

# Gene-specific optimization of data integration in regression-based Gene Regulatory Network inference

SIMBA seminar

**Océane Cassan** : [oceane.cassan@lirmm.fr](mailto:oceane.cassan@lirmm.fr)

Charles-Henri Lecellier, Antoine Martin, Laurent Bréhélin, Sophie Lèbre



June 2024

# Regression-based GRN inference paradigm

## Modelling assumption

The expression of regulators hold predictive and descriptive power over the expression of their target genes

Ex : GENIE3 [Huynh-Thu et al., 2010], The Inferelator [Gibbs et al., 2022]



# Regression-based GRN inference paradigm

## Modelling assumption

The expression of regulators hold predictive and descriptive power over the expression of their target genes

Ex : GENIE3 [Huynh-Thu et al., 2010], The Inferelator [Gibbs et al., 2022]



## Limitations

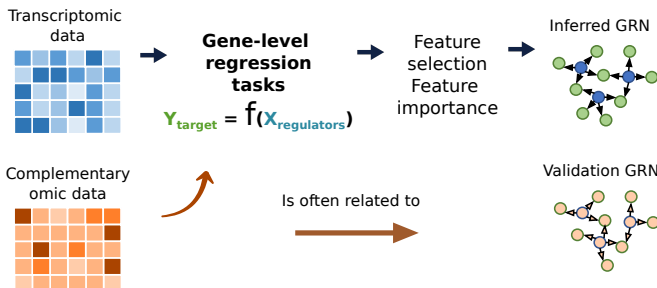
- High dimension
- High correlation among predictors
- Incomplete view of the regulation process

# Data integration in regression-based GRN inference

## Modelling assumption

Complementary omics can bring more causality to GRN inference

Ex : iRafNet [Petrálie et al., 2015], MEN [Greenfield et al., 2013], LASSO-Stars [Miraldi et al., 2019]

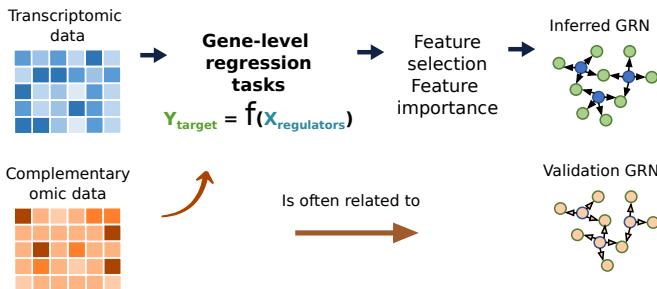


# Data integration in regression-based GRN inference

## Modelling assumption

Complementary omics can bring more causality to GRN inference

Ex : iRafNet [Petrálie et al., 2015], MEN [Greenfield et al., 2013], LASSO-Stars [Miraldi et al., 2019]



## Current limitations

Omics contributions are **rarely finely tuned**.  
They usually **rely on a close gold standard**.

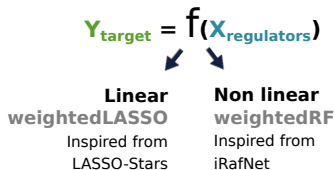
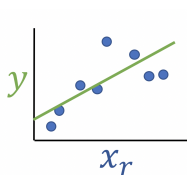
# Objectives

- 1 Can we define a new criterion to **robustly estimate the optimal strength of data integration** based on available data?
- 2 What is the benefit of optimising data integration at the **gene level**?

# Objectives

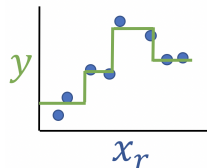
- 1 Can we define a new criterion to **robustly estimate the optimal strength of data integration** based on available data?
- 2 What is the benefit of optimising data integration at the **gene level**?

↳ Investigated for **two common forms of integrative regression**



[Miraldi et al., 2019]

[Petràlia et al., 2015]



# Modelling the root response to nitrate induction in *Arabidopsis thaliana*

RNASeq data :  $Y$ ,  $X$

$N$  conditions



## Temporal response to nitrate induction

1426 genes, 201  
regulators,  $N = 45$   
samples [Varala et al., 2018]

# Modelling the root response to nitrate induction in *Arabidopsis thaliana*

RNASeq data :  $Y, X$

N conditions

Target gene



Regulators



**Temporal response to nitrate induction**

1426 genes, 201

regulators,  $N = 45$

samples [Varala et al., 2018]

TFBM prior matrix :  $\Pi$

PWM occurrence score  
in the target's promoter

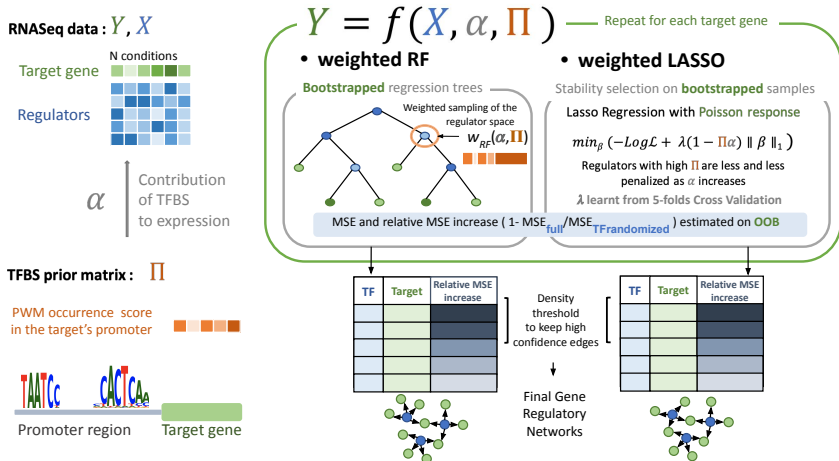


**TF binding sites** from **JASPAR** and the **Plant Cistrome Databases**

[Castro-Mondragon et al., 2021,  
O'Malley et al., 2016]

$$\Pi_{r,t} = \begin{cases} 0 & \text{if the motif of } r \text{ is not in the promoter of } t \\ 1 & \text{if the motif of } r \text{ is in the promoter of } t \\ \frac{1}{2} & \text{if the motif of } r \text{ is missing} \end{cases}$$

# Integrative regression-based GRN inference methods



iRafNet [Petrálie et al., 2015]

LASSO-Stars [Miraldi et al., 2019]

# Models for integrative GRN inference: weightedLASSO

A **generalized linear model** to express RNA-Seq data as counts for target  $t$  in condition  $i$ , based on the expression of  $R$  regulators  $x_{r,i}$ :

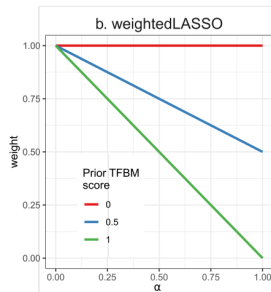
$$Y_{t,i} \sim \mathcal{P}(\mu_{t,i}) \quad \ln(\mu_{t,i}) = \beta_{t,0} + \sum_{r=1}^R \beta_{t,r} x_{r,i}$$

Estimated under the **LASSO** constraint with **differential shrinkage**:

$$\operatorname{argmin}_{\beta_t} \left\{ -\frac{1}{N} \log \mathcal{L}(\beta_t; X, Y_t) + \lambda \sum_{r=1}^R w_{t,r}^{LASSO} |\beta_{t,r}| \right\}$$

$$w_{t,r}^{LASSO} = 1 - \Pi_{t,r} \alpha$$

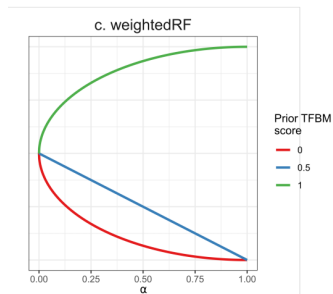
This model is run  $S$  times on bootstrapped samples.



# Models for integrative GRN inference: weightedRF

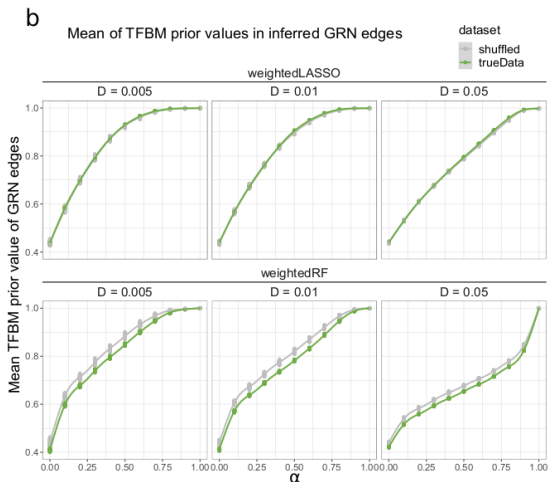
**Collection of regression trees with weighted subsampling of regulators at decision nodes:**

$$Y_{t,i} = \mathcal{RF}(X_i) \quad w_{r,t}^{RF} = \begin{cases} -\sqrt{1 - (\alpha - 1)^2} + 1 & \text{if } \Pi_{r,t} = 0 \\ 1 - \alpha & \text{if } \Pi_{r,t} = \frac{1}{2} \\ \sqrt{1 - (\alpha - 1)^2} + 1 & \text{if } \Pi_{r,t} = 1 \end{cases}$$



# Link functions for prior integration

The link functions  $w_{t,r}^{LASSO}$  and  $w_{t,r}^{RF}$  were calibrated so that TFBM support of inferred sparse GRNs grow smoothly until 1 as  $\alpha$  is increased.



# DIOgene: a gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate **TFBMs while controlling that it improves the prediction of gene expression.**

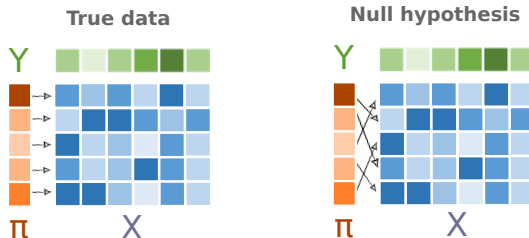
# DIOgene: a gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate **TFBMs** while controlling that it **improves the prediction of gene expression**.

## A simulated null hypothesis

↳ Breaks the link between expression profiles and PWM scores, a case where **data integration is uninformative**



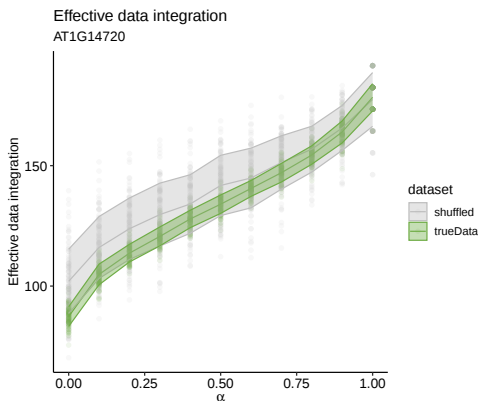
# DIOgene: a gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is still improved over  $H_0$ .**

## Effective data integration

Average rank of PWM-supported regulators based on their importance



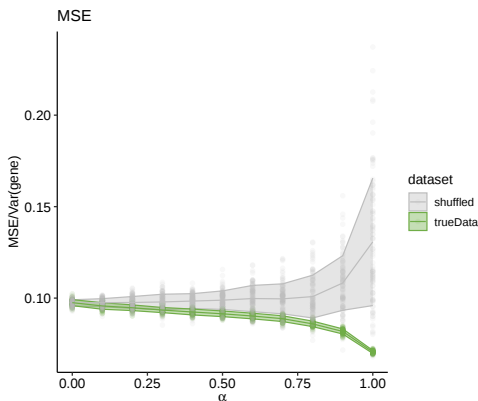
# DIOgene: a gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is still improved over  $H_0$ .**

## Prediction error (MSE)

Error committed by the regression model in predicting the target gene expression on test conditions



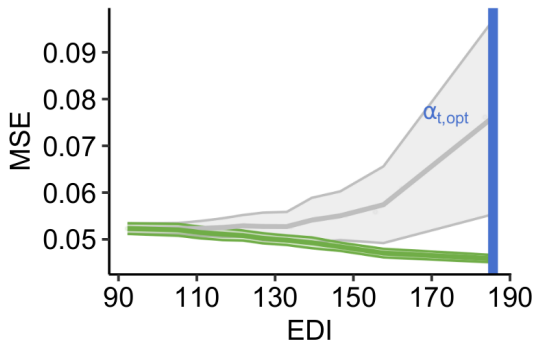
# DIOgene: a gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is still improved over  $H_0$ .**

### Ideal case

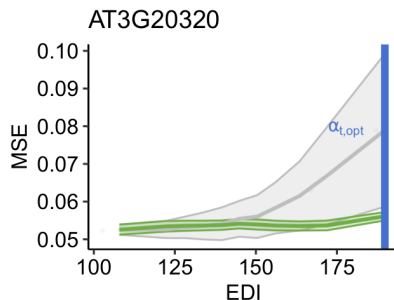
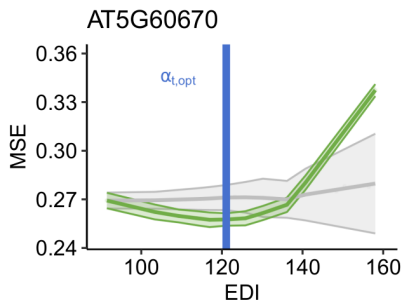
TFBM help selecting robust regulators and improve model generalisation performance



# DIOgene: a gene-wise criterion to optimise data integration

## Intermediate cases

The MSE reaches an optimum, or is improved over  $H_0$

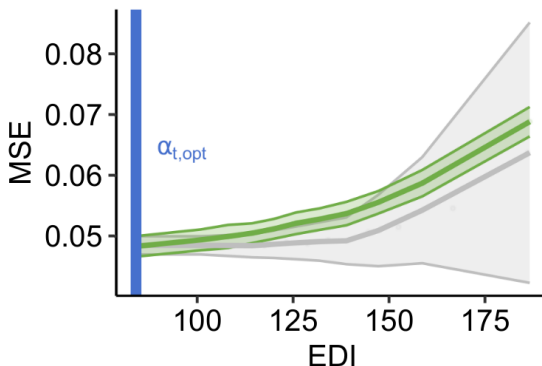


# DIOfgene: a gene-wise criterion to optimise data integration

## Cases where data integration should be avoided

The MSE does not differ from chance or is even higher

AT1G30270



# DIOgene: a gene-wise criterion to optimise data integration

## Formal criterion:

Maximization of the deviation (Student statistic) between the observed and null MSE:

$$T_{t\alpha} = \frac{\mu_{\text{MSE}}(EDI_{t\alpha}) - \mu_{\text{MSE}_0}(EDI_{t\alpha})}{\sqrt{\frac{\sigma_{\text{MSE}}(EDI_{t\alpha})^2 + \sigma_{\text{MSE}_0}(EDI_{t\alpha})^2}{N}}} \quad (1)$$

$T_{t\alpha}$  is then placed on a Student distribution to provide adjusted pvalues and choose  $\alpha$

$$\alpha_{t,\text{opt}} = \begin{cases} 0 & \text{if } \min_{\alpha \in [0,1]} (p_{t\alpha}) > 0.05 \\ \operatorname{argmin}_{\alpha \in [0,1]} (p_{t\alpha}) & \text{otherwise.} \end{cases} \quad (2)$$

# DIOgene offers both low MSE, and good precision/recall

## GRN

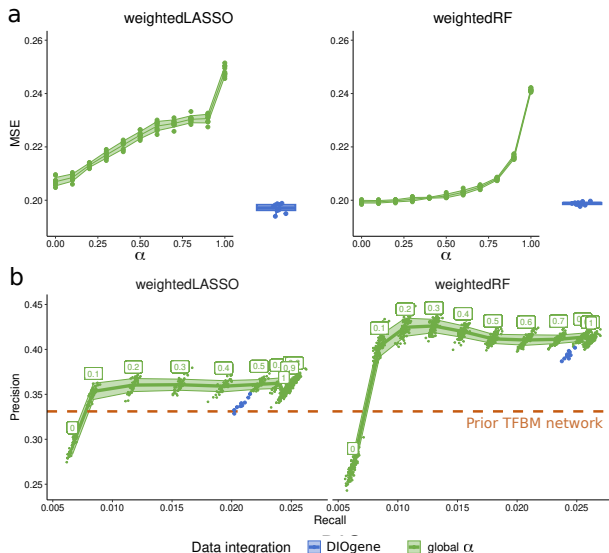
### reconstruction

↳ 0.005 density threshold on importance-ranked edges (1432 edges)

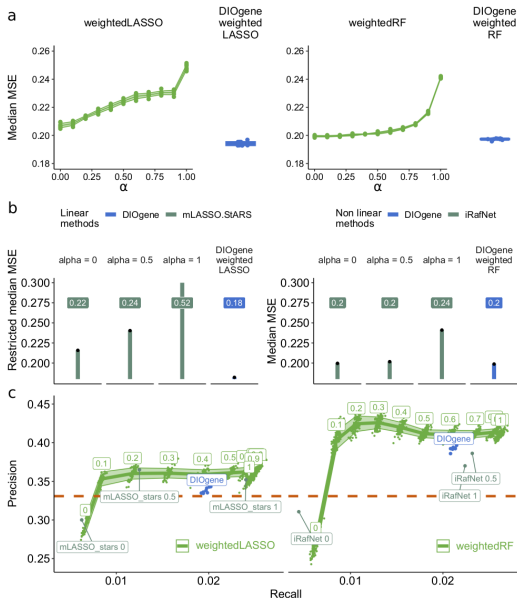
## GRN quality

### metrics

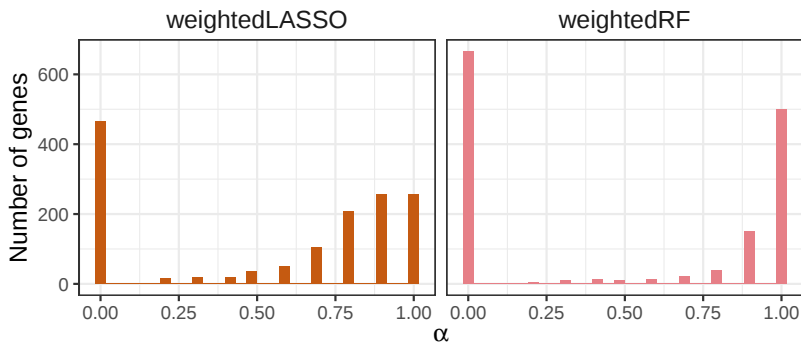
↳ Median MSE, and precision-recall curves against DAP-Seq in vitro TF binding [O'Malley et al., 2016]



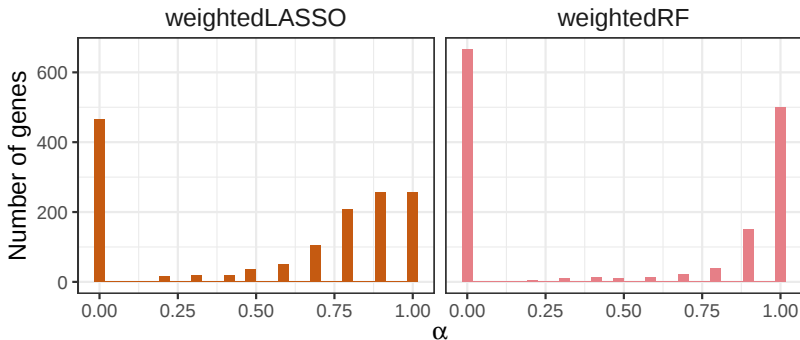
# DIOgene offers both low MSE, and good precision/recall



# Distributions of optimal integration strengths $\alpha_{\text{opt}}$



# Distributions of optimal integration strengths $\alpha_{\text{opt}}$



## TFBM integration is not warranted for all genes

- ↳ Technical or biological causes? Cooperative binding events?
- Post-transcriptional regulations (RNA stability)?

# Perspectives

## Limitations

- **Correlation is still a challenge.**
  - ↳ Developing robust importance metrics to improve feature selection
- **Missing TFBMs.**
  - ↳ Will be reduced as motif databases grow.

# Perspectives

## Limitations

- **Correlation is still a challenge.**
  - ↳ Developing robust importance metrics to improve feature selection
- **Missing TFBMs.**
  - ↳ Will be reduced as motif databases grow.

## Future research directions

- 1 Explore differences between the weightedLASSO and weightedRF and test the impact of **linearity assumptions**.
- 2 Extend it to **other organisms**, potentially with enhancers to scan for TFBMs, and **other types of omics and prior knowledge** to integrate.

# Key message

## Results of our gene-specific hypothesis-driven optimisation scheme

Indiscriminately pushing data integration to its maximal intensity is not always beneficial!

- Provides a desirable **trade-off between MSE and precision/recall**.
- Holds for both our linear and non-linear regression cases.
- Retrieves major players of nitrate nutrition in Arabidopsis.

# Key message

## Results of our gene-specific hypothesis-driven optimisation scheme

Indiscriminately pushing data integration to its maximal intensity is not always beneficial!

- Provides a desirable **trade-off between MSE and precision/recall**.
- Holds for both our linear and non-linear regression cases.
- Retrieves major players of nitrate nutrition in Arabidopsis.

**Code:** [https://github.com/OceaneCsn/integrative\\_GRN\\_N\\_induction](https://github.com/OceaneCsn/integrative_GRN_N_induction)

**Bioinformatics paper:** <https://doi.org/10.1093/bioinformatics/btae415>

# Acknowledgments

**Joint LIRMM - IMAG - IGMM Montpellier**

**Machine Learning for Regulatory Genomics team**

Sophie Lèbre, Laurent Bréhélin, Charles-Henri Lecellier, Mathilde Robin,  
Christophe Vroland



**IPSIM : SIRENE team**

Antoine Martin

**Funding : University of  
Montpellier and CNRS**

# References I

- ▶ Alvarez, J. M., Schinke, A.-L., Brooks, M. D., Pasquino, A., Leonelli, L., Varala, K., Safi, A., Krouk, G., Krapp, A., and Coruzzi, G. M. (2020).  
Transient genome-wide interactions of the master transcription factor NLP7 initiate a rapid nitrogen-response cascade.  
*Nature Communications*, 11(1).
- ▶ Bellegarde, F., Gojon, A., and Martin, A. (2017).  
Signals and players in the transcriptional regulation of root responses by local and systemic n signaling in arabidopsis thaliana.  
*Journal of Experimental Botany*, 68(10):2553–2565.
- ▶ Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Pérez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2021).  
JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles.  
*Nucleic Acids Research*, 50(D1):D165–D173.

# References II

- ▶ Gibbs, C. S., Jackson, C. A., Saldi, G.-A., Tjärnberg, A., Shah, A., Watters, A., Veaux, N. D., Tchourine, K., Yi, R., Hamamsy, T., Castro, D. M., Carriero, N., Gorissen, B. L., Gresham, D., Miraldi, E. R., and Bonneau, R. (2022).

High-performance single-cell gene regulatory network inference at scale: the inferelator 3.0.

*Bioinformatics*, 38(9):2519–2528.

- ▶ Greenfield, A., Hafemeister, C., and Bonneau, R. (2013).

Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks.

*Bioinformatics*, 29(8):1060–1067.

- ▶ 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):1–10.

# References III

- ▶ Marchive, C., Roudier, F., Castaings, L., Bréhaut, V., Blondet, E., Colot, V., Meyer, C., and Krapp, A. (2013).

Nuclear retention of the transcription factor nlp7 orchestrates the early response to nitrate in plants.

*Nature communications*, 4(1):1–9.

- ▶ Miraldi, E. R., Pokrovskii, M., Watters, A., Castro, D. M., De Veaux, N., Hall, J. A., Lee, J.-Y., Ciofani, M., Madar, A., Carriero, N., Littman, D. R., and Bonneau, R. (2019).

Leveraging chromatin accessibility for transcriptional regulatory network inference in T helper 17 cells.

*Genome Res.*, 29(3):449–463.

- ▶ O'Malley, R. C., shan Carol Huang, S., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., and Ecker, J. R. (2016).

Cistrome and epicistrome features shape the regulatory DNA landscape.

*Cell*, 165(5):1280–1292.

# References IV

- ▶ Petralia, F., Wang, P., Yang, J., and Tu, Z. (2015).  
Integrative random forest for gene regulatory network inference.  
*Bioinformatics*, 31(12):i197–i205.
- ▶ Ueda, Y., Kiba, T., and Yanagisawa, S. (2020).  
Nitrate-inducible NIGT1 proteins modulate phosphate uptake and starvation signalling via transcriptional regulation of iSPX/i genes.  
*The Plant Journal*, 102(3):448–466.
- ▶ Varala, K., Marshall-Colón, A., Cirrone, J., Brooks, M. D., Pasquino, A. V., Lérán, S., Mittal, S., Rock, T. M., Edwards, M. B., Kim, G. J., Ruffel, S., McCombie, W. R., Shasha, D., and Coruzzi, G. M. (2018).  
Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants.  
*Proceedings of the National Academy of Sciences*, 115(25):6494–6499.
- ▶ Vidal, E. A., Alvarez, J. M., Araus, V., Riveras, E., Brooks, M. D., Krouk, G., Ruffel, S., Lejay, L., Crawford, N. M., Coruzzi, G. M., and Gutiérrez, R. A. (2020).  
Nitrate in 2020: Thirty years from transport to signaling networks.  
*The Plant Cell*, 32(7):2094–2119.

# A gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate TFBS **only if they improve prediction when used jointly with expression data.**

# A gene-wise criterion to optimise data integration

## Modelling assumption

We want to integrate TFBS **only if they improve prediction when used jointly with expression data.**

## Effective data integration

Importance of PWM-supported regulators :

$$\frac{\sum_{\Pi_{r,t}=1} \text{Rank}(\text{Importance}_{r,t,\alpha})}{N_{\Pi_{r,t}=1}}$$

## Prediction error (MSE)

Error committed in predicting the target gene expression :

$$\frac{1}{N_{OOB}} \sum_{i \in OOB} (y_{t,i} - \hat{y}_{m,t,i,\alpha})^2$$

## A simulated null hypothesis

↳ Breaks the link between expression profiles and PWM scores, a case where **data integration is uninformative**

# Modelling of nitrate signalling can be improved

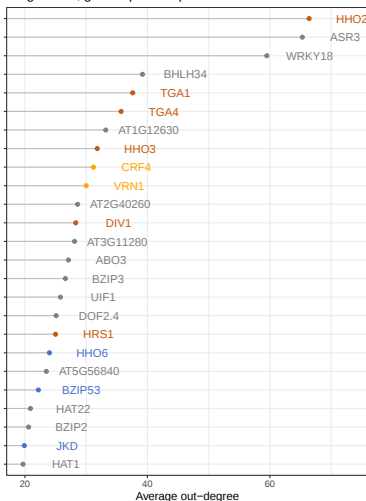
All models capture important nitrate actors

[Bellegarde et al., 2017, Vidal et al., 2020].

Gene specific optimisation of  $\alpha$  uniquely retrieves :

- **NPL7** [Marchise et al., 2013, Alvarez et al., 2020] and **PHL1** [Ueda et al., 2020] for **weightedLASSO**
- New candidate TFs of interest for **weightedRF**

weightedRF, gene-specific optimisation



- Candidate nitrate regulator
- Nitrate regulator
- Retrieved by gene
- specific data integration optimisation