1 minute.

The analyzed results showed the candidate target gene based on the BEST result of two keyword knockout gene name and hypothesis. Furthermore, top 10 DEGs among the candidate genes were selected with integrated DEG analysis result, and generated a network of the genes from three network database. This network consists of TF network, miRNA network, and PPI network, and shows network results by selecting with gene expression values.

The result of networks provide entropy and p-values together to establish the criteria for the significance of network analysis results. All the results data could be saved in user local system, and easily change the hypothesis and re-analyze quickly, and also reconfigure the network by inputting more genes.

Workflow



Step 1. Selection top 10 target genes of regulator and hypothesis

1-1. DEG analysis of miRNA and mRNA

The input mRNA and miRNA data are analyzed using limma (Ritchie et al., 2015) for microarray and DEseq2 (Love et al., 2014) for RNA-sequencing data according to the data format. The result of this analysis is that the log2 fold change value and the p-value are calculated for the expression level of each gene.

1-2. Search target genes related with keywords

The user has to enter two keywords, the regulated gene name and hypothesis (or pathway selection). The entered gene name can be checked if it is an available gene name. When the user enter the hypothesis, the BEST(Lee et al., 2016) program based on the literature search is used to retrieve all candidate genes for the keywords in API format and store them as the result values by setting the knockout gene name and hypothesis inputted by the user as keywords. When the user selects the pathway name, associated genes were found through the KEGG (Khanzada et al., 2017) pathway database.

1-3. Select top 10 hypothesis related target DEGs

Using the DEG analysis (step 1-1) results, top 10 DEGs are selected by comparing the expression changes of the searched genes found by BEST tool or KEGG DB (step 1-2) with the p-value.

Step 2. Network generation

2-1. Extract all edges of selected 10 target DEGs in the three network DB

For each of the 10 genes selected in the previous step and the modified genes entered by the user, all the genes linked to each of the three network TF, miRNA and PPI databases are found. The miRNA network database was obtained from TargetScan (Lewis et al., 2005), and the PPI network database was used by STRING (Franceschini et al., 2013). The TF network data-base was created using NARROMI (Zhang et al., 2013).

2-2. Select the edges related with regulator and 10 target genes.

At the same time, we find out the gene network linked to the target genes by the influence of the modified genes in the network found in the previous step.

2-3. Select the edges of DEGs

The change in the expression level of the modified gene affects the expression amount of the gene linked to the network, and the change in the expression level of this gene will also change the expression amount of the target gene. Therefore, and it was removed from the network. Thus, the final network was constructed.

Step 3. Visualization and network validation

3-1: Network visualization.

The network completed in the previous steps visualized the network using Cytoscape (Shannon et al., 2003). The regulated gene is located at the top, and the candidate target genes are located at the bottom, and grouped into TF, miRNA, and PPI networks, and how each genes are connected according to changes in expression level. The expression level of each gene was visualized according to the amount of change in color. Up-regulated genes are red and down-regulated genes are blue. In addition, the coding gene is represented by a circle and the non-coding gene is represented by a diamond shape. The blue colored edge represents the miRNA network, the purple colored edge represents the PPI network, and the yellow colored edge represents the TF network. When a specific gene is clicked, it provides a function to make a list of related genes, and also a convenient function to find a desired gene.

3-2: Network Verification

We verified the network created by our tool to see how meaningful and how it relates to the hypothesis. The regulator gene and the target gene were fixed, and the network was formed through the randomly generated network, which was repeated 1000 times. We counted number of DEGs on the each networks and generated distribution for number of DEGs. Using this distribution, the p-value was calculated for the number of DEGs in the network generated by the in silico experiment. The significance of p-value means that the regulator gene has an effect on the target genes through many mediator DEGs. If there is a high correlation between the regulator gene and the target genes, there is a large number of mediator genes linking the regulator and the target genes. In addition, if the regulator gene relates with given hypothesis, target genes are evenly connected with three-level

networks. To measure this property, normalized entropy of generated network is calculated by using degree information of target genes (Equation (1), (2), (3)).

$$\mathrm{P}(tg_i) = \frac{degree(tg_i) + \beta}{\sum_{j=1}^{10}(degree(tg_i)+\beta)} \quad (1)$$

$$\mathrm{H(TG)} = -\sum_{i=1}^{10} P(tg_i) log_2 P(tg_i) \quad (2)$$

$$E_H = \frac{H(TG)}{H_{max}} = \frac{H(TG)}{log_2 10} \quad (3)$$

where:

degree(tg_i): degree of i-th target gene in the Target Gene Set TG = {tg_1, tg_2, …, tg_{10}}
β: pseudo count (β=0.00001)
H_{max}: maximum entropy of generated network

Optimization

P-value calculation optimization

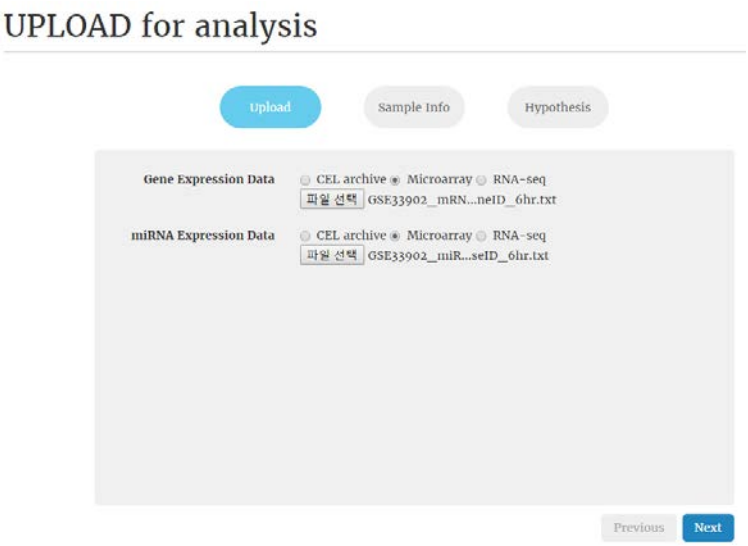
To compute the p-value, a network with 18 million edges and 31897 nodes was randomly generated 5,000 times iteration, and samples of the number of genes mediating between the regulator and the target gene in the each network were generated sample distribution. It took a very long time to create 18 million random edges of the entire network, so instead of rebuilding the entire network, we created a partial network. The edge formation probability is obtained by dividing the total number of edges by the number of the fully connected network. With this 0.0358 probability, a partial network of the regulator-target gene was generated.

Network generation optimization

In the entire network, reading 18 million edges and 30,000 nodes and extracting the sub-network requires a lot of computation. Our system pre-reads the entire network and populates the "network server". When searching regulator and target gene subnetwork, the searching process is optimized by reducing the search space to the limited neighbor gene of the regulator. This reduces the time it takes to compute the network from more than 10 seconds to two seconds, and adds additional functionality.

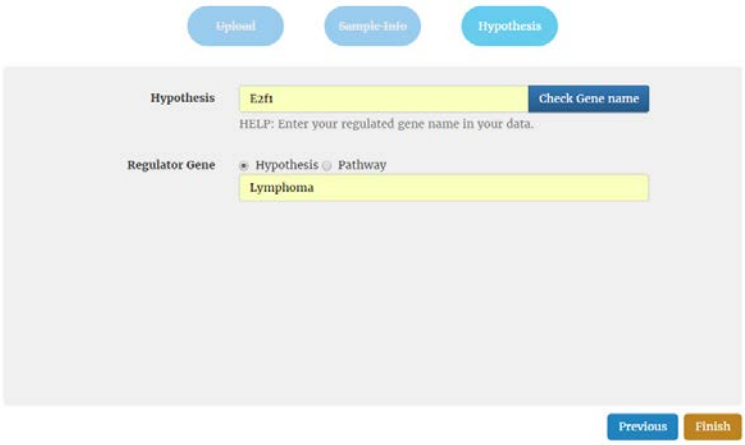
USER UPLOAD

1. Select files in user computer and file type to upload.



2. Check wildtype samples for comparison.

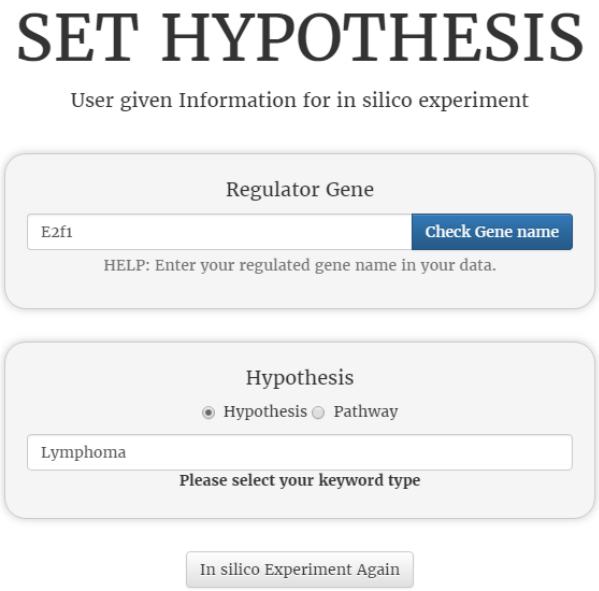
3. Enter the knockout gene name and hypothesis or select pathway.
   Users can check the entered gene name.



4. Click the '**Finish**' button to submit.

RESULT PAGE

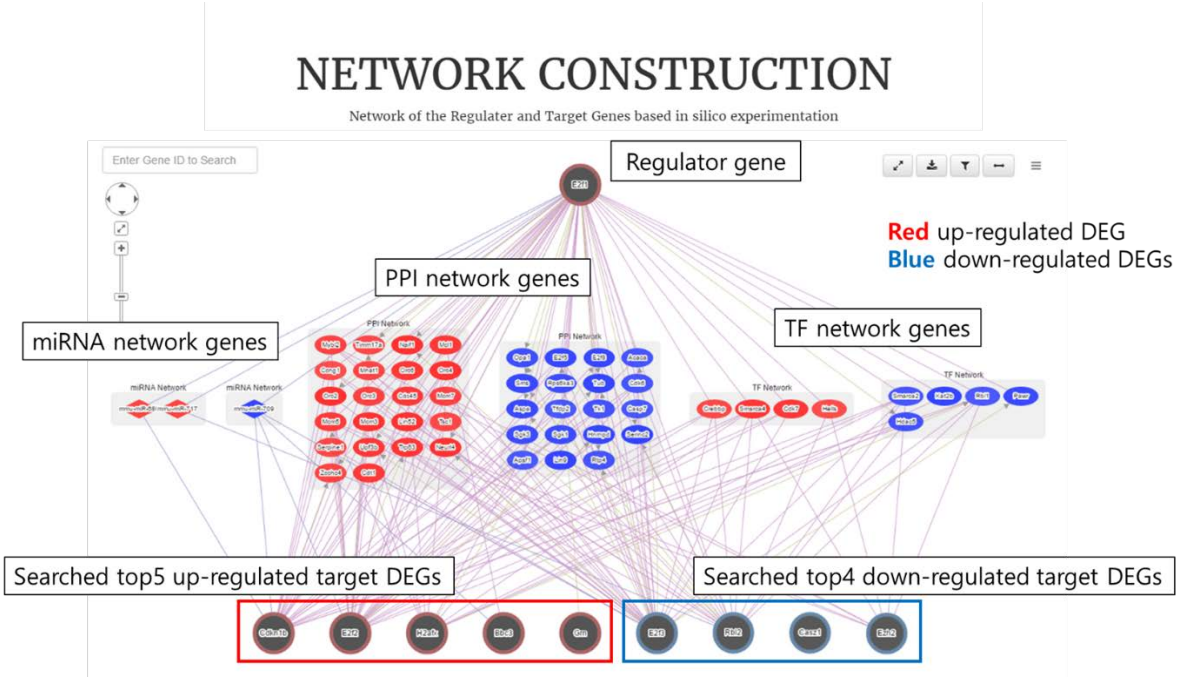
Result has 4 parts, (1) Set hypothesis, (2) Search target genes, (3) Network construction, and (4) Summary and score.



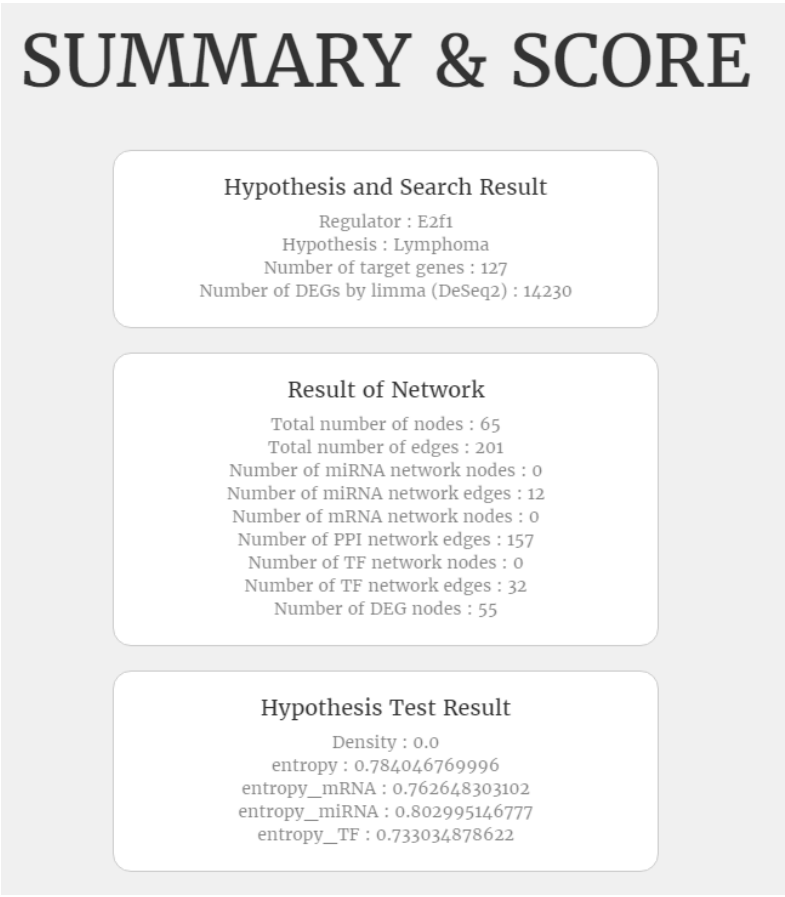
The first part shows the user input information. Users can view the information and immediately rerun other experiment.

The second part shows the top 10 candidate target genes found. The top 10 candidate target genes are selected by integrated all candidate target gene set and result of DEG analysis. User can download the each result by clicking the '**Download'** button.

If user want to add another target genes to the network result, enter the name of custom target genes and click the **'Add new gene'** button.



The third part shows the network result. This network are connected with miRNA, TF and PPI network from the regulator gene to the top 10 target genes. Gene expression values are displayed in red or blue color. The red gene is and up-regulated DEG and blue one is down-regulated DEG. User can download the network image. In the network result, we provide zoom function and users can freely move the gene position.



The final part shows the summary of results and network entropy score.