Automated and genome-scale exploration of the cis-regulatory code involved in neuronal differentiation

Océane Cassan, Christophe Vroland, Julien Raynal, Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin, Masaki Kato, Hazuki Takahashi, Takeya Kasukawa, Robert Lehmann, Vincenzo Lagani [many others from FANTOM6], Chi Wai Yip, Piero Carninci, Laurent Bréhélin, Charles Lecellier

Montpellier Computational Regulatory Genomics group (ML4REGGEN)

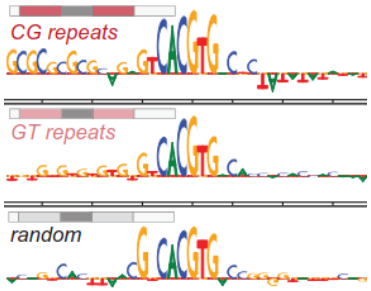
September 26, 2024

Learning the *cis*-regulatory code

What are the sequence features underlying the transcriptional activity of cis-regulatory elements (CREs) during dynamic processes like differentiation?

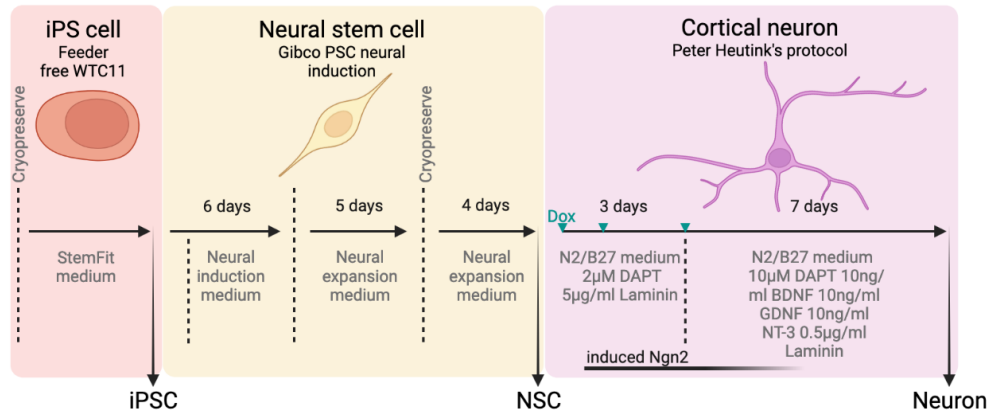
→ Relative impact of

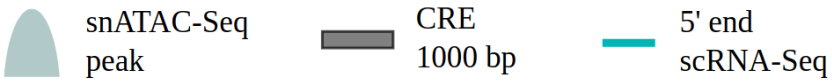
- TF binding motifs (TFBMs)
- k-mer content
- low complexity DNA?

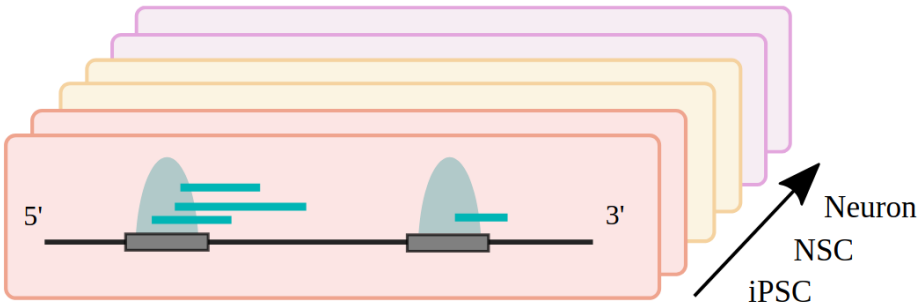


[Horton et al., 2023]

Case study: single cell dataset of neuronal differentiation







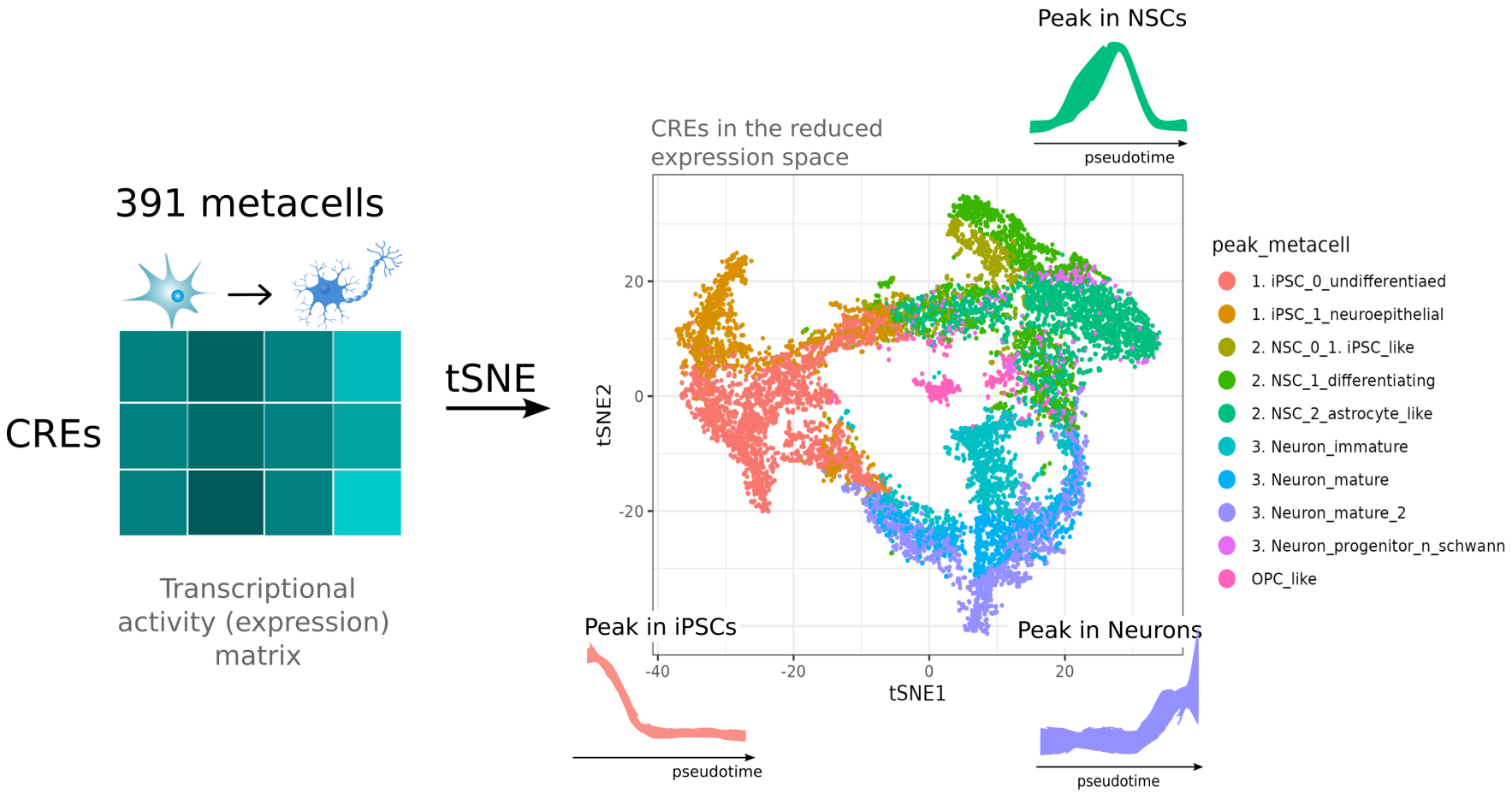
5' end **scRNA-Seq** data & **snATAC-Seq** data

Data from Wallace Yip's lab: Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin

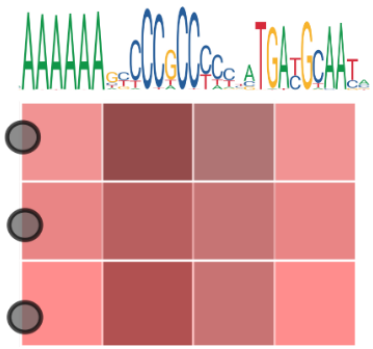
$\rightarrow$ 391 **metacells** inferred by SEACells, and ordered along a differentiation **pseudotime**.

- CRE: ATAC-Seq peak center $\pm$ 500bp
- $N = 10912$ differentially expressed tCREs along differentiation

Representing CRE expression profiles

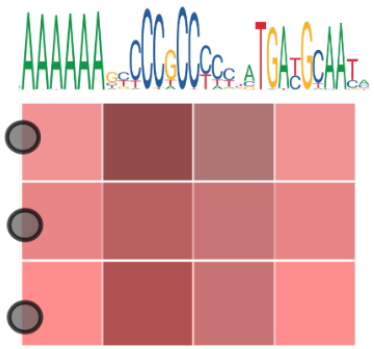


Sequence features of tCREs



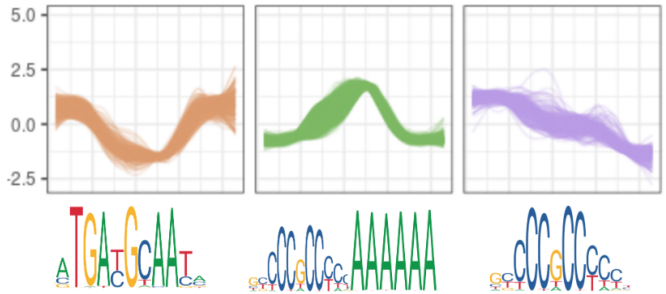
- CRE: ATAC peak center $\pm$ 500bp
- $\sim$ 400 sequences features per CRE: TFBMs scores of expressed TFs from JASPAR 2024 & k-mer frequencies

Sequence features of tCREs



- CRE: ATAC peak center $\pm$ 500bp
- $\sim$ 400 sequences features per CRE: TFBMs scores of expressed TFs from JASPAR 2024 & k-mer frequencies

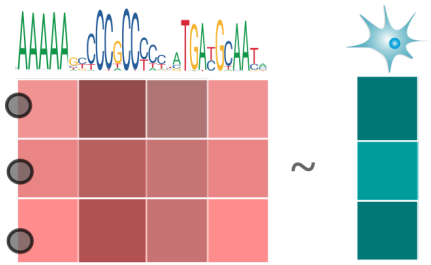
How can we associate CRE sequence features to coordinated activity profiles?



Linking expression profiles to underlying sequence features

- **Supervised models** predicting an expression value from CRE sequence.
  Ex: Enformer, Basenji, Expecto, AI-TAC
  [Avsec et al., 2021, Kelley et al., 2018, Zhou et al., 2018, Maslova et al., 2020]

  → Not well adapted and interpretable for entire expression time-series



Model estimation
(RFs, Deep Learning, linear model, . . .)
↓
Model interpretation to find predictive sequence features

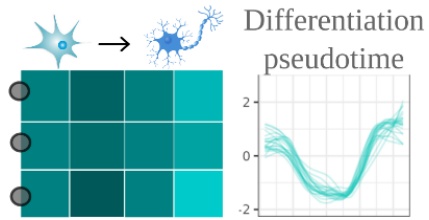
Linking expression profiles to underlying sequence features

- **Supervised models** predicting an expression value from CRE sequence.
  Ex: Enformer, Basenji, Expecto, AI-TAC
  [Avsec et al., 2021, Kelley et al., 2018, Zhou et al., 2018, Maslova et al., 2020]

  → Not well adapted and interpretable for entire expression time-series

- Unsupervised strategies like **motif enrichment analyses** in groups of co-expressed CREs
  Ex: hdWGCNA, cisTopic
  [Morabito et al., 2023, Bravo González-Blas et al., 2019]

  → Sensitive to upstream clustering strategy, K unknown, imperfect correspondence between clusters and sequence, no predictive ability
  [Lajoie et al., 2012]

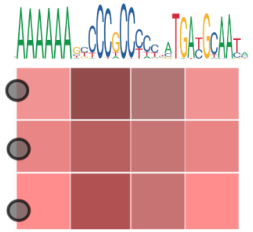


↓

Co-expression clustering (kmeans, dimension reduction, correlation modules . . .)

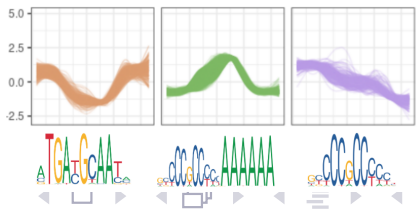
↓

Sequence features enrichment per cluster



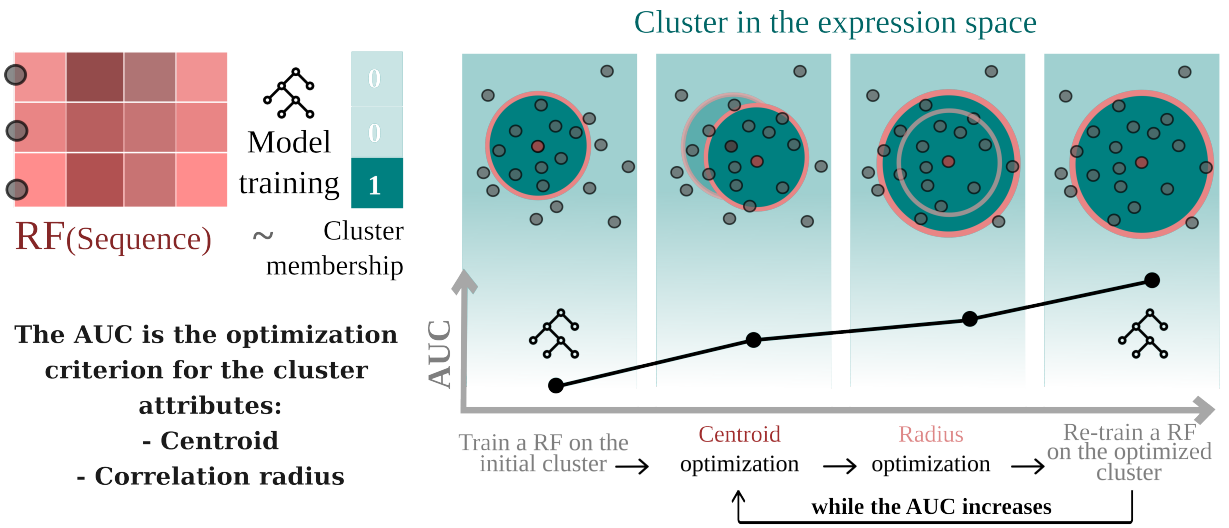
↓

Sequence features & expression profiles associations



Supervised learning To Inform Clustering (STOIC)

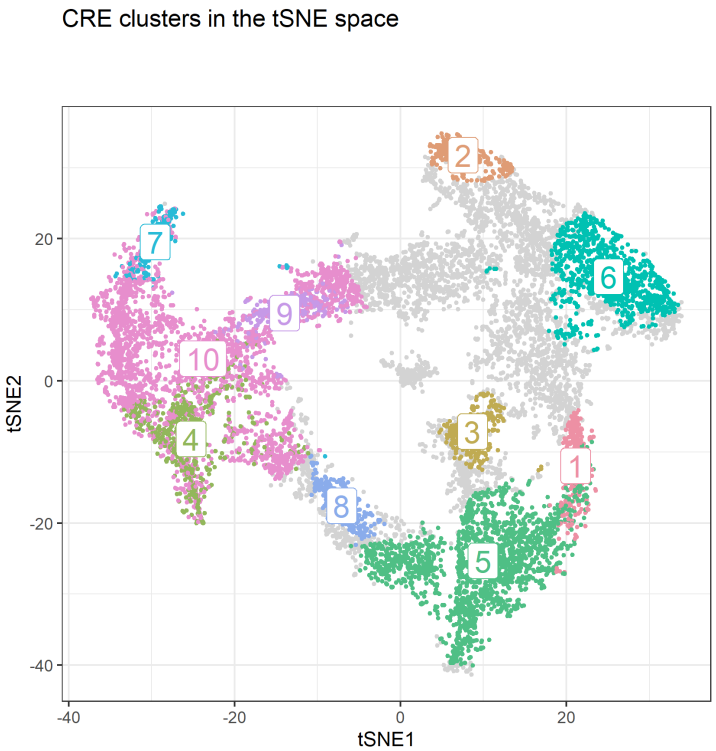
Co-expression clusters are learned to maximize the **AUC of a model using sequence features** to predict cluster expression membership

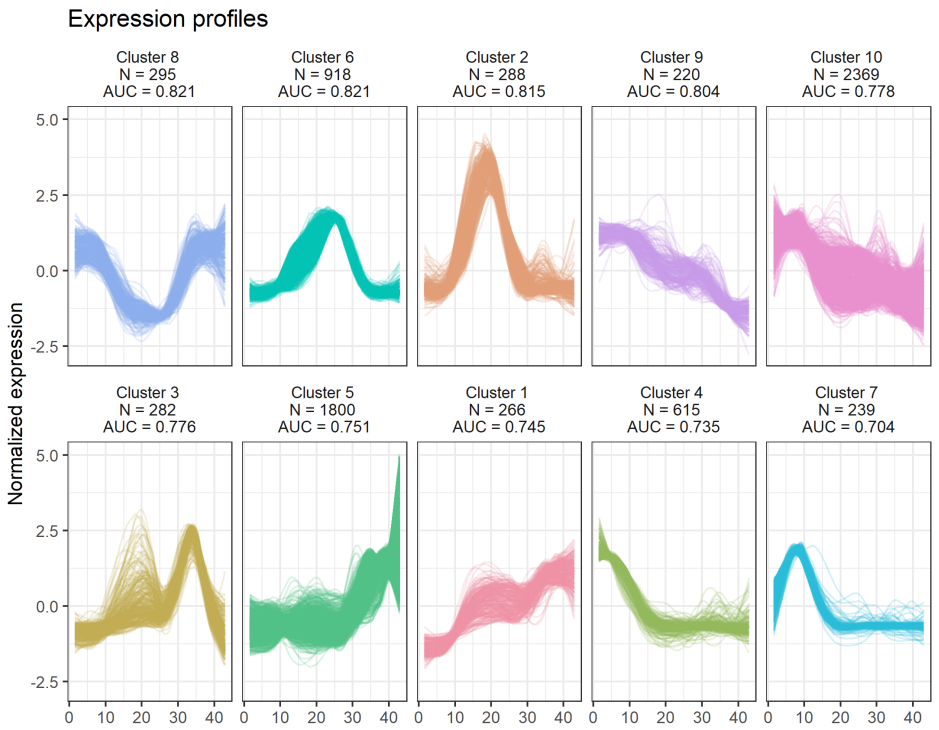


Redundant clusters pruning

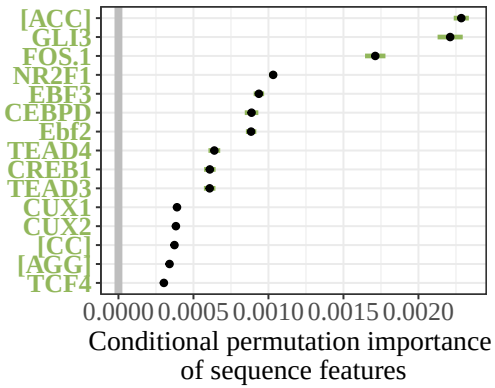
Expression clusters associated with important sequence features

STOIC results on the neuron differentiation study





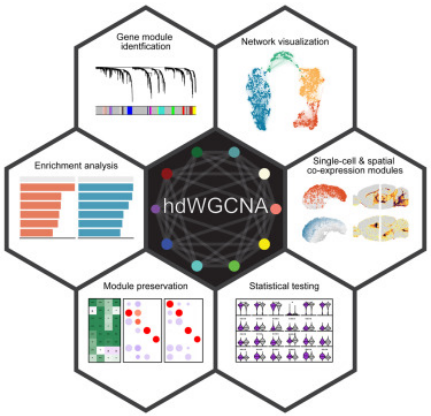
Diverse transcriptional profiles, each associated to distinctive important sequence features



STOIC uncovers strongly associated sequence features

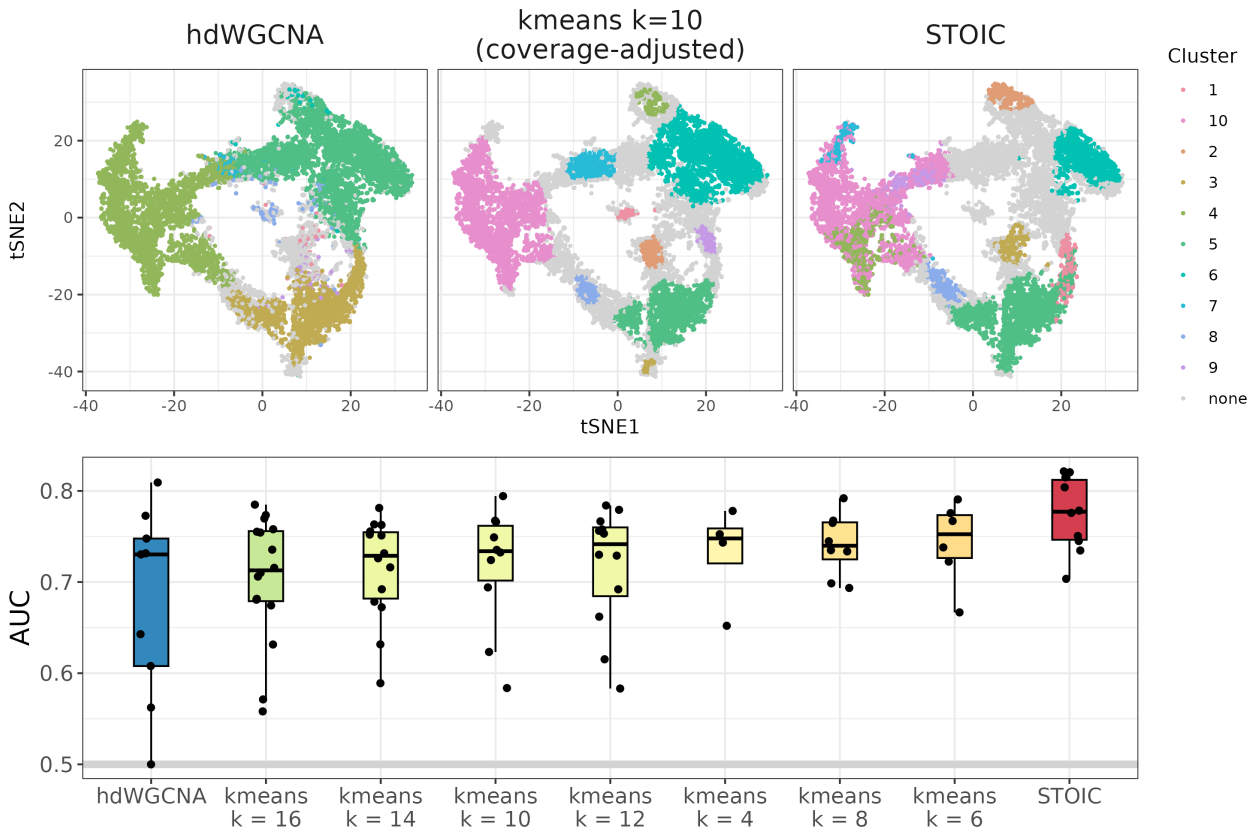
As compared to 2-step approaches:

- **HdWGCNA** (without metacell re-estimation)

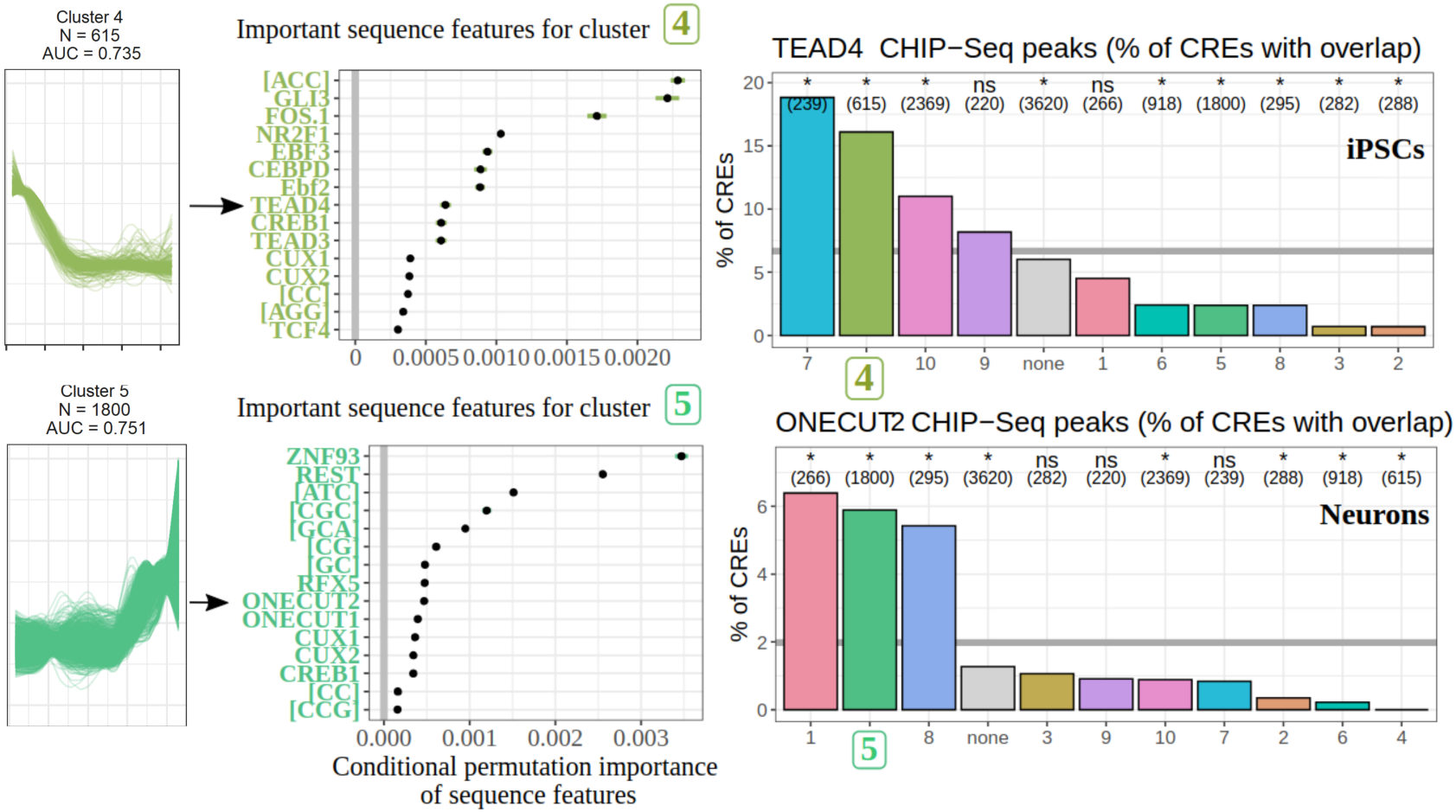


[Morabito et al., 2023] Cell Reports

- **K-means** with same coverage



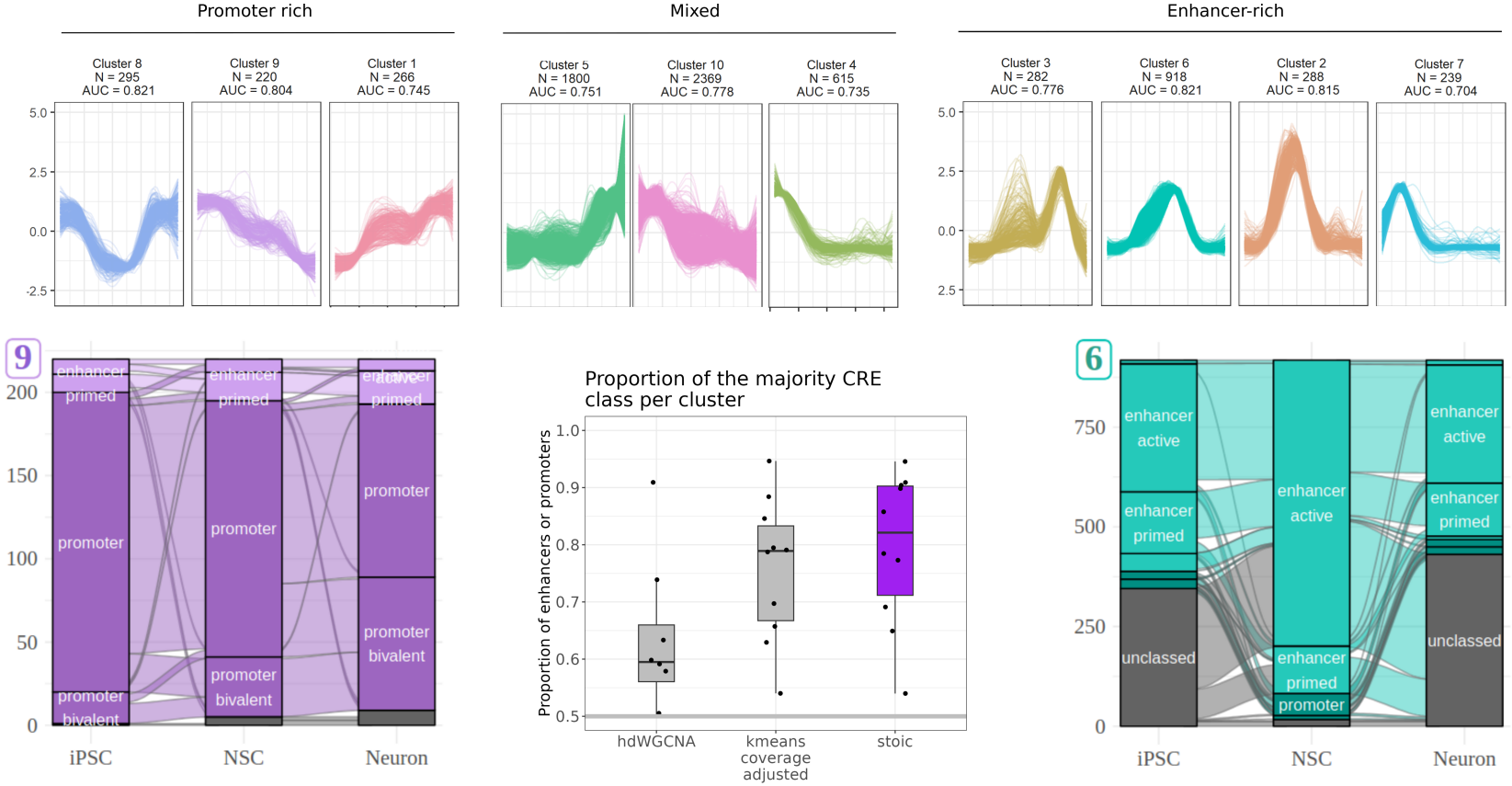
Important sequence features are supported by CHIP-Seq



Data from Wallace Yip's lab: Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin

Homogeneous epigenetic marks within inferred clusters

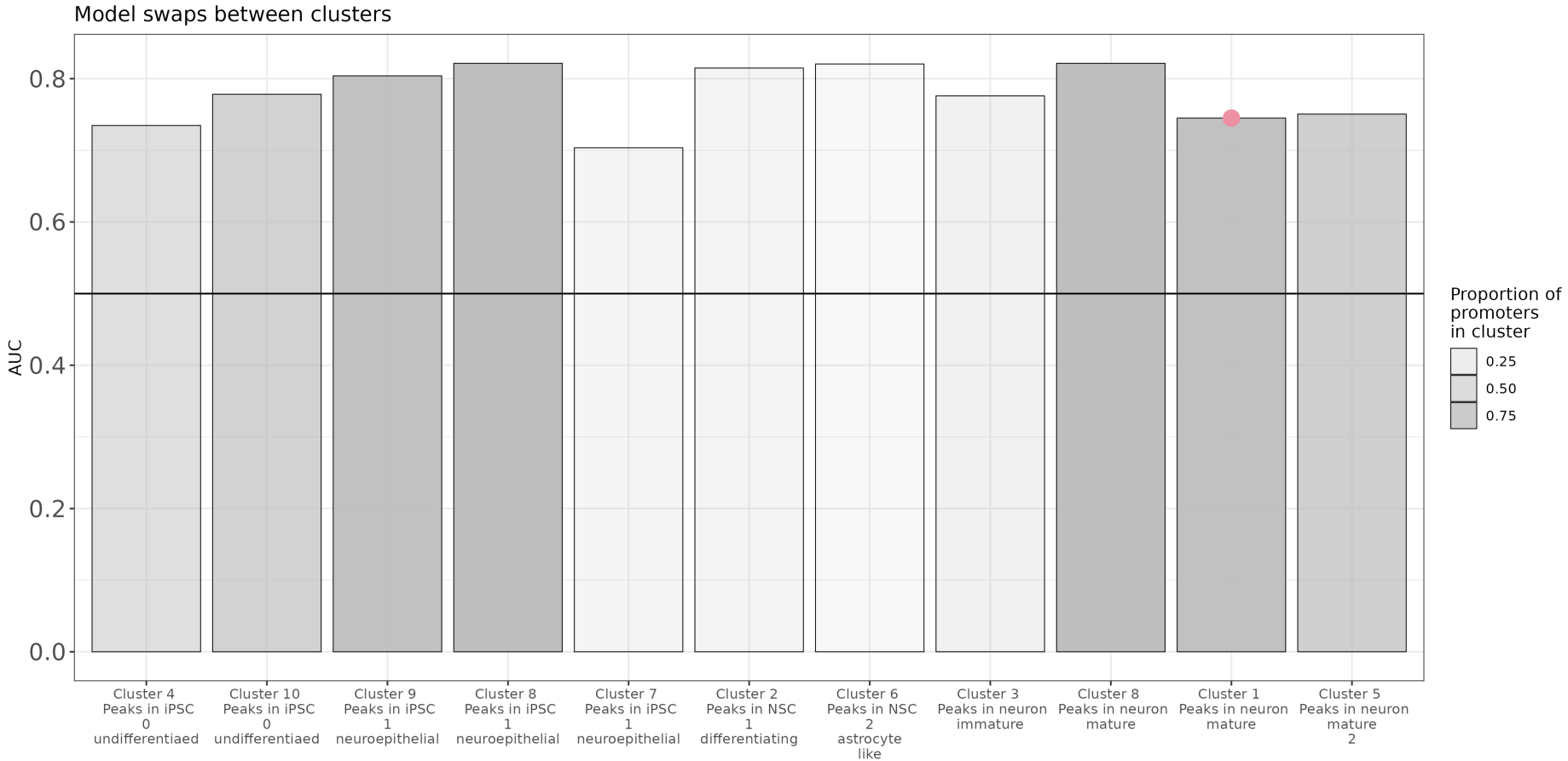
Stoic clusters are very often either **enhancer-rich or promoter-rich**



Derived from chromHMM applied to matched CUT&Tag data from Wallace Yip's lab: Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin

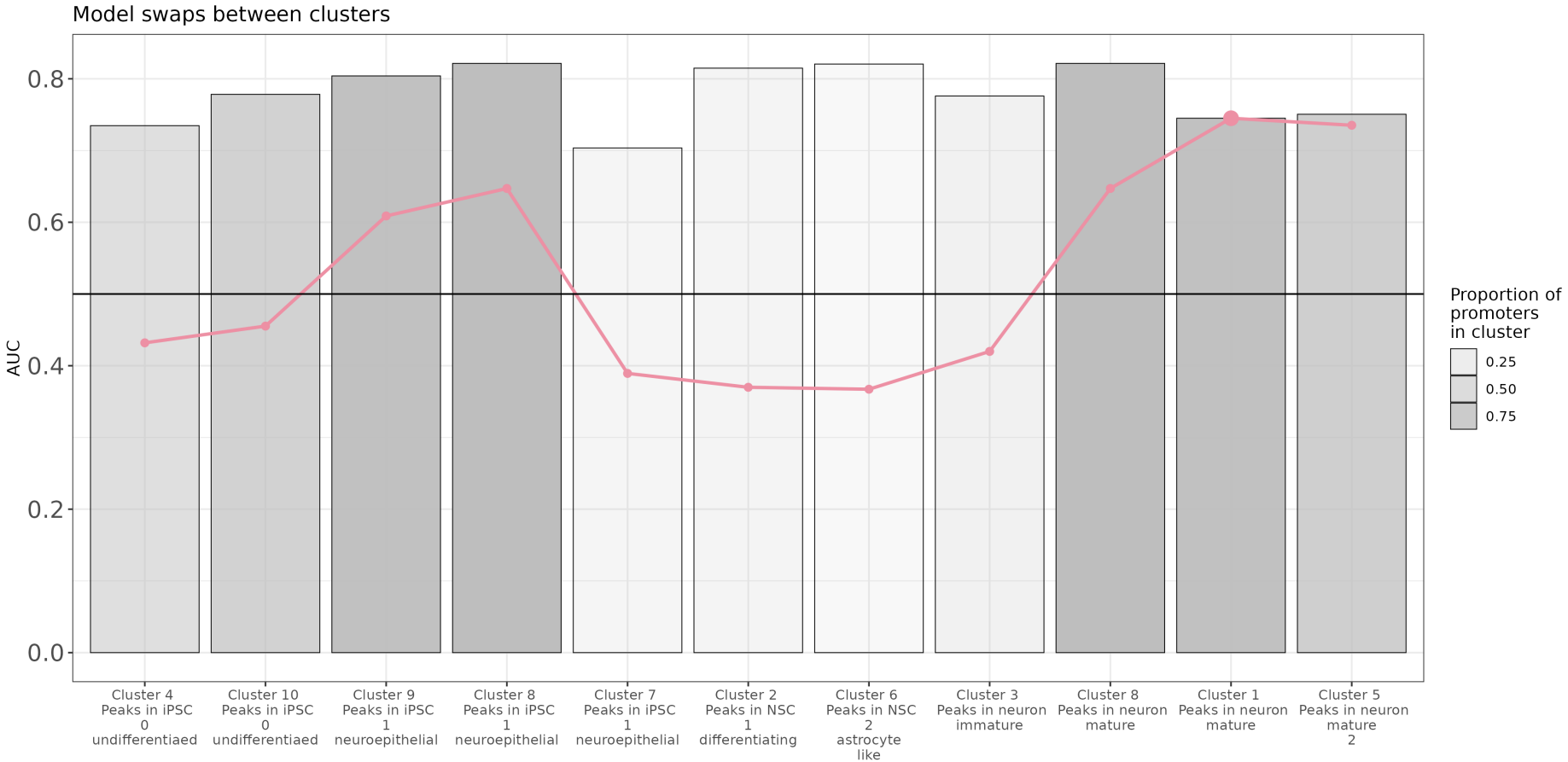
How specific are the learned sequence rules?

Clusters with similar **enhancer-promoter composition** and **expression dynamic** share sequence rules



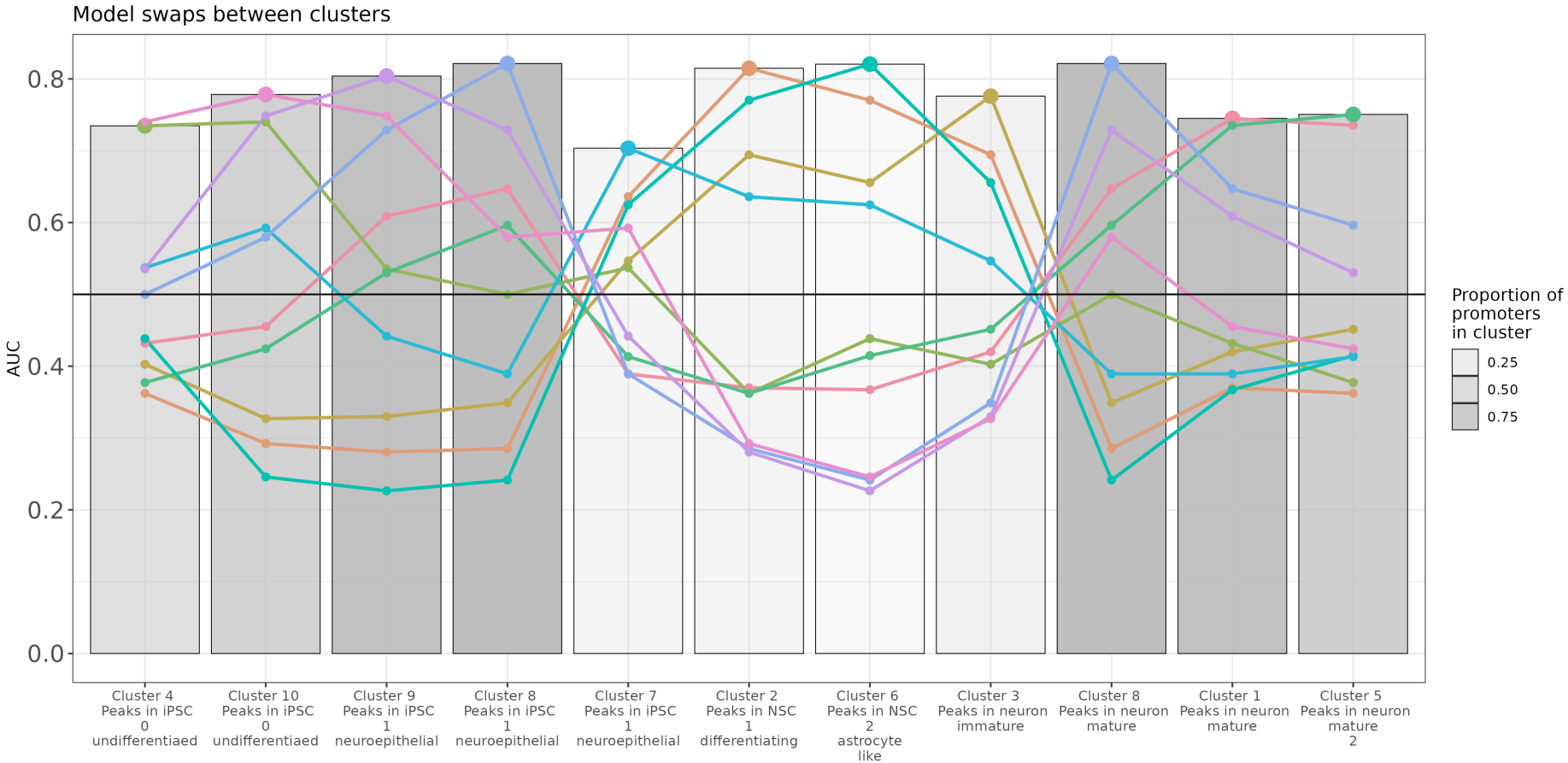
How specific are the learned sequence rules?

Clusters with similar **enhancer-promoter composition** and **expression dynamic** share sequence rules



How specific are the learned sequence rules?

Clusters with similar **enhancer-promoter composition** and **expression dynamic** share sequence rules

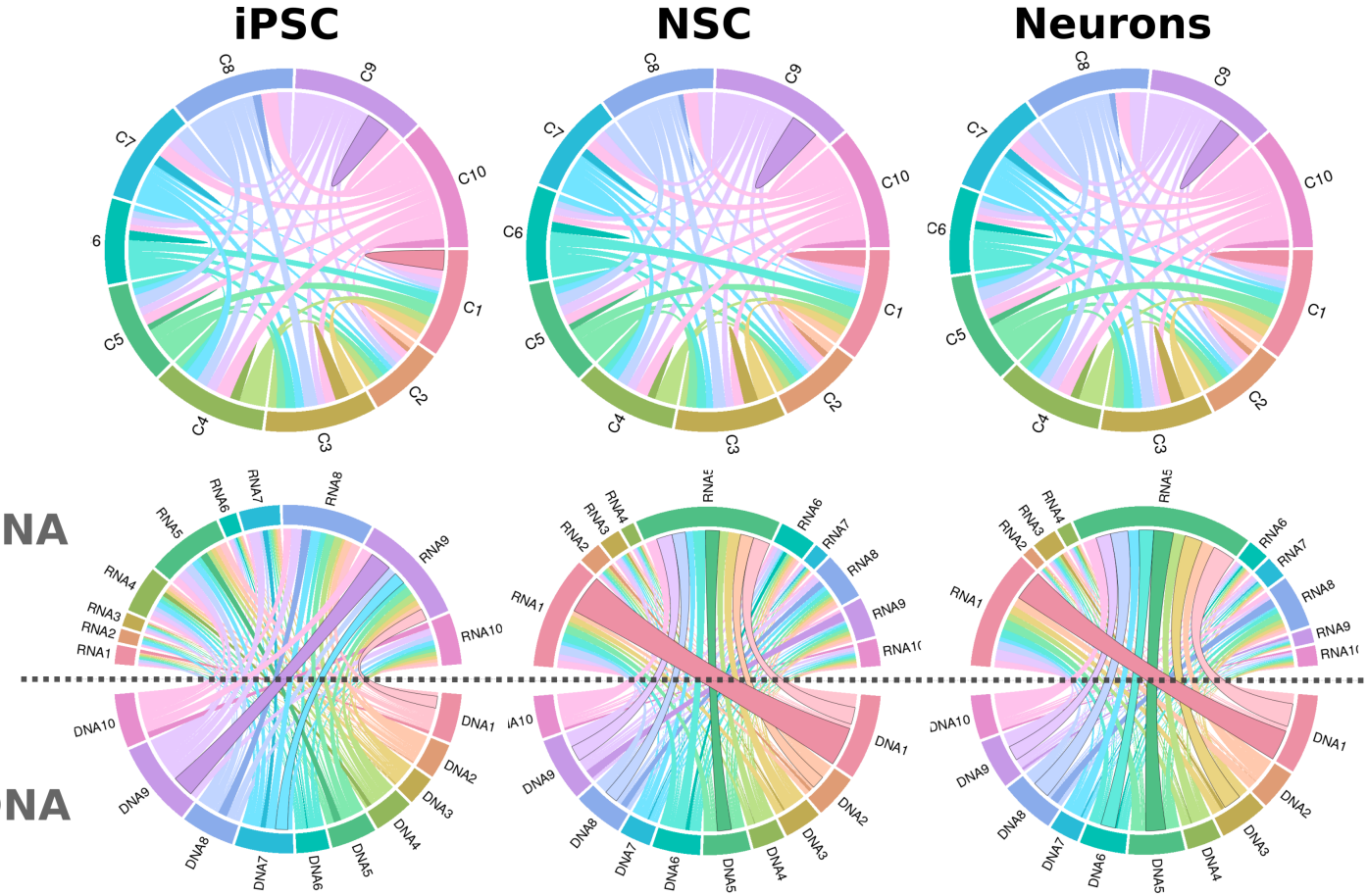


DNA-DNA and RNA-DNA interactions between clusters

Contact rates between clusters (same CRE contacts removed)

Hi-C

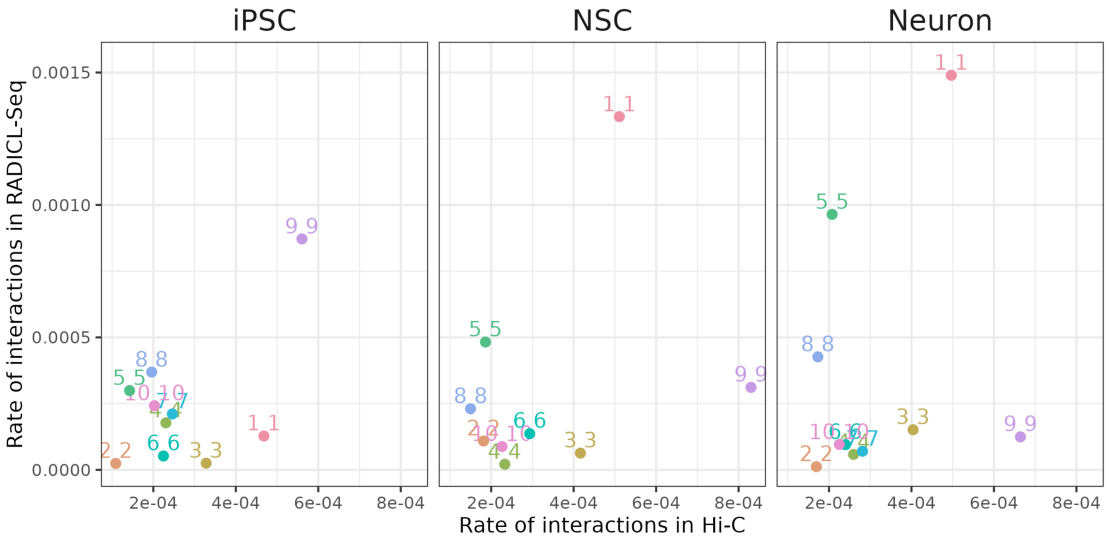
RADICL-Seq

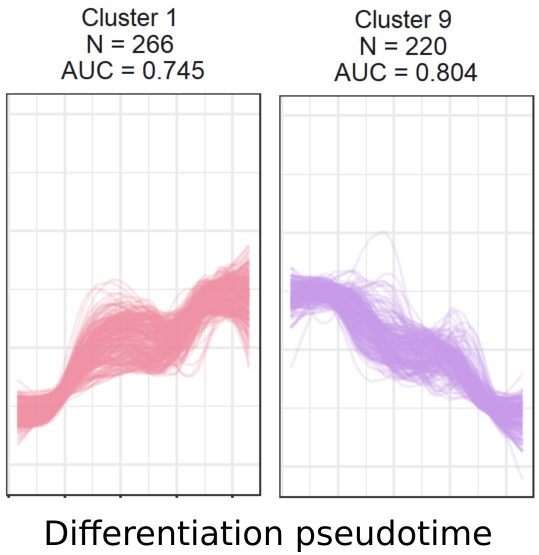


Data from Kayoko Yasuzawa, Lokesh Tripathi, Masaki Kato, Rodi, Wallace Yip's, many others, processed with help from C. Vroland. RADICL-Seq is cell-type specific [Bonetti et al., 2020]

DNA-RNA contacts happen mostly in active promoters

Intra-cluster interactions in promoter clusters seem to support the idea of **transcriptional condensates**





But still many scenarios remain to be explored (enhancer-promoter contacts, non-canonical DNA-RNA structures, lncRNAs . . .)

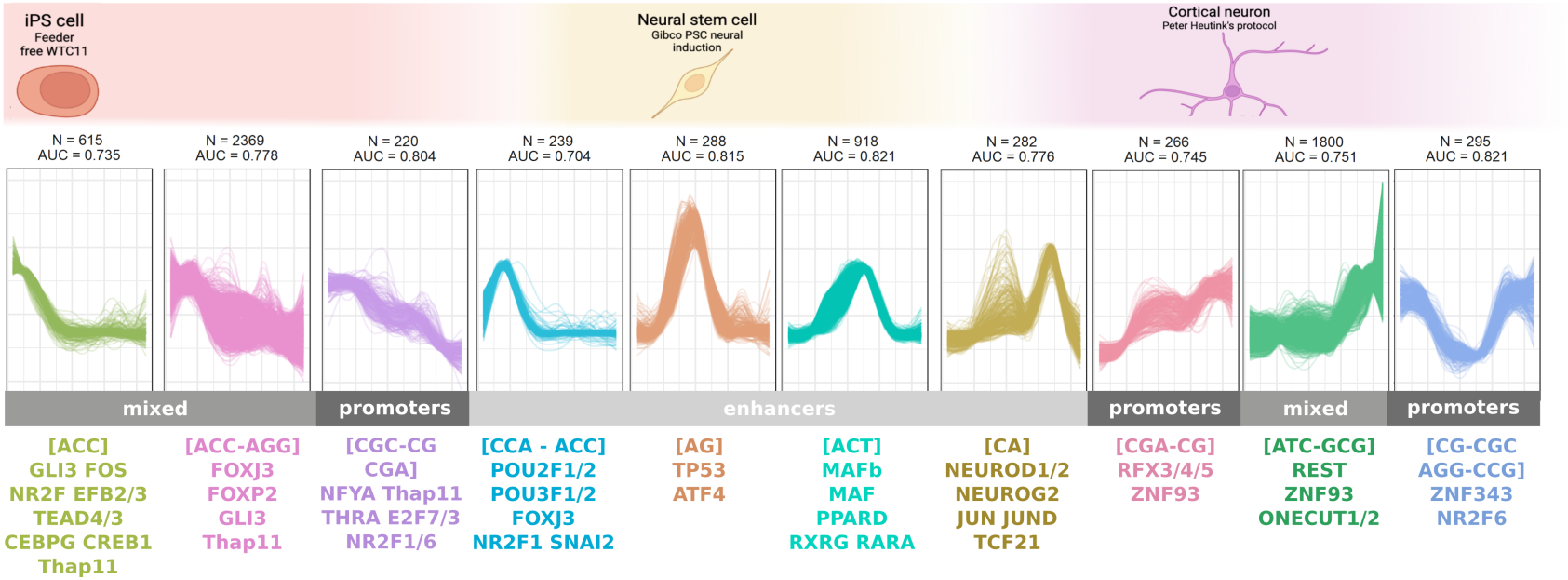
Perspectives

Perspectives for the upcoming article:

- Check the sensitivity of Stoic to the **metacells estimation**
- Apply Stoic to **bulk CAGE kinetics** from FANTOM5: "Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells", Science [Arner et al., 2015]

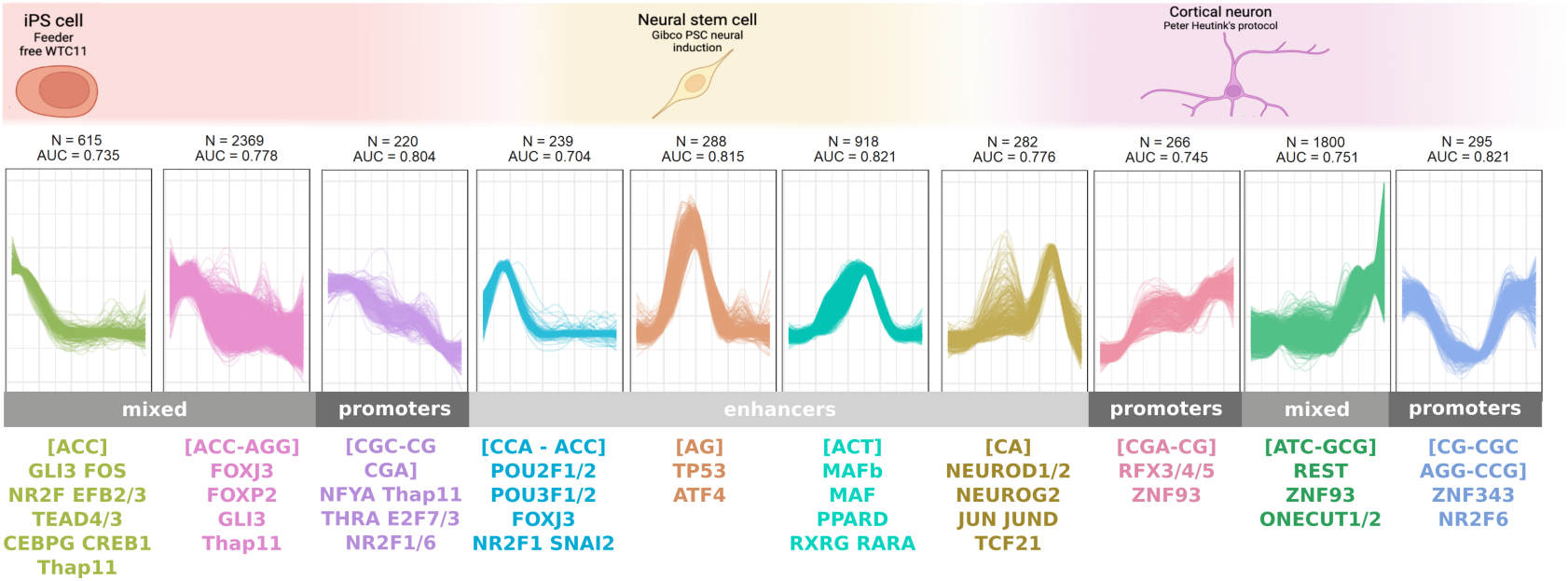
Take home message

STOIC associates a specific expression profile to underlying sequence features which may reflect common and/or **coordinated regulatory processes**



Take home message

STOIC associates a specific expression profile to underlying sequence features which may reflect common and/or **coordinated regulatory processes**



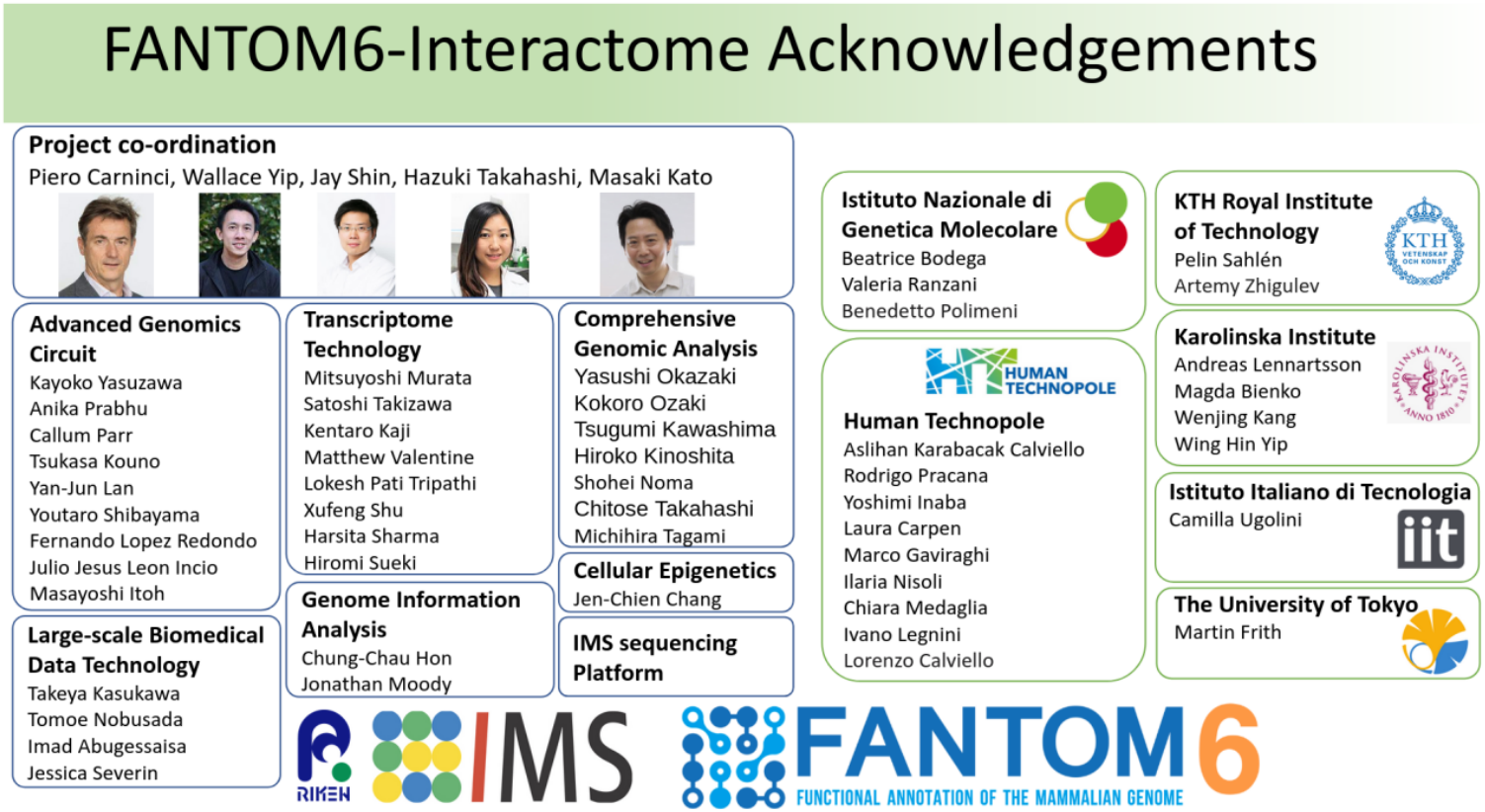
Stoic R package

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")
```

Acknowledgments

- **Wallace Yip** for all the data and help, Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin
- The ML4REGGEN team, Robert Lehmann, Vincenzo Lagani & Vipin Kumar for helpful discussion



References I

- Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., et al. (2015).
  Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells.
  Science, 347(6225):1010–1014.
- Avsec, , Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021).
  Effective gene expression prediction from sequence by integrating long-range interactions.
  Nature Methods, 18(10):1196–1203.
- Bonetti, A., Agostini, F., Suzuki, A. M., Hashimoto, K., Pascarella, G., Gimenez, J., Roos, L., Nash, A. J., Ghilotti, M., Cameron, C. J., et al. (2020).
  Radicl-seq identifies general and cell type–specific principles of genome-wide rna-chromatin interactions.
  Nature Communications, 11(1):1018.

References II

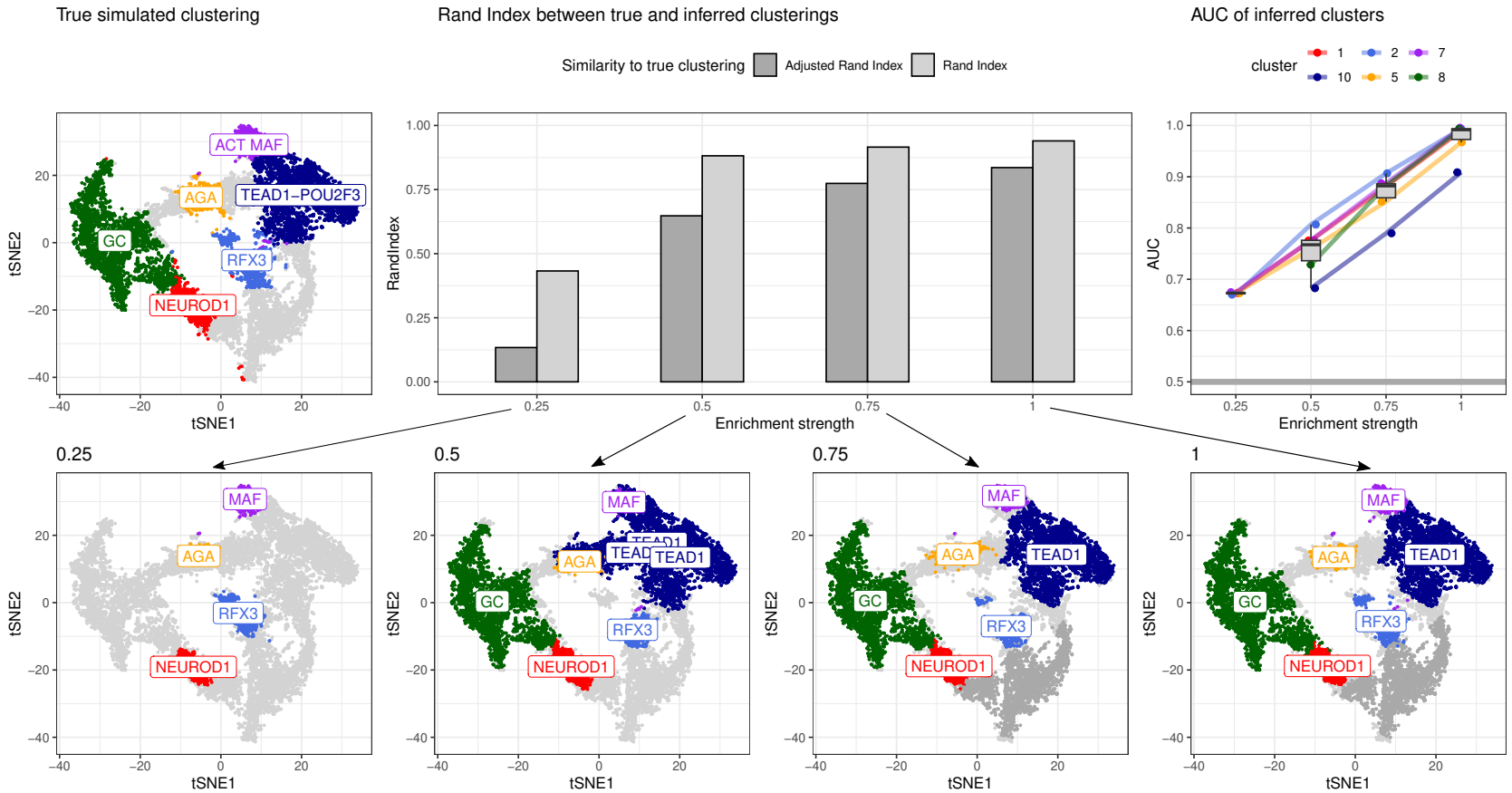
- Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019).
  cistopic: cis-regulatory topic modeling on single-cell atac-seq data.
  Nature Methods, 16(5):397–400.
- Horton et al. (2023).
  Short tandem repeats bind transcription factors to tune eukaryotic gene expression.
  Science, 381(6664):eadd1250.
- Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. (2018).
  Sequential regulatory activity prediction across chromosomes with convolutional neural networks.
  Genome research, 28(5):739–750.
- Lajoie, M. et al. (2012).
  Computational discovery of regulatory elements in a continuous expression space.
  Genome biology, 13:1–17.

References III

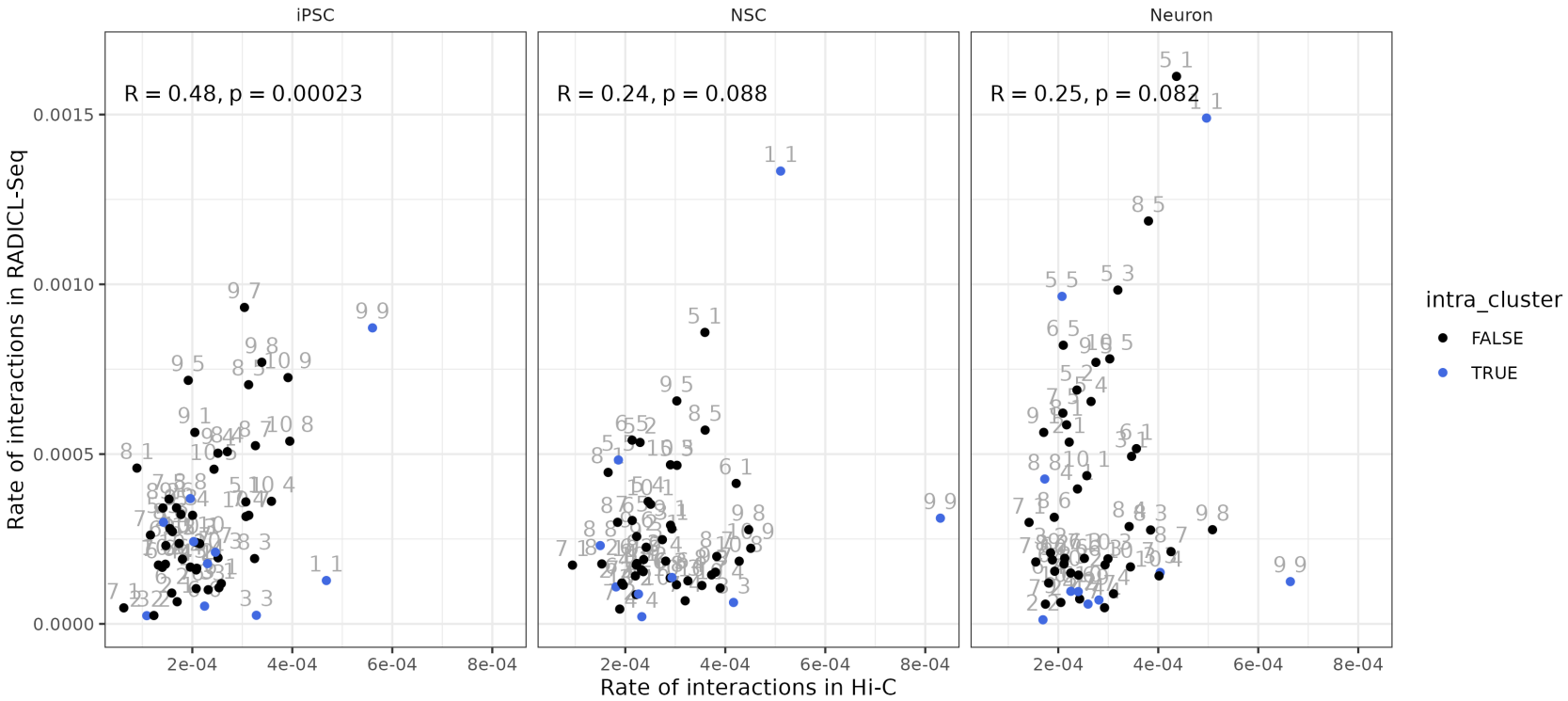
- Maslova, A., Ramirez, R. N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., and Mostafavi, S. (2020).
  Deep learning of immune cell differentiation.
  Proceedings of the National Academy of Sciences, 117(41):25655–25666.
- Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., and Swarup, V. (2023).
  hdwgcna identifies co-expression networks in high-dimensional transcriptomics data.
  Cell Reports Methods, 3(6):100498.
- Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018).
  Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk.
  Nature genetics, 50(8):1171–1179.

Results controlled on simulations

STOIC recovers **artificially enriched sequence features** in co-expression clusters. Its performances increase with enrichment intensity.

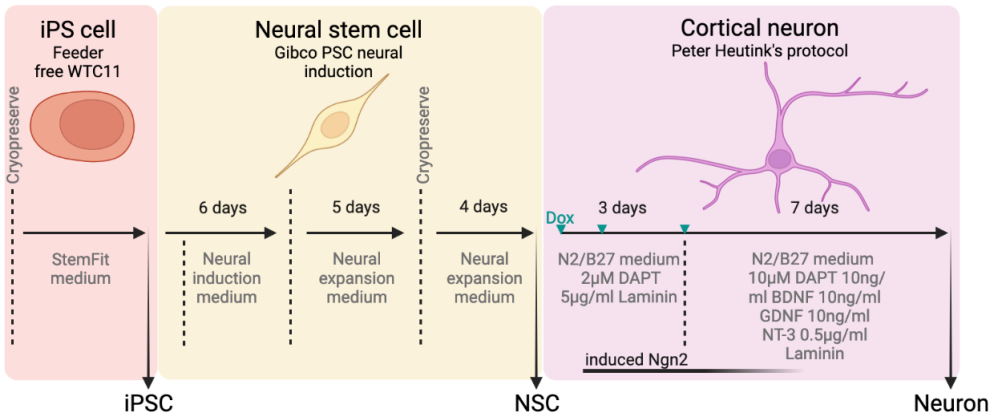


DNA-DNA contacts between inferred clusters are less dynamic than DNA-RNA contacts

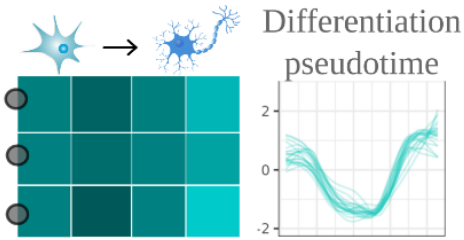


Data from Wallace Yip's, Kayoko Yasuzawa, Lokesh Tripathi, Masaki Kato, Rodi, many other, processed with help from C. Vroland

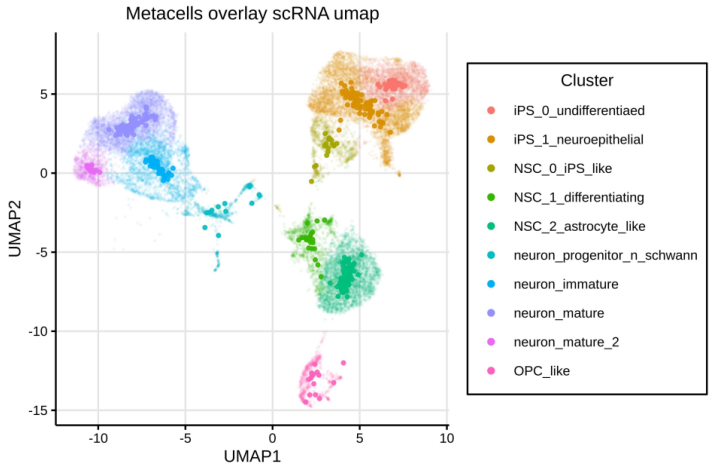
Case study: single cell dataset of neuronal differentiation

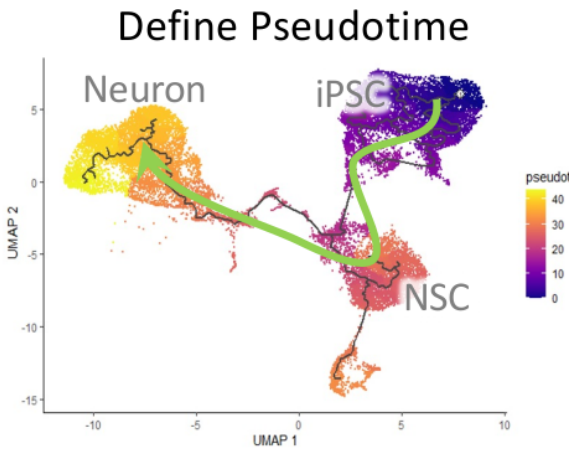


5' end **scRNA-Seq** data & **snATAC-Seq** data



Expression in the 391 **metacells** inferred by SEACells, ordered along the differentiation pseudotime.





Data from Wallace Yip's lab: Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin

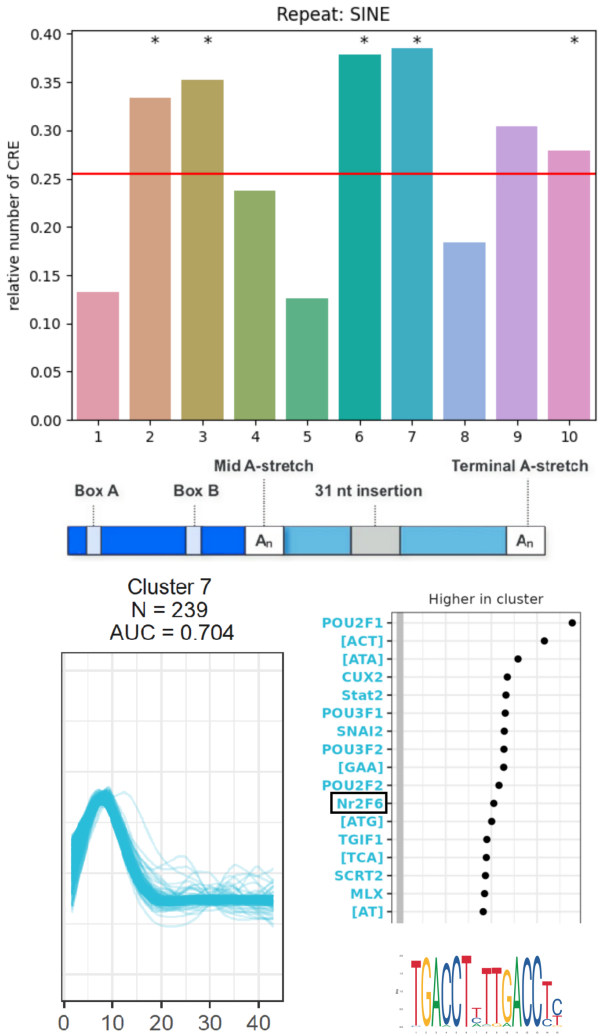
Further interpretations of STOIC models

Additional biological validations:

- TF importance reflects binding affinity for CREs (REMAP)
- Enrichment of GO terms linked to neurological development

Understanding of the cis-regulatory code:

- **SINEs Repeat elements** like AluS and J are enriched in certain enhancer clusters, and co-localize with important TFBMs
- Alteration of important TFBMs by clinically relevant **variants**, enrichment of molecular QTLs within our clusters



Co-occurrence via Poisson binomial tests
Analyses by Julien Raynal.