

# Probing instructions for gene expression regulation in gene nucleotide compositions

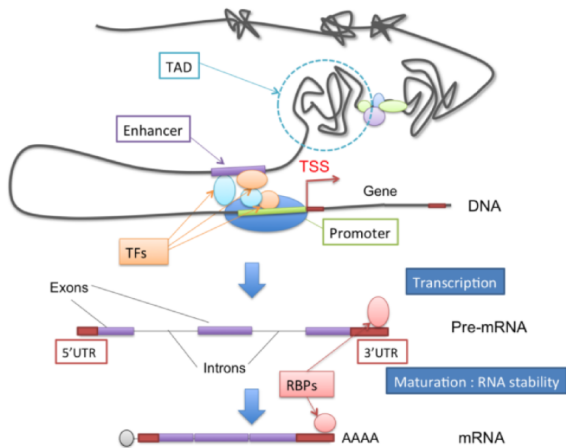
M. Taha, C. Bessière, F. Petitprez, J. Vandell,  
J.-M. Marin, L. Bréhélin, S. Lèbre, C. Lecellier



IMAG  
INSTITUT MONTPELLIEN D'ALEXANDER GROTHENDIECK



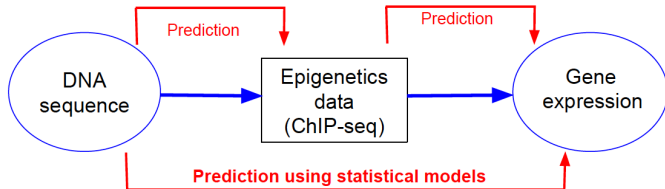
# Gene expression regulation



TFs = Transcription factors  
RBPs = RNA Binding Proteins

Transcriptional  
regulations


Post-transcriptional  
regulations



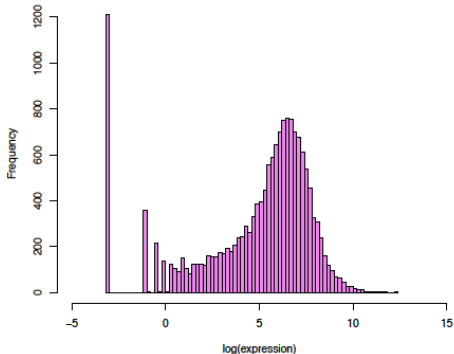
- Predicting Epigenetics data from DNA sequence
  - Whitaker, J. W. et al. Nat. Methods (2015)
  - Zhou, J. et al Nat. Methods (2015)
- Predicting gene expression from epigenetics data
  - RACER : Y. Li and al. PLoS (2014)
  - TEPIC : Schmidt F. et al. Nucleic Acids Res (2017)

↪ Question: Can we identify directly the DNA determinants involved in gene regulation?

- ① Model building
- ② Comparison with experimental data (Chip-Seq)
- ③ Advanced model
- ④ Biological interpretation

- Originality :
  - Modeling gene expression using [DNA sequence data only](#)
  - ONE model per patient (Cancer tumors)
- Data
  - Gene expression measurements for each patient (RNA-Seq)
  - DNA sequence (Genome Reference GRCh38/hg38)
    - Nucleotide and di-nucleotide compositions:  $\%CG = \#CG / (\text{length} - 1)$
    - TF binding motifs : [PWM scores](#) 
    - DNA shapes (computed with the [Bioconductor package DNashaperR](#) )
- N.B.: Similar work on yeast Kasowski et al. Science (2013)  
Sequence variations affect histone modifications

# Response variable : RNA-Seq (log transformed values)



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

- We built a global **linear regression model** to explain the expression of genes using DNA/RNA features associated with their regulatory regions (e.g. nucleotide composition, TF motifs, DNA shapes):

$$Y = X\beta + \varepsilon$$

where

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

$(n \sim 20000)$

# Variable selection with Lasso

- Linear regression with  $\ell_1$ -norm penalty or Lasso (Tibshirani, 1996) applied to standardized data:

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

- The penalty  $\lambda$  is chosen by 10-fold cross-validation to minimize the mean square prediction error.
- Some coefficients  $\beta_i$  are set to 0 exactly ( $\ell_1$ -norm geometry).
- $\lambda$  defines the number of selected variables.



- Criterion :

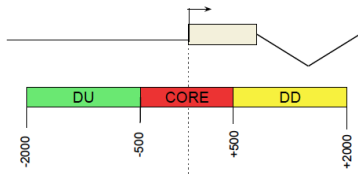
- ① Mean square error (MSE)
- ② Correlation coefficient  $Corr(Y, \hat{Y})$  between the measured expression  $Y$  and the predicted expression  $\hat{Y}$

in a 10-fold cross-validation procedure:

- ① Model is learned in the training data
  - ②  $MSE/Corr(Y, \hat{Y})$  is evaluated in the test data.
- Data shown : RNA-Seq gene expression (TCGA) from 12 cancers types, 20 patients per cancer.

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

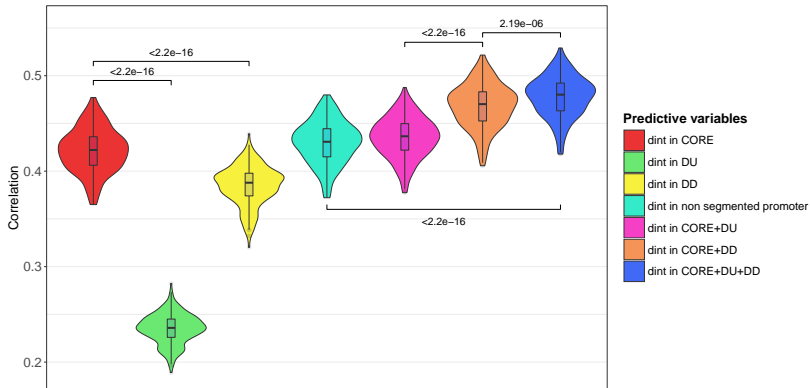
# Promoter definition



- DU** Distal Upstream promoter
- CORE** Core promoter
- DD** Distal Downstream promoter

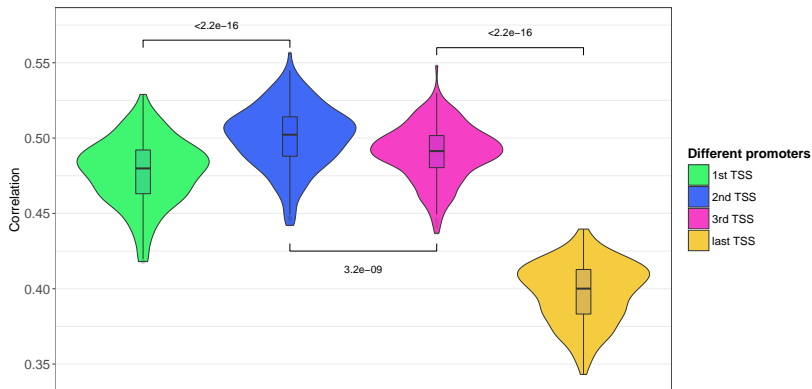
- Nucleotide and di-nucleotide compositions:  $\%CG = \#CG / (\text{length} - 1)$

# Promoter definition



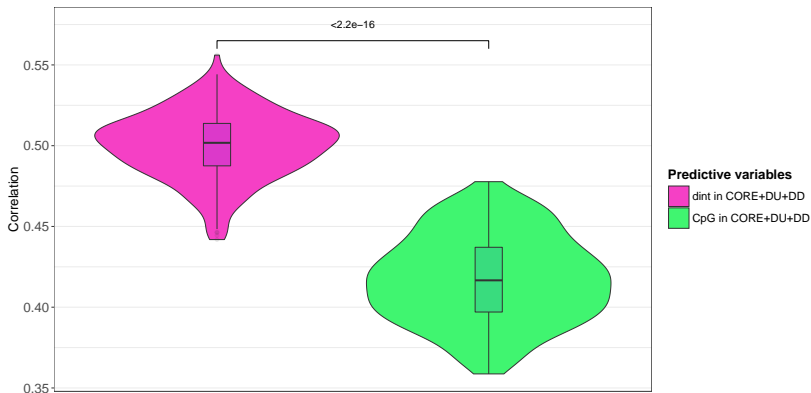
- The highest accuracy was obtained combining the 3 segments.

# TSS choice



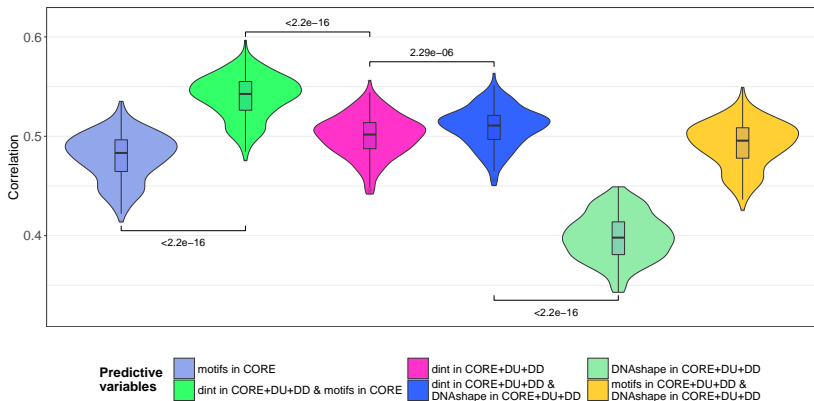
- Our model achieved higher predictive accuracy with the promoters centered around the 2<sup>nd</sup> TSS, in agreement with Cheng et al. (2012).

# All (di-)nucleotides vs CpG only



- Considering all (di-)nucleotides achieved better model performance.

# Contribution of TF motifs and local DNA shapes



- The increase in performance when including TF motifs or DNA shapes is rather modest.

- ① Model building
- ② Comparison with experimental data (Chip-Seq)
- ③ Advanced model
- ④ Biological interpretation

# DNA features vs. experimental data (ChIP-seq)

- Comparison with models integrating:
  - TF-binding signals with Chip-Seq (RACER, Y. Li and al. PLoS, 2014)
  - Open-chromatin signals (TEPIC, Schmidt F. et al. NAR, 2017)
- In both cases, the models were built using the same set of genes:
  - (i) on the original data,
  - (ii) on randomized predictive variables (gene centered shuffling: rand)
  - (iii) on the maximum value of all predictive variables (gene centered maximum: max).

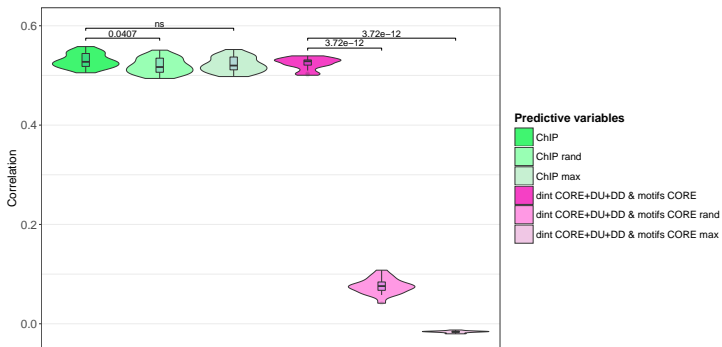
$$\mathbf{X} = \begin{matrix} \text{gene} \rightarrow & \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \end{matrix}$$

Variable

← Shuffle per gene



# Comparison with model integrating TF-binding signals



\*\*\* In cases (ii) and (iii), the links between the predictive variables and expression is broken and a regression model is expected to poorly perform as our model does (Left, light pink). \*\*\*

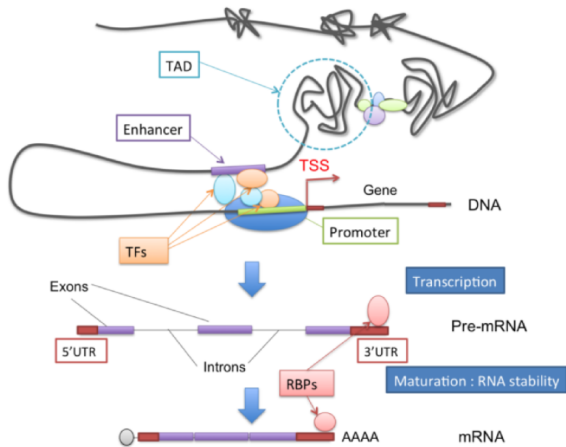
# Comparison with model integrating open-chromatin signals



\*\*\* In cases (ii) and (iii), the links between the predictive variables and expression is broken and a regression model is expected to poorly perform as our model does (Left, light pink). \*\*\*

- 1 Model building
- 2 Comparison with experimental data (Chip-Seq)
- 3 **Advanced model**
- 4 Biological interpretation

# Gene expression regulation

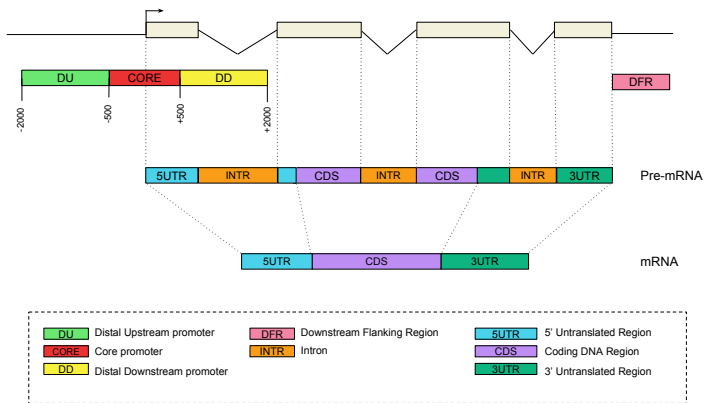


TFs = Transcription factors  
RBPs = RNA Binding Proteins

Transcriptional regulations

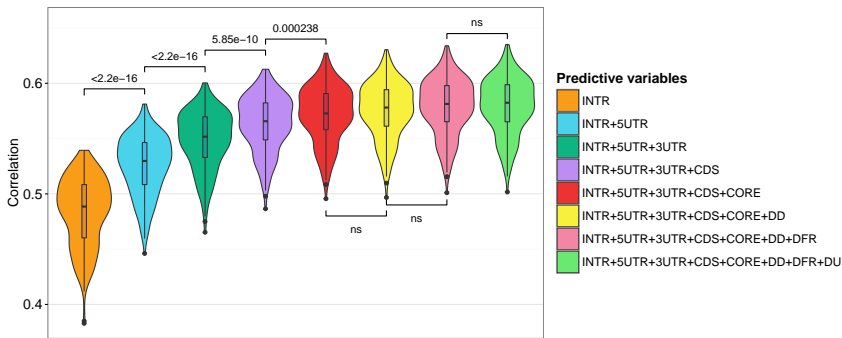
Post-transcriptional regulations

# Contribution of additional genomic regions



↪ Nucleotide and di-nucleotide compositions:  $\%CG = \#CG / (\text{length} - 1)$   
in 8 selected regions (20 variables per region)

# Contribution of additional genomic regions



- DNA regions 'forward-like' selection procedure
- Our model : Nucleotide and di-nucleotide compositions in 8 selected regions (20 variables per region)

# Contribution of 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

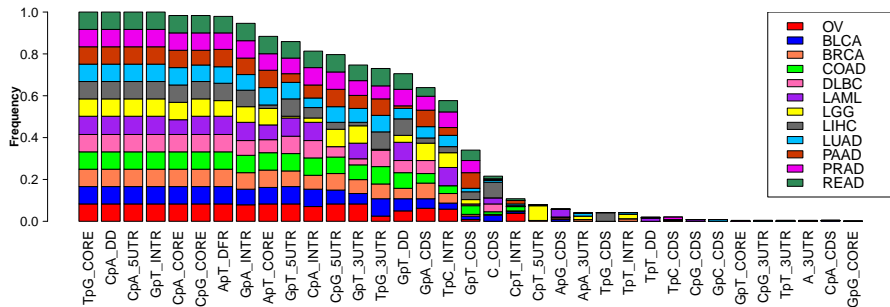
- DNA regions 'forward-like' selection procedure
- Our model : Nucleotide and di-nucleotide compositions in 8 selected regions (20 variables per region)

- **Stability selection** (Meinshausen *et al.*, 2010)
- Lasso inference is repeated 500 times where, for each iteration,
  - (i) only 50% of the 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 `stabse1` from the [R package C060](#) for `glmnet` models.)



# Stable variables selection



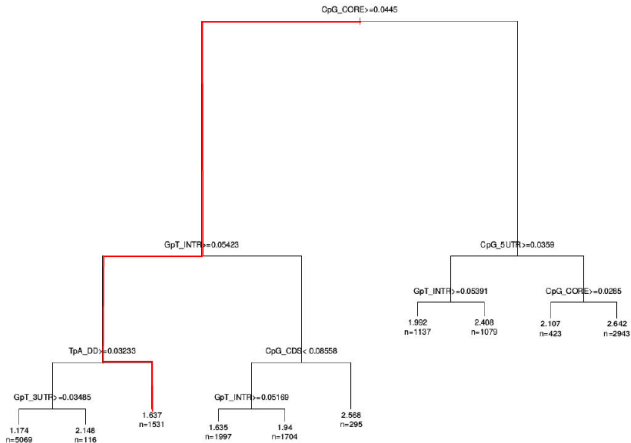
Proportion of samples in which each variable is selected with high consistency ( $> 70\%$  stability)

(Average  $\sim 16$  variables per sample)

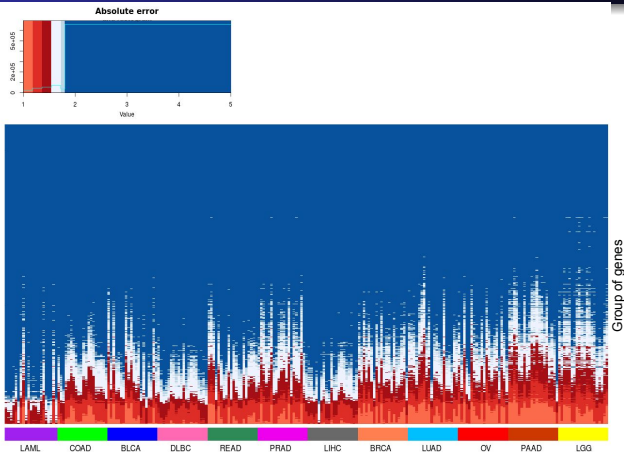
- ① Model building
- ② Comparison with experimental data (Chip-Seq)
- ③ Advanced model
- ④ Biological interpretation

# A) 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



# A) 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

⇒ 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**.

↪ 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.

↪ 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).

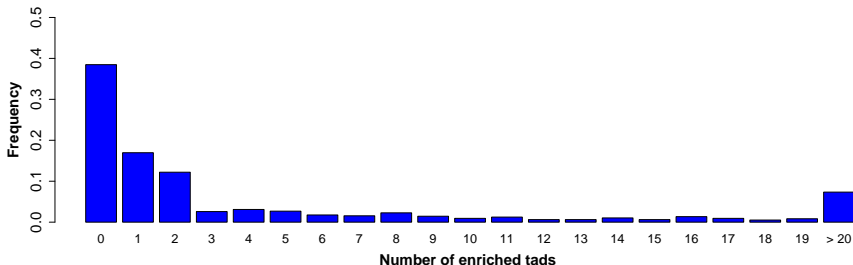
Gene ontology term	Count	Benjamini corrected P-value
Positive regulation of cellular process	528	7.0E-14
Nervous system development	284	1.3E-13
Positive regulation of macromolecule metabolic process	346	3.5E-12
Positive regulation of biological process	565	8.1E-12
Neurogenesis	200	5.9E-11

## B) Link with the genome architecture

- Do the groups of genes identified by the regression trees correspond to specific TADs ?
  - Motivations
    - Genes within the same TAD tend to be coordinately expressed (Nora et al. 2012, Fanucchi et al. 2013).
    - Nucleotide composition along the genome can help define TADs (Jabbari and Bernardi, 2017)
  - Validation :
    - We used the 373 TADs containing more than 10 genes.
    - For each TAD and each (di-)nucleotide, we used a Kolmogorov-Smirnov test to compare the (di-)nucleotide distribution of the embedded genes with that of all other genes (multiple testing controlled with FDR).
- ↪ 87% of the TADs are characterized by at least one specific nucleotide signature.

## B) Link with the genome architecture

- We next considered the 967 groups of genes whose expression is accurately predicted by our model (regression trees).  
↪ 60% of the well predicted groups of genes (top 25% well predicted) were enriched for at least one TAD (p-value < 0.05, hyper-geometric test).



TAD enrichment within groups of genes  
whose expression is accurately predicted by our model.



# Conclusions

- 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.
- We provide evidence that the **genes nucleotide composition** can be **linked** to co-regulations associated with **genome 3D architecture** and to associations of genes within **TADs**.

- Further improve the model
  - Relax linearity assumption ?
  - Include variable interactions ?  
(+ Comparison with deep learning approaches)
  - Integrate TF binding motifs ?
  
- Get more biology
  - TADs
  - methylation
  - ...

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

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

# Thank you for your attention

Preprint available on BioRxiv

Probing instructions for expression regulation in gene nucleotide compositions (under revision) M. Taha, C. Bessière, F. Petitprez, J. Vandel, J.-M. Marin, L. Bréhélin, S. Lèbre, C. Lecellier.



The team!