DIANE

Dashboard for the **I**nference and **A**nalysis of **N**etworks from **E**xpression data

Océane Cassan - oceane.cassan@supagro.fr
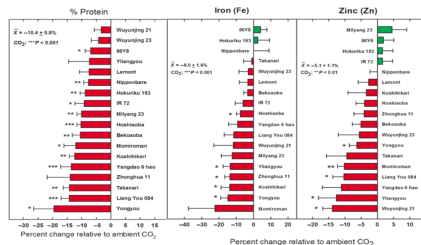Supervised by Antoine Martin and Sophie Lèbre

NETBIO

December 8, 2020

Thesis project
Statistical inference of the gene regulatory network in Arabidopsis thaliana under climate change

- Anthropic CO2 emissions are expected to double until 2100
- Elevated CO2 **depletes plants mineral status**
- Key **regulators** in this response are yet to be identified

Zhu 2018 Science Advances

Gene expression to understand regulatory mechanisms

- Gene expression reprogramming is studied to understand **development**, **adaptation** to environmental constraints in living organisms

- The rise of NGS techniques and **RNA-Seq** made available genome-wide transcripts quantification to researchers
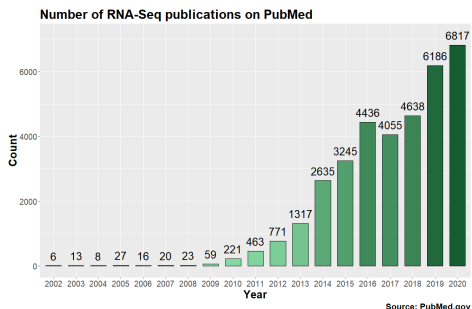
Analyzing RNA-Seq data

To characterize new signalling pathways

- Which genes have their **expression significantly changed** by a perturbation?

- Can I distinguish **groups** of genes with similar behaviors?

- How to model **interactions and dependencies between genes**?

Analyzing RNA-Seq data

To characterize new signalling pathways

- Which genes have their **expression significantly changed** by a perturbation?

- Can I distinguish **groups** of genes with similar behaviors?

- How to model **interactions and dependencies between genes**?

Standard analysis pipelines require **programming** and **methodological** skills:

Bioinformatic pipeline

Quality control, fragments mapping, quantification

Analyzing RNA-Seq data

To characterize new signalling pathways

- Which genes have their **expression significantly changed** by a perturbation?

- Can I distinguish **groups** of genes with similar behaviors?

- How to model **interactions and dependencies between genes**?

Standard analysis pipelines require **programming** and **methodological** skills:

Bioinformatic pipeline

Quality control, fragments mapping, quantification

Statistical pipeline

Normalization, differential gene expression, ontology enrichment, clustering, co-expression or regulatory pathways reconstruction

Current user interfaces for facilitating RNA-seq analysis

## Available satisfactory possibilities

- Normalization

- Exploration and visualization

- Differential Expression

- Gene Ontology Enrichment

See DEBrowser[Kucukural et al., 2019],
DEApp[Li and Andrade, 2017],
iGEAk[Choi and Ratner, 2019],
DEIVA[Harshbarger et al., 2017],
Shiny-Seq[Sundararajan et al., 2019],
IRIS-DEA[Monier et al., 2019], iDEP[Ge et al., 2018],
TCC-GUI[Su et al., 2019], ShinyBN [Chen et al., 2019],
or GeNeck [Zhang et al., 2019]

Current user interfaces for facilitating RNA-seq analysis

Available satisfactory possibilities

- Normalization
- Exploration and visualization
- Differential Expression
- Gene Ontology Enrichment

See DEBrowser[Kucukural et al., 2019],
DEApp[Li and Andrade, 2017],
iGEAk[Choi and Ratner, 2019],
DEIVA[Harshbarger et al., 2017],
Shiny-Seq[Sundararajan et al., 2019],
IRIS-DEA[Monier et al., 2019], iDEP[Ge et al., 2018],
TCC-GUI[Su et al., 2019], ShinyBN [Chen et al., 2019],
or GeNeck [Zhang et al., 2019]

Absent or sub-obtimal features

- Model based clustering
- Statistical models or Machine Learning for network inference
- Oriented toward regulation rather than co-expression
- Advanced exploration of gene clusters or inferred networks

Workflow
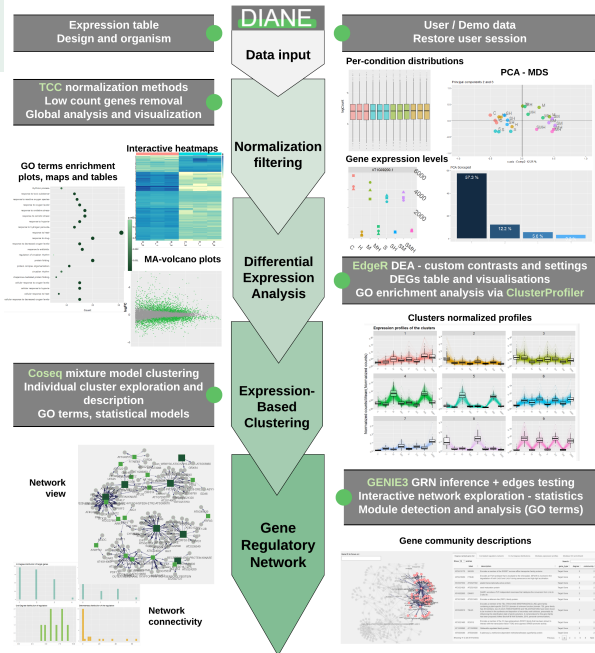
Multifactorial dataset of nutritional starvation and elevated CO2

Complete $3 * 2$ cross design :

- **CO2** (**400ppm** - 900ppm)
- **Nitrate** (low - **high**)
- **Iron** (Starvation - **sufficiency**)

8 conditions with triplicates : 24 root transcriptomes

Arabidopsis thaliana

Normalization and low counts removal

Normalization

- EdgeR TMM strategy [Robinson and Oshlack, 2010]
- DESeq2 Median of ratios [Love et al., 2014]
- Prior removal of DEGs via TCC [Sun et al., 2013]

Filtering

Adjustable threshold for counts sum across samples

Exploratory transcriptome analysis

- **Genes normalized expression levels**
- PCA, Mutli-Dimensional Scaling plot

Exploratory transcriptome analysis

- Gene normalized expression levels
- **PCA, Mutli-Dimensional Scaling plot**

Differential Expression Analysis

Tests performed by EdgeR (Negative Binomial models) [McCarthy et al., 2012]

Differential Expression Analysis

Over-representation tests are performed by ClusterProfiler to detect significantly enriched GO terms [Yu et al., 2012]

Differential Expression Analysis

Venn diagrams to compare lists of differentially expressed genes

Co-expression clustering

Poisson or Gaussian Mixture Models allow gene profiles clustering via Coseq
[Rau et al., 2015, Rau and Maugis-Rabusseau, 2018]

Gene behaviors under CO2 and nitrate perturbations

Elevated CO2 cancels the regulation of nitrate related genes
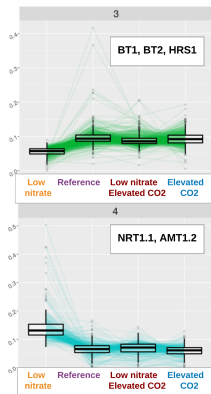
Elevated CO2 reverses the effect of low nitrate

BT1, BT2, HRS1

HHO2, HHO3

NRT1.1, AMT1.2
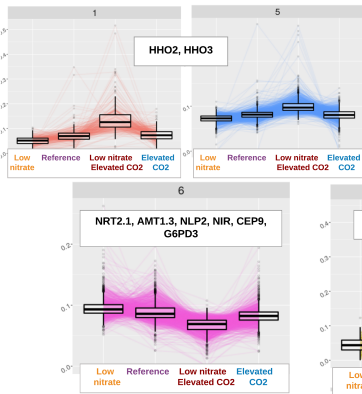
Some genes stay downregulated under low nitrate and elevated CO2

NRT2.1, AMT1.3, NLP2, NIR, CEP9, G6PD3

Ions transport (metal, zinc, iron)
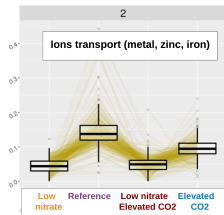
How does elevated CO2 alters the response to nitrate starvation?

Either lessens, cancels or inverses the regulation observed in low nitrate only

Gene Regulatory Network Inference

- How to capture transcriptionnal dependencies and causality between genes ?
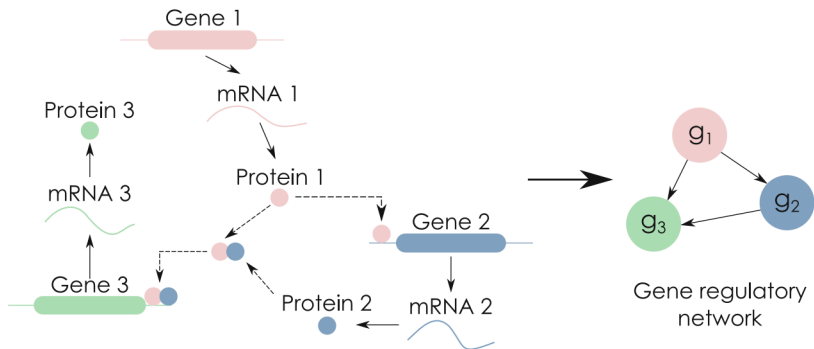
[Sanguinetti and Huynh-Thu, 2019]

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

- **Non-linear** approach, captures regulators **cooperation** for gene regulation

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

- **Non-linear** approach, captures regulators **cooperation** for gene regulation
- **Non-parametric** procedure, with very few parameters and no modelling assumptions

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

- **Non-linear** approach, captures regulators **cooperation** for gene regulation
- **Non-parametric** procedure, with very few parameters and no modelling assumptions
- Easily **scalable** to large datasets

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

- **Non-linear** approach, captures regulators **cooperation** for gene regulation
- **Non-parametric** procedure, with very few parameters and no modelling assumptions
- Easily **scalable** to large datasets
- Best performer on **DREAM4** and **DREAM5** challenges
  [Greenfield et al., 2010, Marbach et al., 2016]

Wisdom of crowds for robust gene network inference, 2012

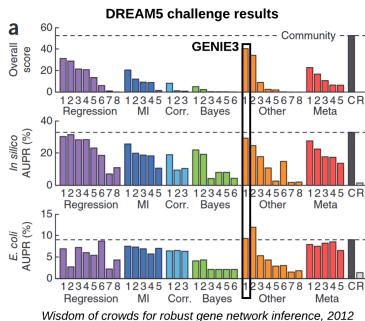# Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

- **Non-linear** approach, captures regulators **cooperation** for gene regulation
- **Non-parametric** procedure, with very few parameters and no modelling assumptions
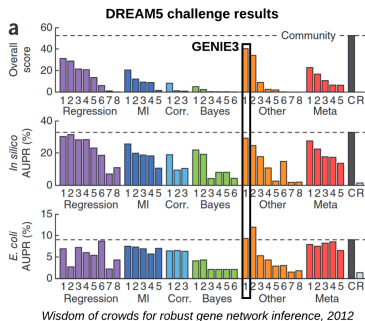- Easily **scalable** to large datasets
- Best performer on **DREAM4** and **DREAM5** challenges
  [Greenfield et al., 2010, Marbach et al., 2016]

Wisdom of crowds for robust gene network inference, 2012

- **Oriented edges** from regulators to other genes

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

Ranking the regulators according to their **relevance for predicting** the other genes expression

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

Ranking the regulators according to their **relevance for predicting** the other genes expression

Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

Ranking the regulators according to their **relevance for predicting** the other genes expression
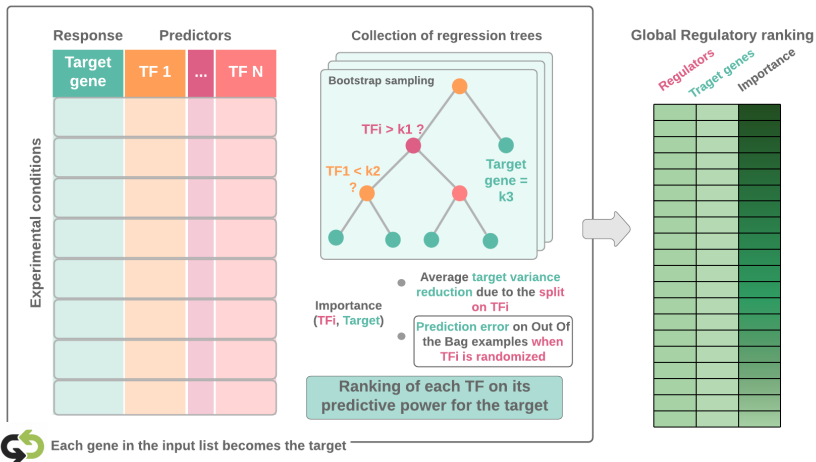
Random Forests inference with GENIE3 [Huynh-Thu et al., 2010]

Ranking the regulators according to their **relevance for predicting** the other genes expression

Each gene in the input list becomes the target

Testing regulator-gene pairs

Selecting meaningful importances from Random Forests with rfPermute [Archer, 2020]

DIANE's interface for network inference

Network exploration

- **Gene connectivity ranking**
- Individual nodes information, targets and regulators
- In and out degree distributions

Network exploration

- Gene connectivity ranking
- **Individual nodes information, targets and regulators**
- In and out degree distributions

Network exploration

- Gene connectivity ranking
- Individual nodes information, targets and regulators
- **In and out degree distributions**

Communities discovery

- Louvain community detection [Blondel et al., 2008]
- Module expression profiles
- Module GO enrichment analysis
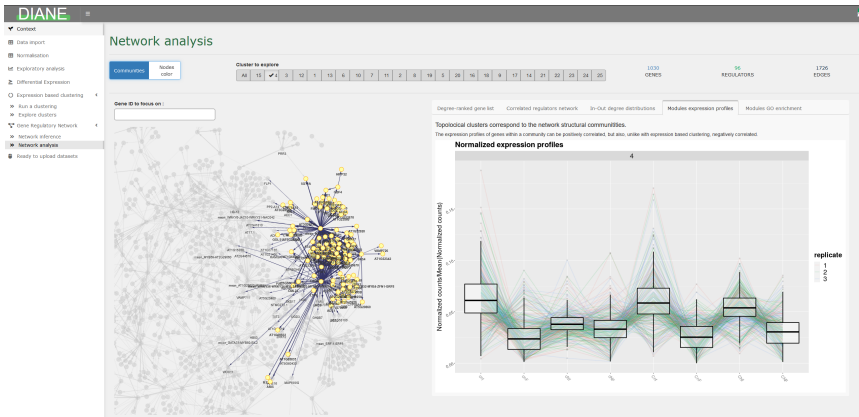
GRN under to elevated CO2 and low nitrate

1550 **differentially expressed genes by elevated CO2 under low nitrate**, 12 experimental samples
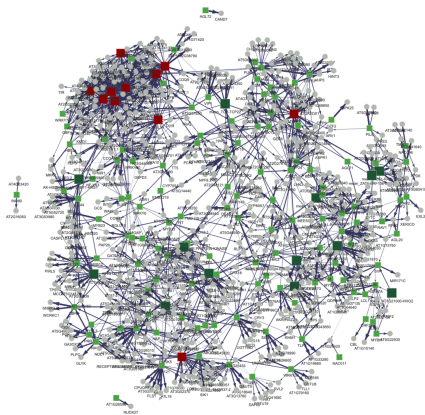
GRN under to elevated CO2 and low nitrate

1550 **differentially expressed genes by elevated CO2 under low nitrate**, 12 experimental samples
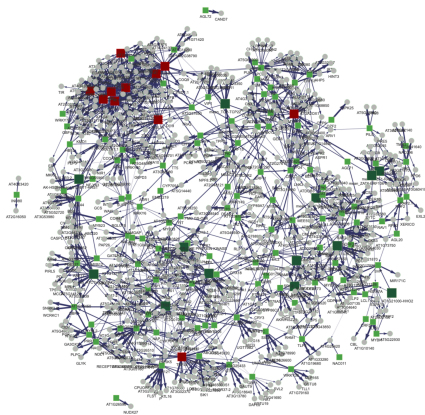
Validation of known genes interactions

- Confirmed associations between nitrate assimilation genes (BT1-BT2-HRS1, NRT2.1-NAR2.1...)

GRN under to elevated CO2 and low nitrate

1550 **differentially expressed genes by elevated CO2 under low nitrate**, 12 experimental samples
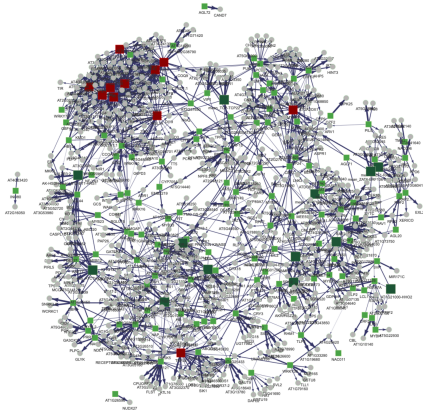
Validation of known genes interactions

- Confirmed associations between nitrate assimilation genes (BT1-BT2-HRS1, NRT2.1-NAR2.1...)

Candidate genes discovery

- The most connected genes are potential **key regulators** of the response
- Mutant plants are programmed for **root system** and **shoots mineral status** phenotyping

GRN under to elevated CO2 and low nitrate

1550 **differentially expressed genes by elevated CO2 under low nitrate**, 12 experimental samples
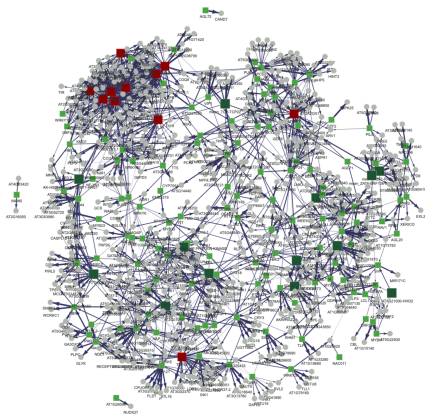
Validation of known genes interactions

- Confirmed associations between nitrate assimilation genes (BT1-BT2-HRS1, NRT2.1-NAR2.1...)

Candidate genes discovery

- The most connected genes are potential **key regulators** of the response
- Mutant plants are programmed for **root system** and **shoots mineral status** phenotyping

Promising leads to explain the depletion of plants mineral status under elevated CO2

Conclusion

Highlights

- **Accessible** online or local user **interface**
- **Interactive**, efficient and **reproducible** analyses
- Strong methodological frameworks with parameterization help for clustering and network inference
- Annotations for **model organisms**, or **custom** files upload for other organism

Conclusion

Highlights

- **Accessible** online or local user **interface**
- **Interactive**, efficient and **reproducible** analyses
- Strong methodological frameworks with parameterization help for clustering and network inference
- Annotations for **model organisms**, or **custom** files upload for other organism

Perspectives

- Adapt inference methodology to **time series** RNA-Seq
- Integrate external databases or other -omics data (Chip-Seq, ATAC-Seq...)

Conclusion

 Project links

Online version : https://diane.bpmp.inrae.fr
GitHub : https://github.com/OceaneCsn/DIANE
Documentation : https://oceanecsn.github.io/DIANE/

Conclusion

 Project links

Online version : https://diane.bpmp.inrae.fr
GitHub : https://github.com/OceaneCsn/DIANE
Documentation : https://oceanecsn.github.io/DIANE/

Merci!

References I

▶ Archer, E. (2020).
rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics.
R package version 2.1.81.

▶ Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008).
Fast unfolding of communities in large networks.
Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.

▶ Chen, J., Zhang, R., Dong, X., Lin, L., Zhu, Y., He, J., Christiani, D. C., Wei, Y., and Chen, F. (2019).
shinybn: an online application for interactive bayesian network inference and visualization.
BMC bioinformatics, 20(1):711.

▶ Choi, K. and Ratner, N. (2019).
IGEAK: An interactive gene expression analysis kit for seamless workflow using the R/shiny platform.
BMC Genomics, 20(1):177.

▶ Ge, S. X., Son, E. W., and Yao, R. (2018).
iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data.
BMC Bioinformatics, 19(1):1–24.

References II

▶ Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. (2010).
DREAM4: Combining genetic and dynamic information to identify biological networks and Dynamical Models.
PLoS ONE, 5(10).

▶ Harshbarger, J., Kratz, A., and Carninci, P. (2017).
DEIVA: A web application for interactive visual analysis of differential gene expression profiles.
BMC Genomics, 18(1):47.

▶ Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010).
Inferring Regulatory Networks from Expression Data Using Tree-Based Methods.
PLoS ONE, 5(9):e12776.

▶ Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., and Garber, M. (2019).
DEBrowser: Interactive differential expression analysis and visualization tool for count data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems.
BMC Genomics, 20(1):6.

References III

► Li, Y. and Andrade, J. (2017).
DEApp: An interactive web interface for differential expression analysis of next generation sequence data.
Source Code for Biology and Medicine, 12(1):10–13.

► Love, M. I., Huber, W., and Anders, S. (2014).
Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.
Genome Biology, 15(12):550.

► Marbach, D., Costello, J. C., Küffner, R., Vega, N., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., Stolovitzky, G., and Performed, G. S. M. (2016).
Wisdom of crowds for robust gene network inference the DREAM5 Consortium HHS Public Access.
Nat Methods, 9(8):796–804.

► McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012).
Differential expression analysis of multifactor rna-seq experiments with respect to biological variation.
Nucleic Acids Research, 40(10):4288–4297.

References IV

▶ Monier, B., McDermaid, A., Wang, C., Zhao, J., Miller, A., Fennell, A., and Ma, Q. (2019).
IRIS-EDA: An integrated RNA-seq interpretation system for gene expression data analysis.
PLoS Computational Biology, 15(2).

▶ Rau, A. and Maugis-Rabusseau, C. (2018).
Transformation and model choice for RNA-seq co-expression analysis.
Briefings in Bioinformatics, 19(3):425–436.

▶ Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015).
Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models.
Bioinformatics, 31(9):1420–1427.

▶ Robinson, M. D. and Oshlack, A. (2010).
A scaling normalization method for differential expression analysis of RNA-seq data.
Genome Biology, 11(3):R25.

▶ Sanguinetti, G. and Huynh-Thu, V. A. (2019).
Gene regulatory networks.

References V

► Su, W., Sun, J., Shimizu, K., and Kadota, K. (2019).
TCC-GUI: A Shiny-based application for differential expression analysis of RNA-Seq count data.
BMC Research Notes, 12(1):133.

► Sun, J., Nishiyama, T., Shimizu, K., and Kadota, K. (2013).
TCC: An R package for comparing tag count data with robust normalization strategies.
BMC Bioinformatics, 14(1):219.

► Sundararajan, Z., Knoll, R., Hombach, P., Becker, M., Schultze, J. L., and Ulas, T. (2019).
Shiny-Seq: advanced guided transcriptome analysis.
BMC research notes, 12(1):432.

► Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012).
clusterprofiler: an r package for comparing biological themes among gene clusters.
Omics: a journal of integrative biology, 16(5):284–287.

► Zhang, M., Li, Q., Yu, D., Yao, B., Guo, W., Xie, Y., and Xiao, G. (2019).
Geneck: a web server for gene network construction and visualization.
BMC bioinformatics, 20(1):1–7.