frontiers in Genetics

METHODS
published: 03 May 2019
doi: 10.3389/fgene.2019.00421

DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists

Liang Sun[1], Sufen Dong[1,2], Yinbing Ge[1], Jose Pedro Fonseca[1], Zachary T. Robinson[1], Kirankumar S. Mysore[1] and Perdeep Mehta[1]*

[1]Noble Research Institute, Ardmore, OK, United States, [2]College of Information Science and Technology, Hebei Agricultural University, Baoding, China

Gene expression data generated from multiple biological samples (mutant, double mutant, and wild-type) are often compared *via* Venn diagram tools. It is of great interest to know the expression pattern between overlapping genes and their associated gene pathways or gene ontology (GO) terms. We developed DiVenn (Dive into the Venn diagram and create a force directed graph)—a novel web-based tool that compares gene lists from multiple RNA-Seq experiments in a force-directed graph, which shows the gene regulation levels for each gene and integrated KEGG pathway and gene ontology knowledge for the data visualization. DiVenn has four key features: (1) informative force-directed graph with gene expression levels to compare multiple data sets; (2) interactive visualization with biological annotations and integrated pathway and GO databases, which can be used to subset or highlight gene nodes to pathway or GO terms of interest in the graph; (3) Pathway and GO enrichment analysis of all or selected genes in the graph; and (4) high resolution image and gene-associated information export. DiVenn is freely available at http://divenn.noble.org/.

Keywords: Venn diagram, visualization, transcriptome data, KEGG, gene ontology, pathogen infection

OPEN ACCESS

Edited by:
Alfredo Pulvirenti,
Università degli Studi di Catania, Italy

Reviewed by:
Fuliang Xie,
Qiagen, United States
Andreas Dräger,
University of Tübingen, Germany

****Correspondence:***
Liang Sun
sunliang@udel.edu;
lsun@noble.org

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *01 February 2019*
Accepted: *17 April 2019*
Published: *03 May 2019*

Citation:
Sun L, Dong S, Ge Y, Fonseca JP, Robinson ZT, Mysore KS and Mehta P (2019) DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists. Front. Genet. 10:421. doi: 10.3389/fgene.2019.00421

INTRODUCTION

With the advance of high-throughput data technologies, huge amounts of gene expression data were generated without in-depth analysis. Several web-based visualization tools—for example, INVEX (Xia et al., 2013), ExAtlas (Sharov et al., 2015), and WebGIVI (Sun et al., 2017)—were successfully used in expression data analysis. However, systematically comparing more than two experiments is still challenging. It is especially challenging to visualize multiple experiments' data along with integrated bioinformatics databases. Venn diagrams are widely used to compare gene lists among multiple experiments. GeneVenn (Pirooznia et al., 2007), Venny (Oliveros, 2015), and InteractiVenn (Heberle et al., 2015) are examples of web-based tools currently being used. However, they have significant limitations: (1) gene IDs cannot be linked to gene functions. No bioinformatics databases such as biological pathway and gene ontology (GO) can be integrated. (2) Gene expression levels cannot be displayed in the graph.

(3) Common or unique genes, which are likely to be interesting in the Venn diagram cannot be extracted with gene expression value and gene function.

We provide an interactive web-based tool that will overcome the above-mentioned limitations and help biologists visualize their gene lists and generate biological hypotheses based on the integrated knowledge from biological pathway and GO databases. Using this tool, researchers cannot only compare and visualize gene lists, but also subset or highlight the gene nodes in the graph based on gene functions of interest. This tool is user friendly and can handle large amounts of input data by using a force-directed focus package.[1] Users can extract and download important gene information from the result/information table and download the high resolution image for their publications.

FEATURES

DiVenn was developed using PHP, JavaScript, R, D3.js (Bostock et al., 2011), and MySQL database. The flow chart of the data visualization is depicted in **Figure 1**. DiVenn currently accepts two types of input data: (1) two-column tab-separated custom data. For example, gene ID and corresponding pathway data, transcription factors and their regulated downstream genes, and microRNAs and corresponding target genes. The second column must be "1" or "2". (2) Gene expression data. The first column is gene IDs and the second column is gene regulation value. The gene regulation value should be obtained from differentially expressed (DE) genes. Users can select the cut-off value of fold change (for example, two-fold change) to define their DE genes. To simplify this gene regulation value, we require users to use "1" to represent upregulated genes and "2" to represent downregulated genes based on their own cut-off value of fold change. If users need to link their genes to the KEGG pathway (Kanehisa et al., 2019) or GO database, 14 model species with KEGG pathway and GO database available are supported in DiVenn. Currently, three types of gene IDs—KEGG gene IDs, Uniprot gene IDs (UniProt, 2008), and NCBI gene IDs (Benson et al., 2018)—are accepted for pathway analysis. All agriGO (Du et al., 2010; Tian et al., 2017) supported IDs are accepted for GO analysis by DiVenn. DiVenn allows users to compare and visualize up to eight gene lists in the network graph.

DiVenn has four major functions:

1. Comparison of gene lists from multiple experiments in the force-directed network graph with more integrated bioinformatics knowledge. As opposed to traditional Venn diagrams, which can show only the total number of overlapping genes among experiments, DiVenn can provide the information of gene expression regulation level, gene description, KEGG pathway, and GO terms.
2. Pathway and GO enrichment analysis of gene lists. DE genes from experiments with similar treatments are more likely to involve in the same pathway or GO terms. DiVenn applied the modified Fisher Exact test to enrich all or selected DE genes into significant KEGG pathways and GO terms similar to what have been used in well-known DAVID software (Jiao et al., 2012).[2] The modified Fisher Exact test is constructed by using an R script (fisher.test).
3. Subset and highlighting of gene nodes of interest. To better visualize the input gene list and also subset the gene list to gene groups, especially to avoid "hairballs" when the gene list is too large, DiVenn has a function to subset gene lists to pathway and GO groups of interest in the force-directed network graph, or change the node shapes of genes of interest to square shapes in the original graph by using the redraw function in the information table. Accordingly, a new modified Fisher Exact test can be performed for the subset gene list.
4. High-resolution image and selected data in the graph export. The graph generated by DiVenn can be downloaded as portable network graphics (PNG) and scalable vector graphics (SVG) files. SVG images can be converted to high-resolution images.

RESULTS AND DISCUSSION

Data Uploading and Processing

DiVenn accepts two-column tab-separated data for each experiment. The input data for each experiment is processed into JSON format and visualized *via* D3.js library. Each experiment is represented as one parent node (experiment node) with an automatically assigned color. All genes corresponding to the experiment are connected to the parental nodes *via* edges. If a gene node is clicked, the edges connecting to this gene node will be colored based on the expression values of this gene in the connected experiment node. GUI functions in the DiVenn graph provide the ability to: (1) Switch the gene label on and off. (2) Download images and data. (3) Change node color. (4) Display gene-related annotation, pathway, and GO information in a sortable table format.

Database Integration

KEGG gene IDs and corresponding UniProt and NCBI gene ID maps, gene function description, and KEGG pathways are captured through KEGG API by self-written PYTHON scripts. GO ID, terms, and categories were downloaded from agriGO (Du et al., 2010; Tian et al., 2017). All these information were stored in the MySQL database and automatically updated in our systems. The KEGG pathway and GO database of 14 model species were integrated to DiVenn. Right-clicking a gene node in the force-directed graph provides options to show the gene name and gene-detailed information (gene description, KEGG pathway, and GO terms).

[1]http://ramblings.mcpher.com/Home/excelquirks/gassites/d3nodefocus

[2]https://david.ncifcrf.gov/content.jsp?file=functional_annotation.html#EXP

Graph Redraw

DiVenn allows users to show all genes and their gene functions, associated KEGG pathways, KEGG maps, and GO terms in an information table. The modified Fisher Exact test is applied to all or selected genes in the graph. The *p*'s for pathways and GO terms are shown in the information table. This table can be sorted based on each column and is searchable by keywords of interest. Users can sort the table based on KEGG pathways or GO terms and can redraw gene nodes with square nodes in the graph when users need to highlight gene nodes

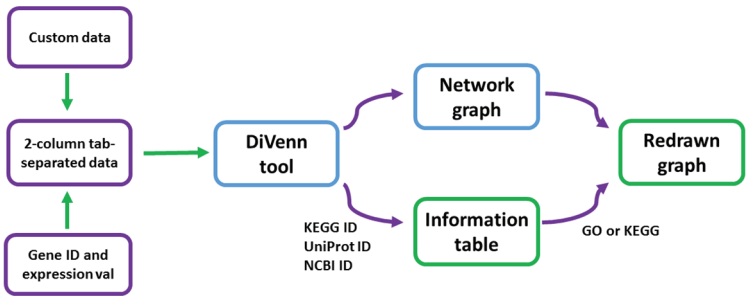
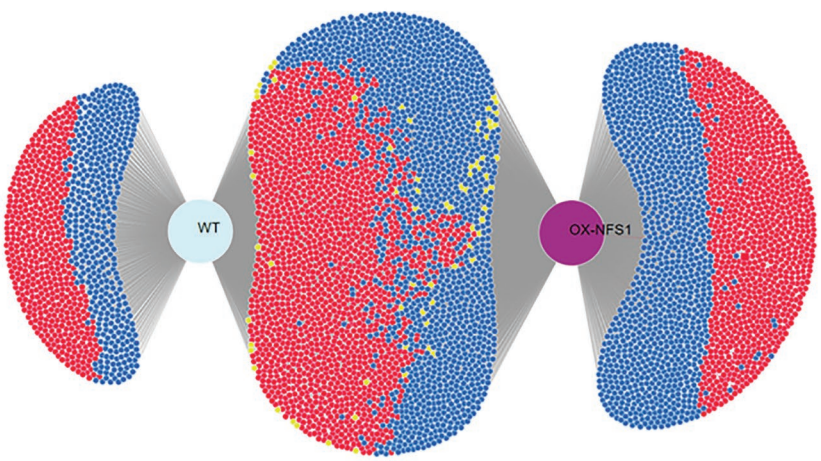


FIGURE 1 | Flow chart of DiVenn. Two-column tab-separated gene ID and expression value data from supported 14 species can be used to redraw the graph based on the KEGG or GO information table.

A



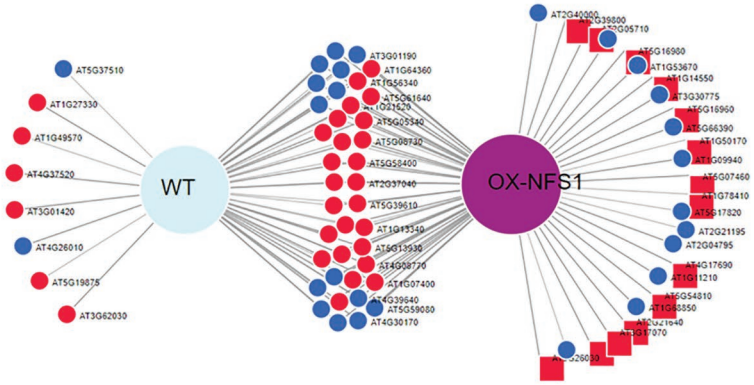
B

Pathway	p value
Biosynthesis of secondary metabolism	2.78E-20
Biosythesis of amino acids	7.00E-03
Carbon metabolism	7.00E-03
Cysteine and methionine metabolism	4.00E-03
Glutathione metabolism	1.00E-04
Phenylpropanoid biosynthesis	6.69E-07
Plant hormone signal transduction	1.78E-07
Plant-pathogen interaction	3.00E-04
Starch and sucrose metabolism	2.17E-05
Sulfur metabolism	9.70E-02
Tryptophan metabolism	2.00E-04

GO term	p value
Response to oxidative stress	2.72E-05
Response to salicylic acid	9.09E-05
response to jasmonic acid	6.41E-10
Response to auxin	3.17E-07
Response to light stimulus	4.00E-03
Oxygen binding	5.93E-07
Kinase activity	9.51E-05
regulation of transcription, DNA-templated	5.66E-06
Response to salt stress	1.00E-03
response to water deprivation	6.00E-04
Oxidoreductase activity	6.00E-03
ATP binding	3.00E-03
Calcium ion binding	9.00E-04
Defense response to bacterium	2.00E-03
Systemic acquired resistance	2.00E-03
Gluthatione transferase activity	4.00E-04
Toxin catabolic process	5.00E-04
Protein tyrosine kinase activity	1.00E-03
Protein phosphorylation	2.00E-04
Response to abscisic acid	1.00E-04
Response to wounding	7.58E-09
Lipid metabolic process	1.00E-03
Response to chitin	7.39E-06
Response to Heat	1.63E-05
Electron carrier activity	3.00E-03
Defense response to fungus	1.00E-03
Defense response	3.37E-05

FIGURE 2 | DiVenn case study showing differentially expressed genes between WT and *NFS1* overexpression line (OX-*NFS1*) in *Arabidopsis* upon bacterial pathogen inoculation. **(A)** Differentially expressed genes, up and downregulated (log2(FC)≤ −2 and ≥2) between 0 and 3 days, after pathogen inoculation in *Arabidopsis* for WT and OX-*NFS1* plants. Blue and red nodes denote downregulated and upregulated genes between different genotypes, respectively. Yellow nodes denote upregulation in one sample but downregulation in another. **(B)** Several pathways and GO terms significantly enriched ($p < 0.05$) in our dataset.

A



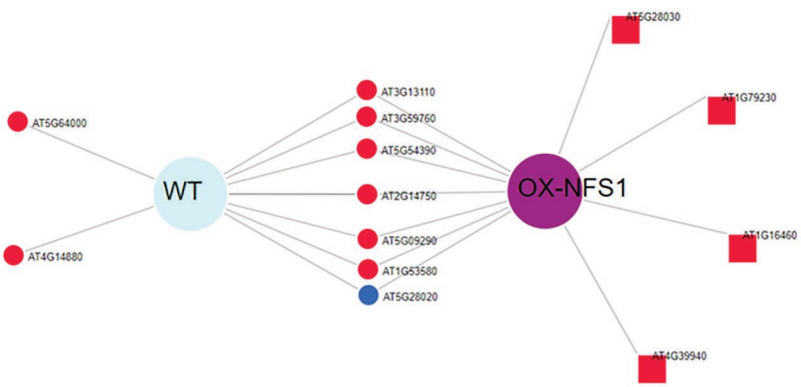
B

Show 100 entries Search: oxidative stress

ID	GO ID	GO term	GO category	OX-NFS1	WT	P Value
AT1G07400	GO:0006979	response to oxidative stress	P	1	1	2.728467e-05
AT1G09210	GO:0006979	response to oxidative stress	P	1	1	2.728467e-05
AT1G09940	GO:0006979	response to oxidative stress	P	2	n/a	2.728467e-05
AT1G11210	GO:0006979	response to oxidative stress	P	2	n/a	2.728467e-05

FIGURE 3 | DiVenn subset analysis of genes obtained using redraw function from gene ontology (GO) search "oxidative stress" displaying 73 DE genes up and downregulated from OX-*NFS1* and WT. **(A)** Genes upregulated only for OX-*NFS1* are highlighted as square nodes *via* the redraw function. **(B)** Table consisting of genes from subset graphic GO search "oxidative stress" displaying DE genes between the two genotypes OX-*NFS1* and WT (just the four first rows out of 73) and GO category *p*'s. The numbers 1 and 2 refer to upregulated and downregulated genes, respectively, for each *Arabidopsis* gene accession number.

A



B

Show 50 entries Search: sulfur metabolism

ID	Description	Pathway	OX-NFS1	WT	P Value
AT1G16460	RDH2; rhodanese homologue 2	Sulfur metabolism - Arabidopsis thaliana (thale cress)	1	n/a	5.093834e-35
AT1G53580	GLY3; glyoxalase II 3	Sulfur metabolism - Arabidopsis thaliana (thale cress)	1	1	5.093834e-35
AT1G79230	MST1; mercaptopyruvate sulfurtransferase 1	Cysteine and methionine metabolism - Arabidopsis thaliana (thale cress)	1	n/a	2.159887e-12
AT1G79230	MST1; mercaptopyruvate sulfurtransferase 1	Sulfur metabolism - Arabidopsis thaliana (thale cress)	1	n/a	5.093834e-35

FIGURE 4 | DiVenn subset analysis of genes obtained using redraw function from KEGG pathway analysis search "sulfur metabolism" showing 14 DE genes up and downregulated from OX-*NFS1* and WT. **(A)** Genes upregulated only for OX-NFS1 are highlighted as square nodes *via* the redraw function. **(B)** Table consisting of genes from subset graphic KEGG pathway search "sulfur metabolite" displaying DE genes between the two genotypes OX-*NFS1* and WT (just first four rows out of 14) and pathway *p*'s. The numbers 1 and 2 refer to upregulated and downregulated genes, respectively, for each *Arabidopsis* gene accession number.

of their interest, or subset gene nodes in a new graph based on the specific KEGG pathway or GO terms of interest, which will simplify the graph especially when a large gene list is being visualized. This knowledge is critical for users making biological hypotheses.

Case Study

In order to demonstrate how DiVenn performs network, GO and pathway analysis using RNAseq data, we obtained a dataset consisting of all DE genes ($\log_2$ fold change ≥ 2 and ≤ -2) **Figure 2A**. This dataset consists of genes differentially expressed between 0 h (basal control) and 3 days after bacterial pathogen *Pseudomonas syringae* pv. tomato DC3000 inoculation in 5 week-old *Arabidopsis* plants for two different genotypes: wild-type (WT) and *NITROGEN FIXATION S (NFS1)-LIKE 1* overexpression line (OX-*NFS1*). We were also able to quickly select several pathways significantly enriched in our dataset from GO and KEGG pathway such as Oxidative stress (GO) and sulfur metabolism (pathway) (**Figure 2B**).

DiVenn allowed us to quickly select only genes involved in the oxidative stress GO terms (GO:0006979) using a keyword search between DE genes from different experiments and generate a subset graph (**Figure 3A**). Oxidative stress genes are involved in several plant stress responses, including biotic and abiotic stresses as well as reactive oxygen species (ROS) generation. We created a subset graph of all 73 oxidative stress-related genes, significantly ($p < 0.05$) DE between both lines using the redraw function and we found more oxidative stress-related genes upregulated in the OX-*NFS1* line in comparison to WT upon bacterial pathogen infection (**Figures 3A,B**). Similarly, searches using "sulfur metabolism" as a keyword under the pathway menu allowed us to visualize all 14 significantly DE genes involved in sulfur metabolism from both genotypes (**Figures 4A,B**). We found more sulfur metabolism genes significantly upregulated in the OX-*NFS1* line compared to WT line upon pathogen treatment such as *MERCAPTOPYRUVATE SULFURTRANSFERASE 1* (*MST1*; AT1G79230) and *L-CYSTEINE DESULFHYDRASE 1* (*DES1*; AT5G28030) that are involved in the sulfur metabolic pathway. The above-mentioned examples are illustrative of the fact that plants under biotic stress go through an extensive transcriptional reprogramming affecting several genes from different pathways and organelles.

CONCLUSION

We have successfully developed DiVenn, a web-based tool to visualize large gene lists from multiple experiments. This tool provides a promising approach for comparing multiple gene expression data sets. The integrated bioinformatics databases and interactive visualization graph will help biologists generate biological hypotheses.

AUTHOR CONTRIBUTIONS

LS conceived the original research plans. LS, YG, SD, and ZR developed this tool. JF and KM performed the case study. LS, JF, ZR, and PM wrote the manuscript. LS, KM, and PM edited the manuscript.

FUNDING

This project was supported by the innovation project from the Department of Enterprise Systems and Informatics at the Noble Research Institute and the Excellent Going Abroad Experts' Training program in Hebei Province.

ACKNOWLEDGMENTS

The authors thank Melanie Davis and Jody Beard for the support from Department of Enterprise Systems and Informatics at the Noble Research Institute. They also thank Andrea Mongler for proofreading the manuscript.

REFERENCES

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., et al. (2018). GenBank. *Nucleic Acids Res.* 46, D41–D47. doi: 10.1093/nar/gkx1094

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D(3): data-driven documents. *IEEE Trans. Vis. Comput. Graph.* 17, 2301–2309. doi: 10.1109/TVCG.2011.185

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 38, W64–W70. doi: 10.1093/nar/gkq310

Heberle, H., Meirelles, G. V., Da Silva, F. R., Telles, G. P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* 16:169. doi: 10.1186/s12859-015-0611-3

Jiao, X. L., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28, 1805–1806. doi: 10.1093/bioinformatics/bts251

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* 47, D590–D595. doi: 10.1093/nar/gky962

Oliveros, J. C. (2015). *Venny. An interactive tool for comparing lists with Venn diagrams*. http://bioinfogp.cnb.csic.es/tools/venny/index.html

Pirooznia, M., Nagarajan, V., and Deng, Y. (2007). GeneVenn—a web application for comparing gene lists using Venn diagrams. *Bioinformation* 1, 420–422. doi: 10.6026/97320630001420

Sharov, A. A., Schlessinger, D., and Ko, M. S. (2015). ExAtlas: an interactive online tool for meta-analysis of gene expression data. *J. Bioinform. Comput. Biol.* 13:1550019. doi: 10.1142/S0219720015500195

Sun, L., Zhu, Y., Mahmood, A., Tudor, C. O., Ren, J., Vijay-Shanker, K., et al. (2017). WebGIVI: a web-based gene enrichment analysis and visualization tool. *BMC Bioinformatics* 18:237. doi: 10.1186/s12859-017-1664-2

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/nar/gkx382

Uniprot, C. (2008). The universal protein resource (UniProt). *Nucleic Acids Res.* 36, D190–D195. doi: 10.1093/nar/gkm895

Xia, J., Lyle, N. H., Mayer, M. L., Pena, O. M., and Hancock, R. E. (2013). INVEX—a web-based tool for integrative visualization of expression data. *Bioinformatics* 29, 3232–3234. doi: 10.1093/bioinformatics/btt562

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.