A machine-learning guided approach to explore the cis-regulatory code involved in neuronal differentiation

Océane Cassan[1*], Christophe Vroland[1,2], Julien Raynal[1,2], FANTOM consortium, Masaki Kato[3], Hazuki Takahashi[3], Takeya Kasukawa[3], Piero Carninci[3], Chi Wai Yip[3*], Laurent Bréhélin[1*] & Charles-Henri Lecellier[1,2*]

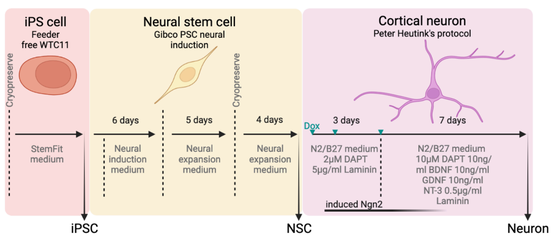
Context

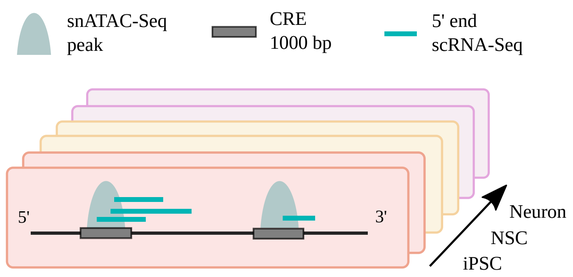
Gene expression is controlled by proximal and distal **cis-regulatory elements (CREs)**, containing DNA motifs bound by various transcription factors (TFs). Other sequence features, such as specific k-mers or low complexity regions, have also been implicated [1–3] .

However, in a dynamic biological process such as cell differentiation, we lack an understanding of how the transcriptional activity of CREs progressively change and what sequence features underlie these transitions, which may reflect common and/or coordinated regulatory processes.

Single cell dataset

Differentiation pseudotime (391 metacells)

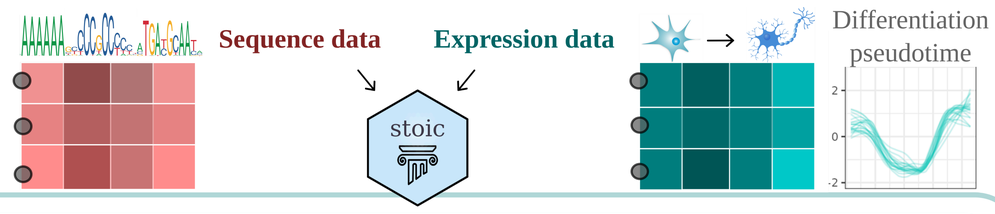




Stoic: a machine-learning guided method to identify the sequence features associated with specific CRE expression profiles

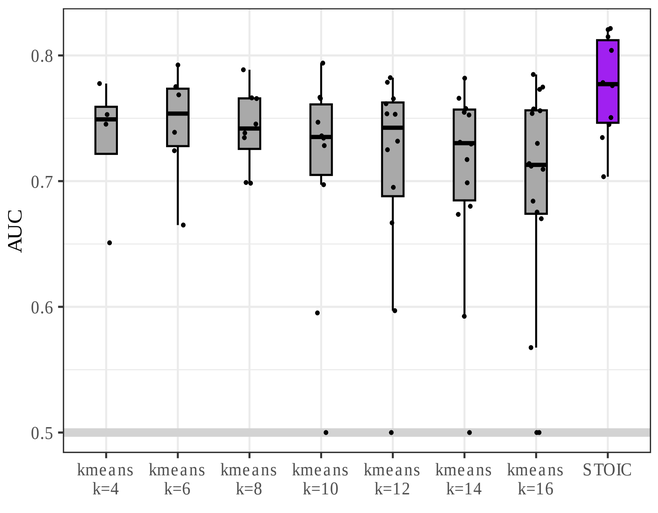
Stoic is an **original approach that aims to identify the sequences features responsible for specific CRE expression profiles** observed in the data. For this, Stoic explores the expression space and delineates the CRE clusters iteratively in order to optimize the performance of a supervised classifier predicting CRE cluster membership using only DNA sequence features.

~400 sequence features for each CRE (k-mers & TF motifs) ~11000 transcriptionally variable CREs



Performances

Stoic was compared to a non-guided approach that first runs a clustering of the CRE expression profiles with a standard k-means algorithm, and then trains k RFs to predict cluster membership using sequence features.

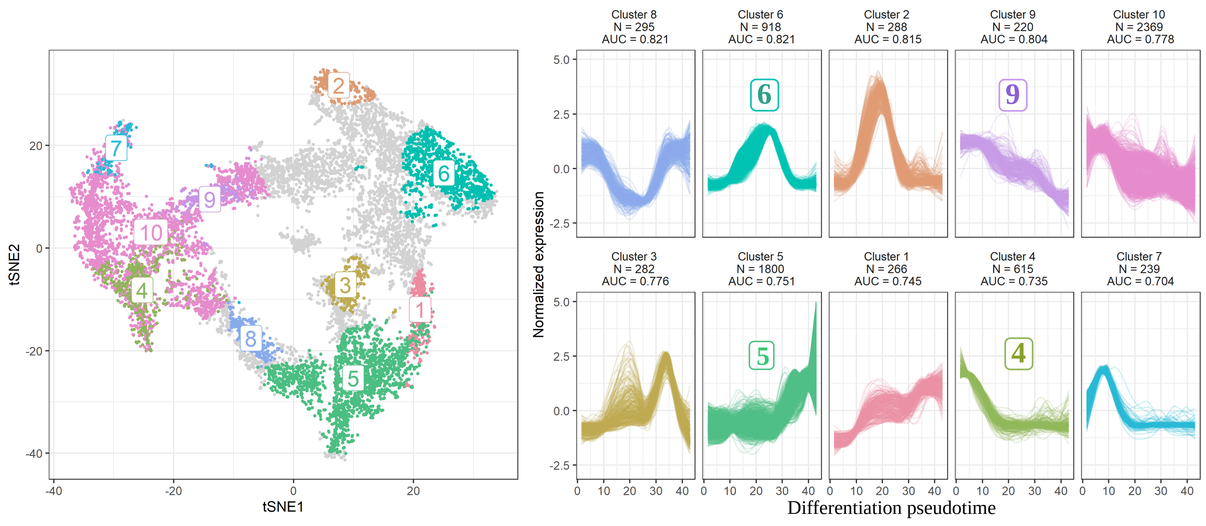


Studies of clustering stability showed that Stoic is comparable to the k-means approach for the same number of clusters and starting points.

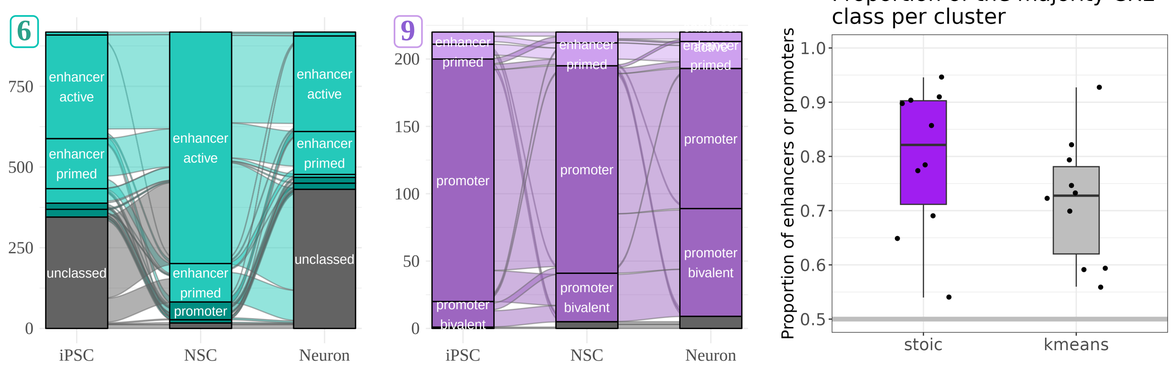
However, Stoic provides higher AUC values for discriminating cluster membership using sequence features, showing that **it recovers stronger sequence to expression associations** than the standard k-means.

Further experiments on simulated data also showed good precision and recall values.

Stoic clusters outline diverse expression profiles and epigenetic signatures

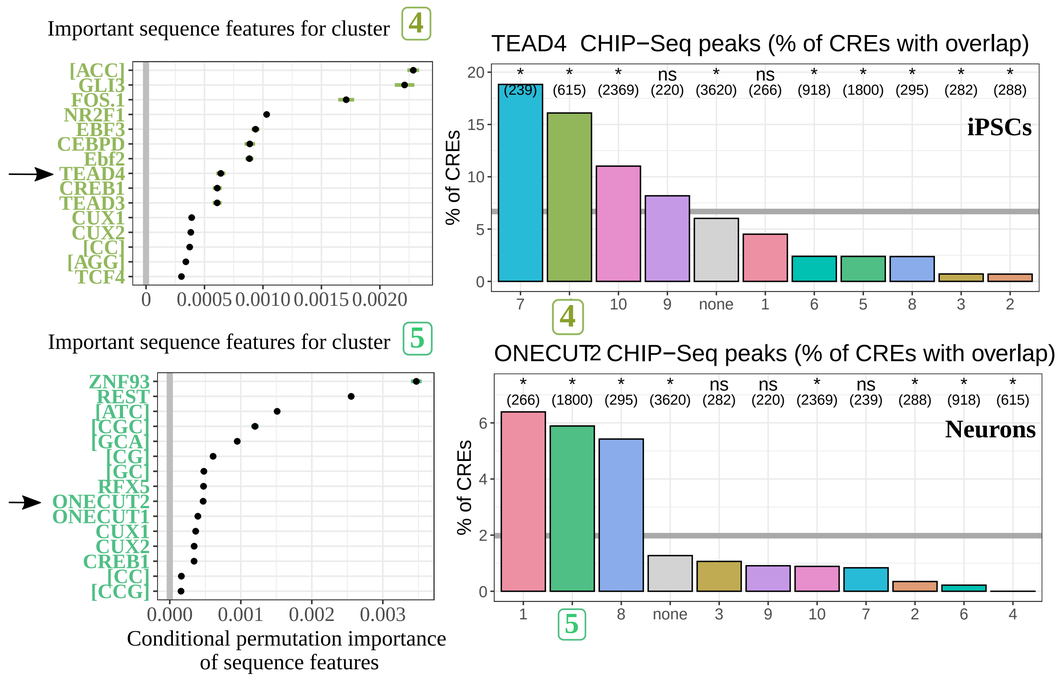


Most inferred clusters are either **enhancer-rich or promoter-rich**



Chromatin states of CREs are predicted from CUT&Tag epigenetic data by chromHMM

Important sequence features in Stoic clusters are supported by CHIP-Seq experiments

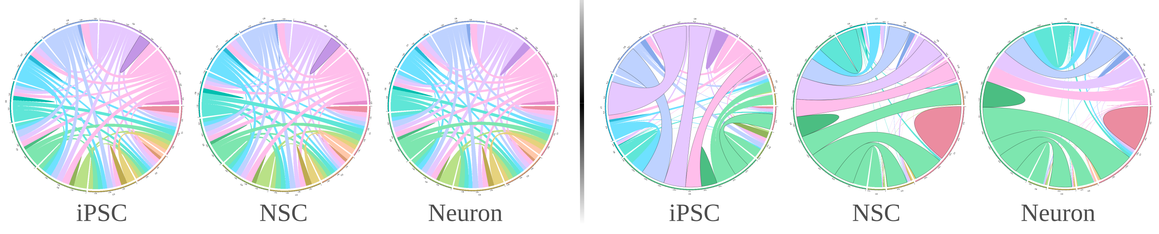


TEAD4 and ONECUT2 are predicted as important sequence features of clusters 4 and 5, and are **enriched in binding events** in the CREs of these clusters.

Transcriptionnally active clusters come into contact in a cell type-specific manner

DNA-DNA contacts between clusters from Hi-C are static during differentiation

DNA-RNA contacts between clusters from RADICL-Seq are dynamic



Further interpretations of Stoic clusters based on **eQTLs, repeat elements, or clinically relevant gene sets** provide an updated perspective on the transcriptional regulations at play during neuronal differentiation.

Stoic R package

Stoic's methodology is available as an R package. The machine-learning guided approach developped in Stoic is applicable to any problem where the clustering of some measurments can be guided by a second matched dataset.

```
library(remotes) # remotes should be installed if it is not
install_gitlab("oceane.cssn/stoic")
```

Affiliations

[1]LIRMM, Univ Montpellier, CNRS, Montpellier, France,
[2]Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, France
[3]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

References

1. C. M. et al. Identification of long regulatory elements in the genome of plasmodium falciparum and other eukaryotes. PLOS Computational Biology, 2021.
2. M. G. et al. Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. Nat. Communications, 2021.
3. Horton et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. Science, 2023.
4. M. Lajoie et al. Computational discovery of regulatory elements in a continuous expression space. Genome biology, 2012.