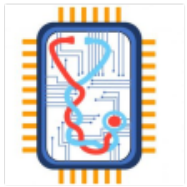
Gene-specific optimization of data integration improves regression-based Gene Regulatory Network inference in Arabidopsis

LEGO

Océane Cassan : oceane.cassan@lirmm.fr
Charles-Henri Lecellier, Antoine Martin, Laurent Bréhélin, Sophie Lèbre


November 23

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

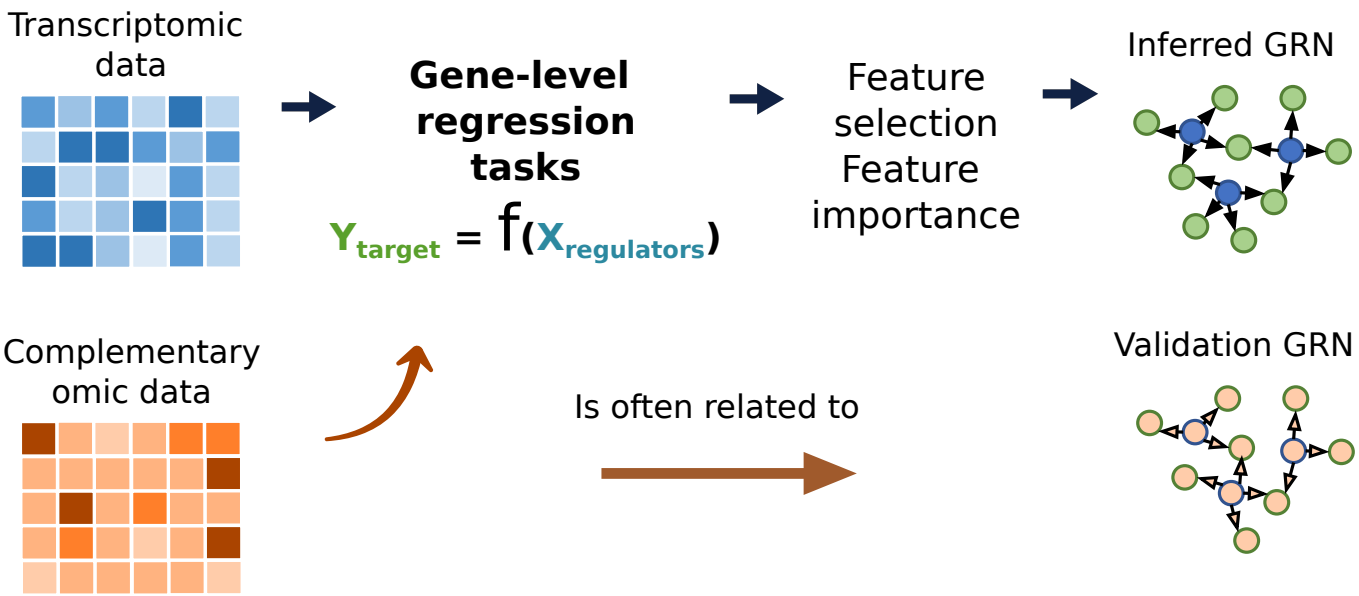
- High dimension
- High correlation among predictors
- Incomplete view of the regulation process

Data integration in regression-based GRN inference

Modelling assumption

Complementary omics can bring more causality to GRN inference

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

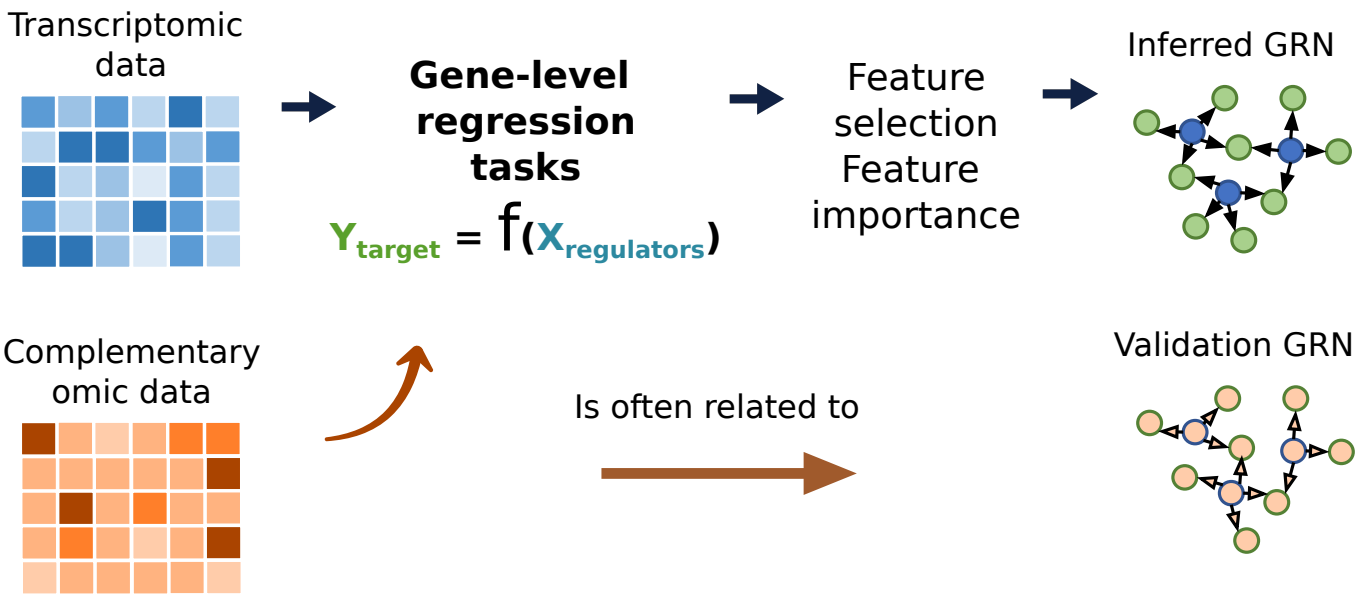


Data integration in regression-based GRN inference

Modelling assumption

Complementary omics can bring more causality to GRN inference

Ex : iRafNet [Petralia 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**.

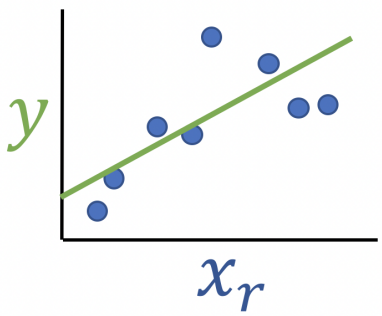
Objectives

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**?

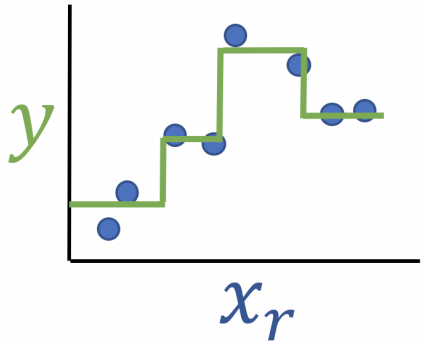
↳ Investigated for **two common forms of integrative regression**



$$Y_{target} = f(X_{regulators})$$

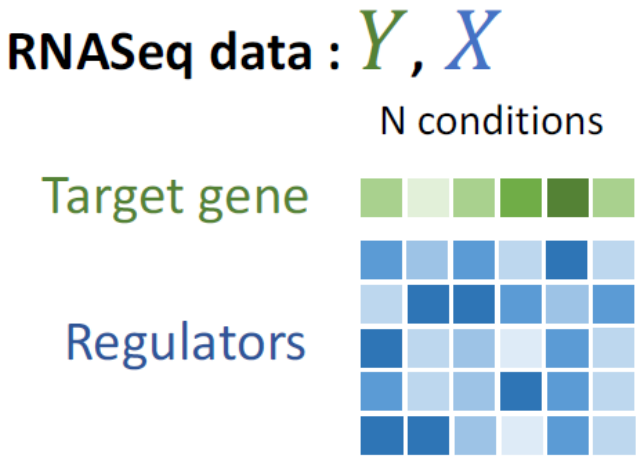
Linear
weightedLASSO
Inspired from
LASSO-Stars

Non linear
weightedRF
Inspired from
iRafNet



[Miraldi et al., 2019] [Petralia et al., 2015]

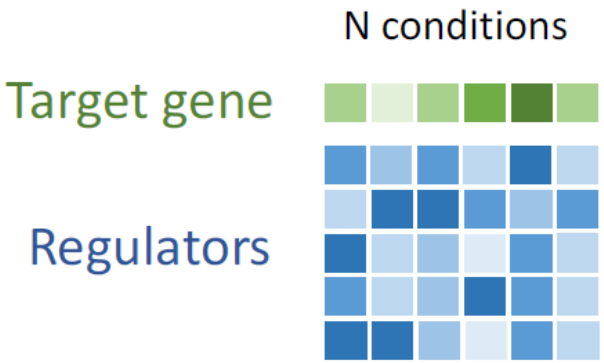
Modelling the root response to nitrate induction in *Arabidopsis thaliana*



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



Temporal response to nitrate induction
1426 genes, 201 regulators, $N = 45$ samples [Varala et al., 2018]

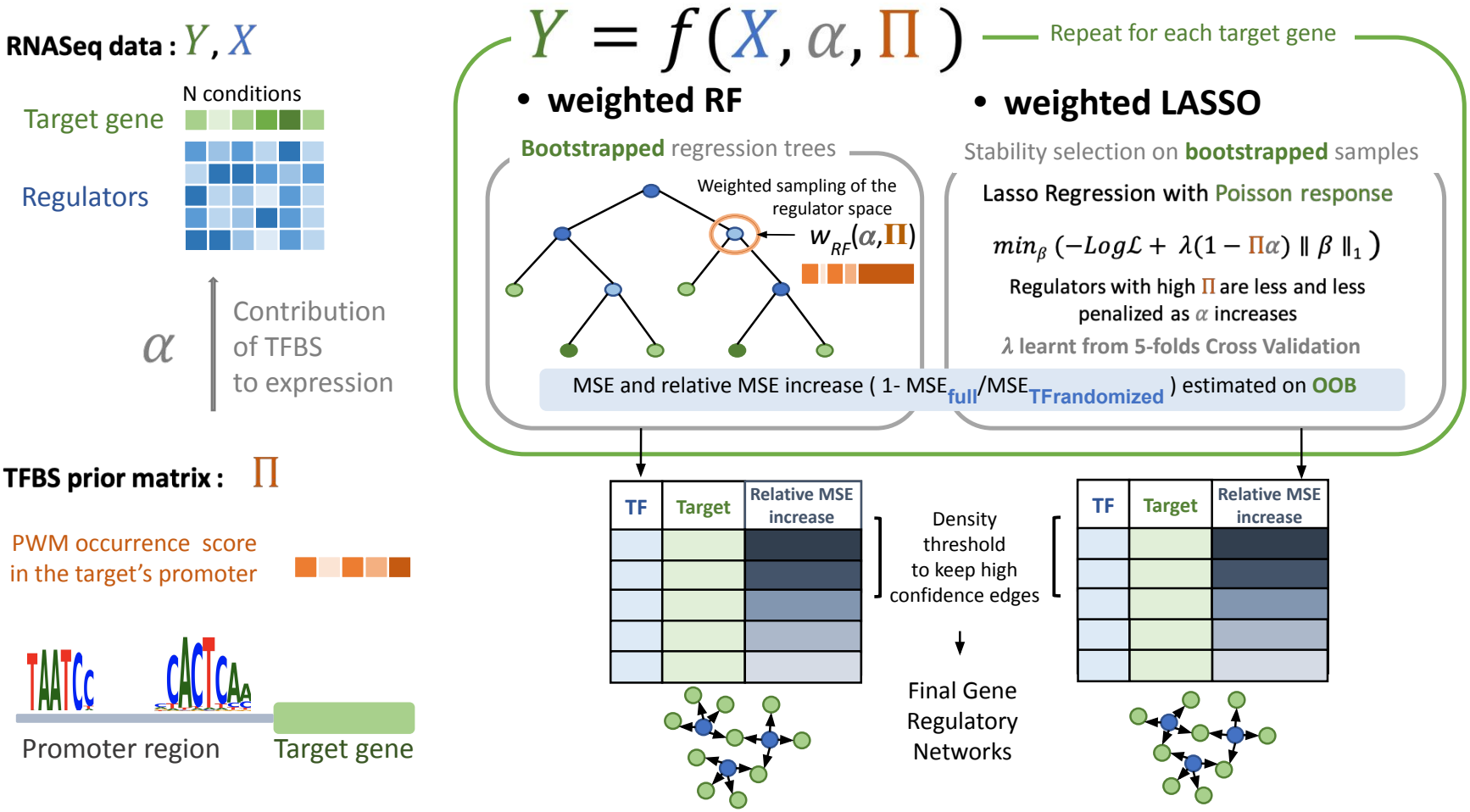
TFBM prior matrix : Π

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 [Petralia et al., 2015] LASSO-Stars [Miraldi et al., 2019]

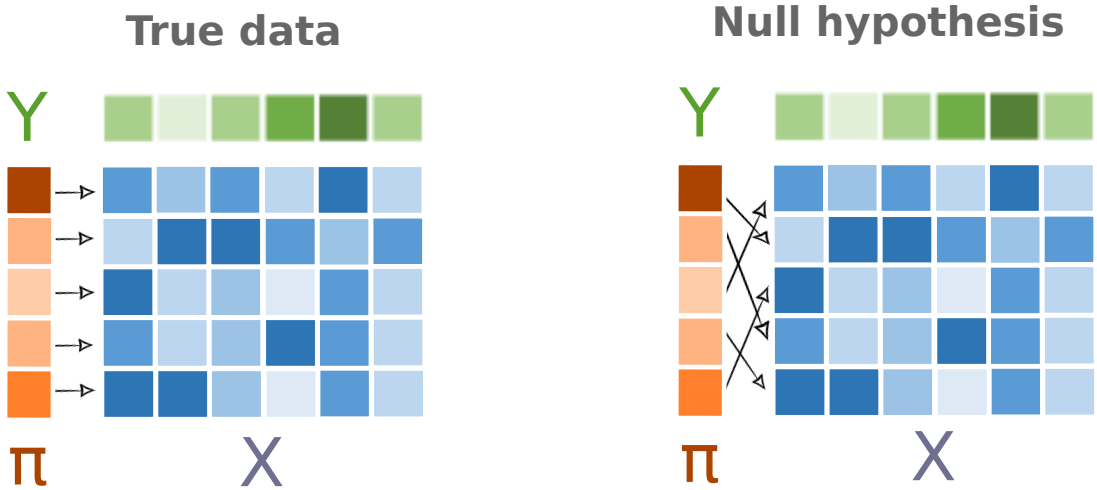
DIOgene: a gene-wise criterion to optimise data integration

Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is not deteriorated**.

A simulated null hypothesis

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



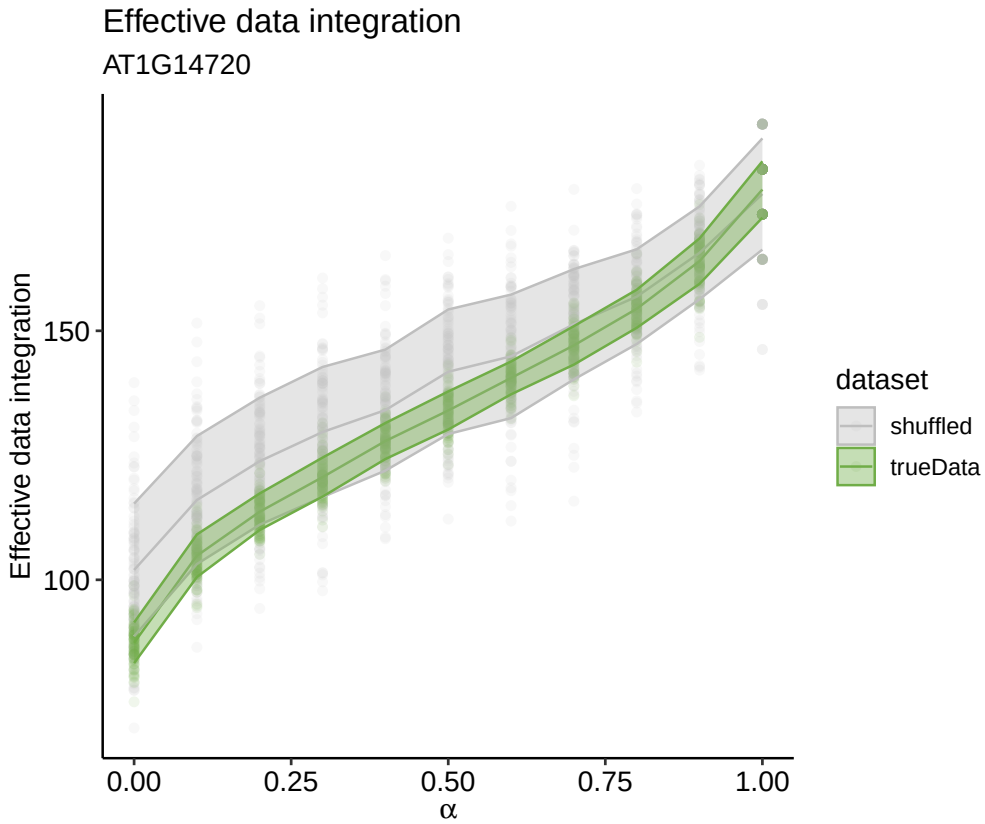
DIOgene: a gene-wise criterion to optimise data integration

Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is not deteriorated**.

Effective data integration

Average rank of PWM-supported regulators based on their importance



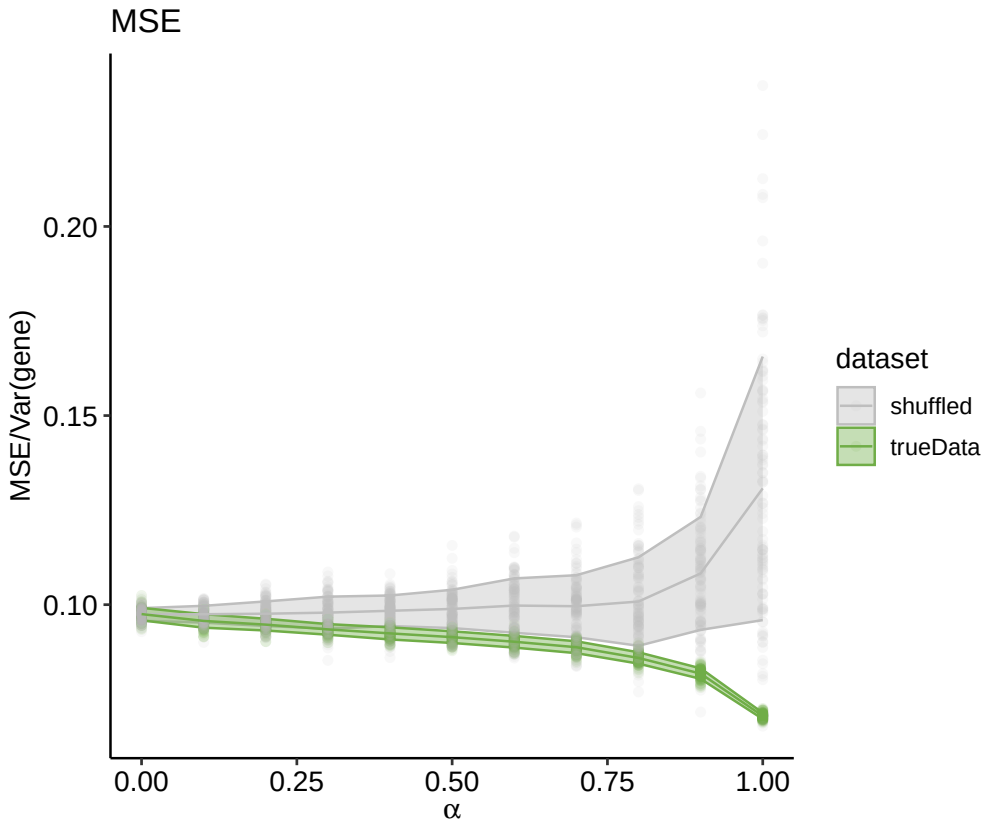
DIOgene: a gene-wise criterion to optimise data integration

Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is not deteriorated**.

Prediction error (MSE)

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



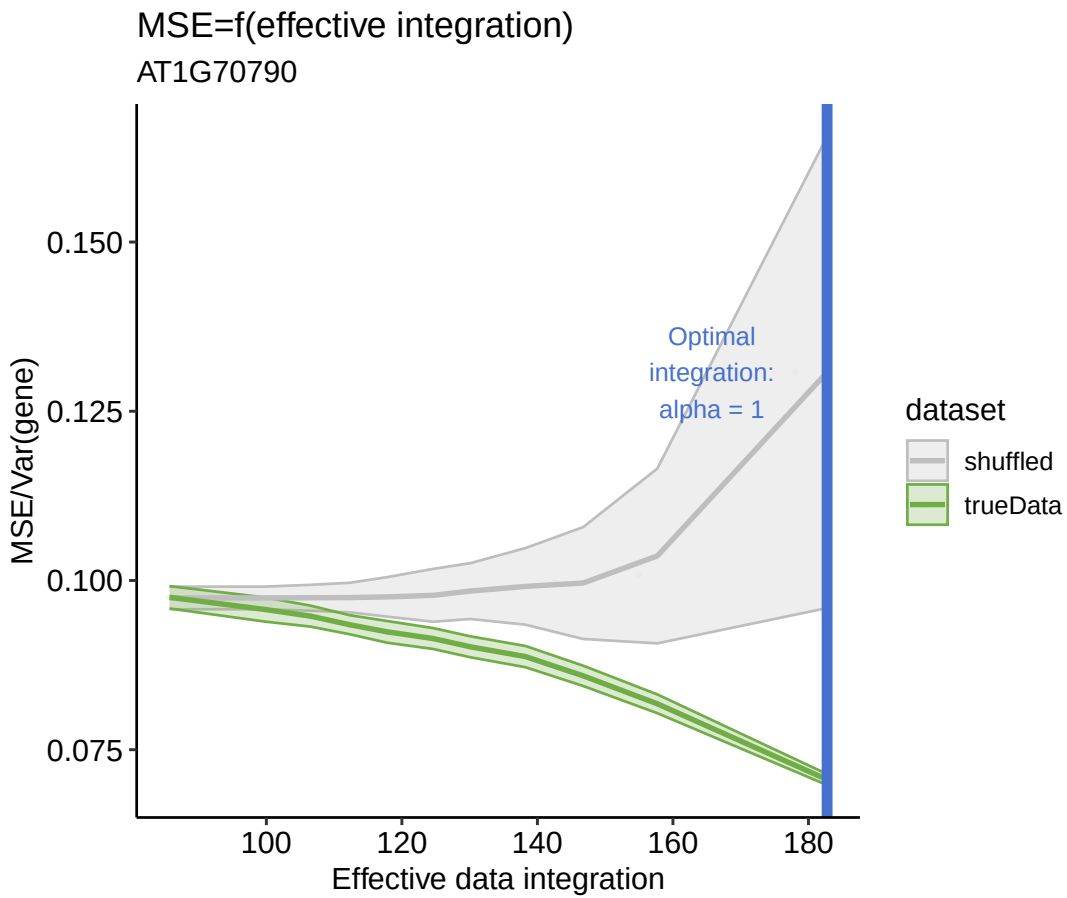
DIOgene: a gene-wise criterion to optimise data integration

Modelling assumption

We want to integrate **TFBMs while controlling that the prediction of gene expression is not deteriorated**.

Ideal case

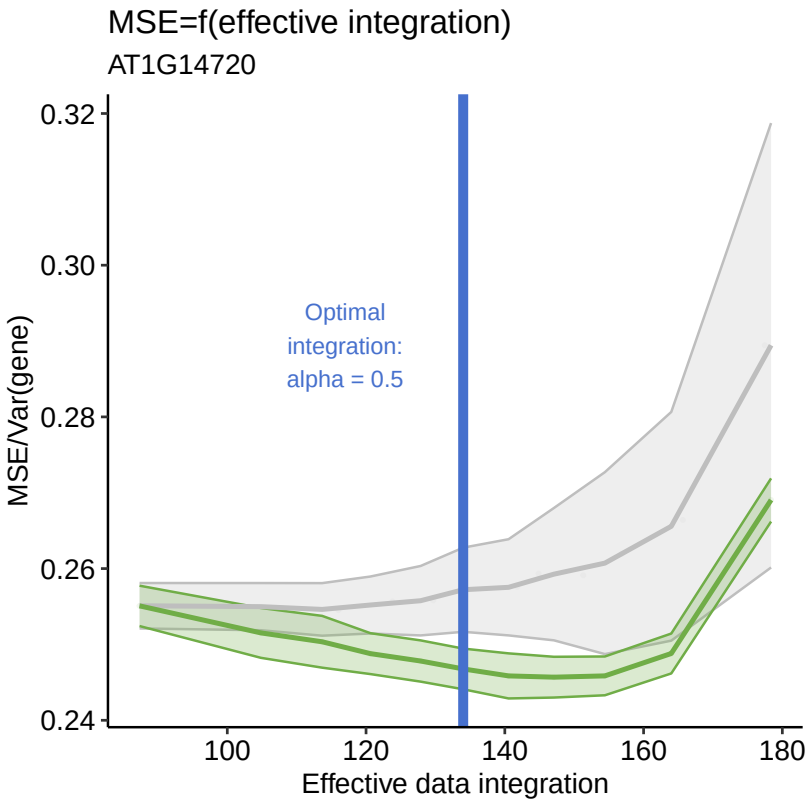
TFBM help selecting robust regulators and improve model generalisation performance

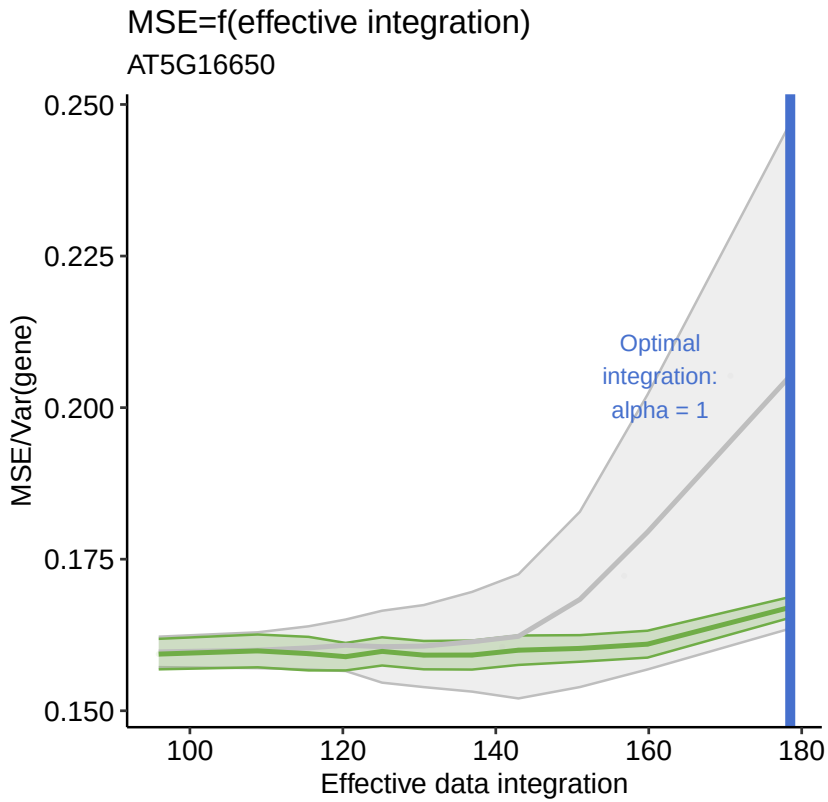


DIOgene: a gene-wise criterion to optimise data integration

Intermediate cases

The MSE reaches an optimum, or is improved over chance

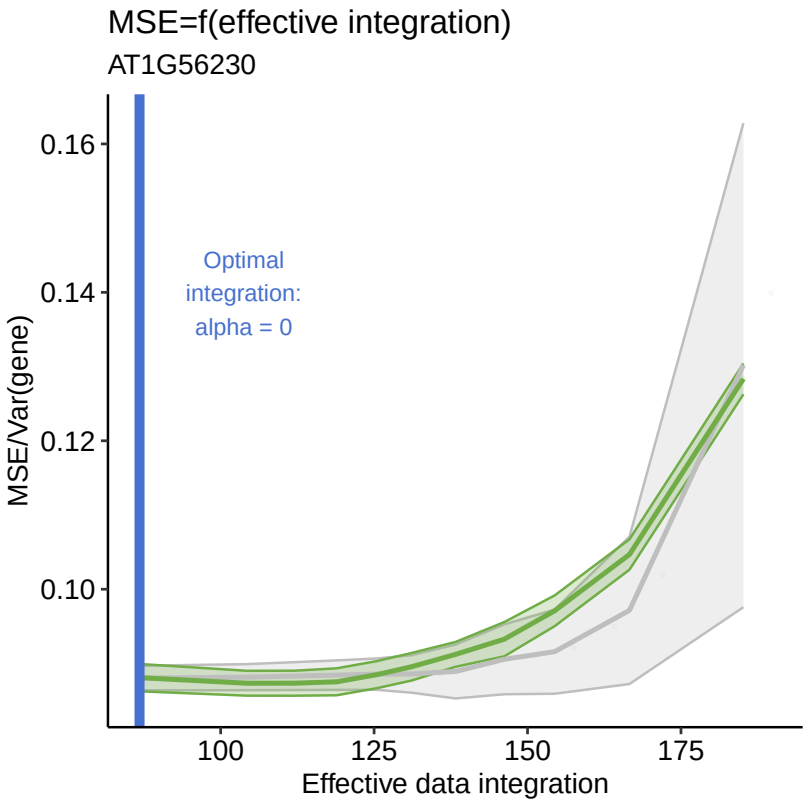


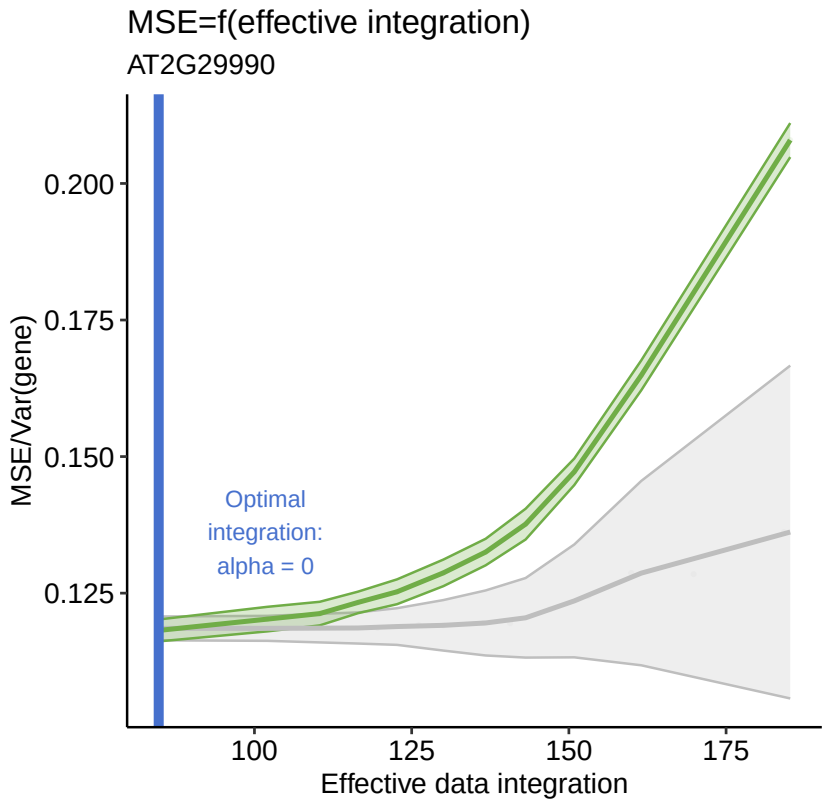


DIOgene: a gene-wise criterion to optimise data integration

Cases where data integration should be avoided

The MSE does not differ from chance or is even higher





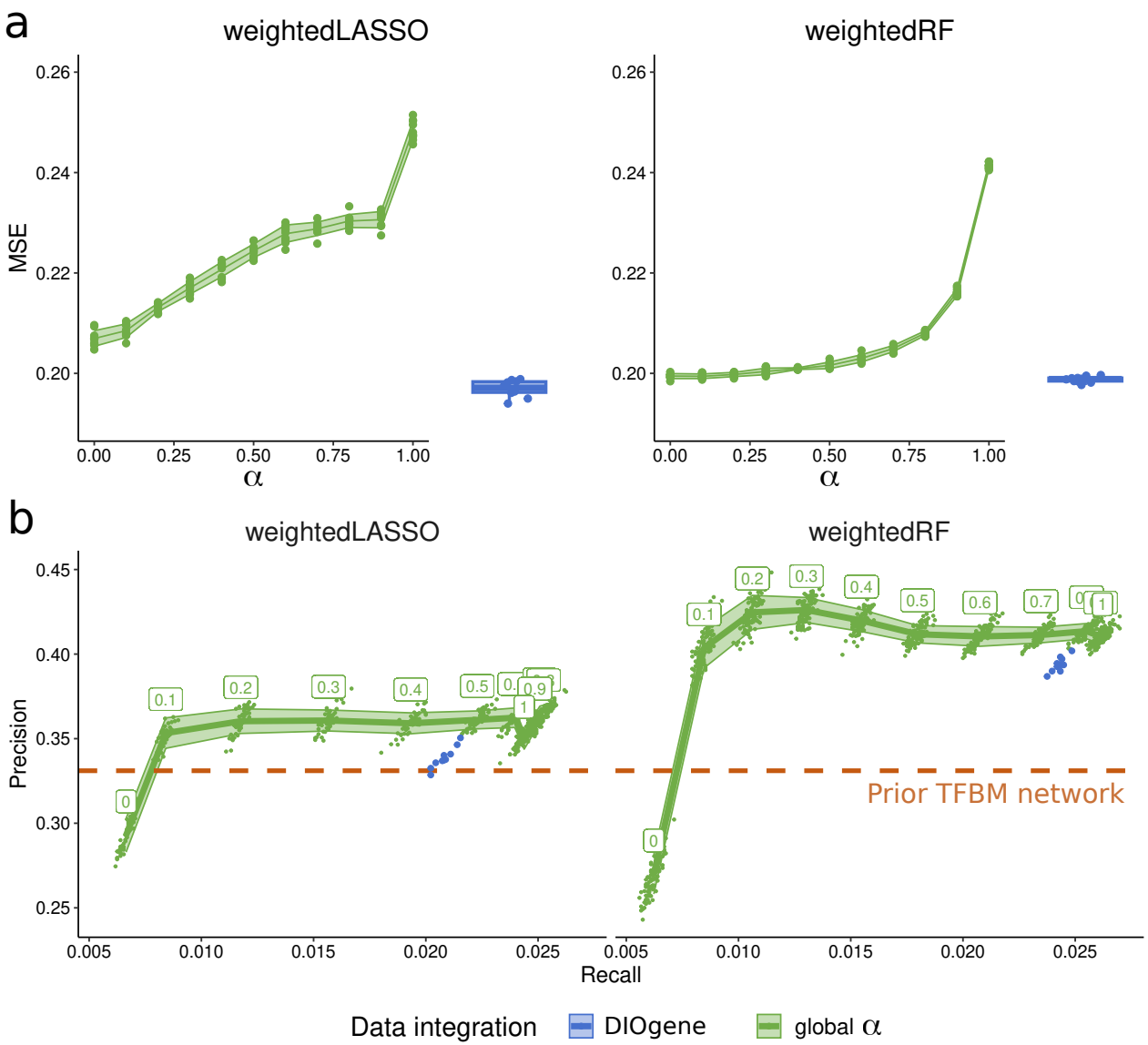
DIOgene offers low MSE, and good precision/recall

GRN reconstruction

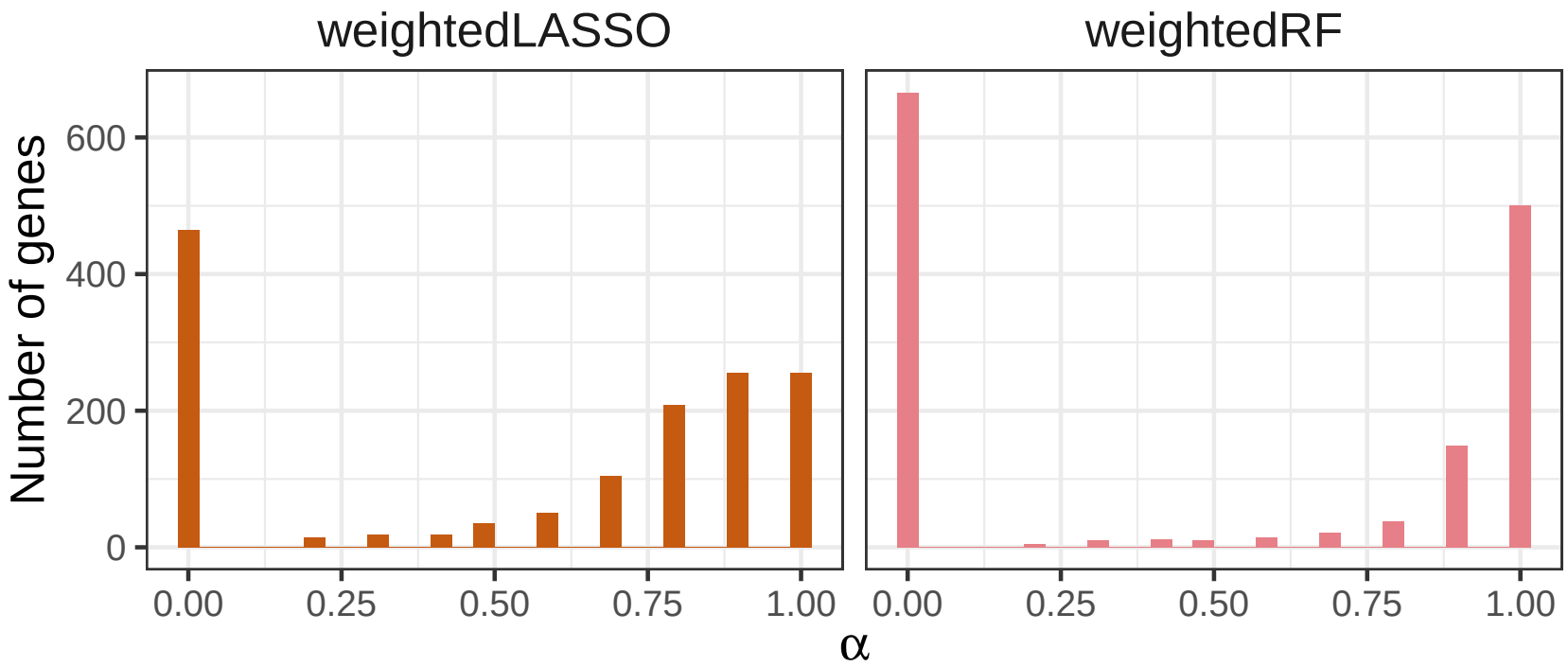
↳ 0.005 density threshold on importance-ranked edges (1432 edges)

GRN quality metrics

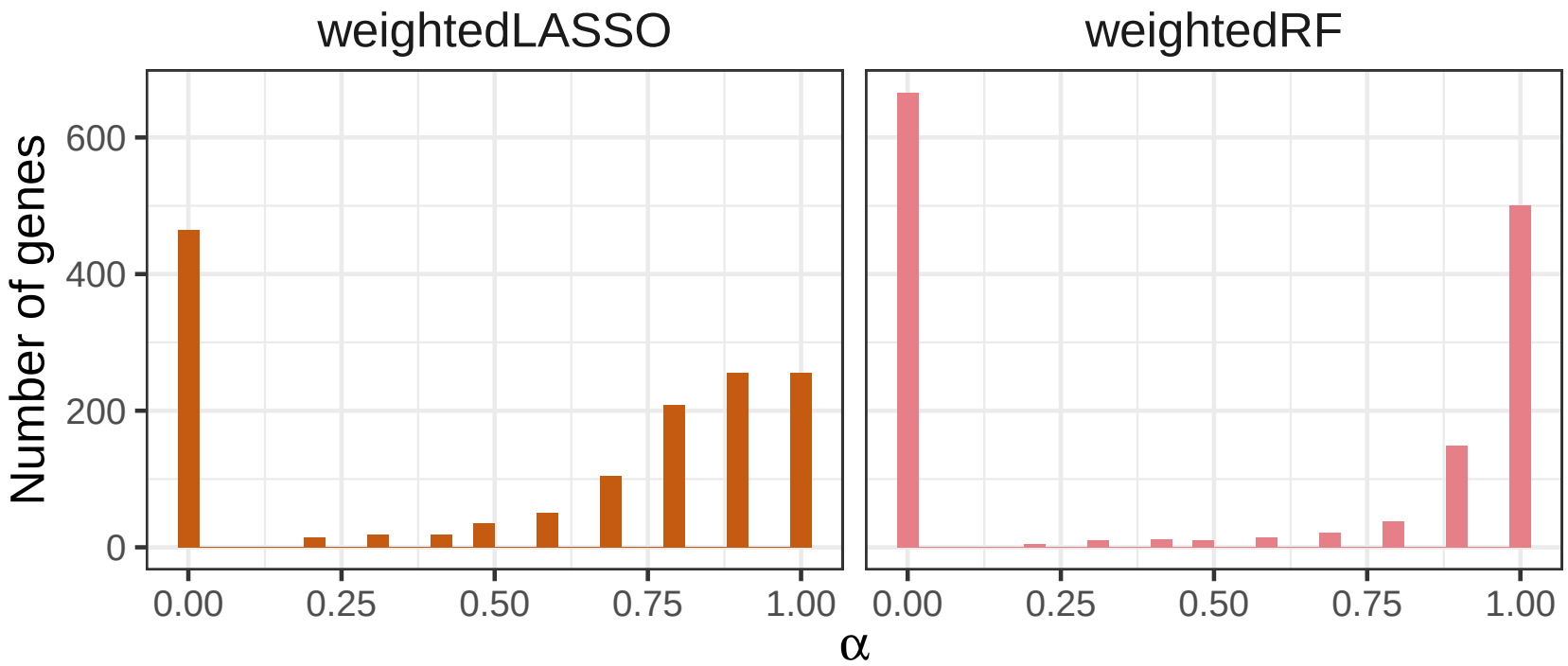
↳ **Median MSE**, and precision-recall curves against **DAP-Seq** in vitro TF binding [O'Malley et al., 2016]



Distributions of optimal integration strengths α_{opt}



Distributions of optimal integration strengths α_{opt}



TFBM integration is not warranted for all genes

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

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`

Preprint: `https://doi.org/10.1101/2023.09.29.558791`

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`

Preprint: `https://doi.org/10.1101/2023.09.29.558791`

Leveraging a problem-specific synthetic baseline

Importance of **in-silico controls** for causal discovery in genomic analyses (Jingyi Jessica Li's Keynote talk at ISMB 2023)

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.

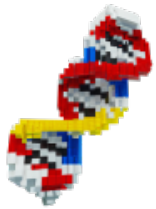
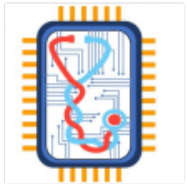
Future research directions

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.

Acknowledgments

The organizing committee of LEGO 2023

Joint LIRMM - IMAG - IGMM Montpellier
Computational Regulatory Genomics team
Sophie Lèbre, Laurent Bréhélin, Charles-Henri Lecellier, Mathilde Robin, Christophe Vroland, Quentin Bouvier


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., Léran, 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 :

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

Prediction error (MSE)

Error committed in predicting the target gene expression :

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

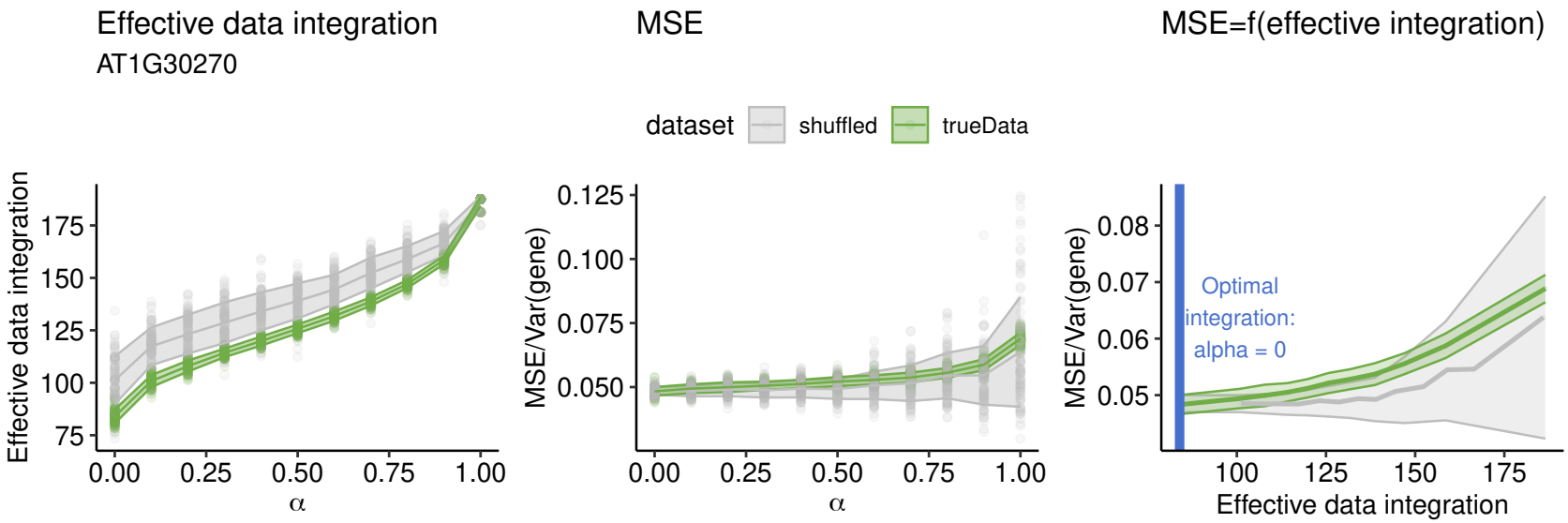
A simulated null hypothesis

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

A gene-wise criterion to optimise data integration

Modelling assumption

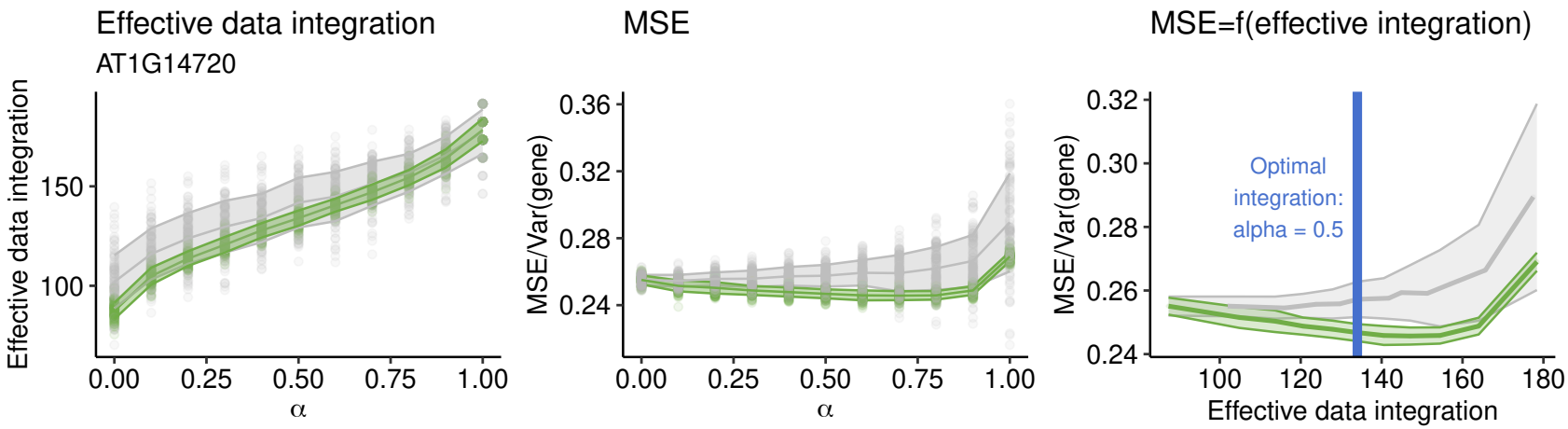
We want to integrate TFBS **only if they improve prediction when used jointly with expression data**.



A gene-wise criterion to optimise data integration

Modelling assumption

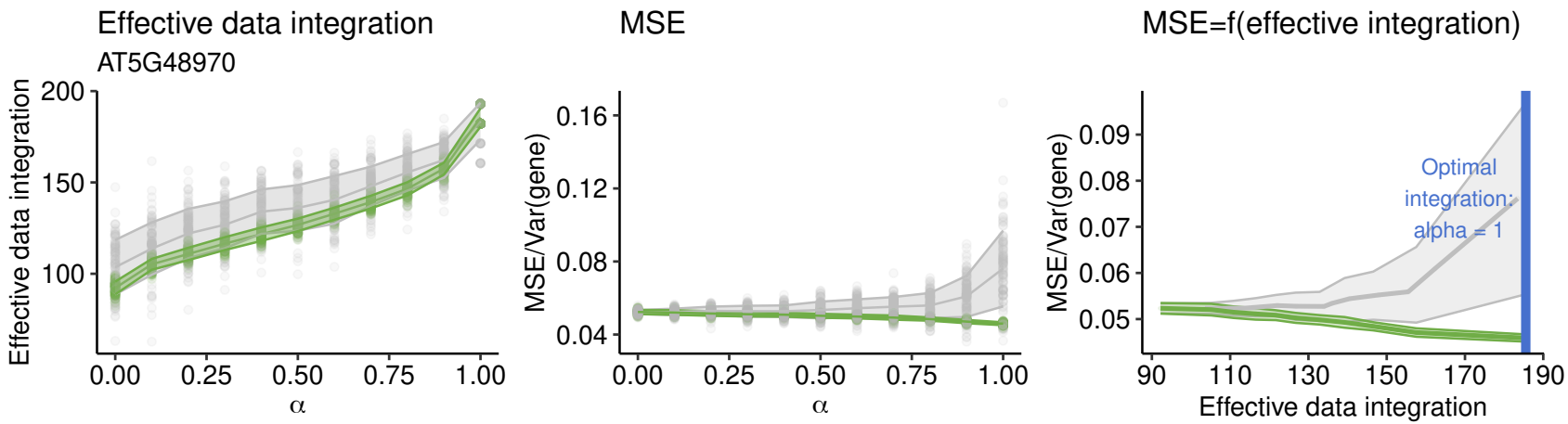
We want to integrate TFBS **only if they improve prediction when used jointly with expression data**.



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**.



Criterion formal definition

$$\Delta_\alpha = \frac{\mu_{\text{shuffle},\alpha} - \mu_{\text{true},\alpha}}{\sigma_\alpha} \qquad \alpha_{\text{opt}} = \begin{cases} 0 & \text{if } \max\limits_{\alpha\in[0,1]} (\Delta_\alpha) \leq 1 \\ \operatorname{argmax}\limits_{\alpha\in[0,1]} (\Delta_\alpha) & \text{otherwise} \end{cases}$$

Deviation measure choice (σ_α) can modulate integration stringency:

- $\sigma_\alpha = \sigma_{\text{shuffle},\alpha} \rightarrow$ Low, stringent data integration
- $\sigma_\alpha = \frac{1}{2}(\sigma_{\text{shuffle},\alpha} + \sigma_{\text{true},\alpha}) \rightarrow$ Moderate data integration
- $\sigma_\alpha = \sigma_{\text{true},\alpha} \rightarrow$ Strong, permissive data integration

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 α uniquely retrieves :

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