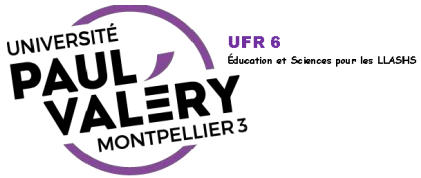
Statistical modeling and inference to identify DNA sequence elements involved in transcription regulation

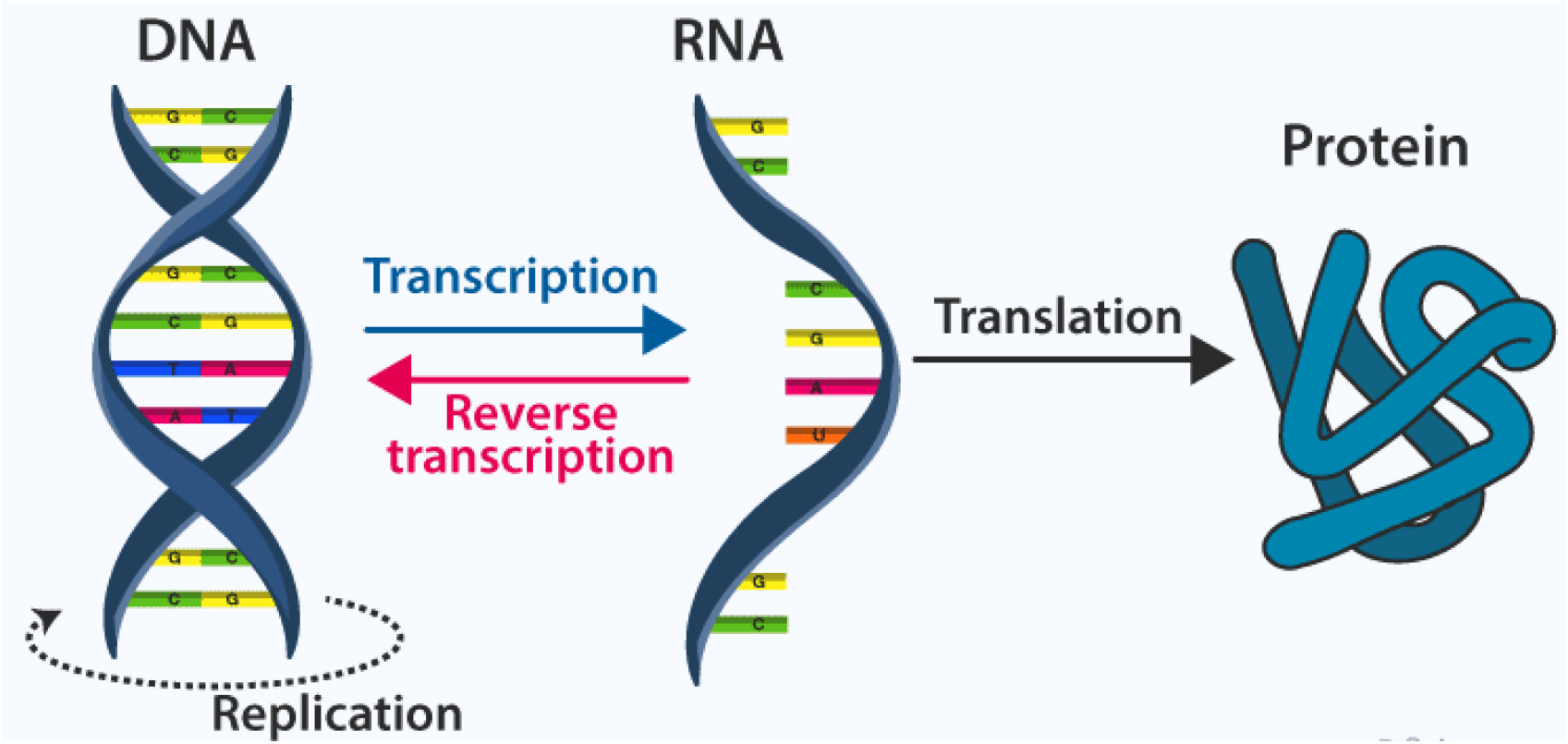
Laurent Bréhélin, Sophie Lèbre, Charles Lecellier



Journée Statistique et Sciences de la Santé
Lille, 27 Juin 2022

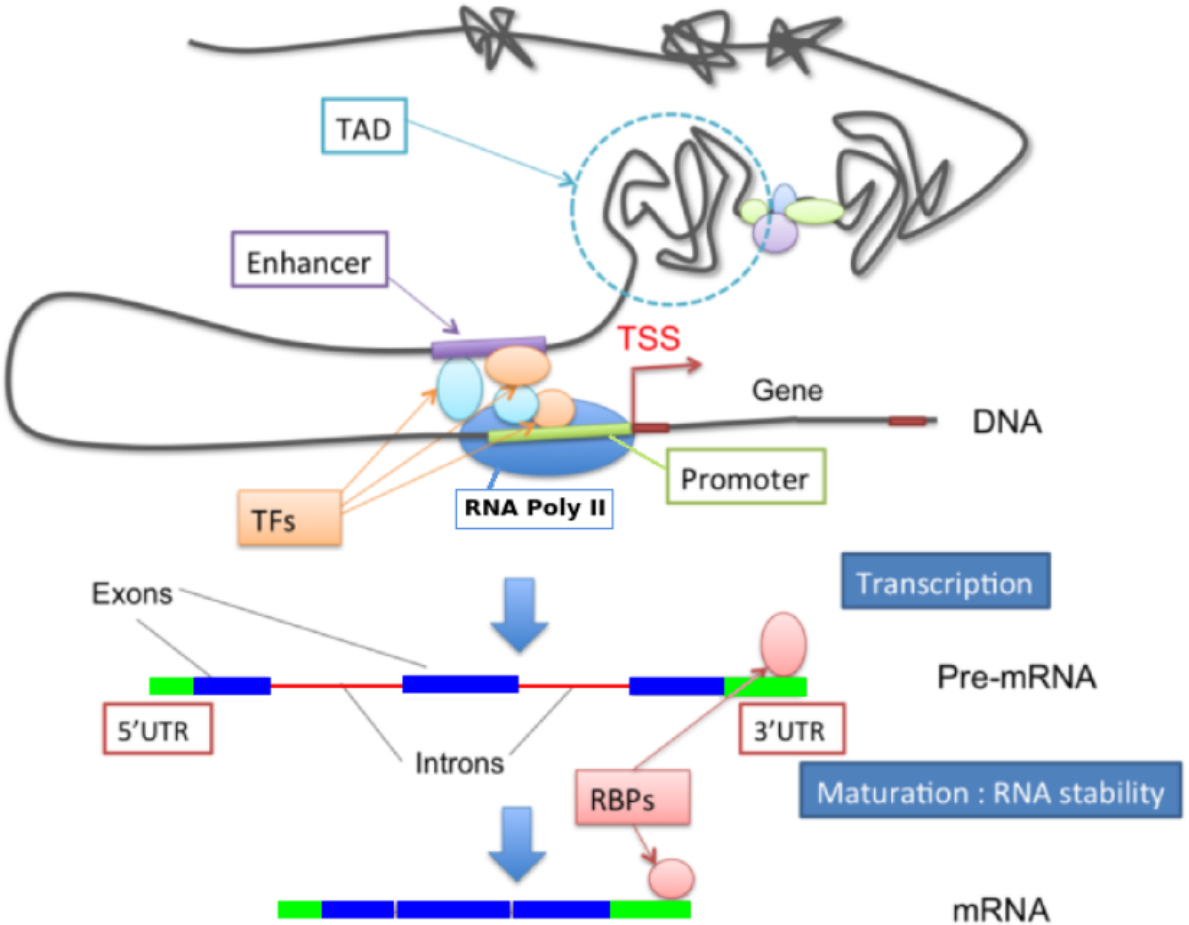
Outline

Central Dogma of molecular biology



First proposed in 1958 by Francis Crick, discoverer of the structure of DNA.

Gene expression regulation machinery



TFs = Transcription factors
RBPs = RNA Binding Proteins

Transcriptional regulations

Post-transcriptional regulations

Health & environment

- Eric Soler (IGMM, CNRS, Montpellier)

  Human, Dynamique Chromatinienne et Hématopoïèse

  => Dynamics of transcription factors (GATA1 and Lmo2) during cell differentiation.

- Diego Tosi (ICM, Montpellier)

  Human cancer, Early Phase Testing

  => Regulators of the over/under expressed genes (TEAD, ...)

- Juan-Jose Lopez-Rubio (LPHI, INSERM, Montpellier)

  Plasmodium falciparum (malaria parasite) nuclear biology

  => Molecular mechanisms underlying gene expression

- Antoine Martin (IPSiM, INRAE, Montpellier)

  Arabidopsis thaliana

  => Regulators involved in plant mineral nutrition adaptation in response to climate change (drought, CO_2, nitrate supply)

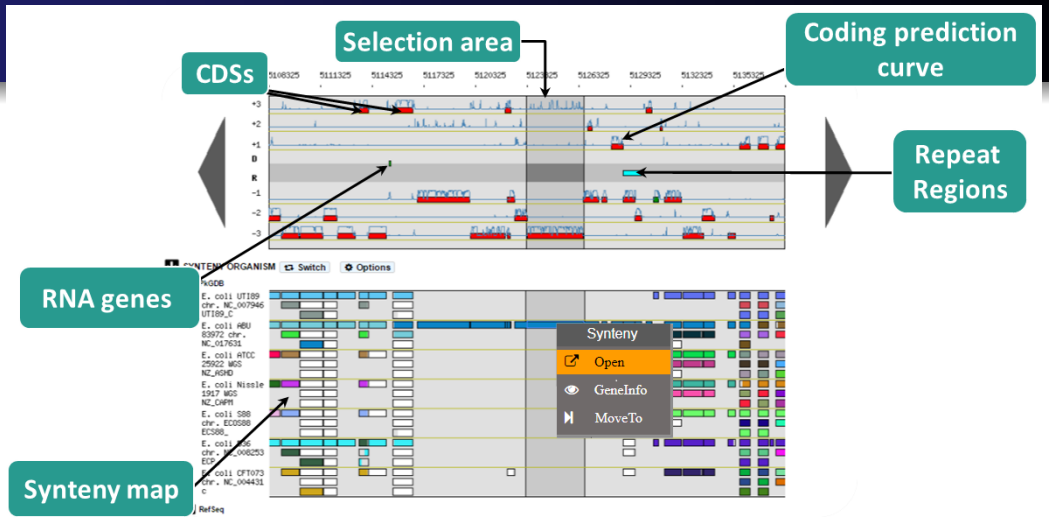
Decoding the architecture of regulatory interactions

'To understand biology at the **system level**, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism.'

'However, many breakthroughs in **experimental devices, advanced software, and analytical methods** are required before the achievements of systems biology can live up to their much-touted potential.'

(Kitano 2002, Science)

Data



- Genomic data
  - Complete **DNA sequence**
    $\rightsquigarrow$ Human : Genome Reference hg38
  - **Annotations** : gene location on the DNA sequence

- **Open databases**:
  - **Fantom** project,
  - Encyclopedia of DNA Elements (**ENCODE**) -> GENCODE for protein coding genes,
  - **Ensembl** genome browser, ...

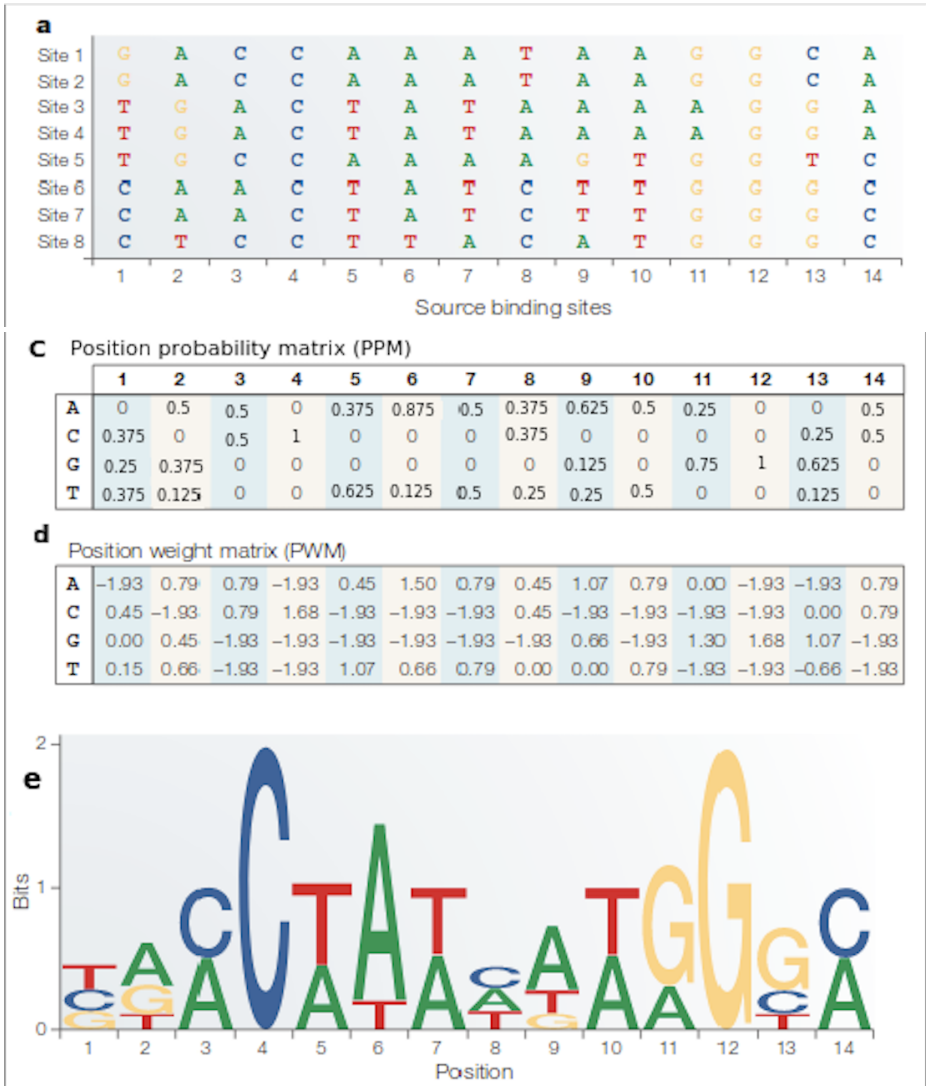
=> Project : identify all **functional elements** in most species genome.

Data

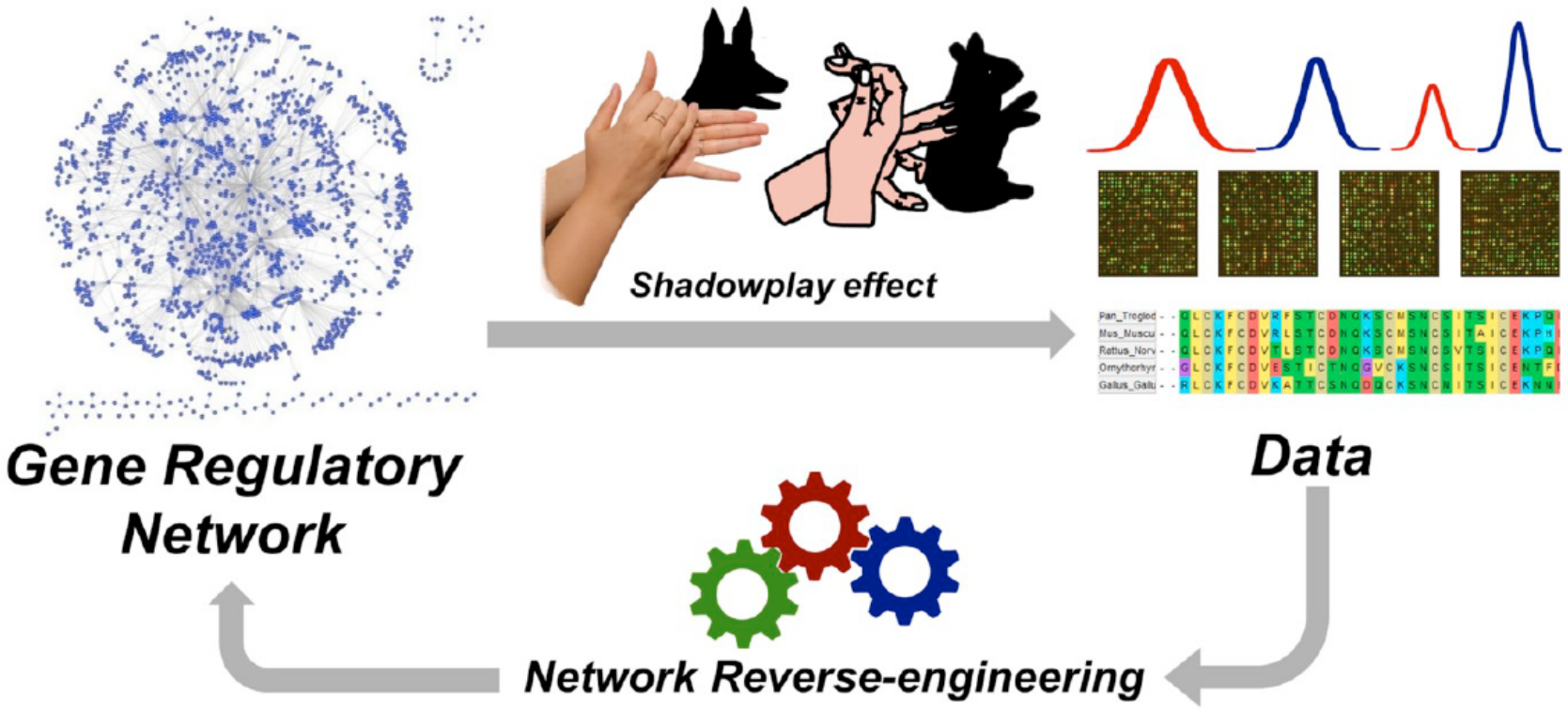
- **DNA transcription measurement : RNA-seq**
  Count data associated with mRNA, micro RNA, long ncRNA
  Human : about 20 000 coding genes (mRNA)
- **Transcription Factor (TF) binding : ChIP-seq**
  Count data associated with sequence location (peak)
  Human : about 1500-2000 known TFs
  $\rightsquigarrow$ **Open databases : ReMap, GTRD, Unibind...**
- Other epigenetics data (Experimental)
  - Chromatin accessibility (ATAC-seq, DNAse-seq, ...)
  - DNA methylation, histone modification,
  - 3D genome sequencing (3C, 4C, 5C, ChIA-PET, Hi-C, ...),
  - Genetic polymorphism (copy number variation (CNV), microsatellite)
  - ...
- **Open databases : TCGA, GEO, GTEx, ENCODE, ...**

Motifs for TF binding sites

- From ChIP-seq data
- Simple statistical models known as Position Weight Matrices (PWMs) used to compute binding affinities and to identify potential binding sites in genomes
- Open databases : JASPAR, HOCOMOCO, ...
- Software : HOMER, the MEME suite (MEME, FIMO,) ...



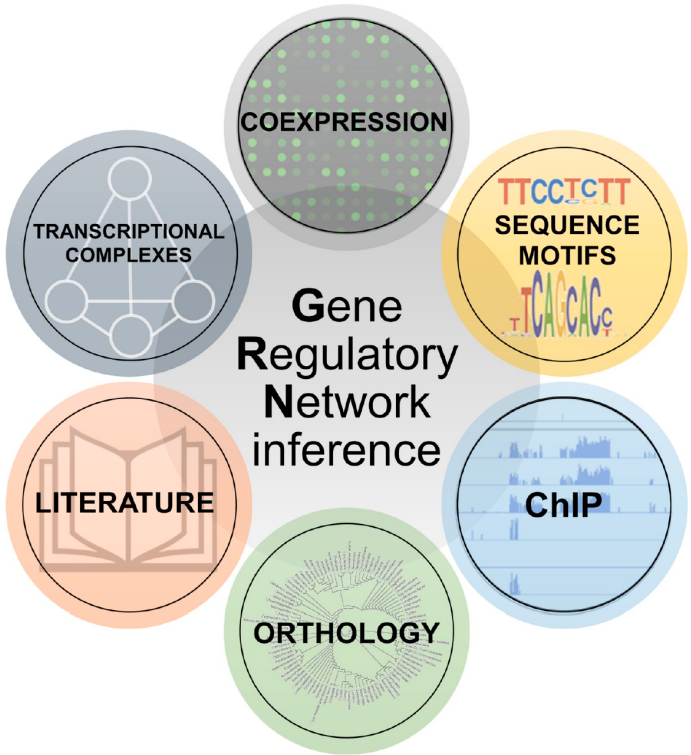
Learning Gene network (Starting from about 2005)



(Source : Mercatelli et al. 2020, BBA - Gene Regulatory Mechanisms)

Gene network inference

- **Various models/inference:** Gaussian Graphical Model (GGM) + lasso
  Poisson Log normal model for count data (Chiquet et al. , 2019),
  Random forests (Genie3, 2010), LARS Regression (TIGRESS, 2012), ...
- **Curse of dimensionality**
  $p \times p$ putative edges
  n measurements
  (in most cases : $n << p$)
- **Data integration**
  See Mercatelli et al. 2020,
  BBA - Gene Regulatory Mechanisms,
  for a review (+ Figure source).

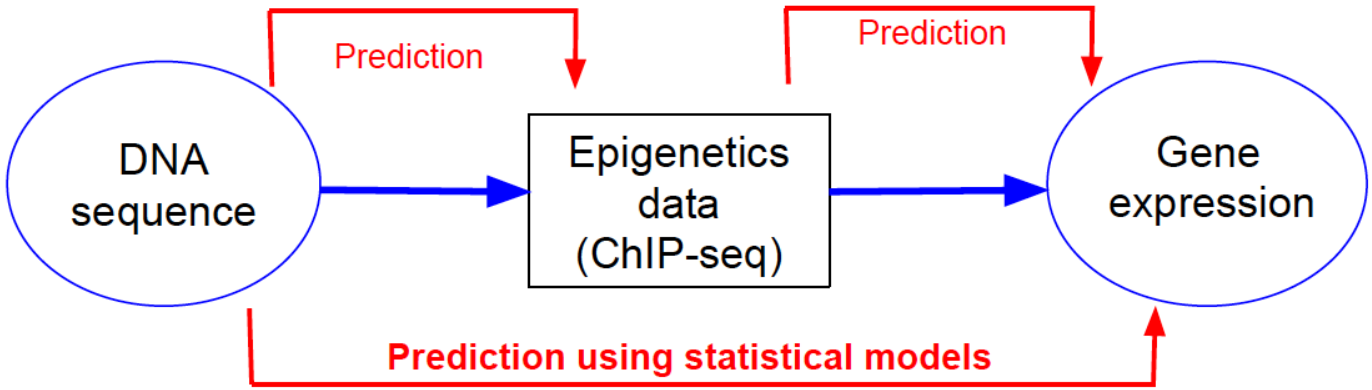


- **Still an open question:** 'Decoding the architecture of regulatory interactions has become one of the main tasks of modern biology' (Mercatelli et al. 2020)

Outline

1. Learning the regulatory code of gene expression (DNA sequence, gene annotations, binding site motifs (PWM))
2. **Modeling DNA transcription from DNA sequence only**
   1. **Building (and validating) the model**
3. Modeling DNA binding
4. Future directions

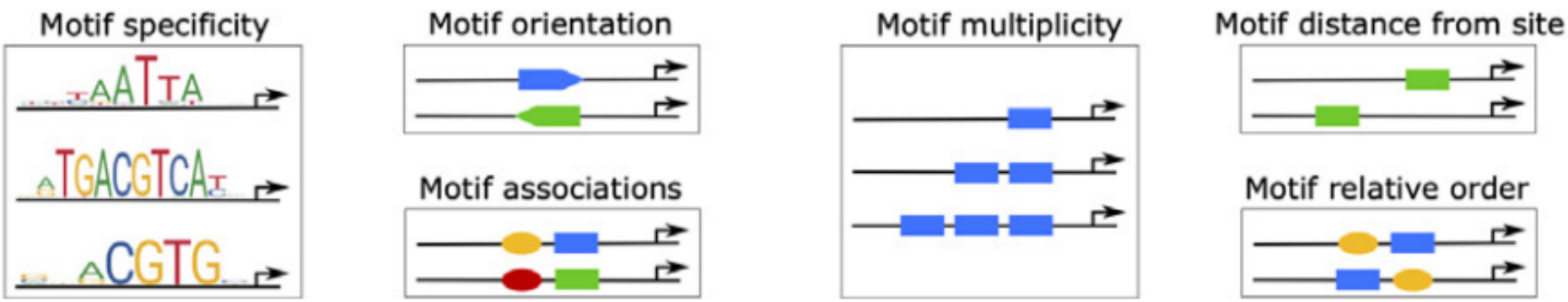
Learning the Regulatory Code from DNA sequence only?



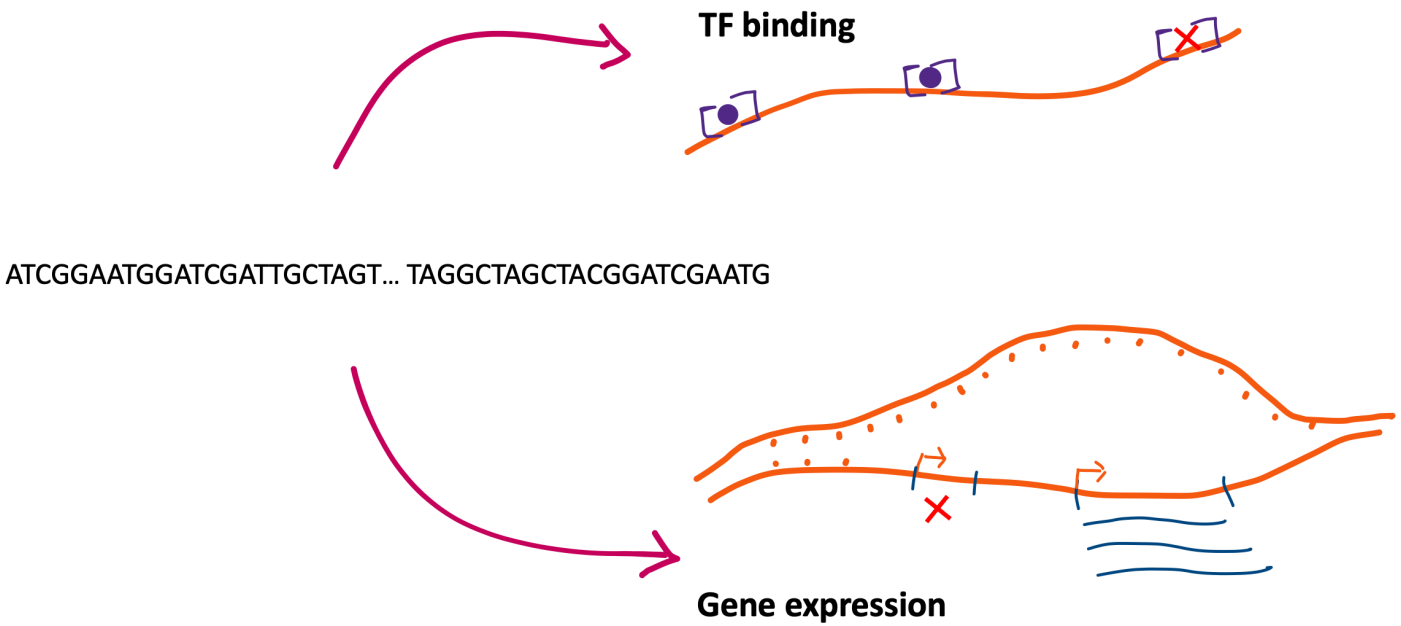
- Predicting gene expression from epigenetics data (ChIP-seq, methylation, ...)
  - RACER : Y. Li and al. PLoS (2014)
  - TEPIC : Schmidt F. et al. Nucleic Acids Res (2017)
- Predicting Epigenetics data from DNA sequence
  - Whitaker, J. W. et al. Nat. Methods (2015)
  - Zhou, J. et al Nat. Methods (2015)

=> Can we straightly identify the DNA determinants involved in gene regulation?

Deciphering TF binding motifs grammar

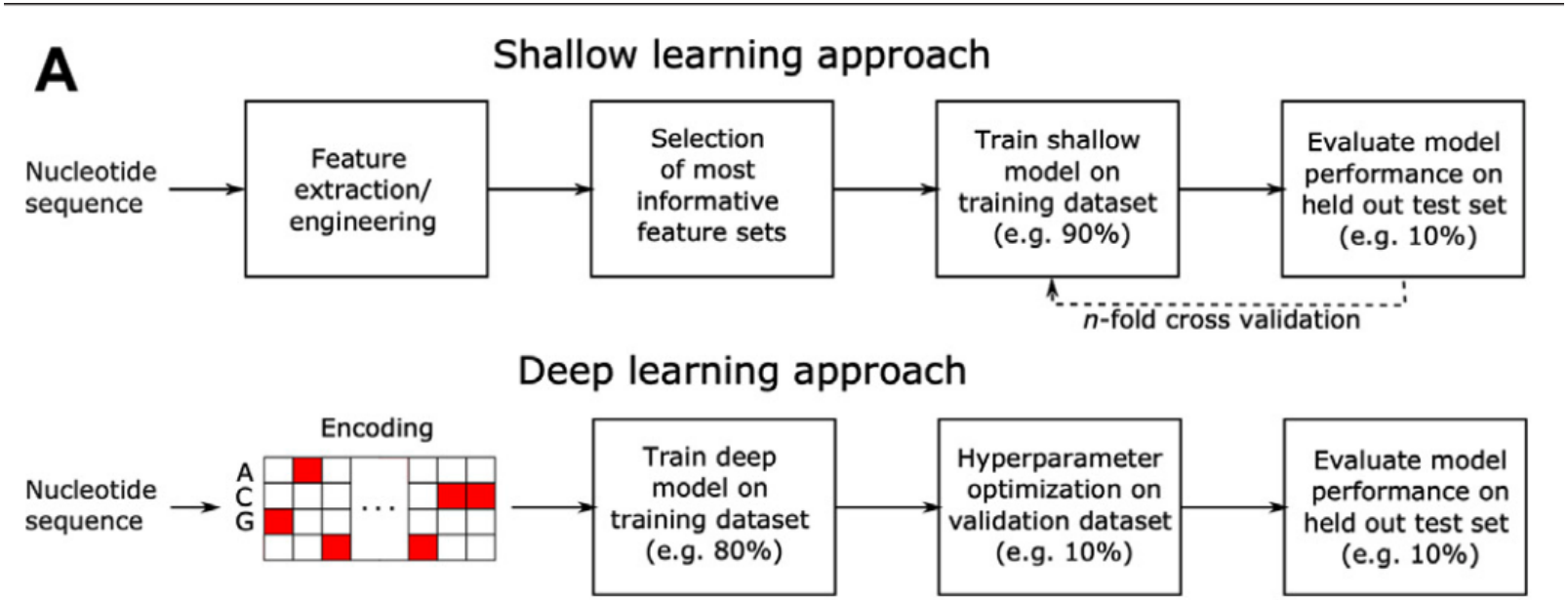


(Zrimec et al. , Frontiers in molecular Biosciences, 2022)



Learning the Regulatory Code from DNA sequence only

- Starting from 2015...
- Shallow learning (Need for Engineered Features) or Deep learning



See Zrimec et al. (Frontiers in molecular Biosciences, 2022) for a review (+ Figure source)

Learning the Regulatory Code from DNA sequence only

2 PhD thesis (2018): May Taha (Applied stat) & Chloé Bessière (Bioinfo)

- Paradigm shift :

  one model for gene expression (whatever the gene is)

  Linear model

$$y_i = \beta_0 + \Sigma_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad (1)$$

  with

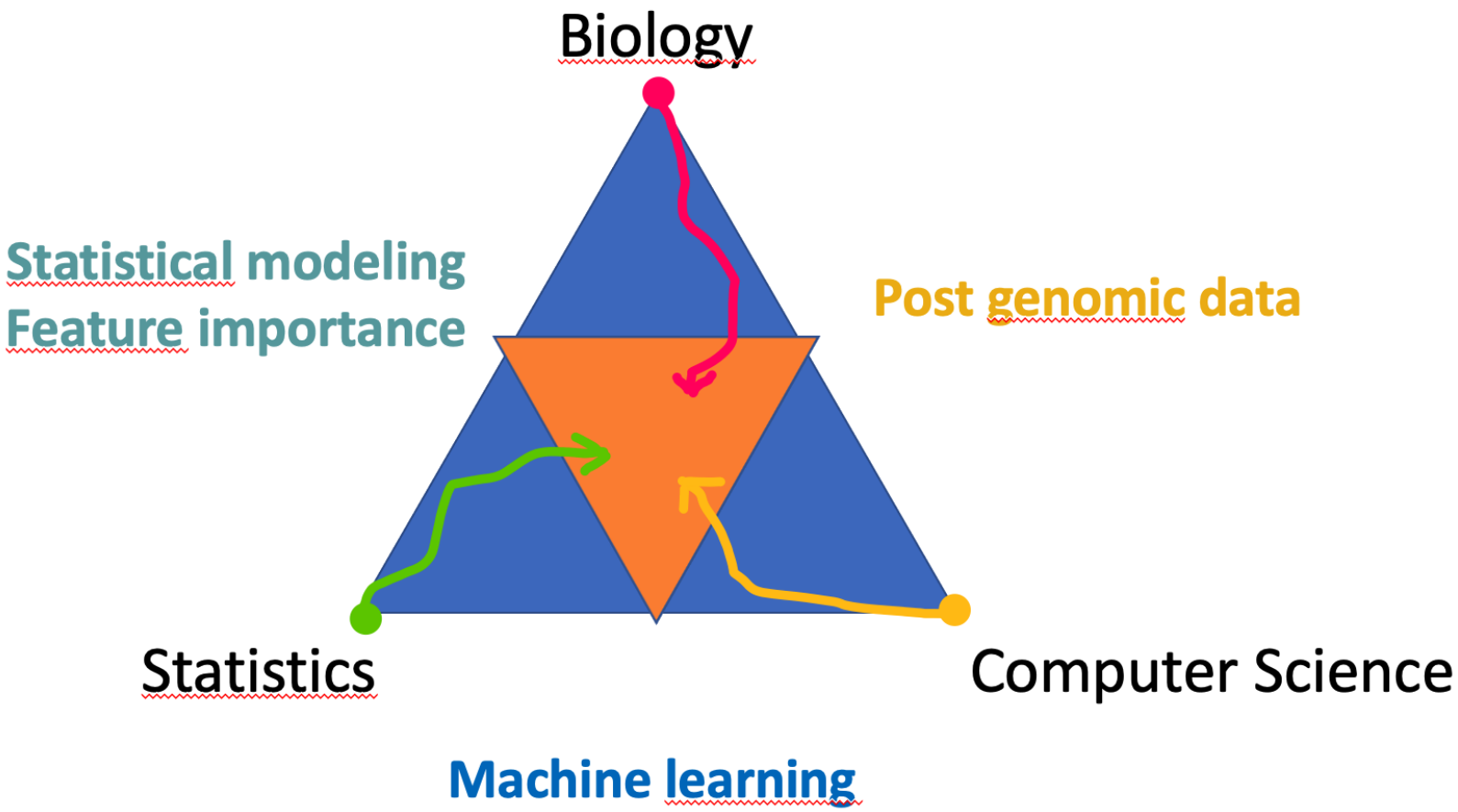
  $Y_{[n\times 1]} = (y_1, \ldots, y_n)'$ vector of observed genes expression level,
  $X_{[n\times p]} = (x_{ij})$ DNA features matrix (x_{ij} is feature j for gene i),
  $\beta_{[p\times 1]} = (\beta_0, \beta_1, \ldots, \beta_p)'$ vector of regression coefficients
  $\varepsilon_{[n\times 1]} = (\varepsilon_1, \ldots, \varepsilon_n)'$ vector of the residual errors.

- n measurements = number of genes ($n \approx 20000$ for humans)

  $=>$ *much more confortable for statistical learning !*

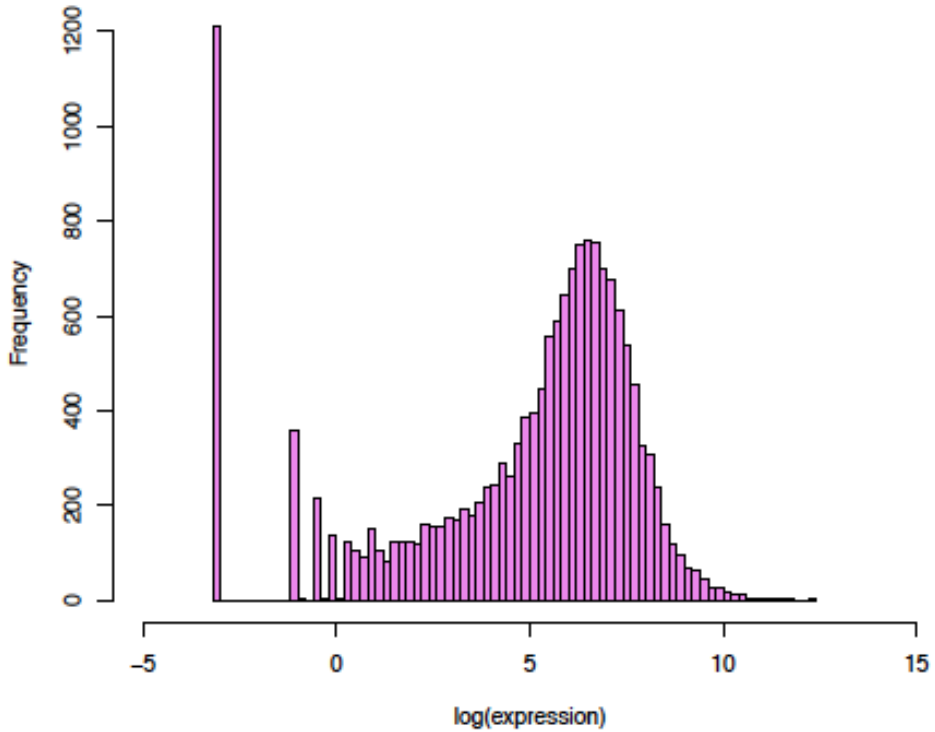
- p explanatory variables (e.g. nucleotide composition, k-mers, DNA motifs, PWMs, chromatin configuration, DNA structural variables)

Pluri-disciplinary research



Response variable

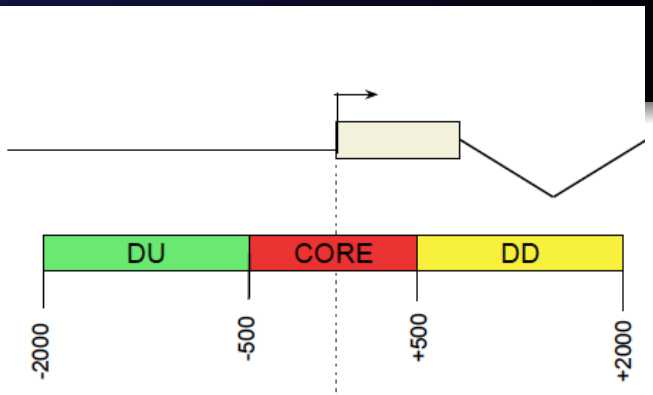
RNA-seq (log transformed values) from the TCGA database



- Gene expression measured by RNA-seq (reads count)
- 12 different types of cancer from **TCGA**: Breast, Leukemia, Liver...

$n \approx 20000$ genes

Feature engineering



from DNA sequence

- TF binding motifs : PWM scores 

  471 motifs from JASPAR CORE 2016 database
  => Maximal score in the CORE promoter region for each gene i

- Nucleotide (A, C, G, T) and di-nucleotide (AA, AC, ... TT)
  => Relative frequencies in 3 promoter regions (CORE, Distal Upstream, Distal Downstream)

$$\text{e.g. } \%CG = \frac{\#CG}{length - 1}$$

$p \approx 500$ DNA features

All sequences were mapped to the hg38 human genome reference.

Model inference with the Lasso

- Linear regression with ℓ_1-norm penalty or Lasso (Tibshirani, 1996) applied to standardized data:

$$\hat{\beta}_{LASSO} = \underset{\beta}{\text{argmin}} \left(\sum_{i=0}^{n} (Y - X\beta)^2 + \lambda \sum_{i=0}^{p} |\beta| \right)$$

  with n genes, p DNA features.

- The penalty λ is chosen by 10-fold cross-validation to minimize the mean square prediction error.

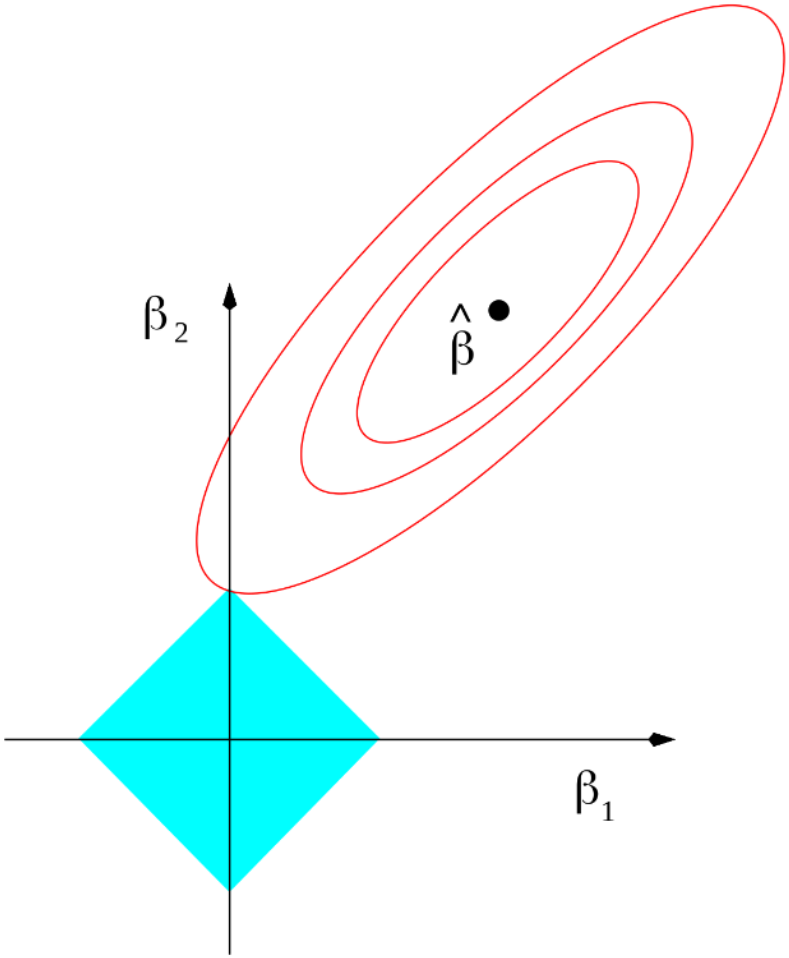
- Some coefficients β_i are set to 0 exactly (ℓ_1-norm geometry).

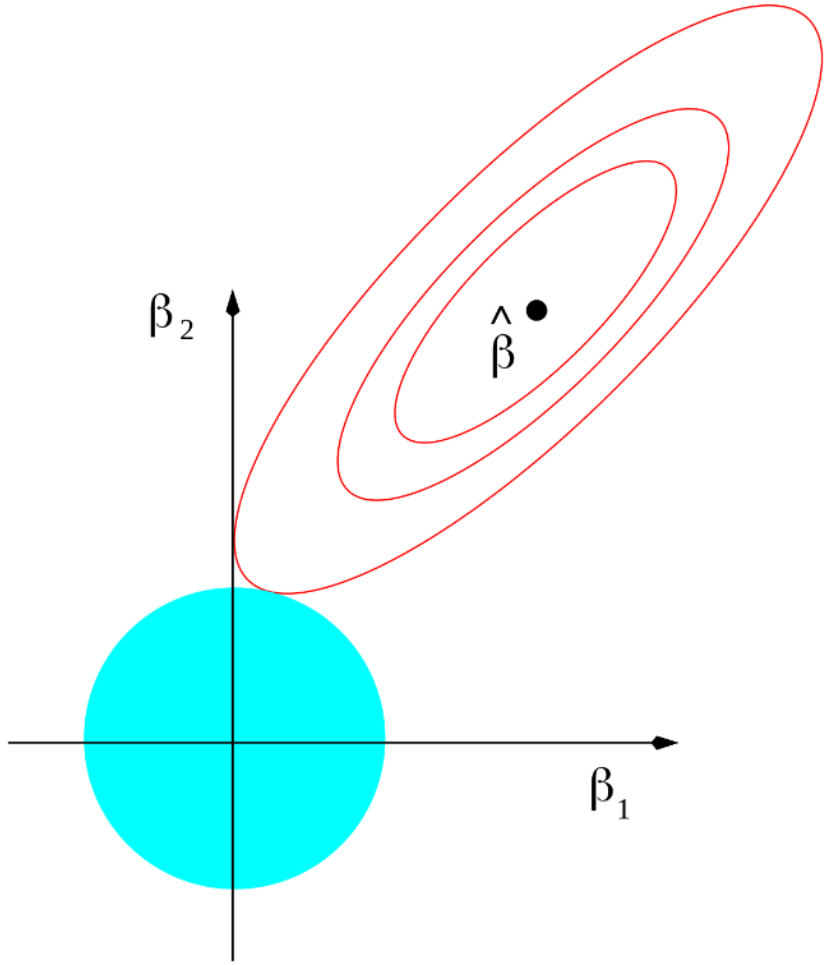
- The larger λ, the fewer variables are selected.

R package `glmnet` by (Hastie, Qian, Tay), option `family = 'gaussian'` for linear models.

Geometry of the lasso

Least Absolute Shrinkage and Selection Operator (LASSO) : L1 norm





Ridge : L2 norm

Model evaluation

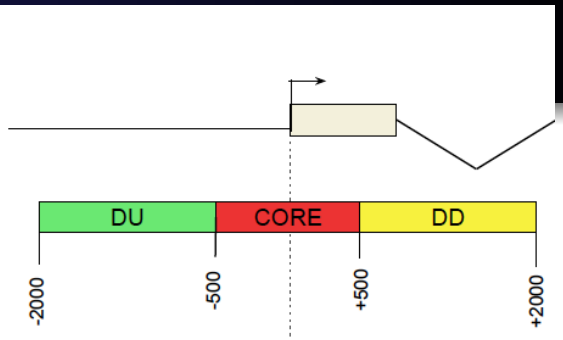
- Criterion :
  1. Mean square error (MSE)
  2. Correlation coefficient $Corr(Y, \hat{Y})$ between the measured expression Y and the predicted expression $\hat{Y}$.

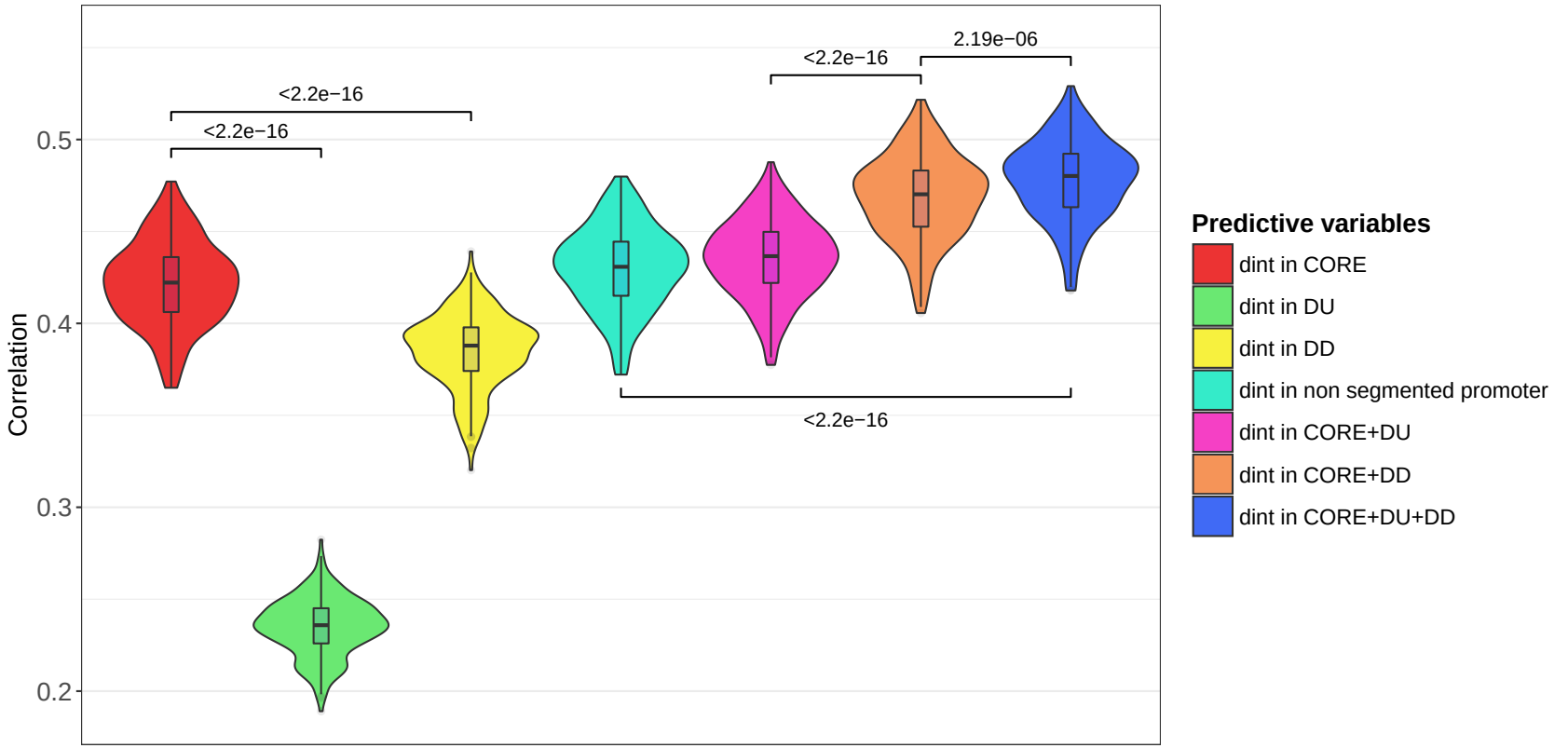
  in a 10-fold cross-validation procedure: Model is
  1. learned in the training data
  2. evaluated in the test data.

- Results shown : RNA-seq gene expression from the TCGA database **(12 cancers types, 20 patients per cancer)**.

  (Further evaluation not shown: 1,270 RNA-seq samples and 582 microarrays datasets.)

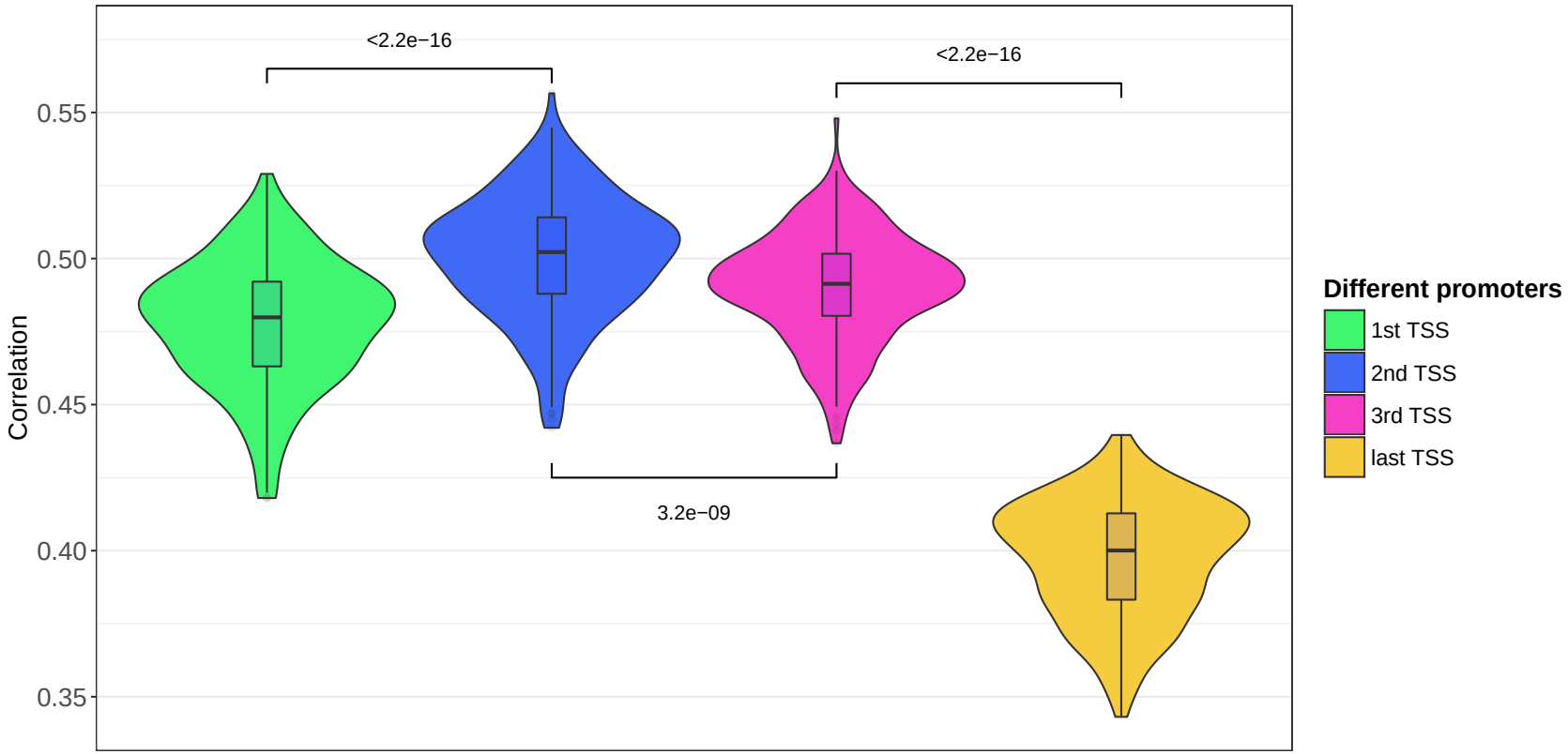
Definition of the promoter region





=> The highest accuracy is obtained **combining the 3 segments**.

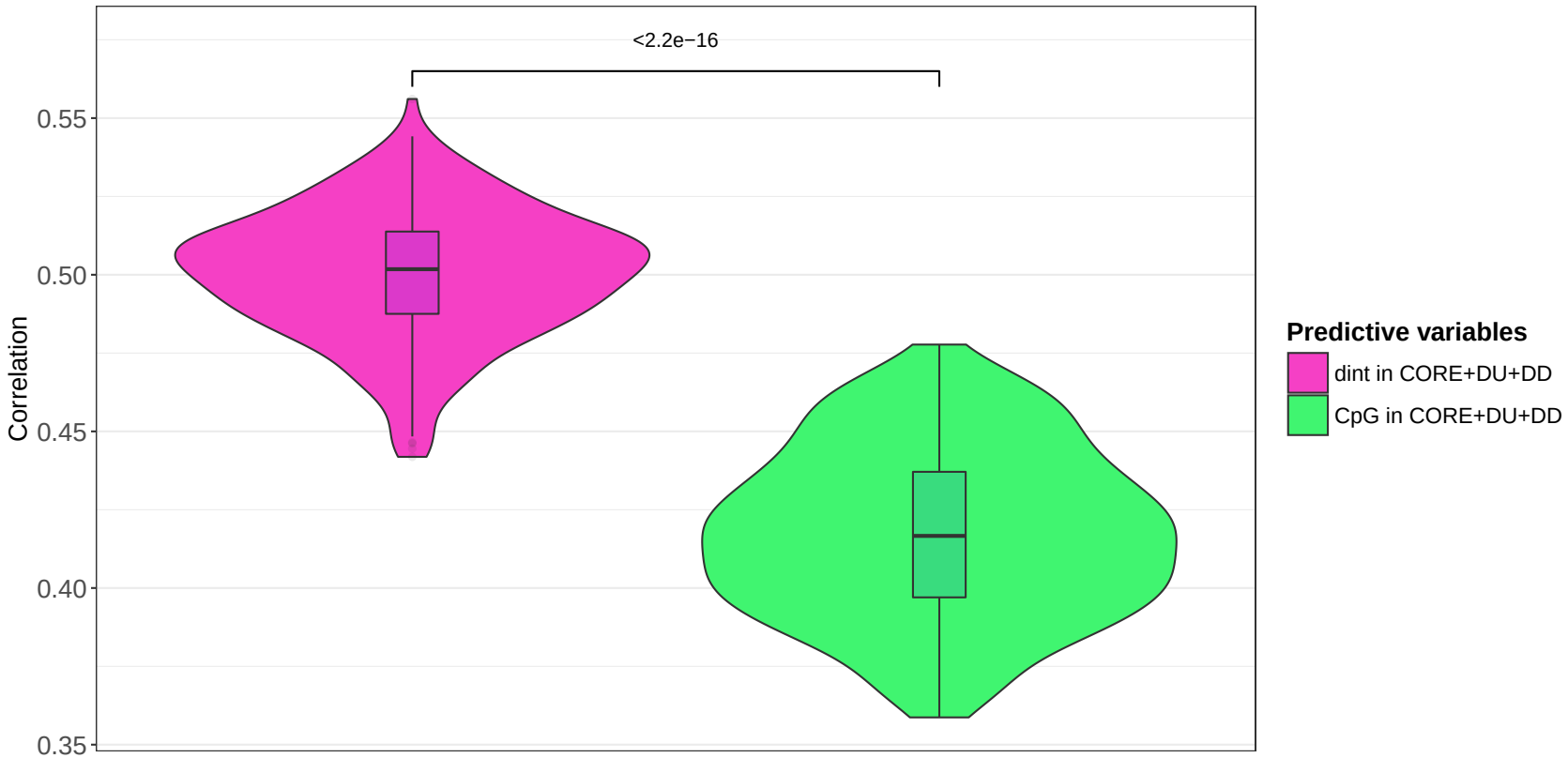
Choice of the Transcription Starting Site (TSS)



=> The highest accuracy is obtained with the promoter region centered around the 2^{nd} TSS (blue) [in agreement with Cheng et al. (2012)].

Gene TSS positions were extracted from GENCODEv24.

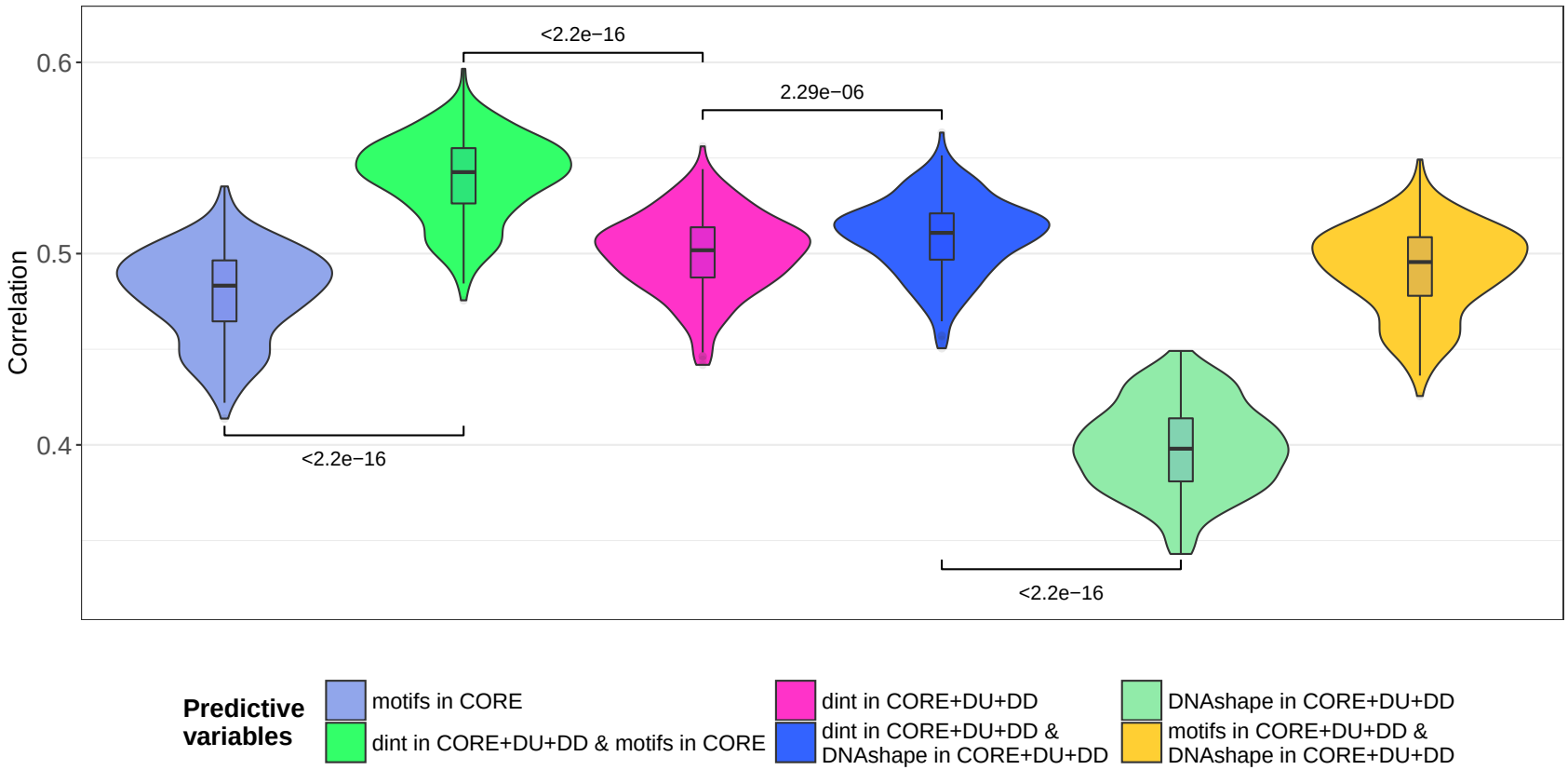
All di-nucleotides vs CpG only



=> Considering all (di-)nucleotides leads to higher accuracy than CG only

=> Indicating that dinucleotides other than CG contribute to gene expression regulation.

TF binding motifs and local DNA shapes

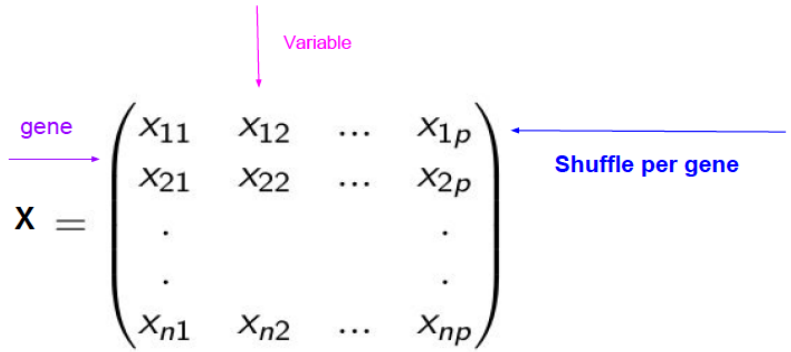


- TF motifs: Position Weight Matrix (PWM) from the JASPAR database
- DNA shapes: computed with the Bioconductor package `DNAshapeR`.

=> Small improvement with TF motifs (green) vs. dinucleotides only (pink), even less improvement with DNA shapes (dark blue).

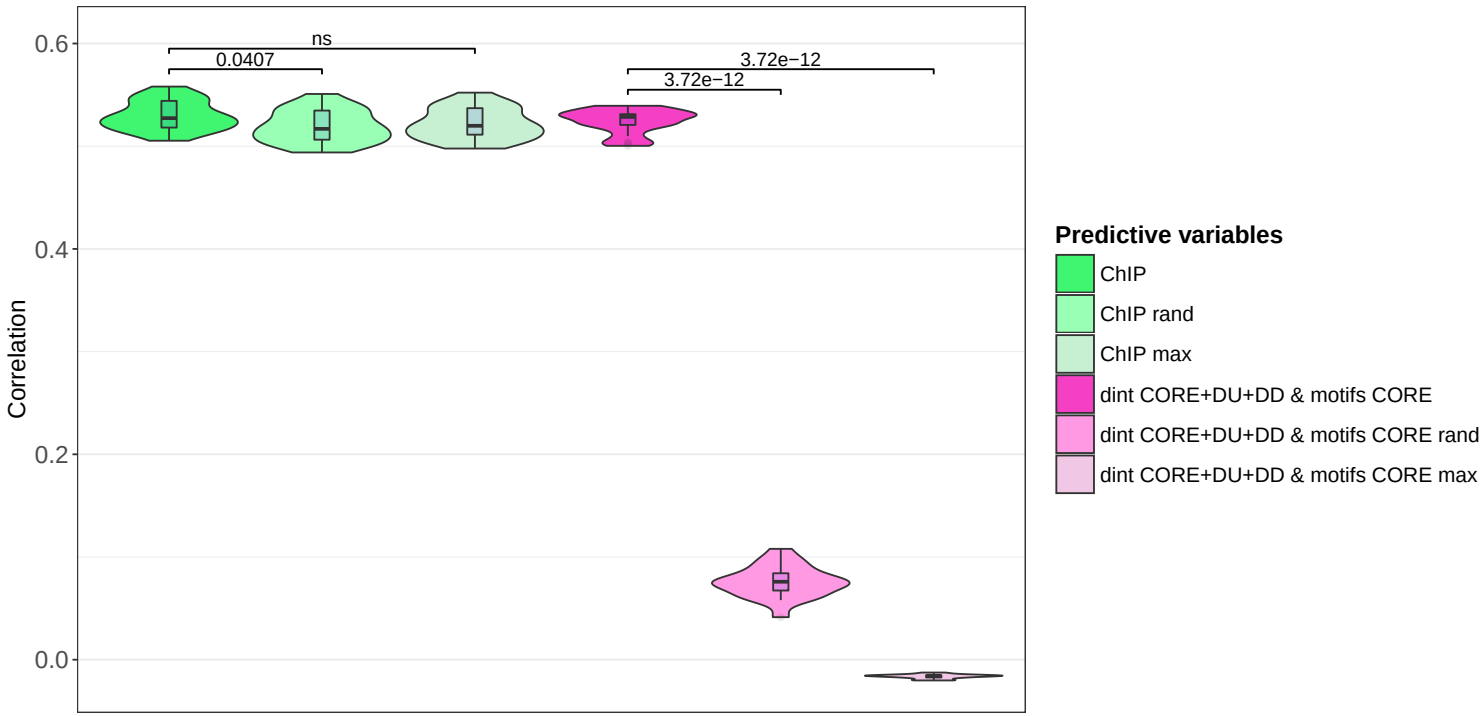
Selected variables relevance

- Comparison with models integrating epigenetics features (ChIP-seq):
  - TF binding signal with ChIP-seq (RACER, Y. Li and al. PLoS, 2014)
  - Open-chromatin signal (TEPIC, Schmidt F. et al. NAR, 2017)
- Feature randomization to measure selected variables relevance.
  Using DNA or epigentics data, models built using either:
  (i) the original features
  (ii) **randomized features** => horizontal shuffling (response centered)
  (iii) only one feature: the **maximum value** of all predictive variables



$\star\star\star$ For (ii) and (iii), the links between each predictive variables (column) and the response variable (gene expression) is broken : the regression model is expected to perform poorly. $\star\star\star$

Comparison with model integrating TF binding signals



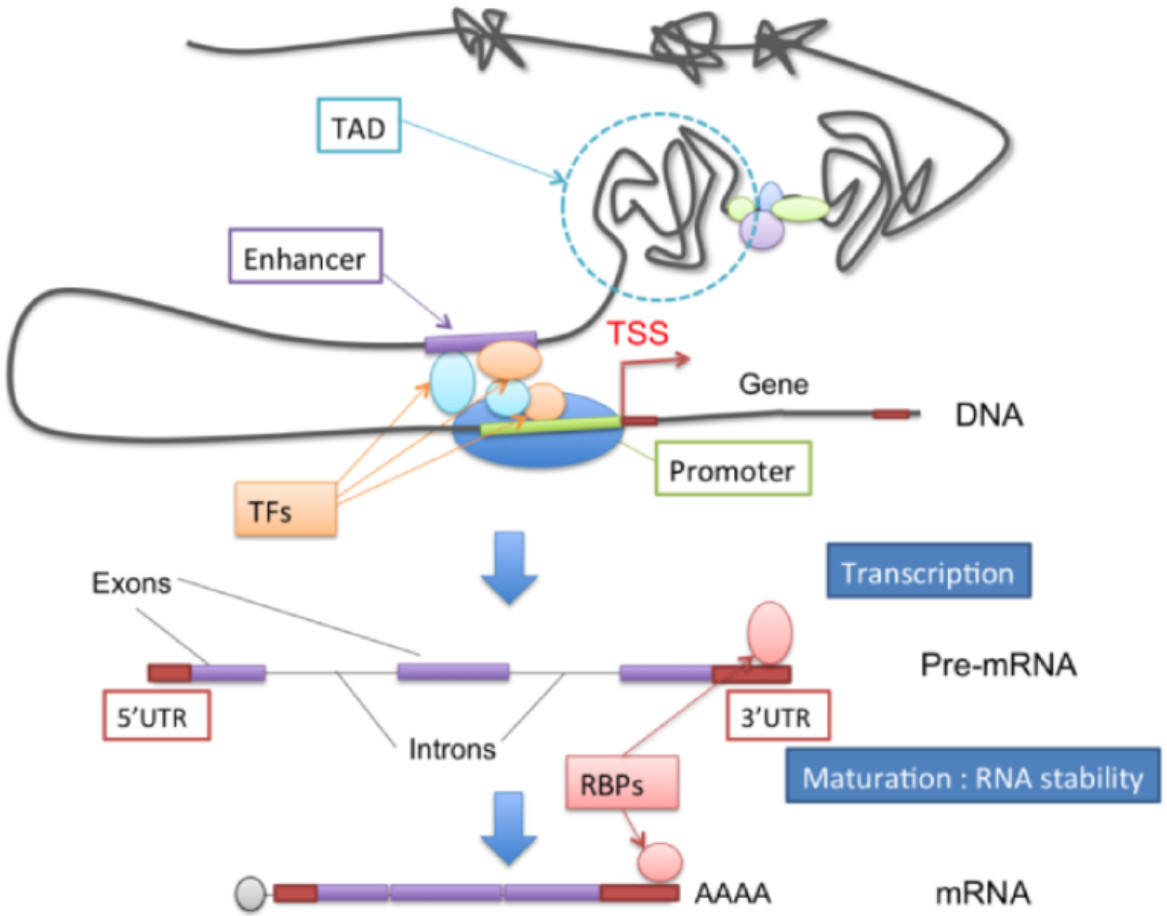
(i) original features, (ii) gene centered randomization, (iii) only one feature (maximum predictive variable).

⋆⋆⋆ **For (ii) and (iii), the links between the predictive variables and expression is broken and a regression model is expected to perform poorly, as our DNA model does (pink).** ⋆⋆⋆

Comparison with model integrating open-chromatin signals

⋆⋆⋆ **For (ii) and (iii), the links between the predictive variables and expression is broken and a regression model is expected to perform poorly, as our DNA model does (pink).** ⋆⋆⋆

Back to the gene expression regulation mechanism

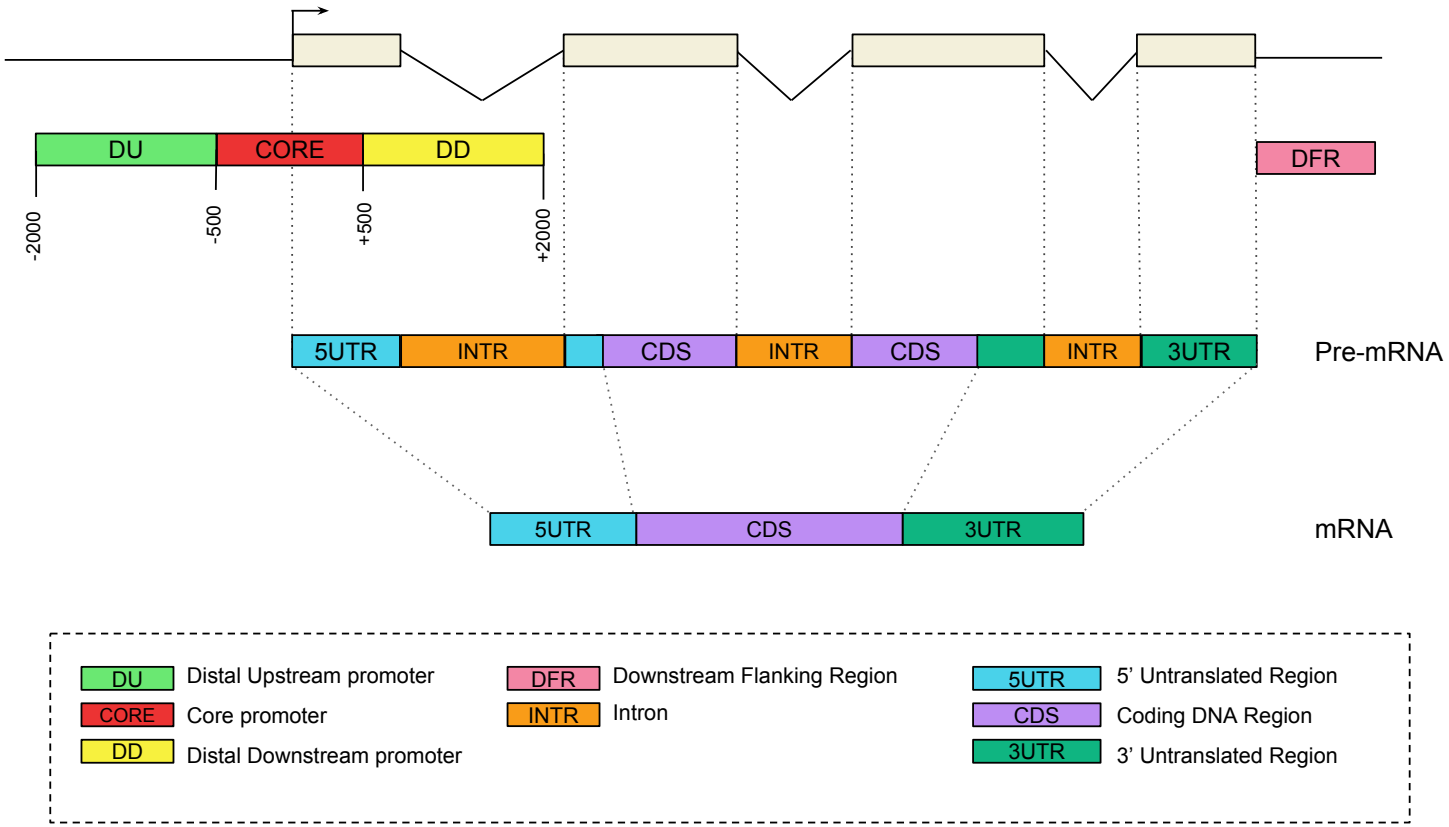


TFs = Transcription factors
RBPs = RNA Binding Proteins

Transcriptional regulations

Post-transcriptional regulations

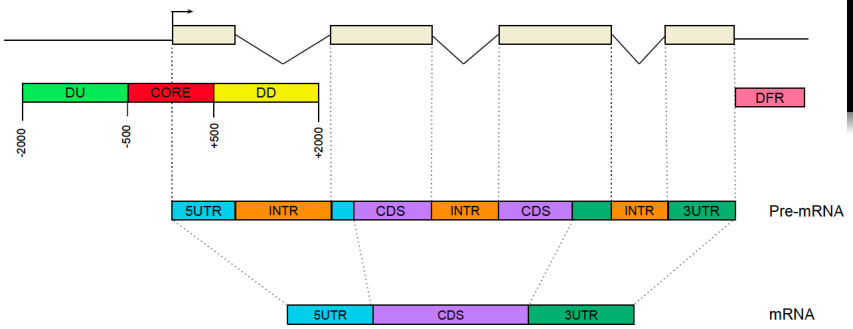
Additional genomic regions

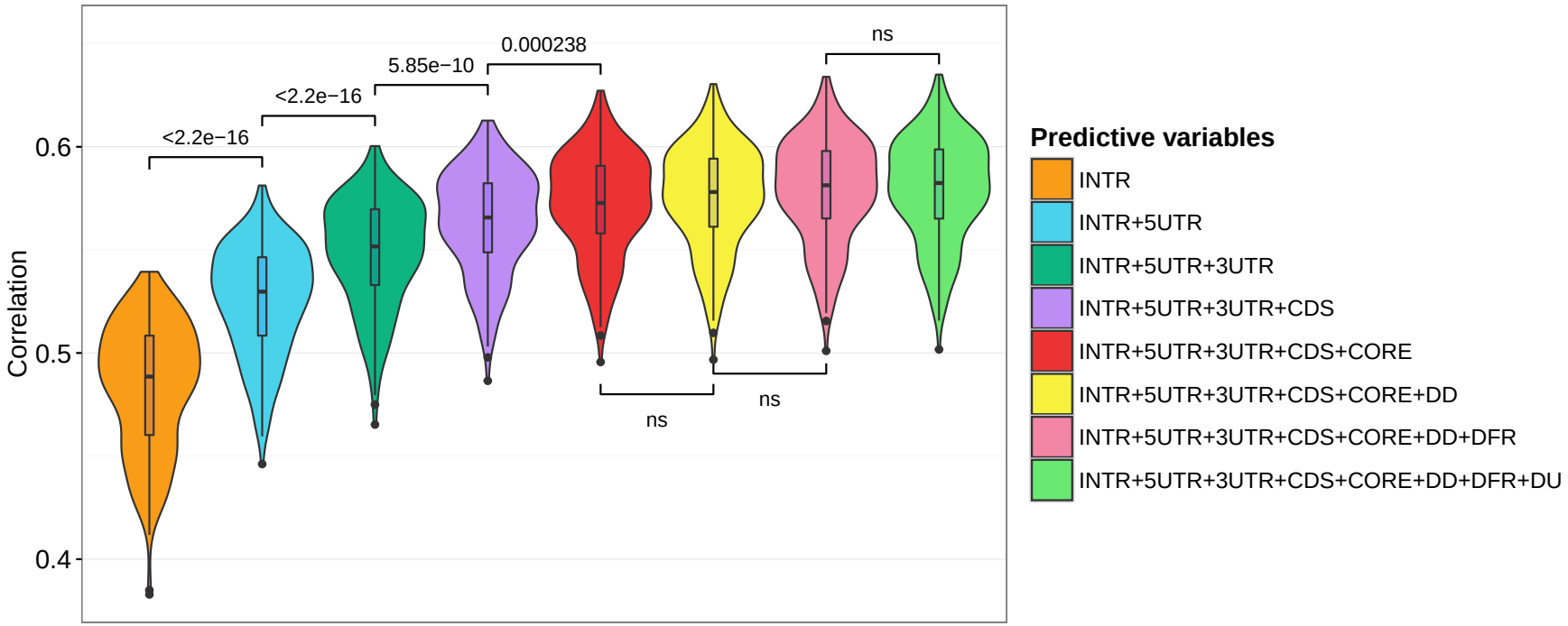


=> Features : Nucleotide and di-nucleotide compositions in 8 carefully selected DNA regions (20 variables per region)

UTR and CDS coordinates were extracted from `ENSEMBL Biomart`.
To assign only one 5UTR (resp. 3UTR) sequence to one gene, all annotated 5UTRs associated with the gene of interest were merged using `Bedtools`.

Additional genomic regions

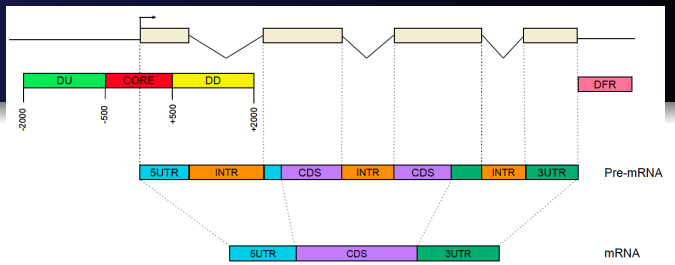




- Forward-like selection procedure among the 8 DNA regions

(Features: Nucleotide and di-nucleotide compositions in 8 carefully selected DNA regions, 20 variables per region)

Additional genomic regions



	INTR	5UTR	3UTR	CDS	CORE	DD	DFR	DU
STEP 1	0.4885	0.3771	0.358	0.2688	0.3996	0.3562	0.2369	0.2279
STEP 2		0.5298	0.5242	0.5069	0.5211	0.5037	0.4929	0.4887
STEP 3			0.5517	0.5488	0.5397	0.5391	0.5368	0.5306
STEP 4				0.5657	0.5587	0.5583	0.5575	0.553
STEP 5					0.5728	0.5718	0.5693	0.567
STEP 6						0.5781	0.5779	0.5733
STEP 7							0.5813	0.5786
STEP 8								0.5824

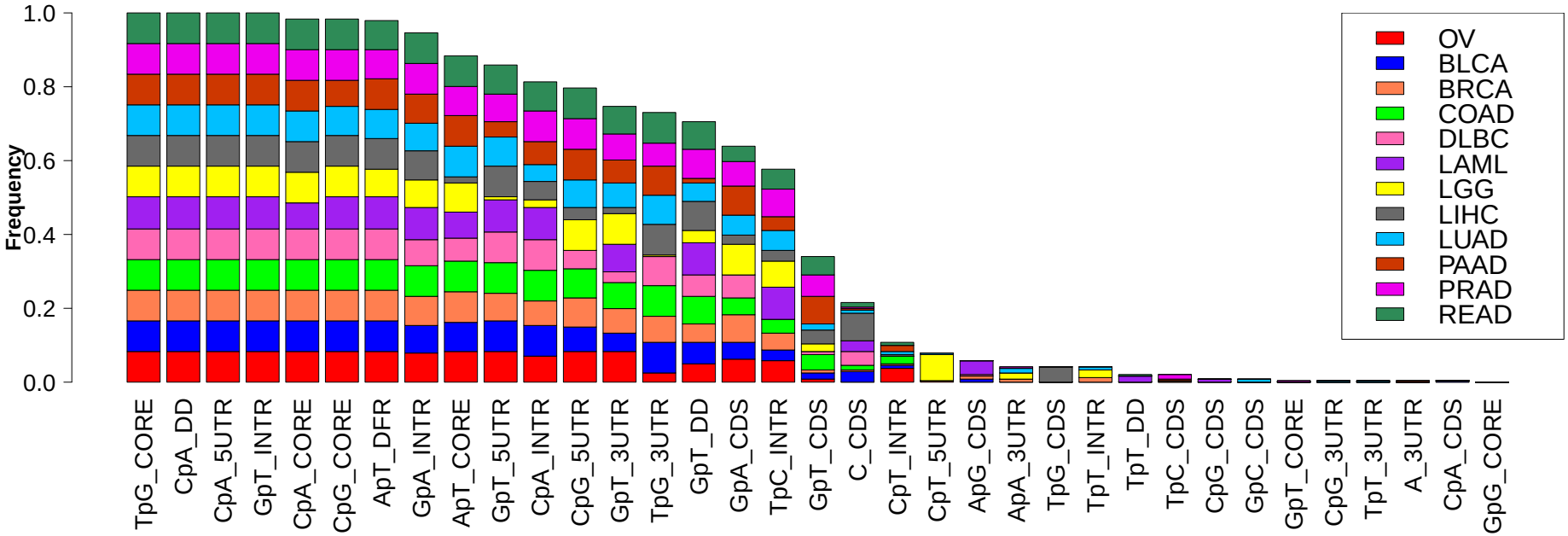
- Features : Nucleotide and di-nucleotide compositions in 8 carefully selected DNA regions (20 variables per region)
- Forward-like selection procedure among the 8 DNA regions

Stable variables selection

- Consistently selected variables were identified using **Stability selection** (Meinshausen *et al.*, 2010)
  - Lasso inference is repeated 500 times where, for each iteration,
    - (i) only 50% **of the individuals** (genes) is used (uniformly sampled)
    - (ii) a **random weight** (uniformly sampled in $[0.5; 1]$) is attributed to each predictive variable.
  - A variable is considered as **stable if selected in more than** 70% **of the iterations.**

(Functions `stabpath` and `stabsel` from the `R` package `C060` for `glmnet` models.)

Stable variables selection

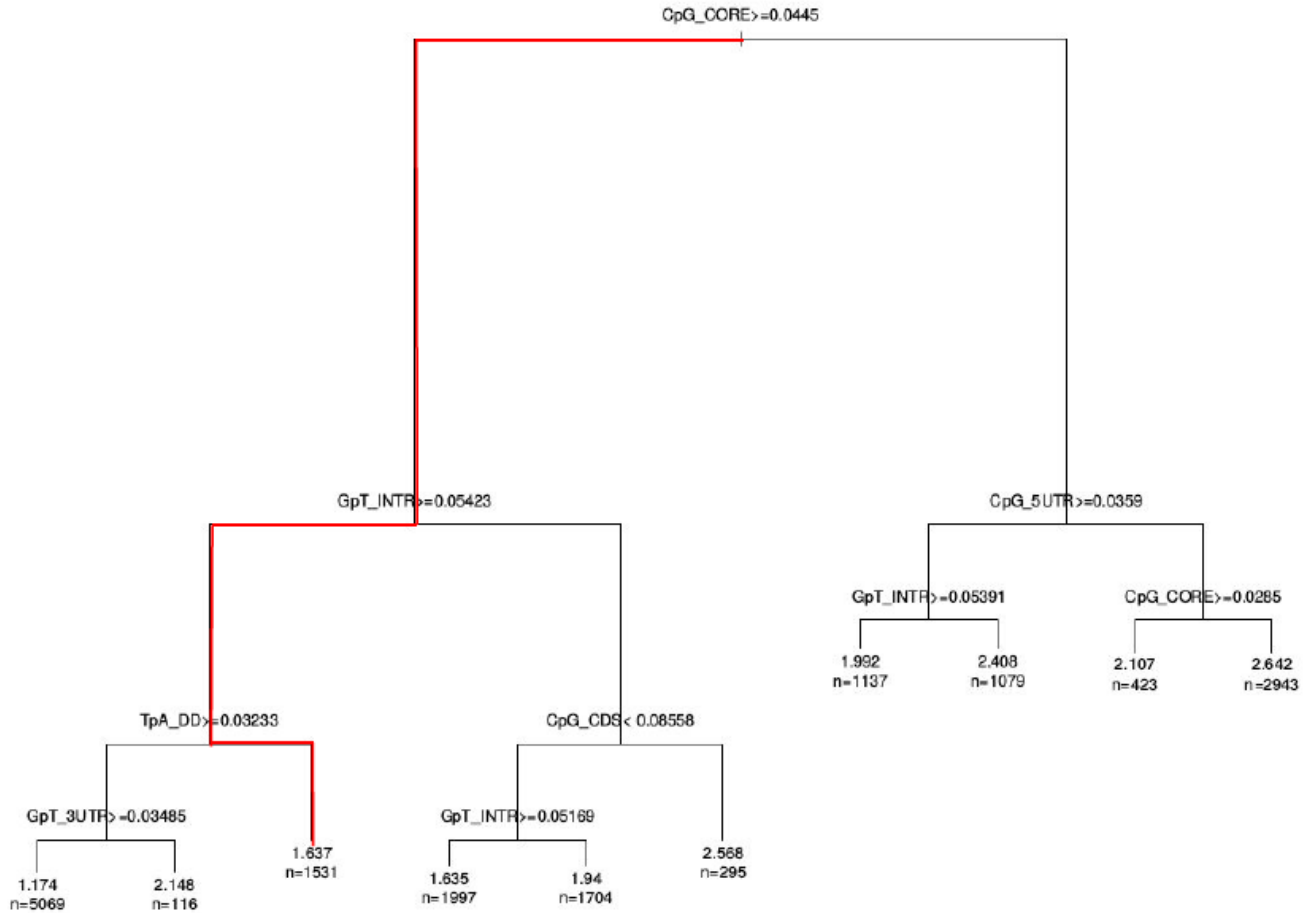


Proportion of samples in which each variable is selected with high consistency ($>$ 70% stability)

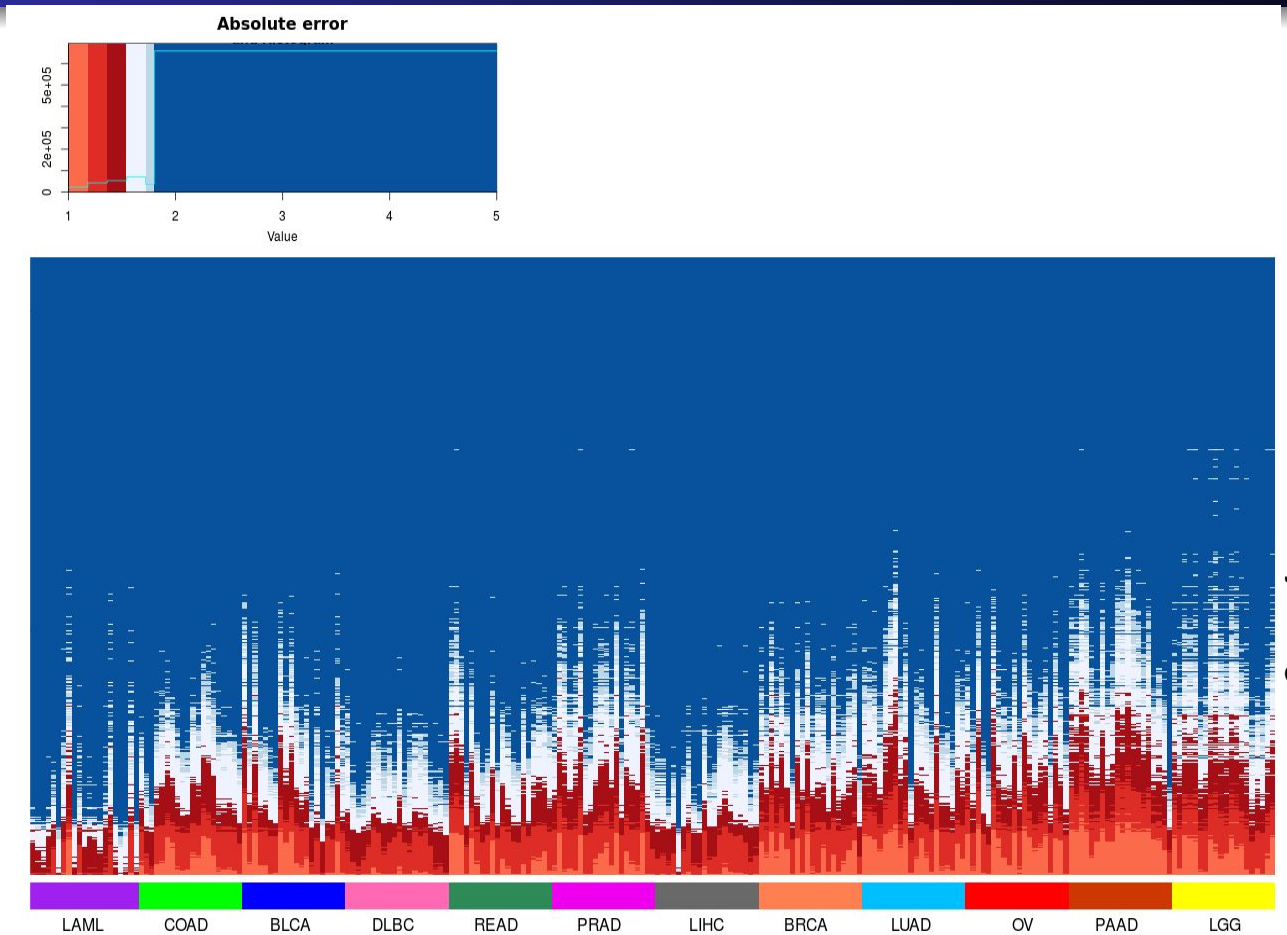
(Average $\sim$ 16 variables per sample)

DNA features associated with good predictions

- We characterized best predicted genes with regression trees (CART) which performs sequentially binary splits (minimizing RSS)
- Response variable is the prediction error of our linear model.
- (di-)nucleotide compositions are used as classifiers



DNA features associated with good predictions



- Columns : samples gathered by cancer type, ranked by decreasing error
- Lines : the 3,680 groups of genes ranked by decreasing error
- Red and light blue: Top 25% well predicted groups of genes

$\rightsquigarrow$ Our model mainly fits certain classes of genes with specific genomic features

Groups well predicted in all cancers

- Groups of genes well predicted in all cancers (low prediction error) seems to correspond to ubiquitously expressed and housekeeping genes.

$\rightsquigarrow$ Functional enrichment for general and widespread biological processes:

Gene ontology term	Count	Benjamini corrected P-value
Cellular macromolecule metabolic process	612	1.8E-23
Cellular metabolic process	681	1.2E-16
Cellular protein metabolic process	390	2.8E-16
Macromolecule metabolic process	624	4.0E-16
Nucleic acid metabolic process	404	4.0E-16

Groups well predicted in only certain cancer types

- In contrast, groups well predicted in only certain cancers are associated to specific biological function.

  $\rightsquigarrow$ For instance, a regression tree learned in one PAAD sample identified a group of 1,531 genes, which has:
  - Low prediction error in LGG and PAAD but high error in LAML, LIHC and DLBC.
  - Functional enrichment for specific biological processes (brain).

Gene ontology term	Count	Benjamini corrected P-value
Positive regulation of cellular process	528	7.0E-14
Nervous system development	284	1.3E-13
Positive regulation of macromolecule metabolic process	346	3.5E-12
Positive regulation of biological process	565	8.1E-12
Neurogenesis	200	5.9E-11

Main results

- We confirm the existence of sequence-level instructions for gene expression by developing a method able to explain the expression of different genes using only DNA sequence.
- Our model is as accurate as methods based on experimental data but its biological interpretation appears more straightforward.
- Rather small contribution of TF binding motifs (PWM).

Probing instructions for expression regulation in gene nucleotide compositions
Bessière C., Taha M., Petitprez F., Bréhélin L. , Marin J.-M., Lèbre S. , Lecellier C.-H. PLOS Computational Biology (2018)

Outline

1. Learning the regulatory code of gene expression
   (DNA sequence, gene annotations, binding site motifs (PWM))

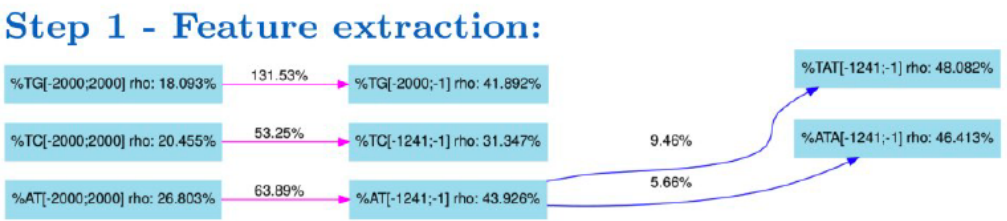
2. Modeling DNA transcription (Linear model) (RNA-seq)
   1. Building (and validating) the model
      [`feature engineering` + `lasso` + `stability selection`]
   2. **Improving features extraction by scanning the sequence**
   3. Looking for variable interactions

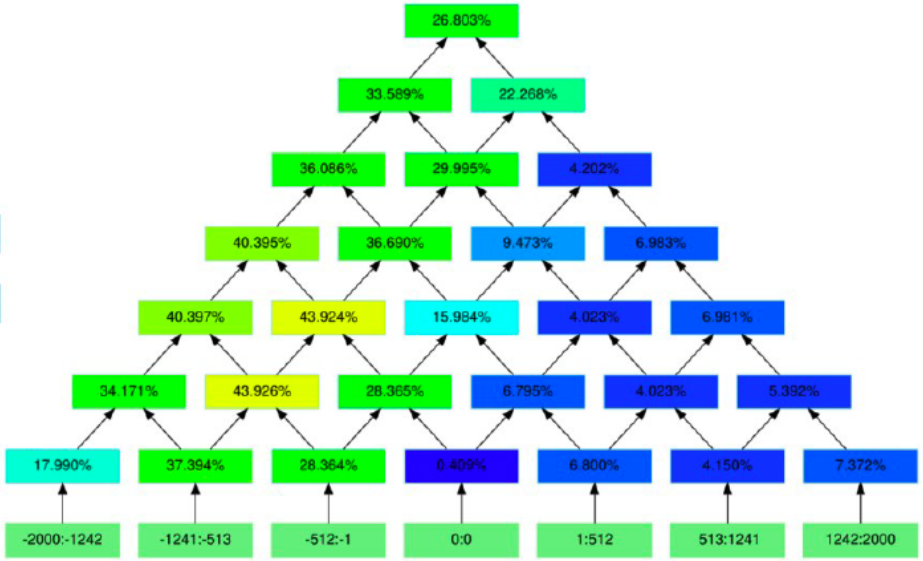
3. Modeling DNA binding

4. Future directions

Step 1 : Heuristic search for features extraction

Christophe Menichelli, part of PhD thesis (2020)





- Select best descriptors (x_j) among {k-mers frequencies[DNA region]}
  For each di-nucleotide ($k = 2$),
  - select the region maximizing elementwise correlation(x_j, y)
  - increase k-mer length , while correlation rises.
- => Improved features for the same linear model

$$y_i = \beta_0 + \Sigma_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

with $Y_{[n\times 1]} = (y_1, \ldots, y_n)'$ vector of observed genes expression level,
and $X_{[n\times p]} = (x_{ij})$ DNA features matrix (x_{ij} is feature j for gene i),..

Source code

Christophe Menichelli, part of PhD thesis (2020)

- Method named DExTER (Domain Exploration To Explain gene Regulation)
- Available source code (python) : `https://gite.lirmm.fr/menichelli/dexter`

Step 2 : Variable selection with group lasso

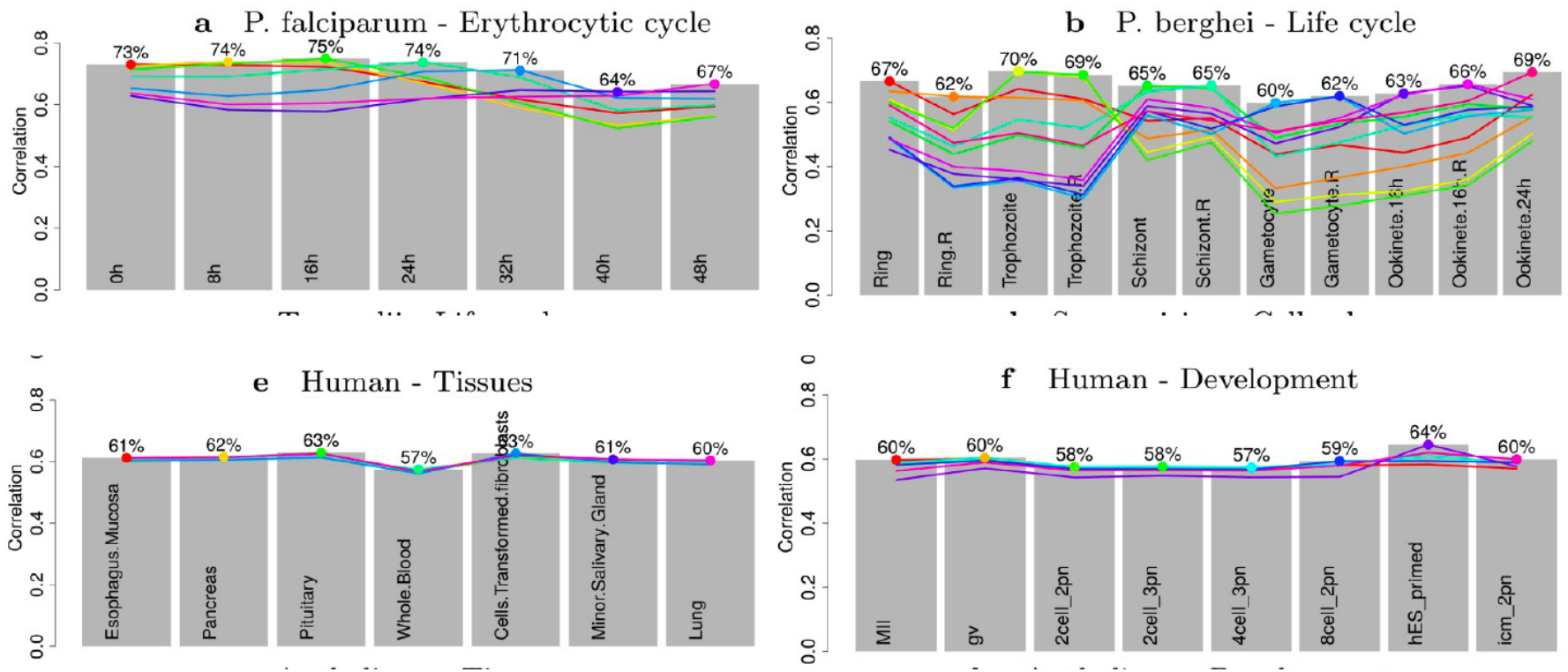
- In order to stabilize variable selection, we encourage that each feature is either selected in all samples or never selected.
- We use the group LASSO (Yuan M. & Lin Y., 2006, JRSS Serie B) for multi-task learning (m conditions)
  For each predictive variable, a group is formed by this variable for all the m models: each variable is either included in the m models or excluded for all.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(||Y - \sum_{j=1}^{p} X^{(j)} \beta^{(j)}||_2^2 + \lambda \sum_{j=1}^{p} ||\beta||_2 \right) \quad (2)$$

  where Y is a $[n \times m]$ matrix, $X^{(j)}$ a vector of size $[n \times 1]$, $\beta^{(j)}$ a vector of of size $[1 \times m]$

- R package `glmnet` by (Hastie, Qian, Tay), option `family = 'mgaussian'` for multi-response.

Biological analysis

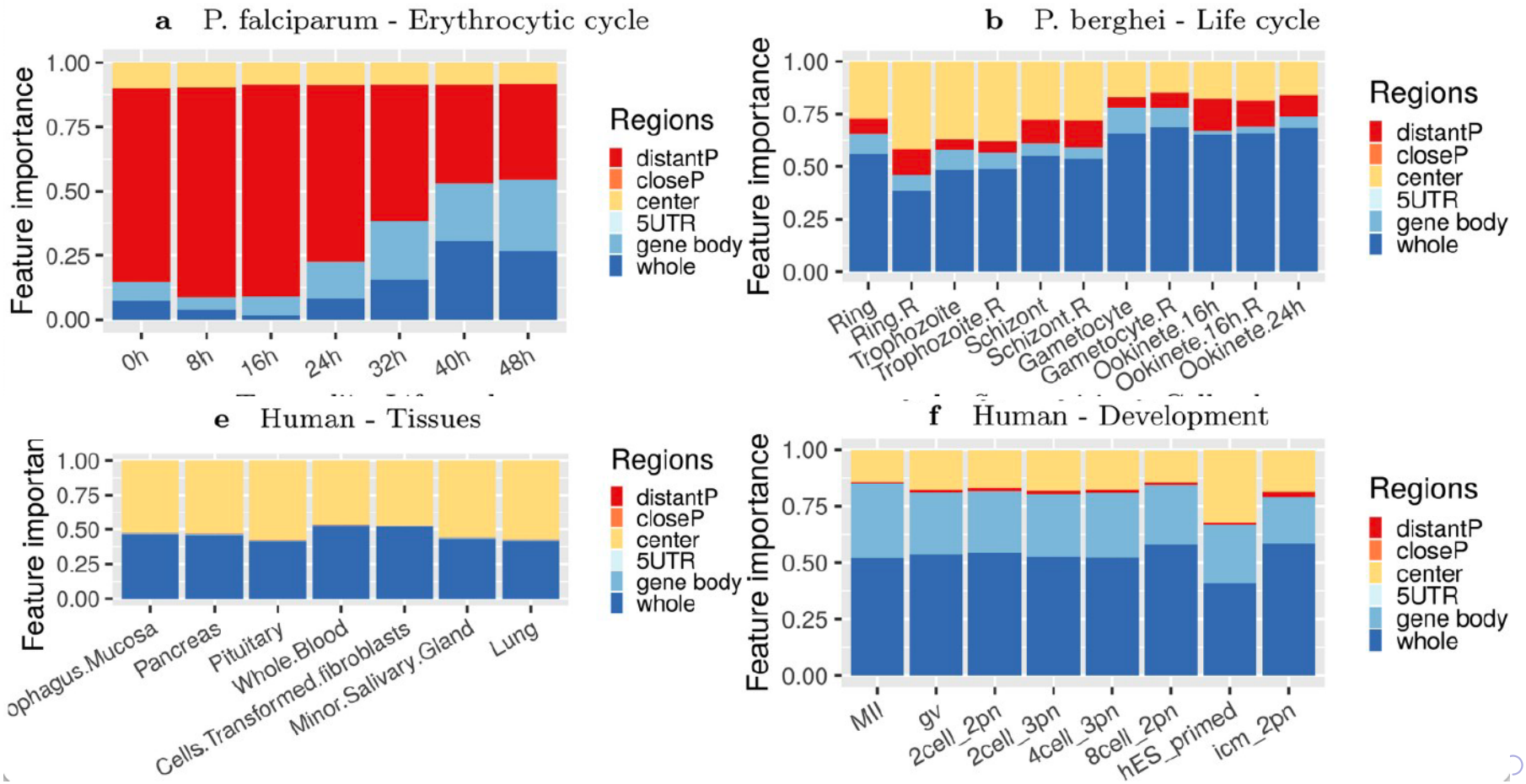


- Long sequences with specific composition are predictive of expression for several Eukaryotes and especially for P. falciparum (malaria parasite)

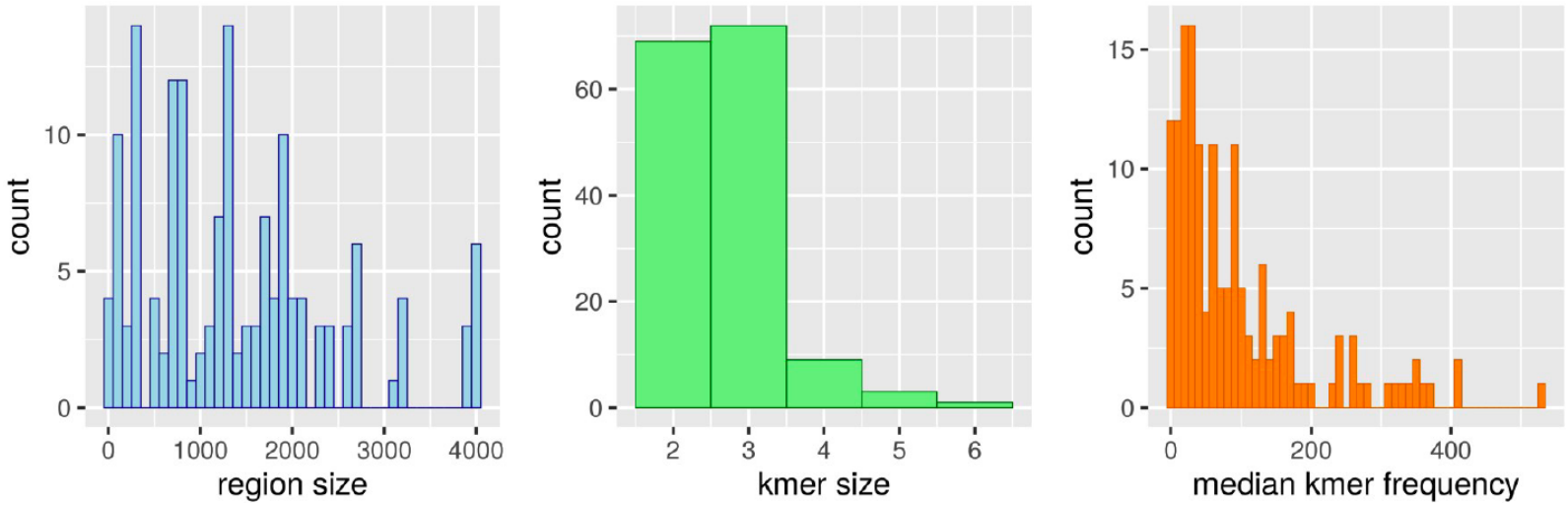
Gene expression data : **a.** Otto et al. 2010 (Molecular Microbiology); **b.** Otto et al 2014 (BMC Biology); **e.** The Genotype-Tissue Expression (GTEx) project. Lonsdale et al. 2013 (Nature Genetics). **f.** Wu et al. 2018 (Nature).

Feature importance

- Model with the first 15 variables selected by the lasso.
- Variable X_j importance in condition l is estimated by the Mean Square Error (MSE) difference between the complete model and the model obtained by setting $\beta_l^{(j)}$ to 0 .



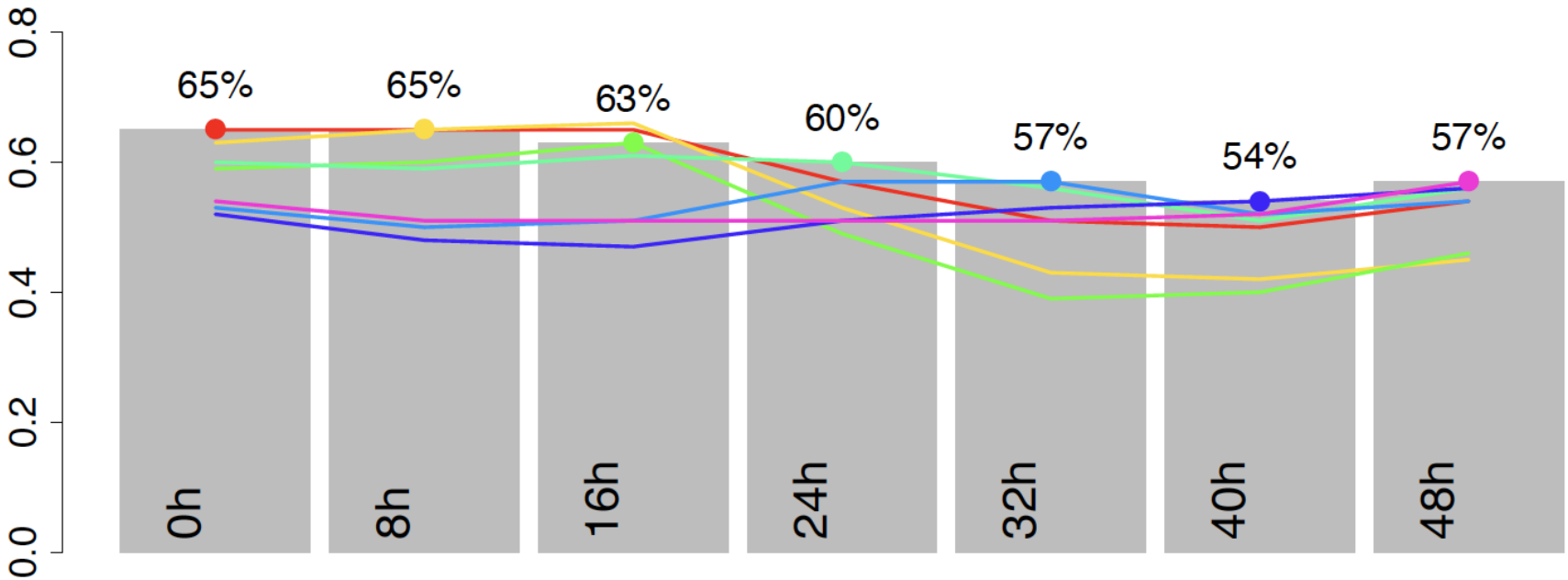
Biological outputs summary



- Most selected variables :
  - large size regions (hundreds base pairs)
  - short k-mers (k = 2 or 3)
  - high number of occurrences (median number > 20)
- => incompatible with classical TFBSs, (usually a dozen base pairs)
  N.B. : From the 154 studied variables, we estimate that less than a dozen may correspond to traditional TFBS motifs.

Comparison with CNN (DeepSea like)

- Convolutional neural networks using the keras implementation (Chollet, F. et al. , 2015) with an architecture similar to the architecture proposed in DeepSea (Zhou et al. 2015)
- => Slightly lower accuracy, but the same dynamic behavior.



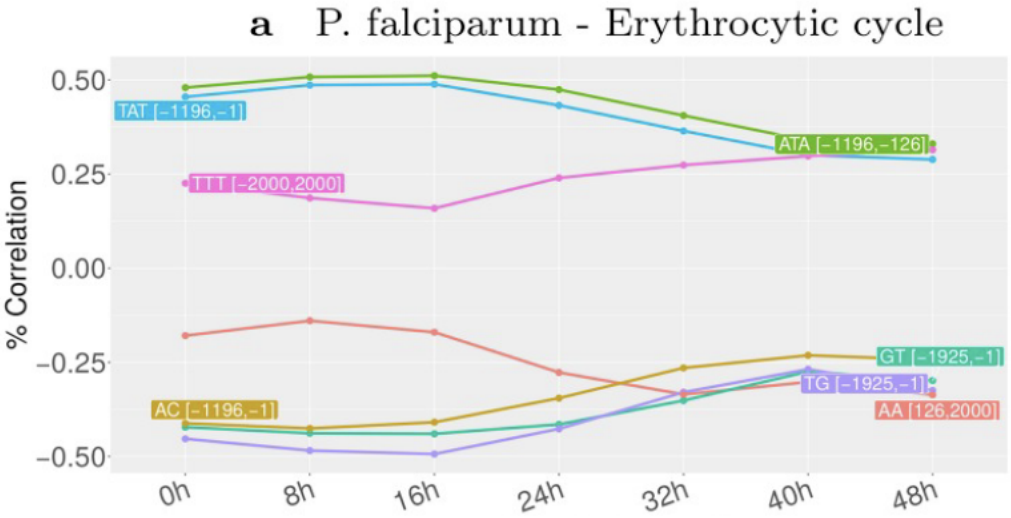
Prediction of P. falciparum gene expression during erythrocytic cycle.

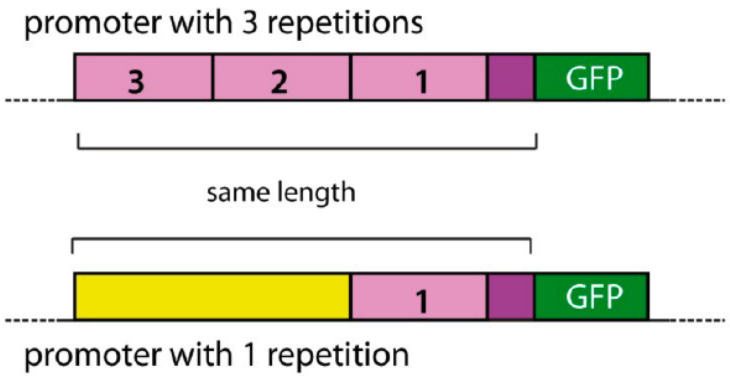
GFP reporter assay: ATA frequency in region [-1196,-126]

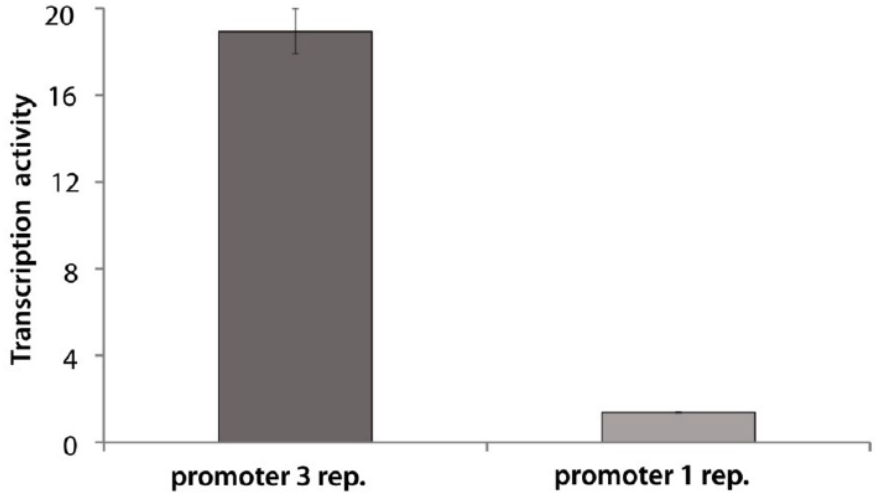
Juan-Jose Lopez-Rubio, Laboratory of Pathogen-Host Interactions (LPHI), Montpellier

- Two chimeric promoters
- Prediction:

  higher activity for the promoter with 3 repetitions







Reporter assay validates a Long Regulatory Element (LRE) controlling gene expression in P. falciparum

Main results

- Gene expression can be predicted with rather high accuracy on the basis of k-**mer relative frequencies in long regions only**.
- Regulation by candidate Long regulatory elements (cLREs) exhibits very **different behaviours** depending on species and conditions.
- The highest accuracy ($cor(Y, \hat{Y}) \approx 0.7$) is obtained in P. falciparum, whose genome is strongly **depleted of transcription factors**.
- The biological relevance of one LRE in P. falciparum was assessed using an **in vivo reporter assay**.

(DExTER) ***Identification of long regulatory elements in the genome of Plasmodium falciparum and other eukaryotes***
Menichelli C. , Guitard V., Martins R. M., Lèbre S., Lopez-Rubio J-J., Lecellier C.-H. and Bréhélin L. PLOS Computational Biology (2021)

Outline

1. Learning the regulatory code of gene expression
   (DNA sequence, gene annotations, binding site motifs (PWM))

2. Modeling DNA transcription (Linear model) (RNA-seq)
   1. Building (and validating) the model
      [`feature engineering` + `lasso` + `stability selection`]
   2. Improving features extraction by scanning the sequence
      [`Iterative search` + `group lasso`]
   3. **Looking for variable interactions**

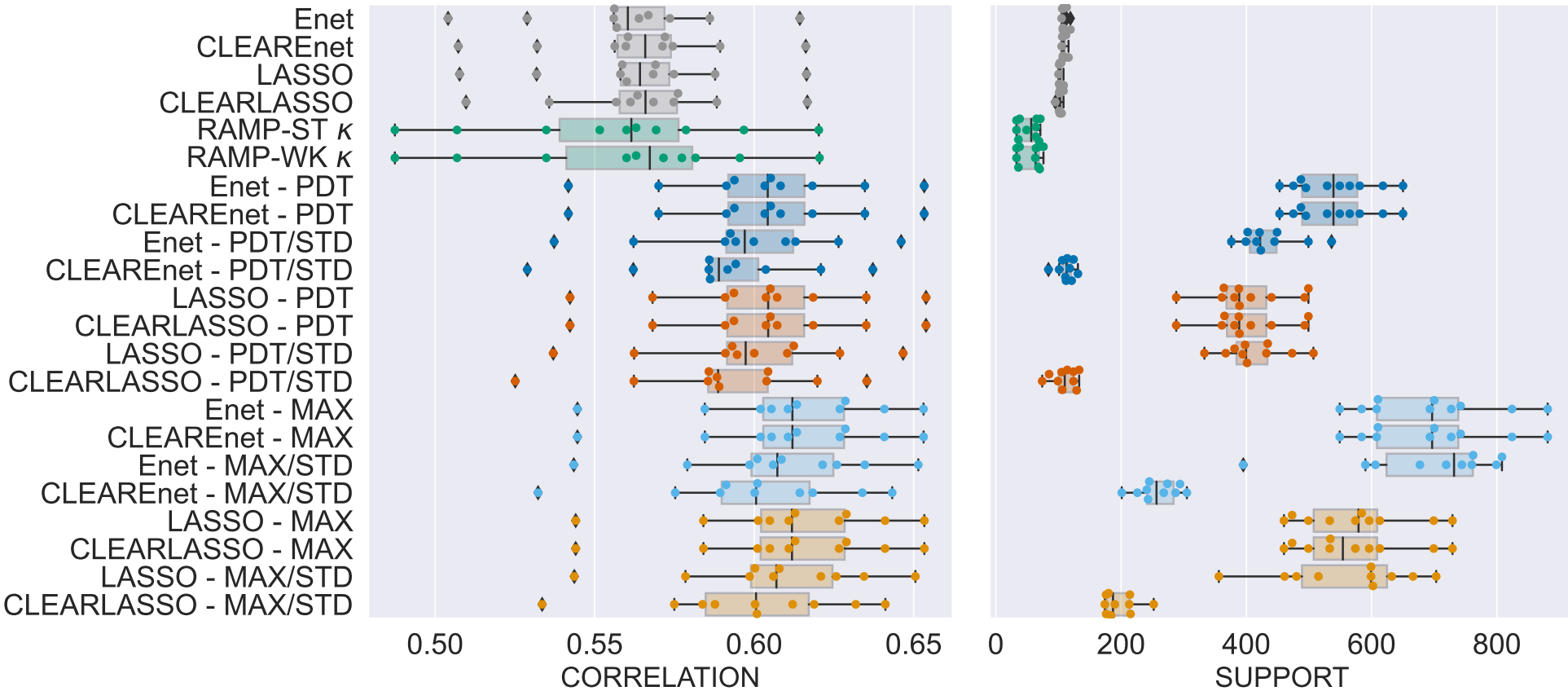
3. Modeling DNA binding

4. Future directions

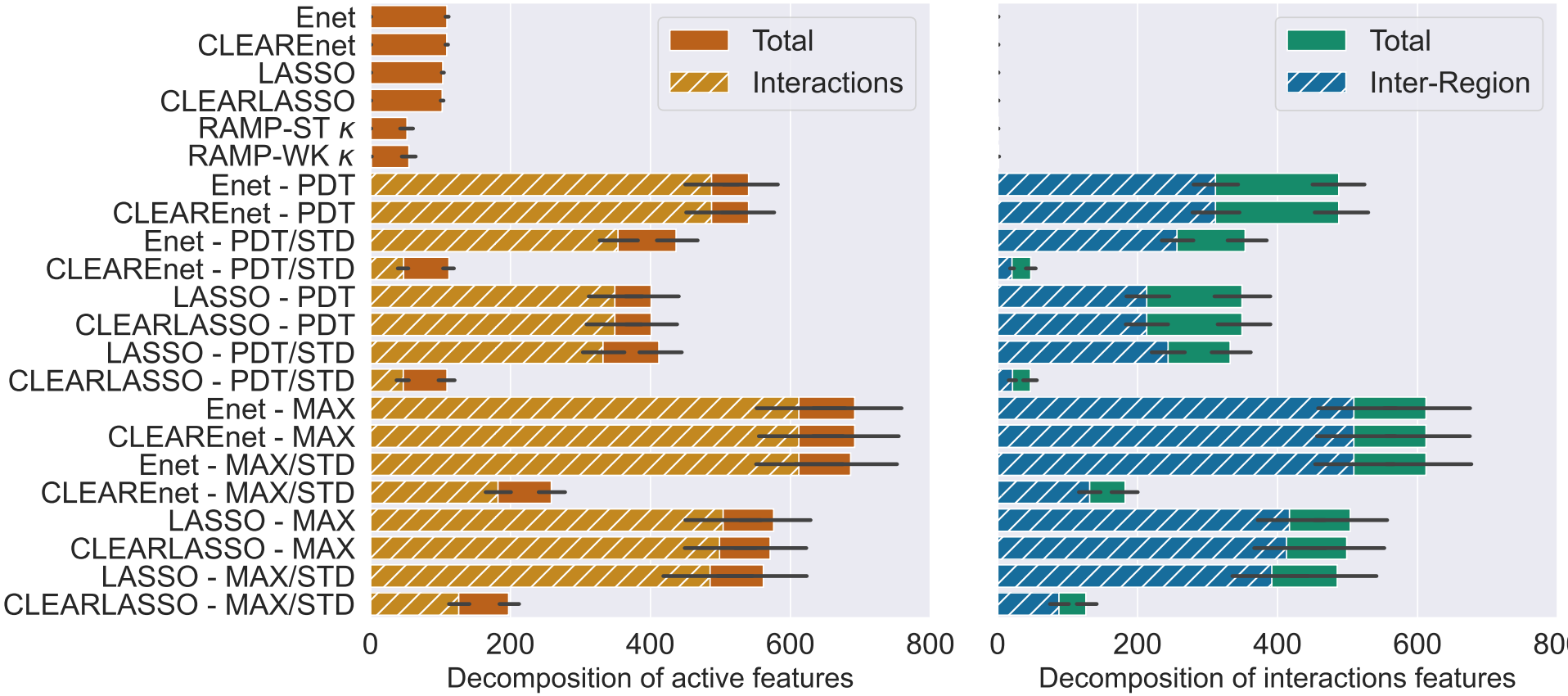
CLEAREnet : accelerated and de-biased Elasticnet for interaction selection

Florent Bascou PhD thesis, with Joseph Salmon superviser (defense in September)

$n \approx 20\,000$, $p = 8x20 + 12880$ => Coordinate descent algorithm (python)



CLEAREnet : accelerated and de-biased Elasticnet for interaction selection



(CLEAREnet) ***An accelerated and de-biased estimator for linear model with quadratic interactions.***
F. Bascou, S. Lèbre, J. Salmon, (In progress.)

Outline

1. Learning the regulatory code of gene expression
   (DNA sequence, gene annotations, binding site motifs (PWM))

2. Modeling DNA transcription (Linear model) (RNA-seq)
   1. Building (and validating) the model
      [`feature engineering` + `lasso` + `stability selection`]
   2. Improving features extraction by scanning the sequence
      [`Iterative search` + `group lasso`]
   3. Looking for variable interactions
      [`Elastic Net` + `coordinate descent` + `debiasing`]

3. **Modeling DNA binding** (Supervised classification) (ChIP-seq)

4. Future directions

Predicting TF binding with cooperating TFs

Jimmy Vandel Post doc (now member of the Bilille platform)

- Contrary to bacterial DNA Binding Domains (DBDs), **most eukaryotic DBDs recognize short binding motifs (around 10bp)** that are not sufficient for specific the usually large ($\approx 10^9$bp) eukaryotic genomes
- Most TFs only associate with **a small subset of their potential genomic sites** in vivo.
- => DBD motifs as resumed by PWMs are not sufficient to completely determine TF binding
- => Predict TF binding using cooperating TFs with a logistic model

$$P(1|s) = S\left(a + \sum_j b_j \cdot \mathrm{Din}_j(s) + \sum_k c_k \cdot \mathrm{CF}_k(s)\right), \quad (3)$$

where $P(1|s)$ is the probability that sequence s is bound by the TF (ChIP-seq experiment), S is the sigmoid function, $\mathrm{Din}_j(s)$ is the relative frequency of the jth di-nucleotide for sequence s, $\mathrm{CF}_k(s)$ is the maximal score of the kth PWM in sequence s

(TFcoop) ***Probing transcription factor combinatorics in different promoter classes and in enhancers***
Vandel J., Cassan O., Lèbre S. , Lecellier C.-H. and Bréhélin L. BMC Genomics (2019)

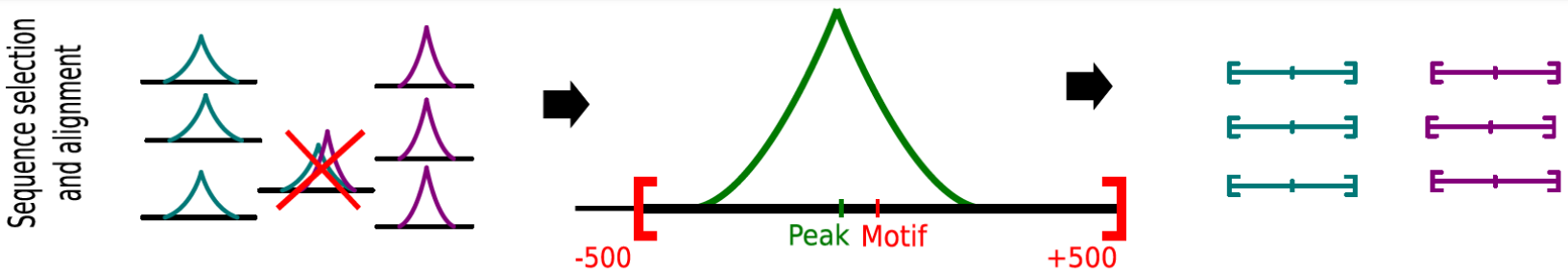
Predicting TF binding with TFscope

Raphaël Roméro PhD thesis (2021)

- Consider cooperating TFs motif location with a lattice iterative search.

- Binding sites of a given TF often vary substantially between cell types and conditions
  => Discrimimant model

Background definition



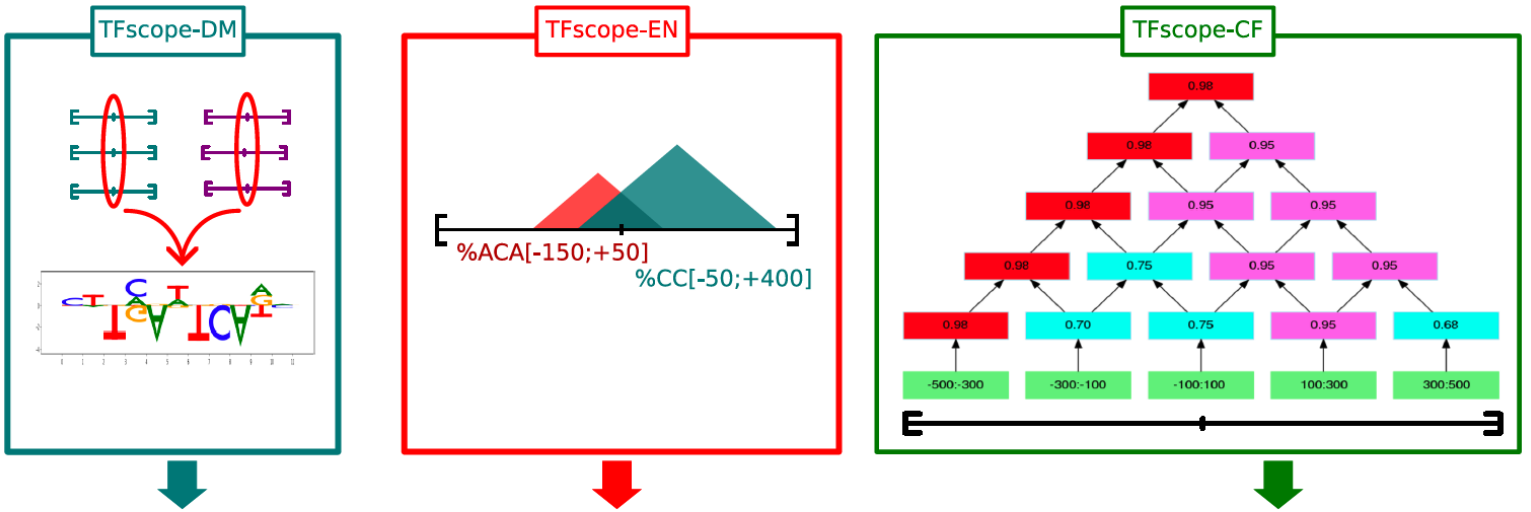
- The definition of the background [which aims to distinguish bound (foreground) versus unbound (background) genomic regions in a given cell type] is often challenging and strongly influences the results and the conclusions.

- The binding sites of a given TF often vary substantially between cell types and conditions.

  => Characterize TF binding specificity **between 2 cell types (or 2 experimental conditions)**

Predicting TF binding with TFscope

Given two ChIP-seq data, we investigate the importance of 3 sources of information:



	DM	%ACA[-150;+50]	...	%CC[-50;+400]	PWM_1	...	PWM_N
seq1	**95**	12%	...	**25%**	5.2	...	**2.4**
seq2	**98**	8%	...	**35%**	3.1	...	**8.6**
seq3	**91**	9%	...	**31%**	3.1	...	**1.2**
...	...	...	...	...	...	...	...

Given two ChIP-seq data, we investigate the importance of 3 sources of information with a logistic model

$$P(1|s) = S\left(a \cdot \mathrm{DM}(s) + \sum_i b_i \cdot \mathrm{NE}_i(s) + \sum_j c_j \cdot \mathrm{CF}_j(s)\right), \quad (4)$$

where $P(1|s)$ is the probability that sequence s belongs to the first class, S is the sigmoid function, $\mathrm{DM}(s)$ is the score of the discriminative motif for sequence s, $\mathrm{NE}_i(s)$ is the value of the ith nucleotidic-environment variable for sequence s, $\mathrm{CF}_j(s)$ is the value of the jth co-factor variable for sequence s,.

Predicting TF binding with TFscope

Given two ChIP-seq data, we investigate the importance of 3 sources of information:

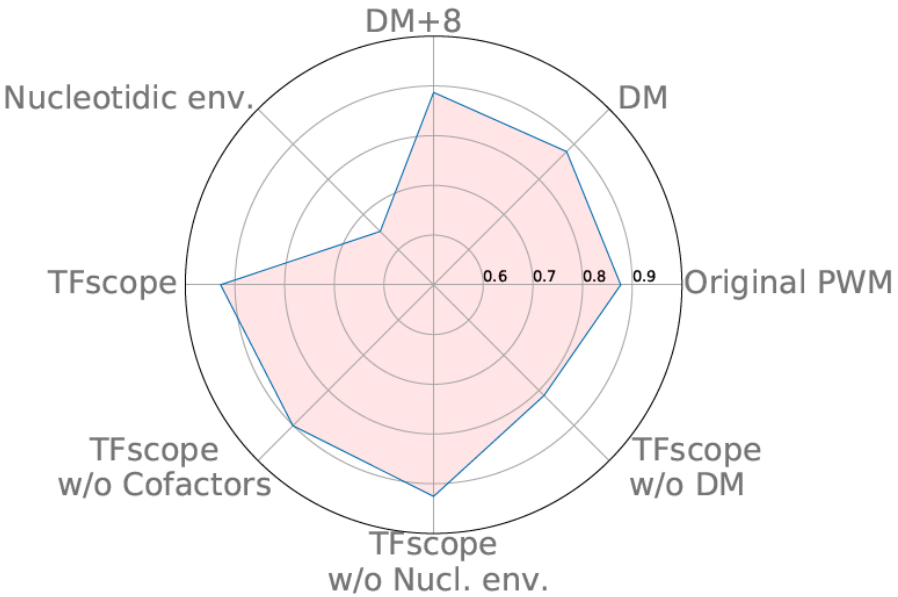
- the core motif => learns discriminative PWMs with logistic lasso
  (Previous work : DREME/STREME, DAMO, Homer)
- the genomic environment => captures the nucleotidic environment in the form of short k-mers (2-4 bps) enriched in specific regions around the core motif with (DExTER method, adapted for logistic regression)
- the cooperative TFs => refinement of the TFcoop method that identifies TF combinations via associated PWM maximal score in a specific region.

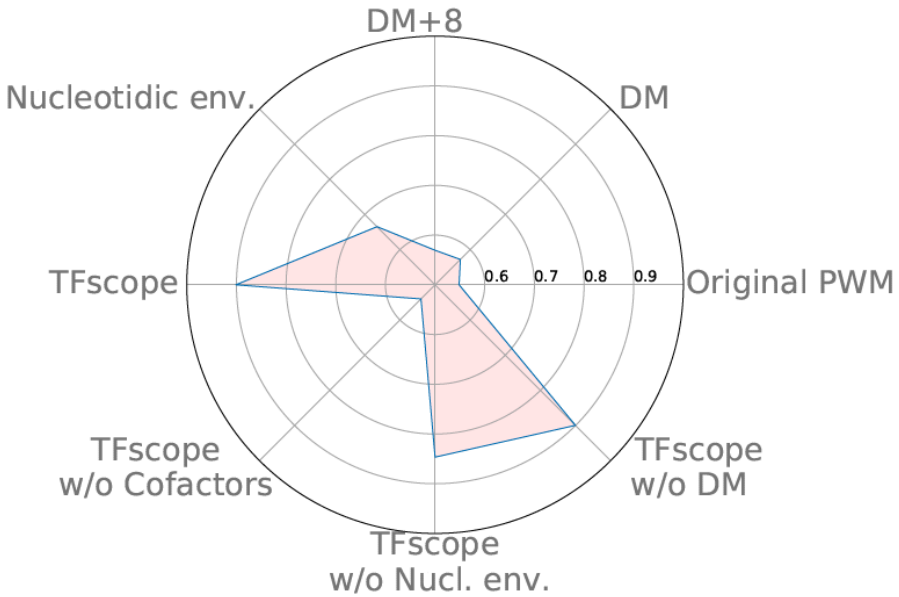
with a logistic model:

$$P(1|s) = S\left(a \cdot \mathrm{DM}(s) + \sum_i b_i \cdot \mathrm{NE}_i(s) + \sum_j c_j \cdot \mathrm{CF}_j(s)\right),$$

where $P(1|s)$ is the probability that sequence s belongs to the first class, S is the sigmoid function, $\mathrm{DM}(s)$ is the score of the discriminative motif for sequence s, $\mathrm{NE}_i(s)$ is the value of the ith nucleotidic-environment variable for sequence s, $\mathrm{CF}_j(s)$ is the value of the jth co-factor variable for sequence s,.

Biological output



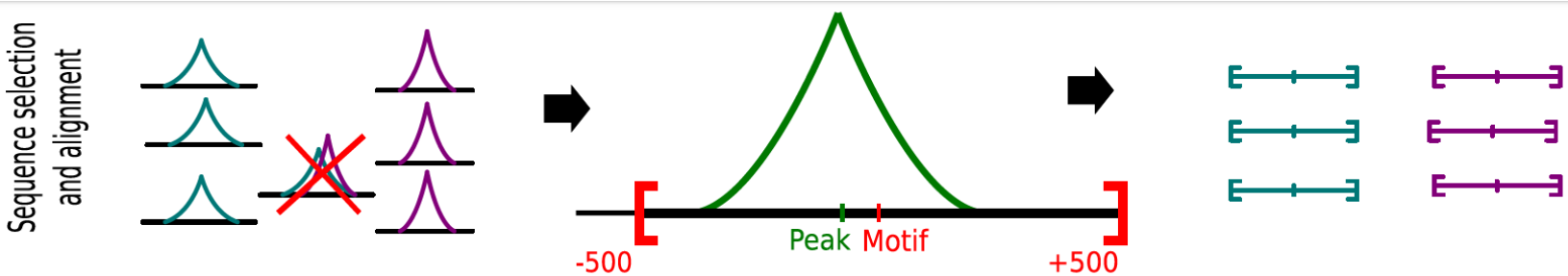


Outline

1. Learning the regulatory code of gene expression
   (DNA sequence, gene annotations, binding site motifs (PWM))
2. Modeling DNA transcription (Linear model) (RNA-seq)
   1. Building (and validating) the model
      [feature engineering + lasso + stability selection]
   2. Improving features extraction by scanning the sequence
      [Iterative search + group lasso]
   3. Looking for variable interactions
      [Elastic Net + coordinate descent + debiasing]
3. Modeling DNA binding (Supervised classification) (ChIP-seq)
   1. Considering cooperating TFs
      [Logistic lasso]
   2. Background definition
      [Discrimiant classifier]
   3. 3 sources of information (Bound TF discriminant motif, genomic environnement, binding motif of the cooperative TFs)
      [Logistic lasso + Iterative search + lattice]
4. **Future directions**

Summary & Future directions

- We proposed explicative models (and learning procedure) for gene expression / TF binding
  - exploiting 'biologically' engineered features, from DNA sequence only
  - that allows to quantify the relative importance of each source information (easier than CNN based methods)
- Elements of biological validation (P. falciparum reporter assay, Breast cancer: Fra-1/Fra-2, ...)
- Future research directions
  - Relax the motif anchor/sequence alignement [New model & inference]
  - Extend the Elastic Net optimization procedure to logistic model
  - Novel 'wet' biological experiments (P. falciparum) and analysis (Human cancer, Early Phase Testing)



Thank you for your attention

The team!

Contributed references

(TFscope) A systematic analysis of the genomic features involved in the binding specificity of transcription factors
R. Roméro, C. Menichelli, J.-M. Marin, S. Lèbre, C.-H. Lecellier, L. Bréhélin, (In progress).

(CLEAREnet) An accelerated and de-biased estimator for linear model with quadratic interactions.
F. Bascou, S. Lèbre and J. Salmon, (In progress.)

Fra-1 regulates its target genes via binding to remote enhancers without exerting major control on chromatin architecture in triple negative breast cancers
Bejjani et al., Nucleic Acid Research (2021)

(DExTER) Identification of long regulatory elements in the genome of Plasmodium falciparum and other eukaryotes
C. Menichelli, V. Guitard, R. M. Martins, S. Lèbre, J.-J. Lopez-Rubio, C.-H. Lecellier, L. Bréhélin, Plos Computational Biology, (2021).

Debiasing the Elastic Net for models with interactions
F. Bascou, S. Lèbre and J. Salmon, Journées de Statistique de la SFDS, (2020).

(TFcoop) Probing transcription factor combinatorics in different promoter classes and in enhancers
J. Vandel, O. Cassan, S. Lèbre, C.-H. Lecellier and L. Bréhélin, BMC Genomics, (2019).

Probing instructions for expression regulation in gene nucleotide compositions
C. Bessière, M. Taha M., F. Petitprez, J. Vandel, J.-M. Marin, L. Bréhélin, S. Lèbre and C.-H. Lecellier, PLoS Computational Biology, (2018).

Other references

Shallow v.s deep learning for learning review (mostly classification) Learning the Regulatory Code of Gene Expression Zrimec et al. (Frontiers in molecular Biosciences, 2021)

Predicting gene expression with CNN

(Xpresso) Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks V. Agarwal, J. Shendure. Cell Reports (2020)

(DeepSea) Predicting effects of noncoding variants with deep learning-based sequence model. Zhou J, Troyanskaya OG. Nature Methods. (2015)

Predicting TF binding with CNN

(Deepbind) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning B. Alipanahi, A. Delong, M. T Weirauch, B. J. Frey. Nature Biotechnology (2015)

Predicting gene expression from epigenetics data

(TEPIC) Schmidt F. et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Res. 2017.

(RACER) Li Y., Liang M., Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS Comput Biol. 2014.