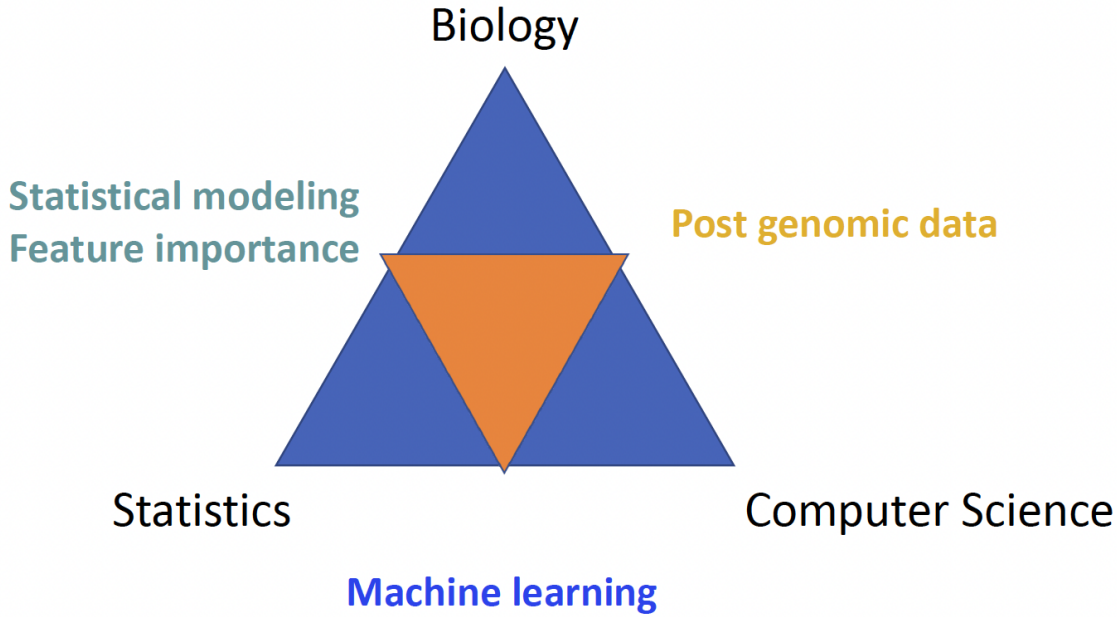
# Optimizing data integration improves Gene Regulatory Network inference

O. Cassan, C.-H. Lecellier, A. Martin, L. Bréhélin, S. Lèbre





Journées NetBio - 2026

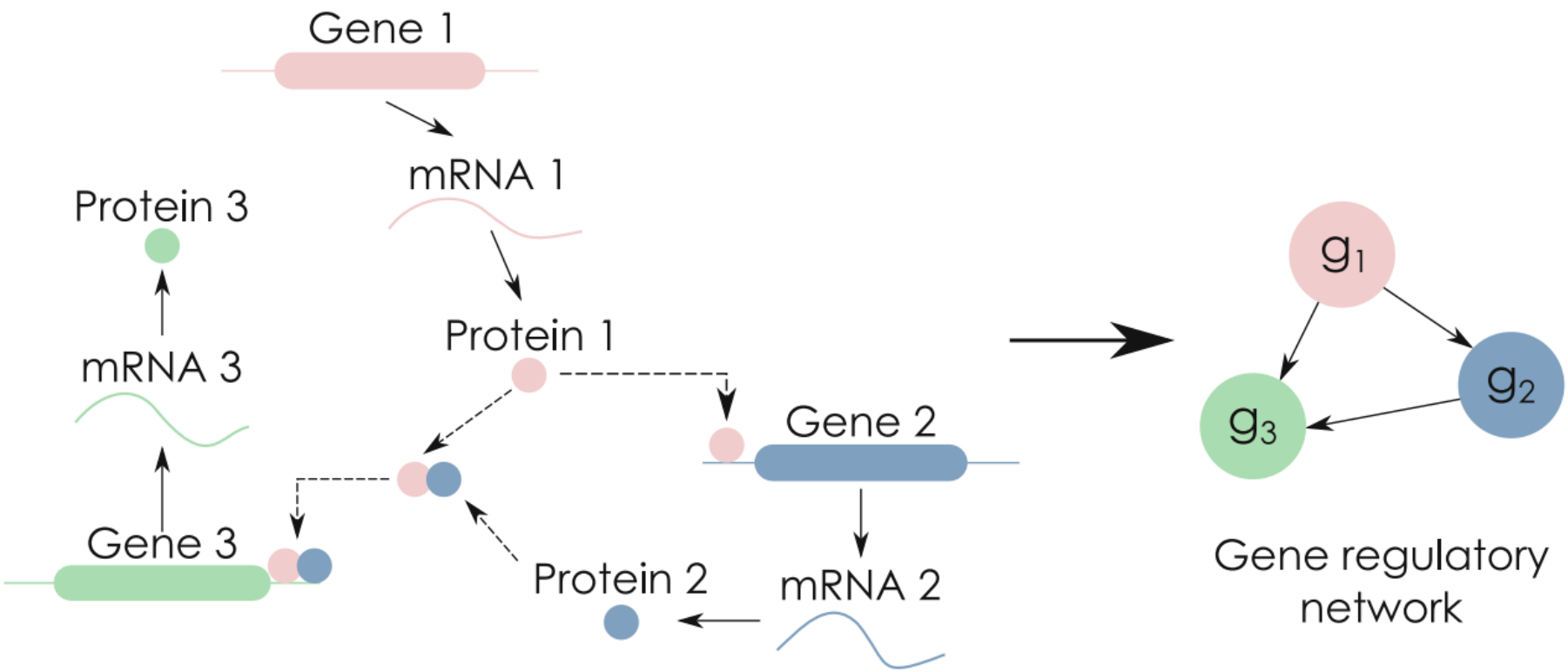



- PhD thesis by **Océane Cassan**
  - Co-supervised with **Antoine Martin (IPSIM)**

- Machine Learning for Regulatory Genomics group
  - **Charles Lecellier (IGMM)**
  - **Sophie Lèbre (IMAG)**
  - **Laurent Bréhélin (LIRMM)**

`https://www.lirmm.fr/ml4reggen/`

# Gene regulation plays a critical role in the cell



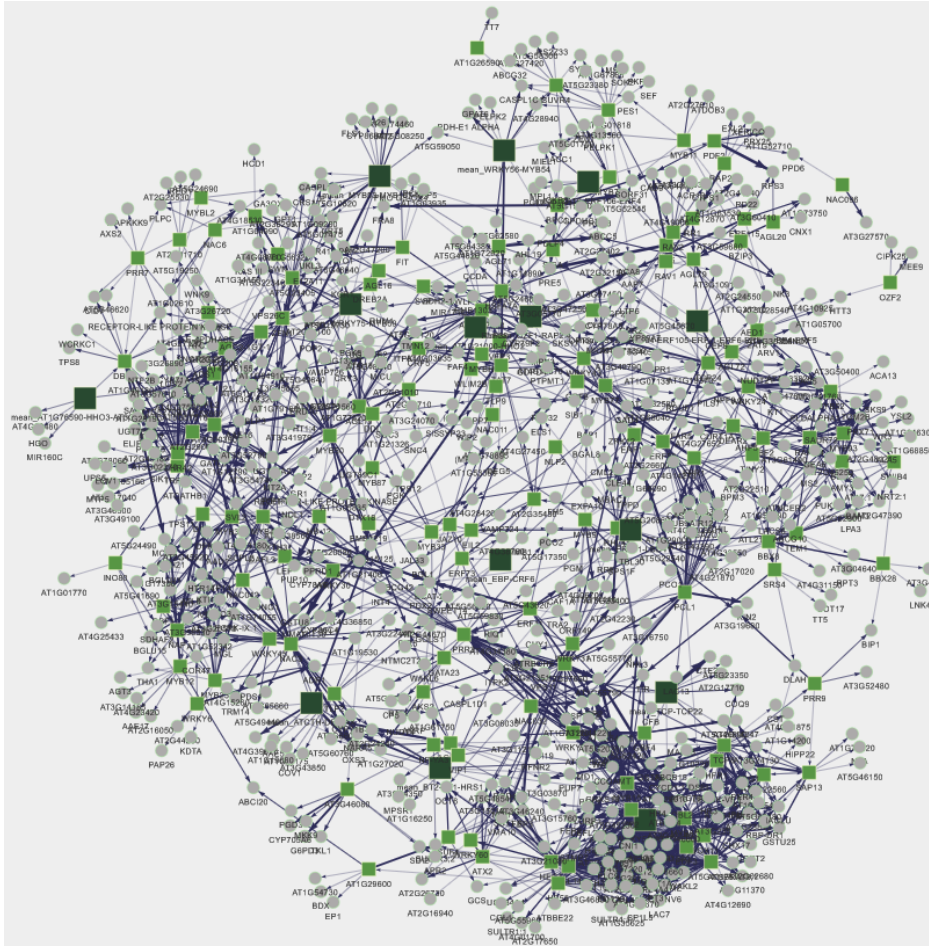

(Figure: Océane Cassan)

- The **regulation of gene transcription** plays a fundamental role.
- **Dysregulation** of such programs can contribute to the development of **numerous diseases**.

# Expression based GRN inference

**System level**

- Quantitative measurement of mRNA abundance.
  - bulk RNA-Seq
  - single cell RNA-Seq
- every gene is measured simultaneously



**But ...**

- **Large dimension**:
  $p \approx 20\,000$ coding genes in human
  $p \approx 30\,000$ in plant *A. thaliana*.
  and number of replicates $n << p$

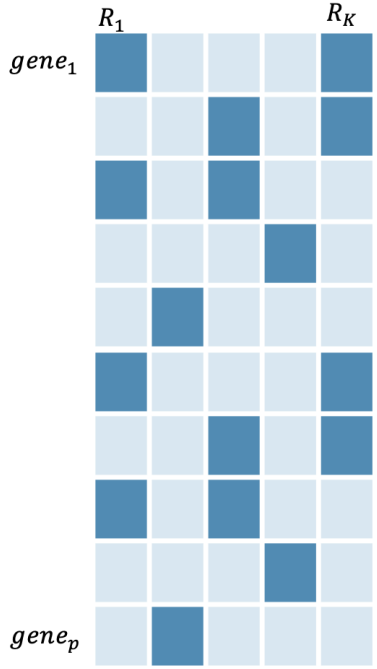
# Outline

1. GRN statistical inference from RNA-seq data alone
2. Build a prior GRN for integrative inference
3. Optimizing data integration with null hypothesis simulation

# Multivariate statistical models

- Multiple TFs within a single model.
- Capture their relative contributions and combinatorial effects.
- **Graphical models**
  - **Gaussian Graphical Models (GGM)** (Schäfer and Strimmer, 2005; Meinshausen and Bühlmann, 2006; Castelo and Roverato, 2006; Ambroise et al., 2009)
  - **Bayesian Networks** (Friedman and Koller, 2003; Grzegorczyk and Husmeier, 2008; Ellis and Wong, 2008); RNA-seq: (Gallopin et al., 2013; Chiquet et al., 2019)
  - **Dynamic Bayesian Networks** (Murphy and Mian, 1999; Kim et al., 2003; G1DBN (Lèbre, 2009); ARTIVA (Lèbre et al., 2010); EDISON (Dondelinger et al. , 2013); RNA-seq: (Thorne, 2018)
- Refined statistical models
- but **limited number of replicates** ($n < \#$ parameters)

# Large scale inference ($n << p$)




- Reduce to **one regression per target gene** ($p$ models)
  - Steady State model :

    $y_t = \beta_{t,1}x_1 + \ldots + \beta_{t,r}x_r + \ldots + \beta_{t,R}x_R + \epsilon_t$
    with:

    $y_t$ the expression of target gene $t$

    $x_r$ the expression of candidate TF $r$

  - Time-step model :

    $\forall 1 \leq i \leq p,\ \forall 1 \leq k < n, \quad X_i(t_{k+1}) = f\left(X(t_k)\right) + \varepsilon_i$

  - ODE model:
    $\forall 1 \leq i \leq p,\ \forall 1 \leq k < n, \quad \frac{X_i(t_{k+1}) - X_i(t_k)}{t_{k+1} - t_k} + \alpha_i X_i(t_k) = f\left(X(t_k)\right) + \varepsilon_i$

    with $\alpha_i$ mRNA degradation rate for gene $i$

- non-zero coefficients $\beta_{t,r}$ determine the active regulations

  $\rightsquigarrow$ Dimension reduction inference methods.

# Linear models

$$y_t = \beta_{t,1} x_1 + \ldots + \beta_{t,r} x_r + \ldots + \beta_{t,R} x_R + \epsilon$$

- **Variable selection** with the **Lasso** ($L_1$-norm regularisation)

  $\rightsquigarrow$ Joint optimization: Find $\hat{\beta}_{Lasso}$ which minimizes

$$||y_t - \sum_{r=1}^{R} x_r \beta_{t,r}||^2 + \lambda \sum_{r=1}^{R} |\beta_{t,r}|$$

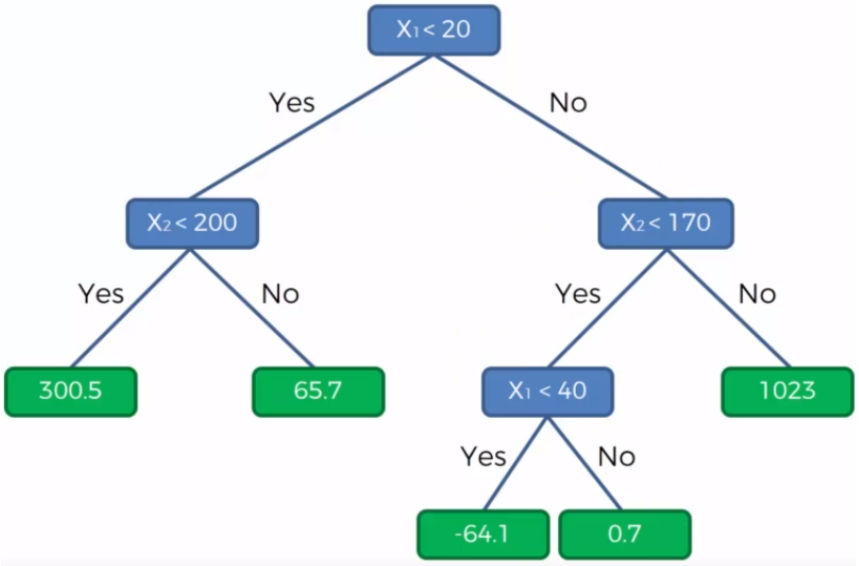
  $\rightsquigarrow$ Weight penalty $\lambda$ is estimated by cross validation.

- Selection of a **small subset of (non correlated)** candidate TFs

- Software:
  `The Inferelator` (Bonneau et al., 2006), `TIGRESS` (Haury et al. 2012), `MERLIN` (Roy et al. 2013), `BBSR` (Greenfield et al., 2013), `StARS-lasso` (Miraldi et al. 2019), `The Inferelator 3.0` (Skok Gibbs et al. 2022), `CellOracle` (Kamimoto et al., 2023)

# Non linear models: Random Forests (RFs)

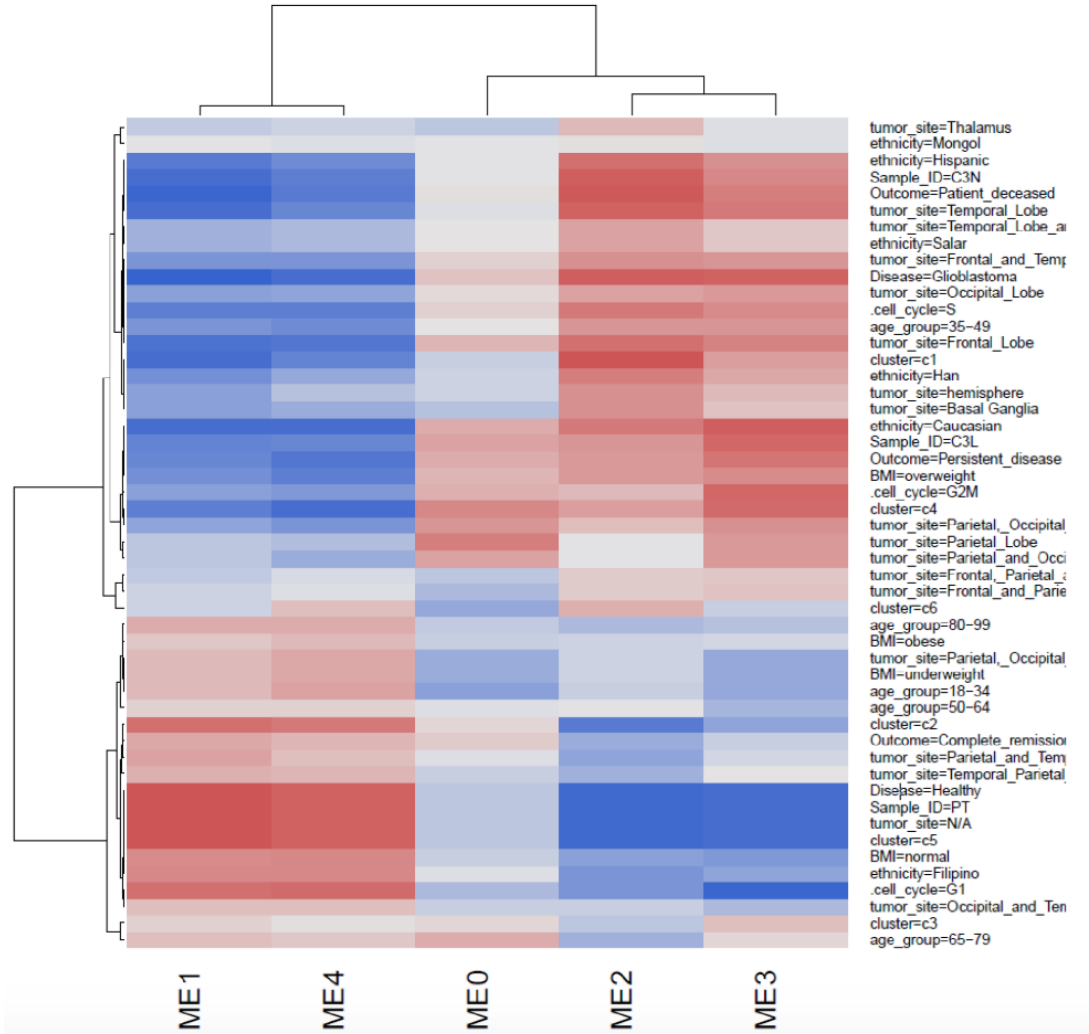
- Ensemble of **regression trees**.
- At each node, the variable and split threshold are chosen to **minimize prediction error for the target gene**.
- **Order candidate TFs** by importance.




- Software:

  `GENIE3` (Huynh-Thu et al. 2010), `iRafNet` (Petralia et al. 2015), `dynGENIE3` (Huynh-Thu et al. 2018), `SCENIC` (Aibar et al., 2017); `SCENIC+` (González-Blas et al., 2022);

# mRNA-only inference: limitations

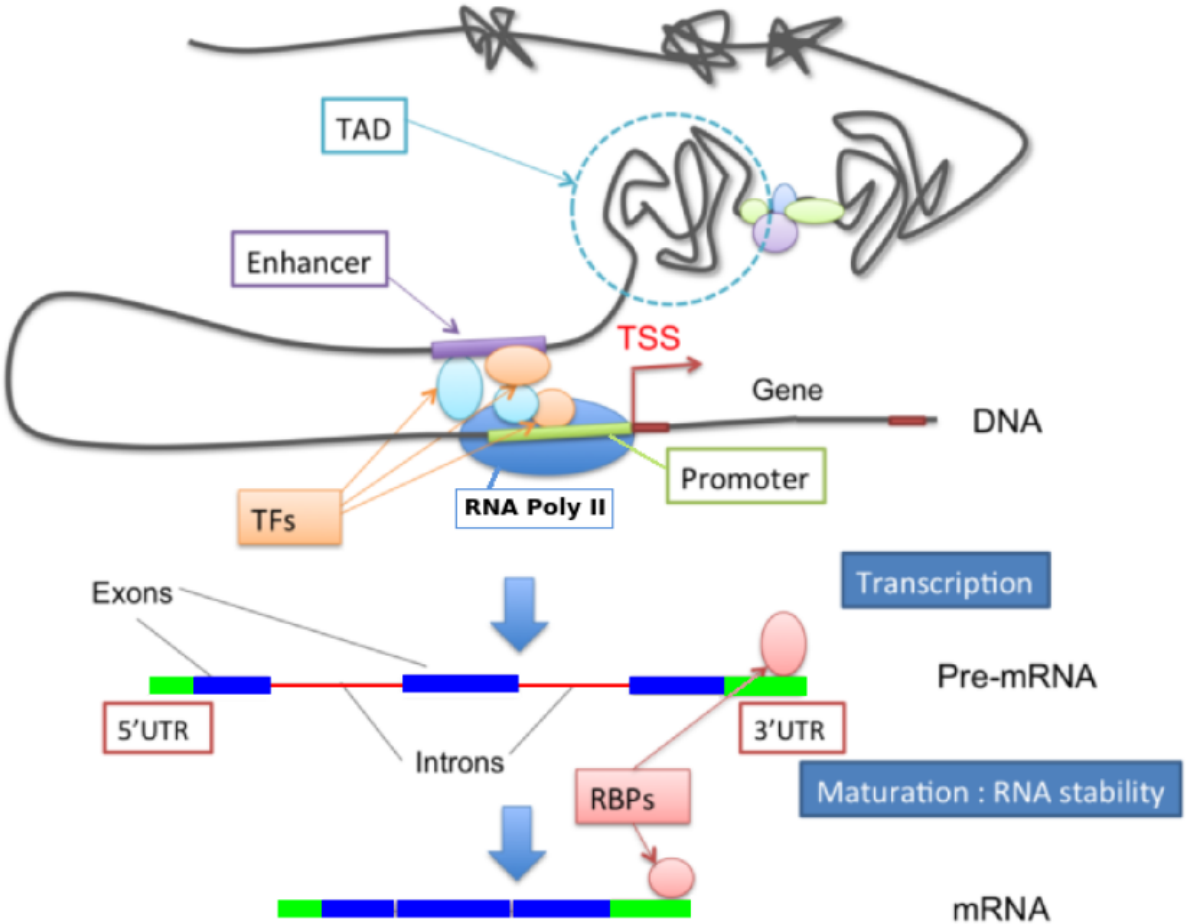



- **Highly correlated TF** expression patterns can be difficult to disentangle in predictive models.

# Outline

1. GRN statistical inference from RNA-seq data alone
2. Build a prior GRN for integrative inference
3. Optimizing data integration with null hypothesis simulation

# Gene expression regulation is highly complex



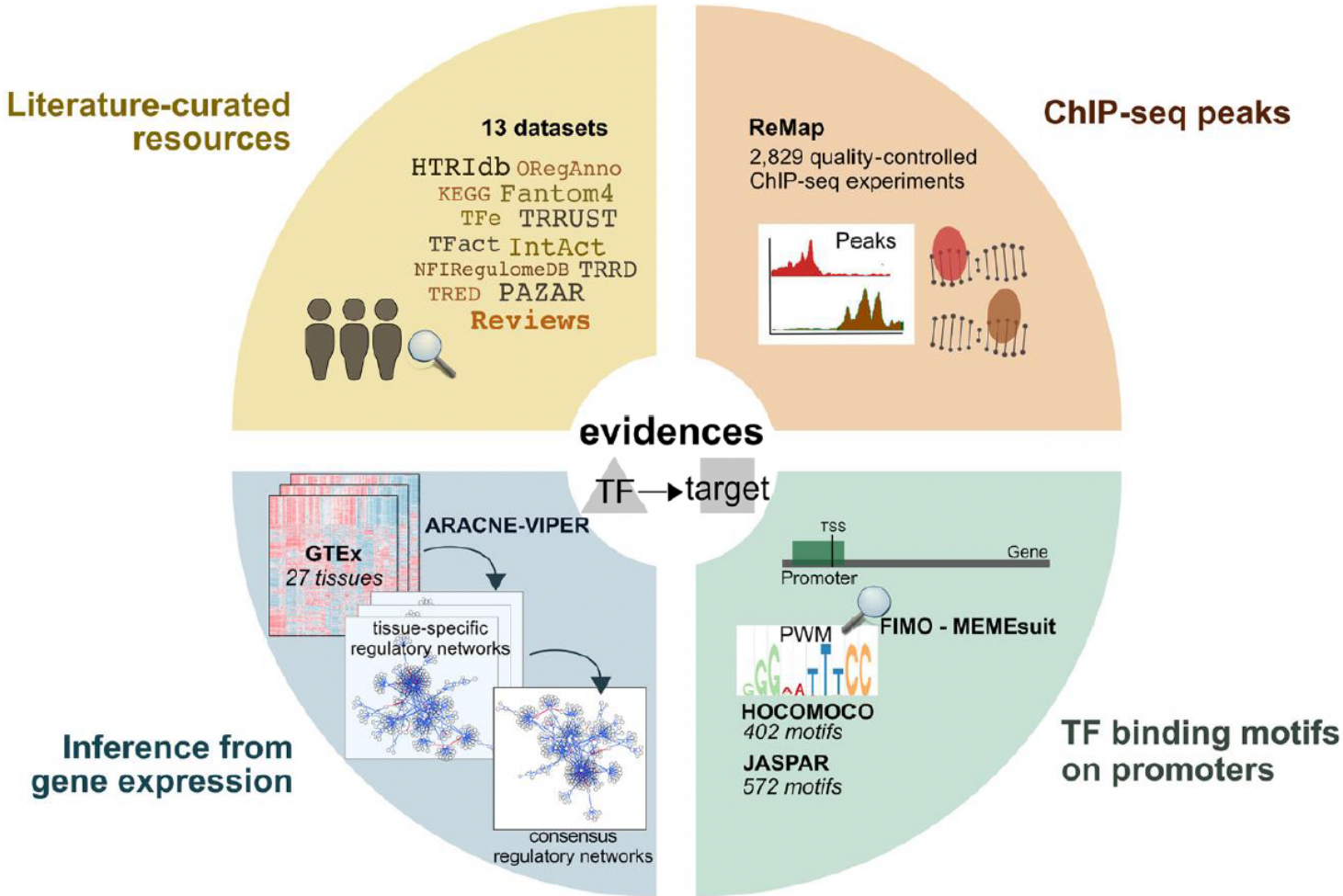

TFs = Transcription factors
RBPs = RNA Binding Proteins

**Transcriptional regulations**

**Post-transcriptional regulations**

(Figure: Chloé Bessière)

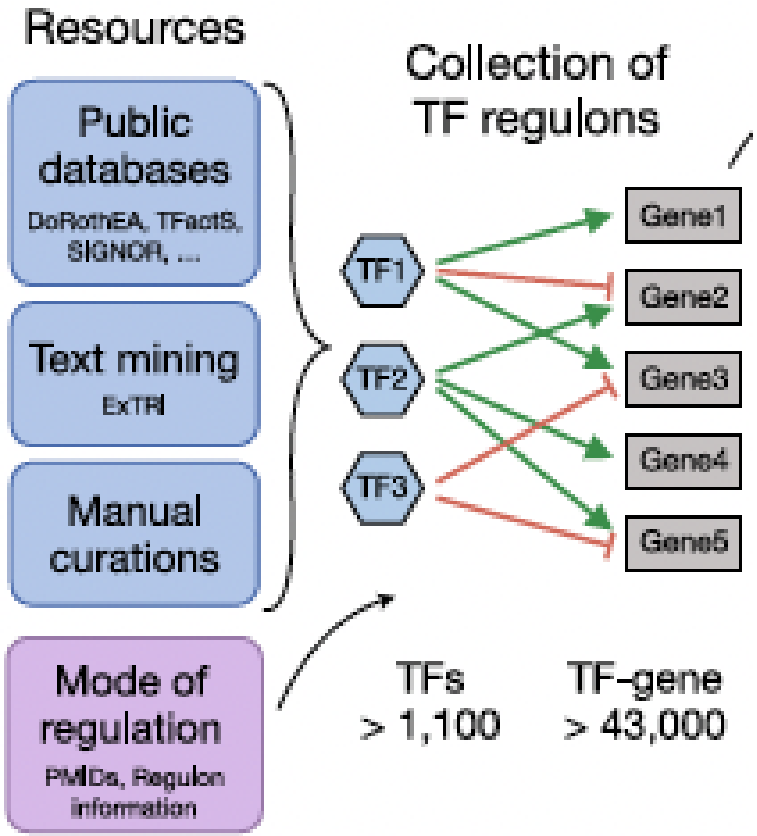
# Classical methods for identifying a prior GRN




(Garcia-Alonzo et al. 2019)

- Capture consensus regulatory patterns

# Compendium databases



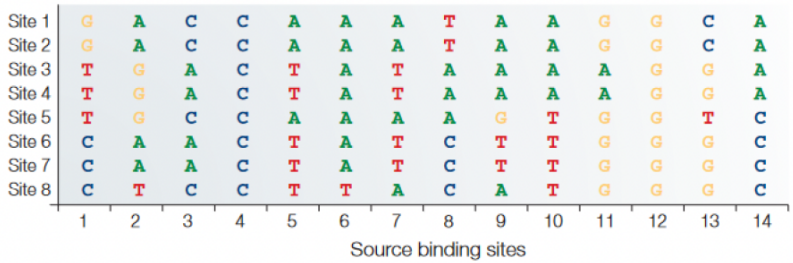

ChEA3 (Keenan et al. 2019)

DoRothEA (Garcia-Alonso et al. 2019)

Pathway Commons (Rodchenkov et al. 2019),

CollecTRI (Müller-Dott et al. 2023)

# TF binding motifs construction (PWM)




Alignment of TF binding sites

Position frequency matrix (PFM)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

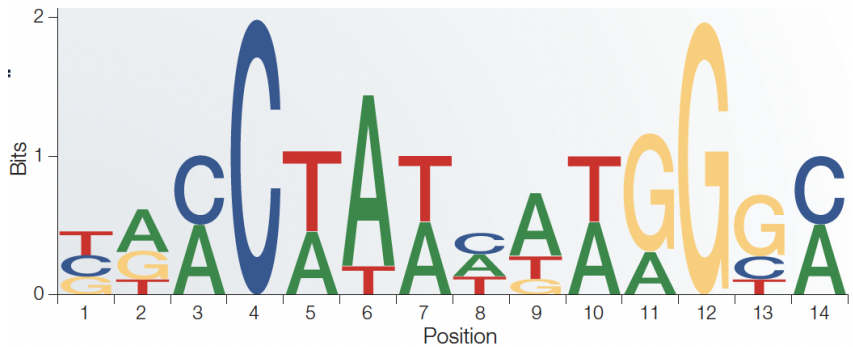
Count the number of occurrences of each nucleotide at each position

Position weight matrix (PWM)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| C | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| T | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

$p(b,i)$: probability of having base $b$ at position $i$
$p(b)$: expected probability of base $b$




(Wasserman & Sanderlin 2004)

# Scoring TF motif in a DNA sequence

C T A C A T A A G T A G T C C A G ...

Position weight matrix (PWM)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| C | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| T | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

$\sum$ = [9.19

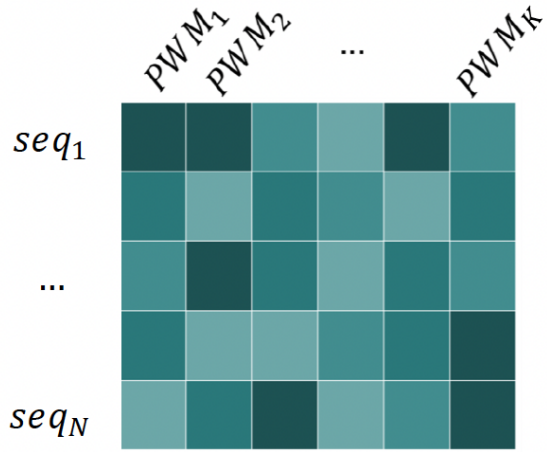
# Scoring TF motif in a DNA sequence

C A T A A G T A G T C C A G ...

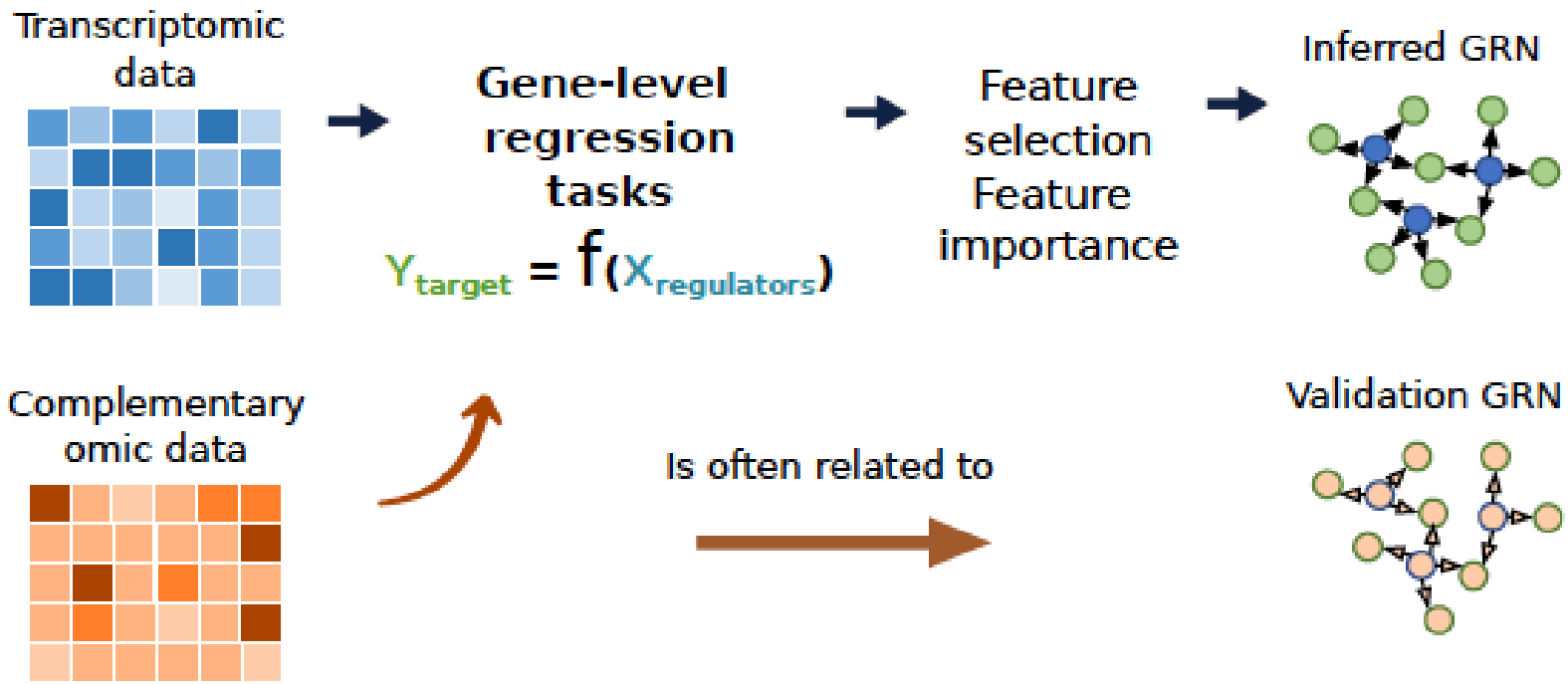
Position weight matrix (PWM)

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | –1.93 | 0.79 | 0.79 | –1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | –1.93 | –1.93 | 0.79 |
| **C** | 0.45 | –1.93 | 0.79 | 1.68 | –1.93 | –1.93 | –1.93 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | 0.00 | 0.79 |
| **G** | 0.00 | 0.45 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | –1.93 | 0.66 | –1.93 | 1.30 | 1.68 | 1.07 | –1.93 |
| **T** | 0.15 | 0.66 | –1.93 | –1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | –1.93 | –1.93 | –0.66 | –1.93 |

$\sum = [9.19, 2.18, -8.91, -9{,}13 , \ldots ]$



# Joint inference with a biological prior




- Linear model: weighted Lasso ($L_1$-norm regularisation)
- Non-linear model: weighted Random Forests.

# Linear model: weighted Lasso

- Instead of $\hat{\beta}_{Lasso}$ minimizing

$$\text{Error} + \lambda_t \sum_{r=1}^{R} |\beta_{t,r}|$$

- Find $\hat{\beta}_{weighted\ Lasso}$ minimizing

$$\text{Error} + \sum_{r=1}^{R} \lambda_{t,r} |\beta_{t,r}|$$

  $\rightsquigarrow$ with a lower penalty $\lambda_{t,r}$ when positive prior

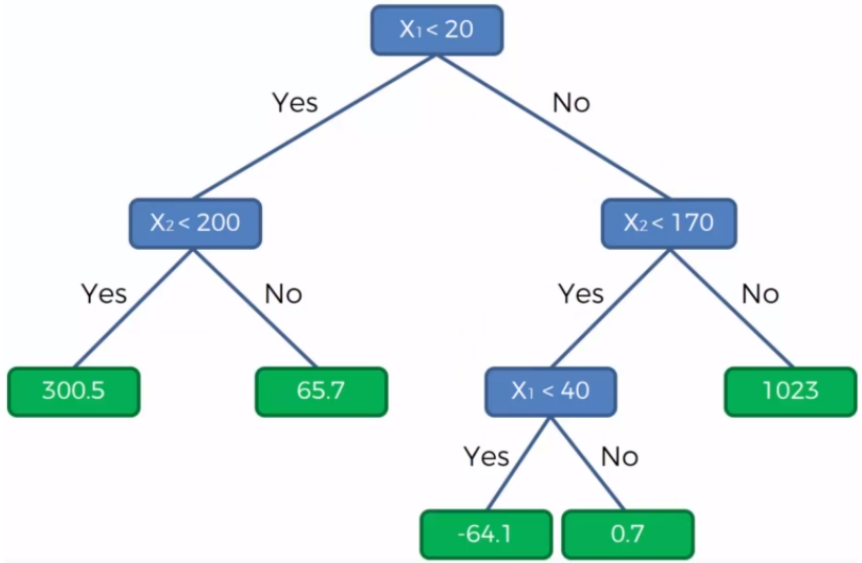
- Software:
  MEN (Greenfield et al.,2013); modified lasso-StaRS (Miraldi et al., 2019) ; The Inferelator (Skok Gibbs et al., 2022).

# Non linear models: Random Forests (RFs)

- Ensemble of regression trees.
- Random Forests
  Random selection of a subset

  of $\sqrt{R}$ TFs with uniform weight:

$$\forall 1 \leq r \leq R,\ w_{t,r} = \frac{1}{R}.$$




- weighted Random Forests: higher weight $w_{t,r}$ when positive prior
- Software:
  `iRafNet` (Petralia et al., 2015), OutPredict (Cirrone et al., 2020).

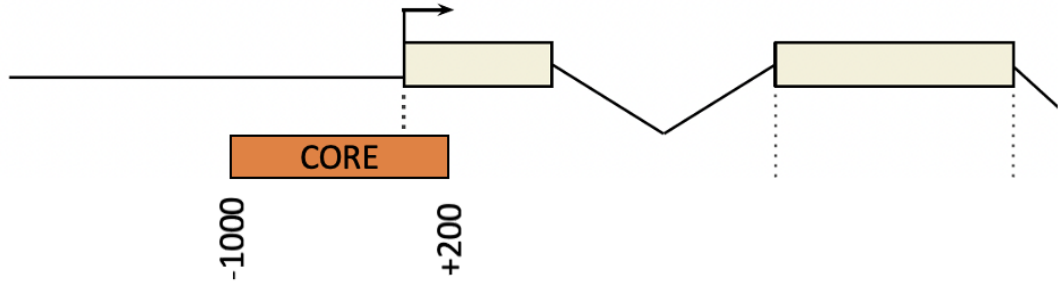
# Outline

1. How can GRNs be (statistically) inferred from RNA-seq data alone?

2. Build a prior GRN for integrative inference approaches

3. Optimizing data integration with null hypothesis simulation
   (O. Cassan, C.-H. Lecellier, A. Martin, L. Bréhélin & S. Lèbre (2024))

# Defining a prior GRN




- Presence of DNA binding motifs (PWM) in target gene core promoter

$$\Pi_{r,t} = \begin{cases} 1 : \text{if TF } r \text{ motif IS in promoter of gene } t \\ \frac{1}{2} : \text{if TF } r \text{ motif is unknown} \\ 0 : \text{if TF } r \text{ motif IS NOT in promoter of gene } t \end{cases}$$

- Our settings
  - Promoter: [-1000; + 200] around TSS (Yu et al. 2016)
  - PWM motifs (JASPAR & The Plant Cistrome Database)
  - Software suite MEME (threshold $10^{-4}$)
  - (only 70 TFs with a known, among 201 TFs)

# Defining a weighted penalty

- Find coefficients $(\beta_{t,1} \ldots, \beta_{t,r}, \ldots, \beta_{t,R})$ minimizing:

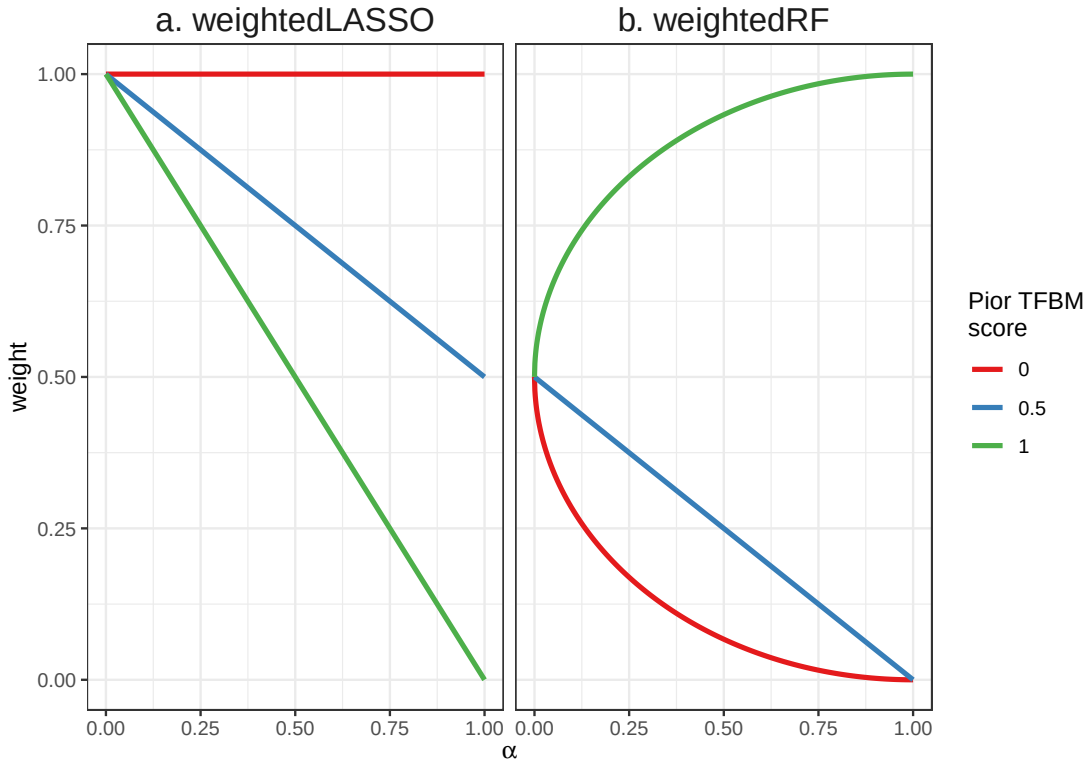
$$\text{Error } + \lambda \, \Sigma_{r=1}^{R} w_{t,r} |\beta_{t,r}|$$

with weight:

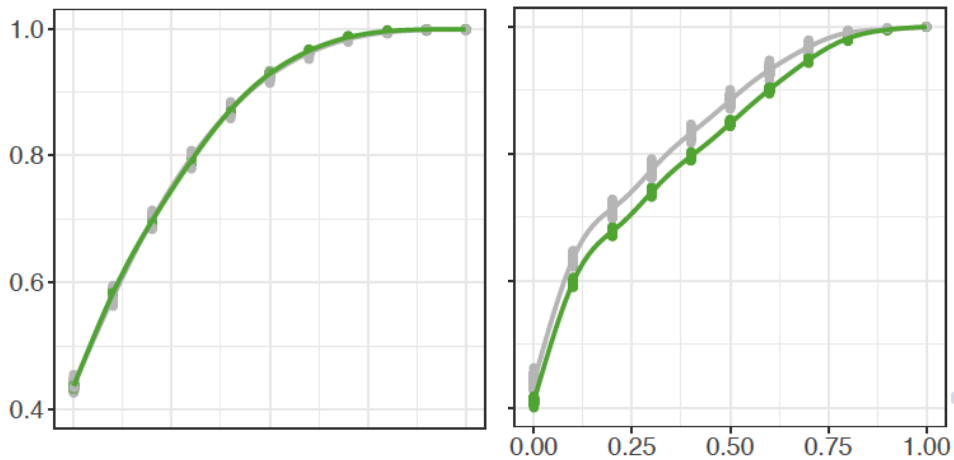
$$w_{t,r} = 1 - \Pi_{r,t} \alpha$$

- $\lambda$ controls global regularisation intensity (R package `glmnet`)
- $\alpha \in [0, 1[$ controls data integration intensity.

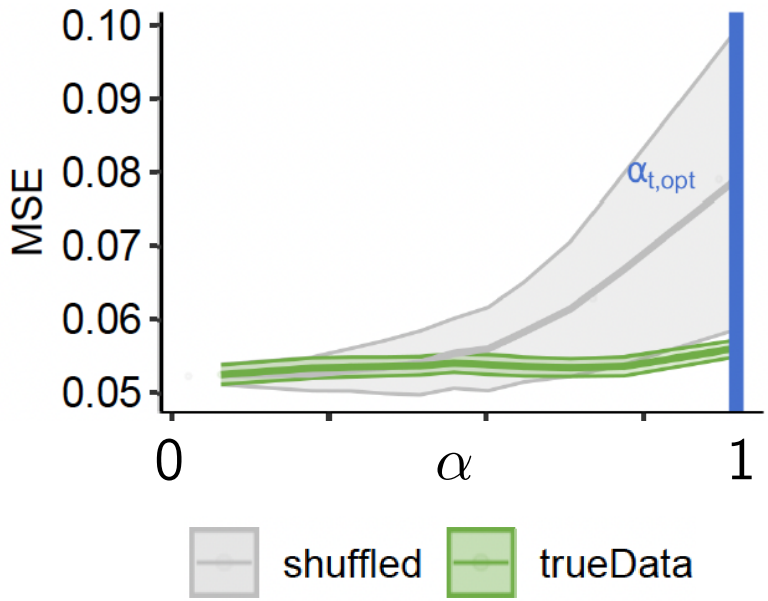
# Weight $w_{r,t}$ and integration strength $\alpha$ functions




Rate of edge with $\Pi_{r,t} = 1$

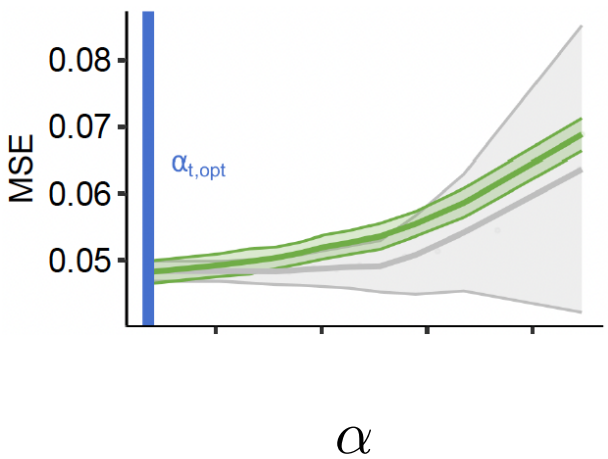


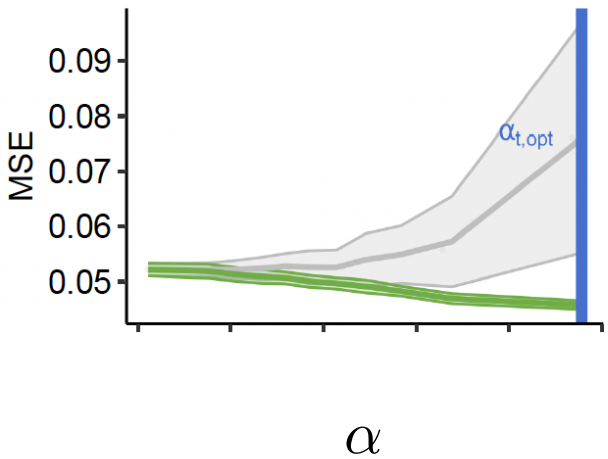
# Choosing optimal integration strength $\alpha \in [0, 1]$

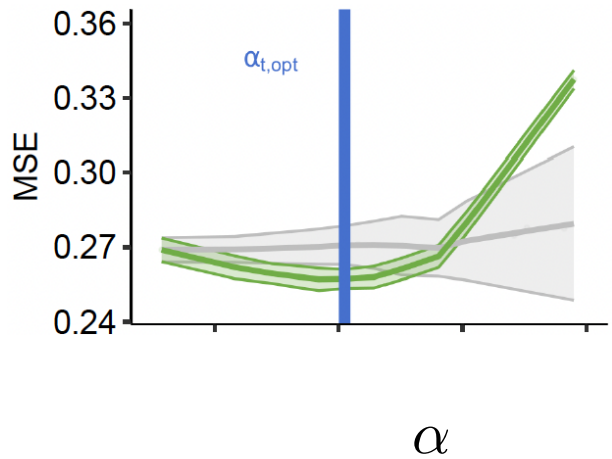



- Based on Mean Square Error (MSE)
- Our aim:
  - Accept a slight loss of performance, to gain biological validation.
  - What is a major deterioration?
- Null hypothesis simulation:
  - random premutation of the TFs expression profile
  - (break the link with the prior $\Pi_{r,t}$)

# Choosing optimal integration strength $\alpha \in [0, 1[$



# Optimizing integration strength $\alpha \in [0, 1[$

- Statistical t-test for mean comparison
  (+ False Discovery Rate correction)

  For each target gene $t$,

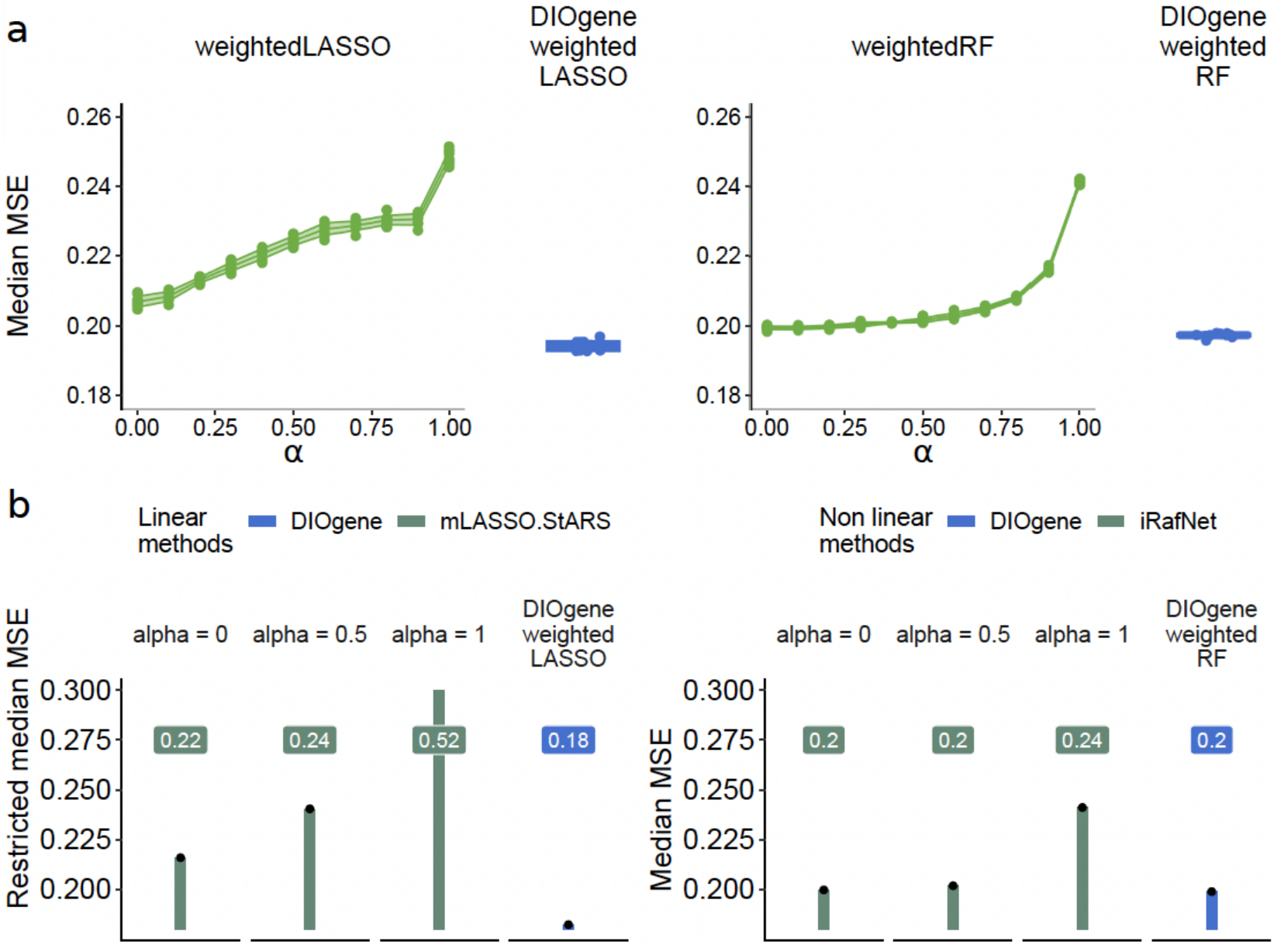
$$T_\alpha = \frac{\mu_{\mathrm{MSE}}(\alpha) - \mu_{\mathrm{MSE}_0}(\alpha)}{\sqrt{\frac{\sigma^2_{\mathrm{MSE}}(\alpha) + \sigma^2_{\mathrm{MSE}_0}(\alpha)}{N}}} \sim t_{(2N-2)}$$

  $N$ is the total number of estimates performed.

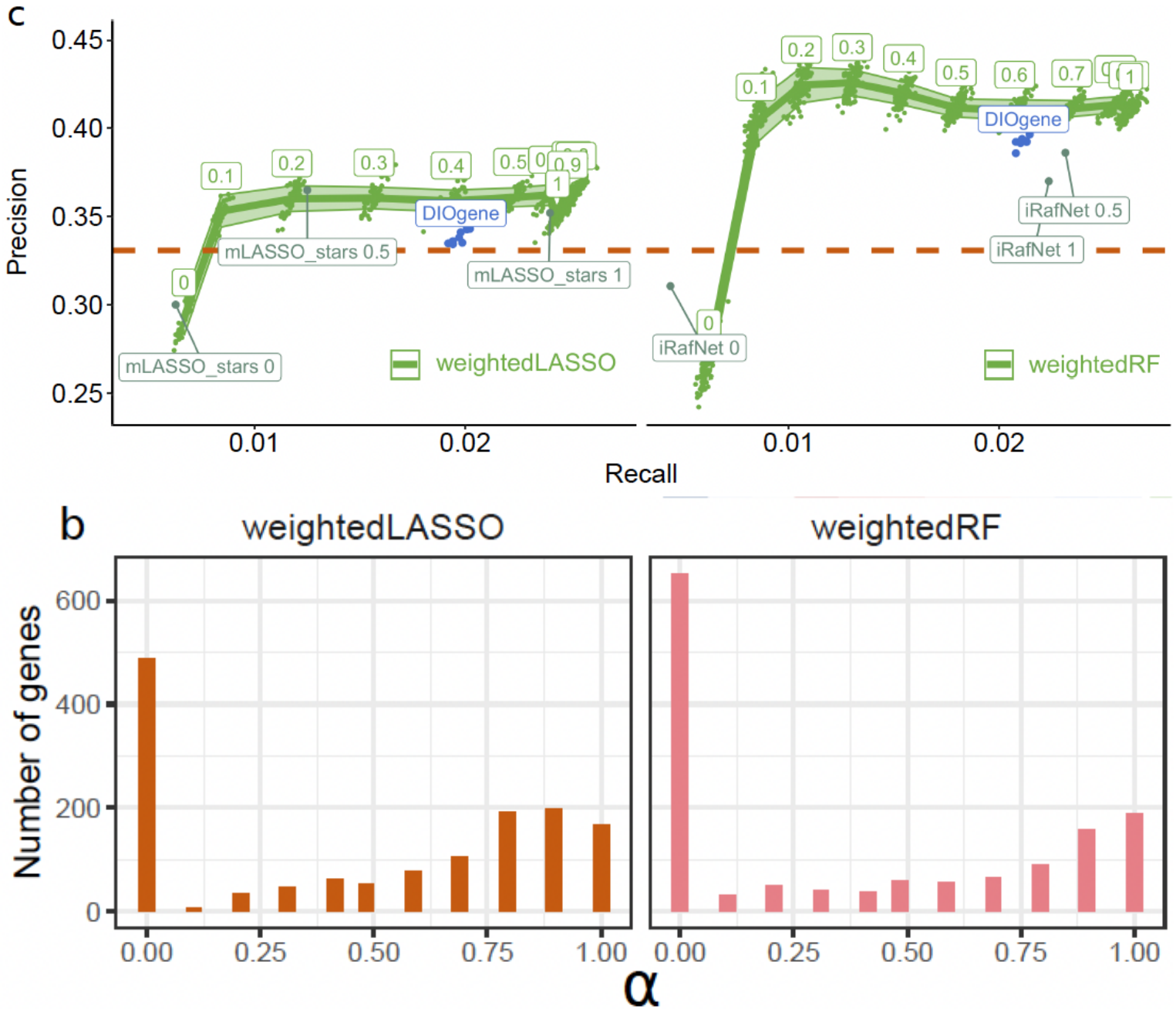
- Maximize the error prediction difference

$$\alpha_t^\star = \arg\min_\alpha \ p\text{-value}(t, \alpha)$$

# Results: 1426 target genes, 201 TFs, $n = 45$ (Varala et al. 2018)



# Validation: DAP-Seq interactions (O'Malley et al. 2016)



# DIOgene: Data Integration Optimization for gene networks

- For 2 classical regression model (linear & non-linear)
  - weighted lasso
  - weighted Random Forests

- Weight definition from biological prior $w_{r,t}(\alpha)$.

- Null hypothesis simulation by random permutation to choose optimal integration strength $\alpha$.

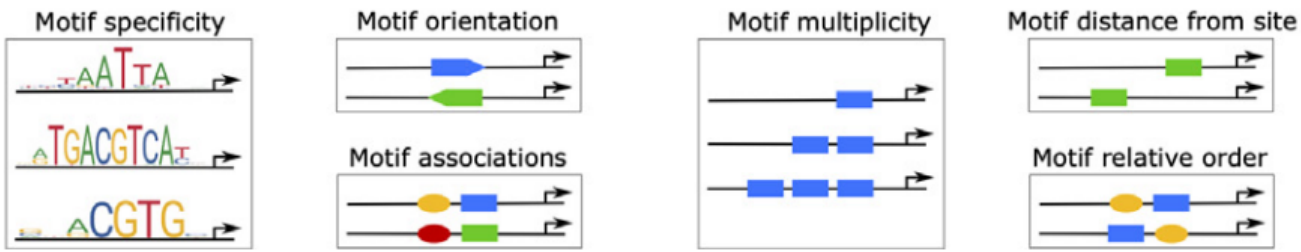
- DIOgene R code & notebooks:
  `https://github.com/OceaneCsn/integrative_GRN_N_induction`

**Gene-specific optimization of data integration improves regression-based Gene Regulatory Network inference in *Arabidopsis thaliana***
O. Cassan, C.-H. Lecellier, A. Martin, L. Bréhélin & S. Lèbre (2024)

# Perspective

- **Null hypothesis testing**
  - Improve simulation scheme (Highly correlated data)
  - Consider statistical testing (for the linear model)

- **Refine prior GRN definition**
  - Include ATAC-seq data to focus on active promoter.
  - Include grammar binding motif




(Zrimec et al. 2022)

  - Use other sources of prior (compendium databases)
    $\rightsquigarrow$ Work in progress

# The team

**Océane Cassan (PhD 2022)** **Antoine Martin (IPSIM)**


**Machine Learning for Regulatory Genomics group**

https://www.lirmm.fr/ml4reggen/