JOBIM 2021 — 06>09 JUIL — INSTITUT PASTEUR | PARIS

Dashboard for the Inference and Analysis of Networks from Expression Data

Océane Cassan[1], Sophie Lèbre[2,3], Antoine Martin[1]

DIANE

DIANE : a graphical user interface to infer and explore regulatory pathways

Gene expression reprogramming is studied to understand development, adaptation to environmental constraints in living organisms. **The rise of NGS techniques and RNA-Seq made genome-wide transcripts quantification available to researchers.** A number of software tools exist to bring to **users with little programming experience** the standard analysis pipelines of these kind data :

- These existing tools allow satisfacory possibilities for normalisation, exploration, differential expression, Gene Ontologie enrichment.
- However, they either do not propose, or do not reflect state of the art methodological advances to perform **gene clustering, gene regulatory network inference, and explore their respective outputs.**

1. Expression data

BIOINFORMATIC PIPELINE

Quality control, mapping and quantification are steps that need to be performed **prior to any analysis in DIANE**

EXPRESSION MATRIX

Gene expression in each experimental conditions are required. **Optional other inputs** : deisgn, gene annotations, GO terms for non-model organisms

	Cond1_1	...	CondJ_3
Gene1			
GeneN			

DEMONSTRATION DATA : ARABIDOPSIS UNDER COMBINED ABIOTIC STRESSES

In the context of climate change, Arabidopsis plants were studied under the multifactorial design of salinity stress, heat stress, and osmotic stress[1].

Data upload

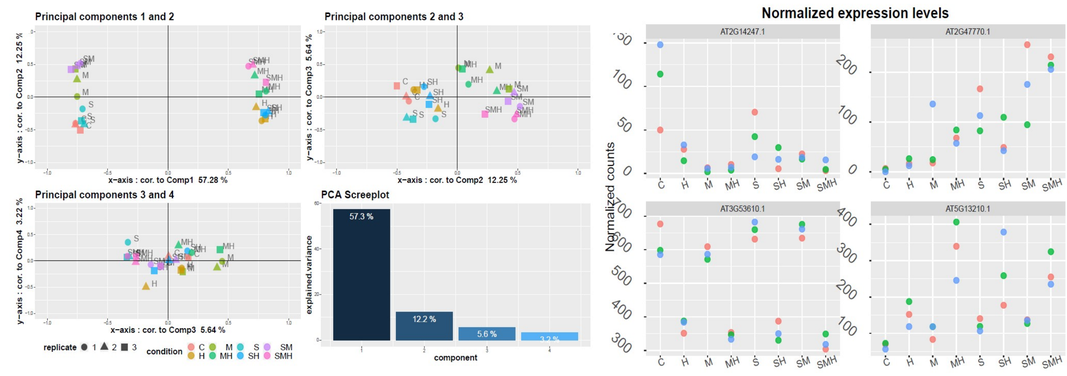
Normalisation Exploration

2. Explore normalized transcriptomes

Normalisation can be performed with 3 standard procedures : TMM, DESeq2 or TCC method.

EXPLORATION OF NORMALIZED EXPRESSION DATA

Dimensionality reduction allows to assess replicates homogeneity, and quantify the impact of experimental perturbations on gene expression variation. Expression levels of genes of interest can be browsed.



Here, **heat stress has a predominant effect** on plant transcriptomes, as the first principal component, linked to heat stress, explains 57 % of gene expression variance.

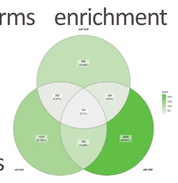
3. Differential expression

DIFFERENTIAL EXPRESSION ANALYSIS

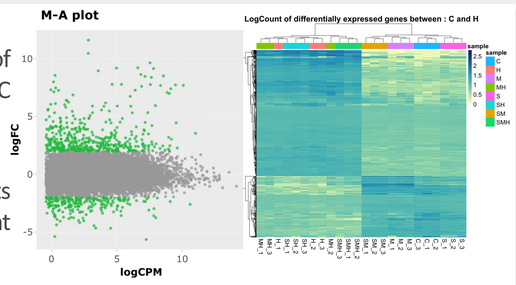
EdgeR is used to compare pairs of transcriptomes, with custom FDR and LFC thresholds.

INTERPRETING LISTS OF GENES

Standard visualisations, gene lists intersections, and GO terms enrichment analyses are proposed.



Venn diagram of 3 lists of differentially expressed genes



MA plots and heatmap of genes differentially expressed under a simple heat stress

Differential expression

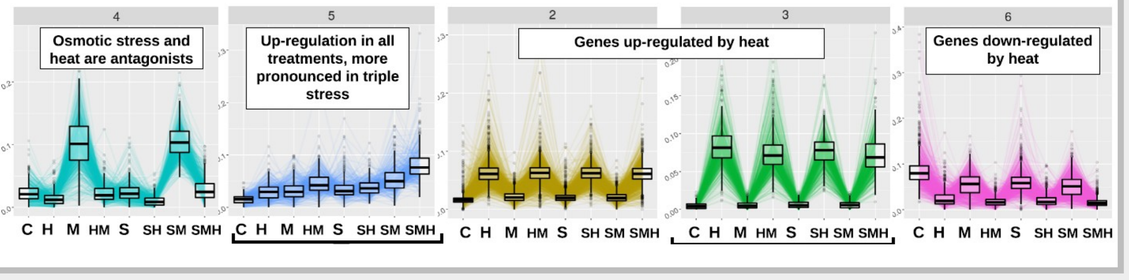
Expression-based clustering

4. Genes clustering

MIXTURE MODELS TO GROUP SIMILAR GENE EXPRESSION PROFILES

The **Coseq** package[2] enables gene clustering using **Mixture Models**, adjusted through an Expectation-Maximisation algorithm. It is possible fit either **Poisson** multivariate distributions to the clusters, or **Gaussian** mixtures after prior transformations on expression values. A penalized model selection criteria (ICL) is then used to determine the **optimal number of clusters**. We applied this method to the genes differentially expressed in at least one of the perturbation in the experiment against the control :

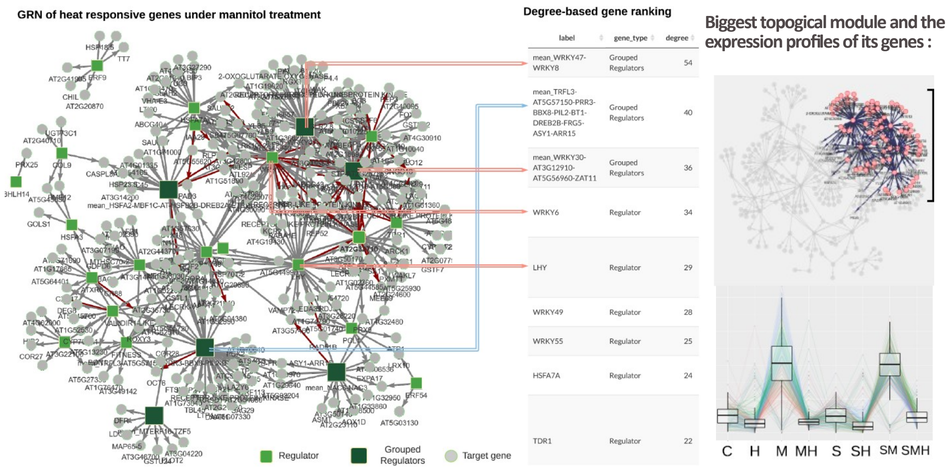
Mean-normalized expression profiles for 5 representative gene clusters



5. Network inference

REGULATORY NETWORK OF THE RESPONSE TO HEAT UNDER OSMOTIC PERTUBATION

The expression profiles of the **genes differentially expressed by heat under osmotic stress** were used to reconstruct the regulatory network involved in this kind of extreme climatic event



Degree-based gene ranking

label	gene_type	degree
mean_WRKY47-WRKY8	Grouped Regulators	54
mean_TRFL3-AT5G57150-PRR3-BBX8-PIL2-BT1-DREB2B-FRG5-ASY1-ARR15	Grouped Regulators	40
mean_WRKY30-AT3G12910-AT5G56960-ZAT11	Grouped Regulators	36
WRKY6	Regulator	34
LHY	Regulator	29
WRKY49	Regulator	28
WRKY55	Regulator	25
HSFA7A	Regulator	24
TDR1	Regulator	22

Biggest topogical module and the expression profiles of its genes :

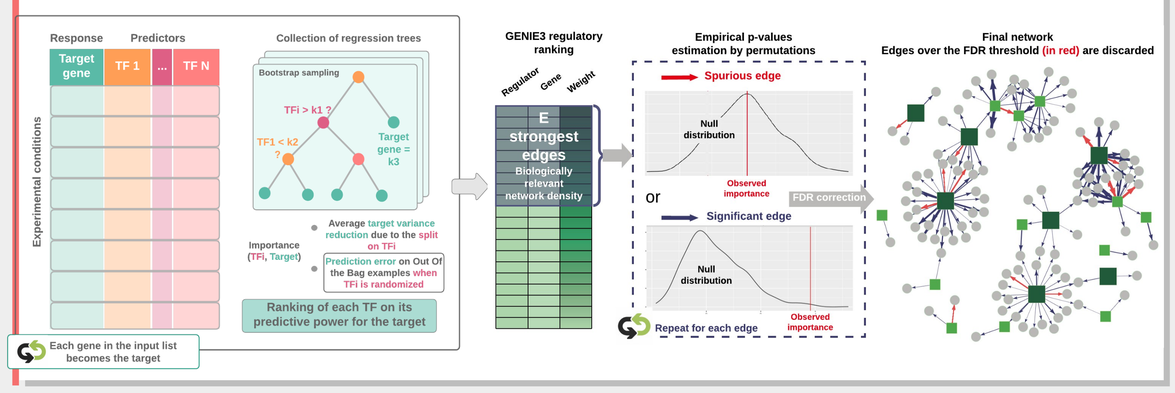
NETWORK INTERPRETATION

Some of the most connected TFs of the network are known to be involved in **drought and heat** responses, the other ones are **promising candidates for crops improvement**

Gene Regulatory Network Inference

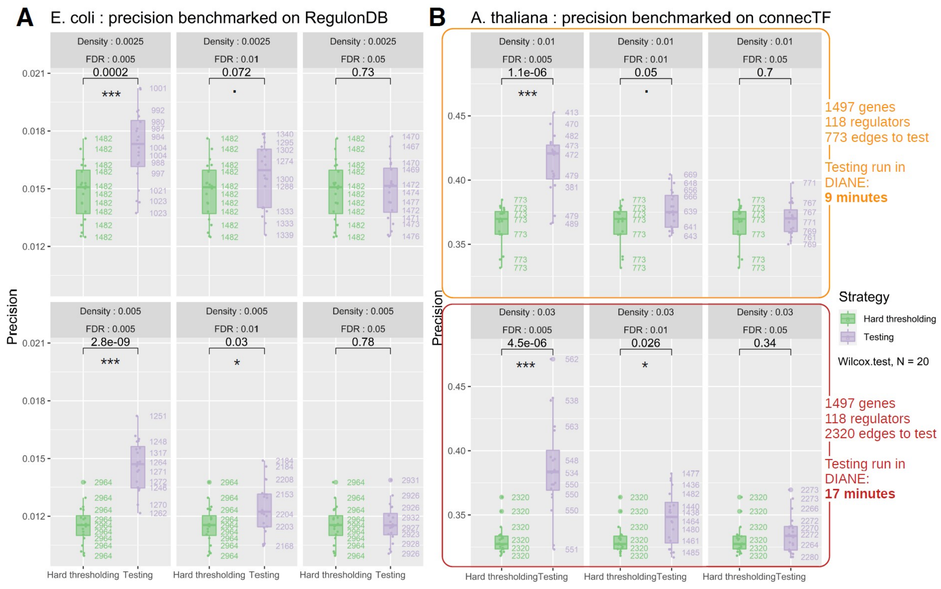
Network inference method

The chosen approach, **GENIE3** [3], uses **Random Forests to infer oriented edges from regulators to targets**, wieghted by the **predictive power** of each regulator on the target gene expression.
To select statistically significant pairs in the final network, we designed a procedure of empirical tests on regulatory edges using **permutations on Random Forests importance metrics via the rfPermute package.**



Benchmarking the added value of emprical tests on edges importance metrics

We applied **our novel edges selection strategy from GENIE3 edges ranking** on two different datasets, for which robust regulator-gene validation information is available. We validated the network of **heat-responsive genes** in our companion data using the **ConnecTF**[4] database. For *E. coli*, conditions from a public compendium of expression data (Faith JJ, 2007) was used to infer a gene regulatory network that we validated via **RegulonDB** (Santos-Zavaleta A, 2019).



For both organisms, a significant increase of precision can be achieved on both datasets when choosing **stringent adjusted p-values for edges removal,** independently of prior density.
This finding supports that **p-values obtained from permutations on Random Forest importance metrics can allow more confidence in the inferred edges than hard thresholding GENIE3's fully connected network.**

Reproducibility, links and references

REPRODUCIBLE RESEARCH

- **Dynamically generated html reports all along the analysis pipeline via Rmarkdown**
- **Seed setting for reproducible results**

Online version of DIANE :
https://diane.bpmp.inrae.fr/

Documentation and get started :
https://oceanecsn.github.io/DIANE/

Github project :
https://github.com/OceaneCsn/DIANE

Article with a detailed description of DIANE :
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07659-2

Use DIANE locally

DIANE relies on R >= 4.0.1, available for all OS at https://cloud.r-project.org/.

Download and install DIANE in your R console as follows (you need the remotes package installed `install.packages("remotes")`):

```
remotes::install_github("OceaneCsn/DIANE")
```

You can then launch the application :

```
library(DIANE)
DIANE::run_app()
```

MAIN REFERENCES

[1] Sewelam N, et AL Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes. J Exp Bot. 2020.

[2] Rau A, Maugis-Rabusseau C. Transformation and model choice for RNA-seq co-expression analysis. Brief Bioinforma. 2018; 19(3):425–36. https://doi.org/10.1093/bib/bbw128

[3] Huynh-Thu VA, et AL Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE. 2010; 5(9):12776. https://doi.org/10.1371/journal.pone.0012776.

[4] Brooks MD et AL ConnecTF: A platform to integrate transcription factor-gene interactions and validate regulatory networks. Plant Physiol. 2020; 185(1):49–66. https://doi.org/10.1093/plphys/kiaa012

[5] Cassan, O., Lebre, S., and Martin, A. (2021). Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. BMC Genomics, 22(1)

AUTHORS AFFILIATIONS

[1] BPMP, CNRS, INRAE, Institut Agro, Univ Montpellier, Montpellier, France
[2] IMAG, Univ Montpellier, CNRS, France
[3] Univ Paul Valéry Montpellier 3, France

CONTACT

oceane.cassan@cnrs.fr

PARTNER INSTITUTIONS