vocado

Deep tensor factorization characterizes the human epigenome through imputation of thousands of epigenomic and transcriptomic experiments

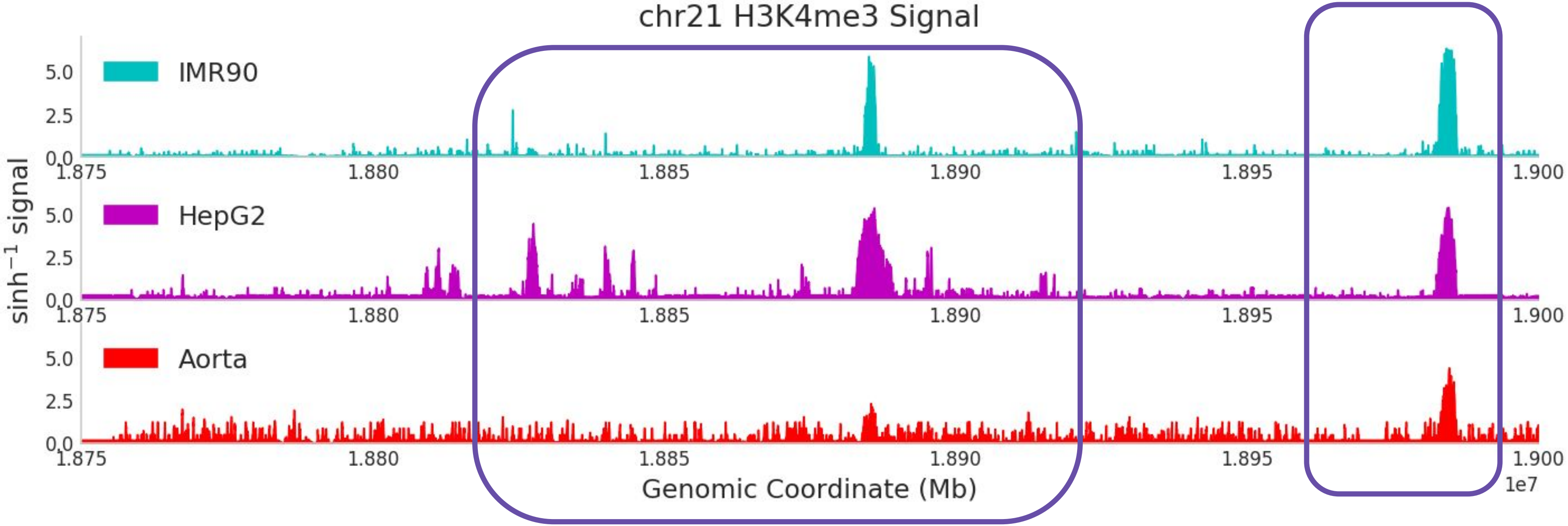
Jacob Schreiber
Paul G. Allen School of Computer Science and Engineering
University of Washington

in jmschreiber91

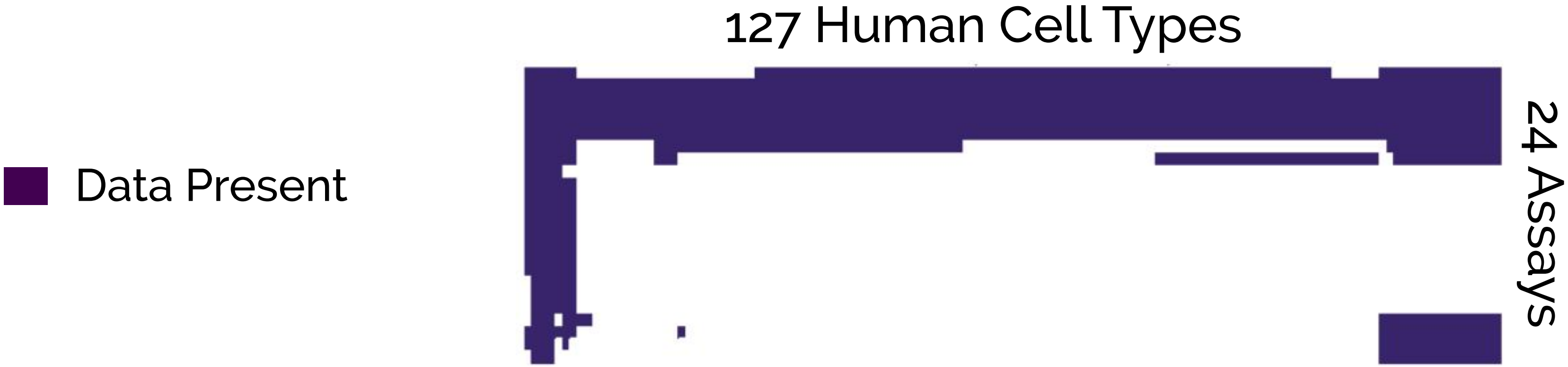
@jmschrei

@jmschreiber91

The signal of epigenomic assays vary across cell types



The Roadmap Compendium includes over a thousand epigenomic experiments

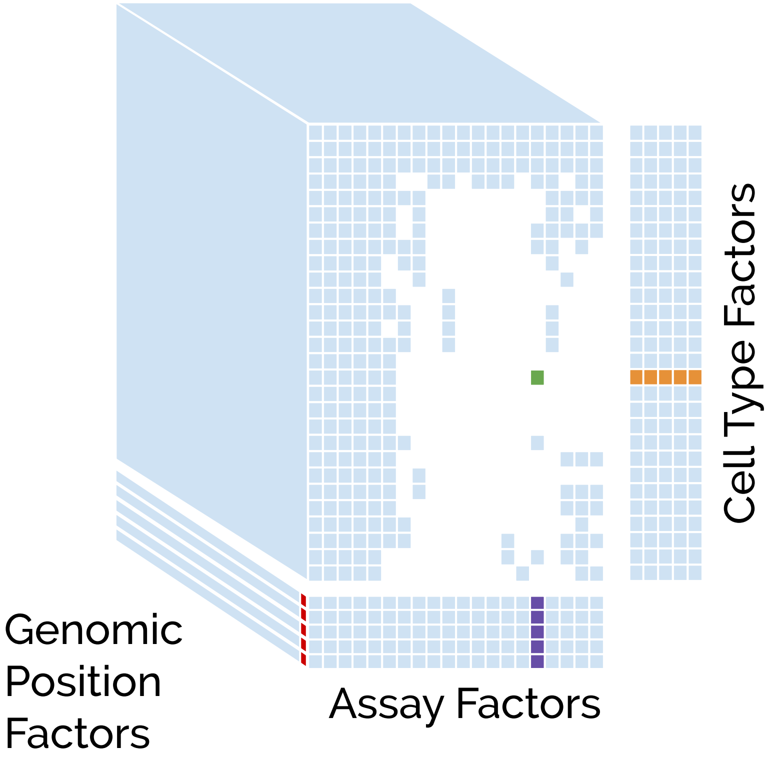


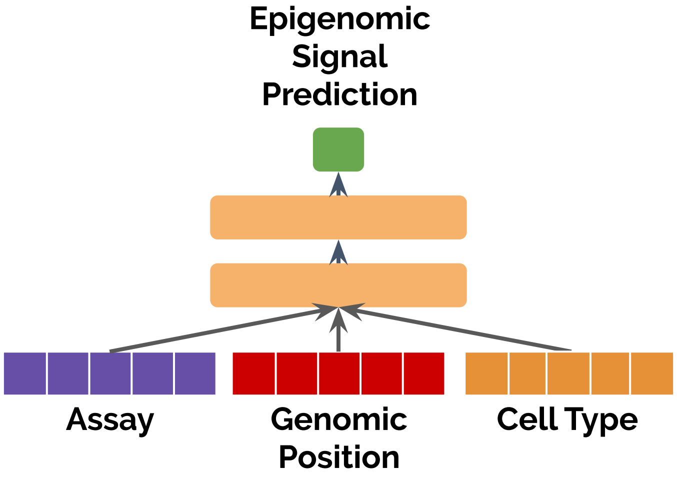
Unfortunately the Roadmap compendium is incomplete. Previous work sought to fill in the matrix through imputing all potential experiments (ChromImpute[1], PREDICTD[2])

[1] Ernst, et al. *Nature Methods, 2015*

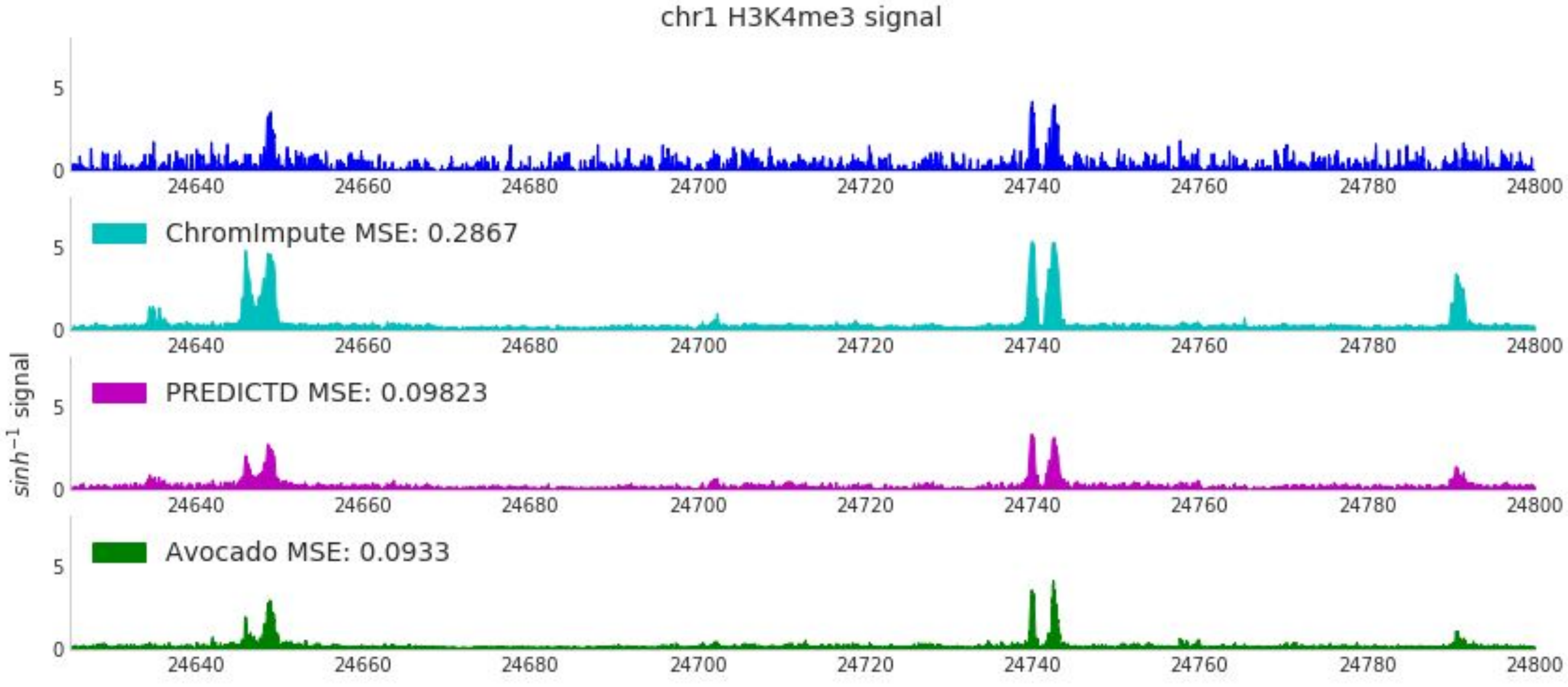
[2] Durham, et al. *Nature Communications, 2018*

Avocado is a deep tensor factorization approach



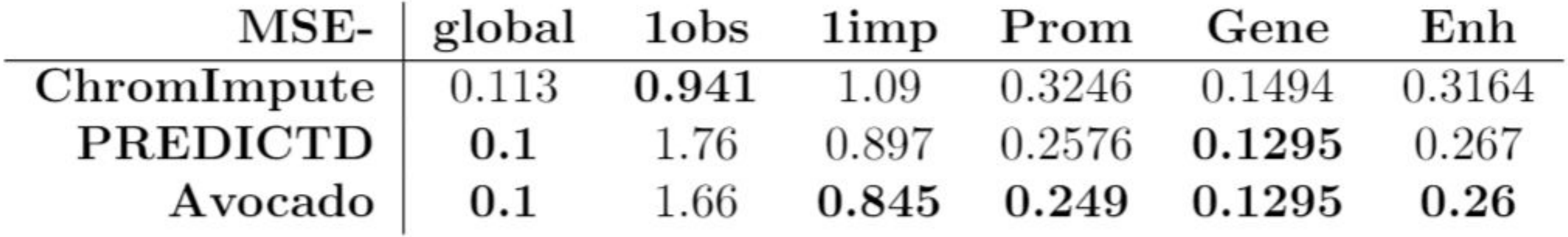


Initial inspection of the imputations suggest that Avocado performs well



Genomic Coordinate (kb)

Avocado performs well genome-wide



MSE-	global	1obs	1imp	Prom	Gene	Enh
ChromImpute	0.113	**0.941**	1.09	0.3246	0.1494	0.3164
PREDICTD	**0.1**	1.76	0.897	0.2576	**0.1295**	0.267
Avocado	**0.1**	1.66	**0.845**	**0.249**	**0.1295**	**0.26**

MSE-global: Mean squared error (MSE) across the full length of the genome

MSE-1obs: MSE at the top 1% of genomic positions ranked by experimental signal

MSE-1imp: MSE at the top 1% of genomic positions ranked by imputed signal

MSE-Prom: MSE at promoter regions defined by GENCODE

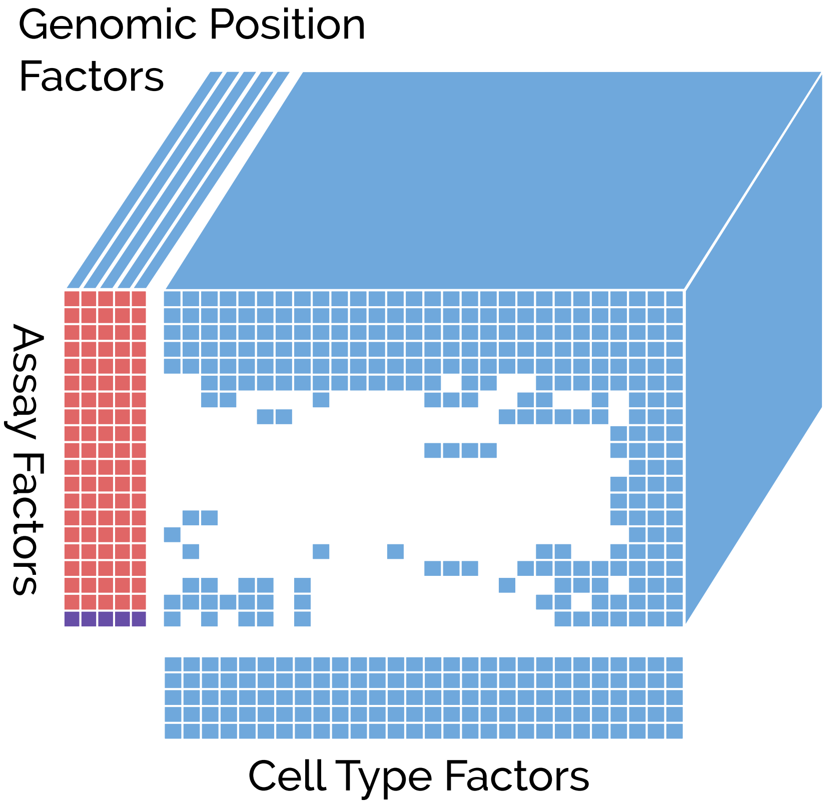
MSE-Gene: MSE at gene bodies defined by GENCODE

MSE-Enh: MSE at enhancer regions defined by FANTOM5

Okay, so have we characterized human epigenomics now?

Histone Modification ChIP-seq

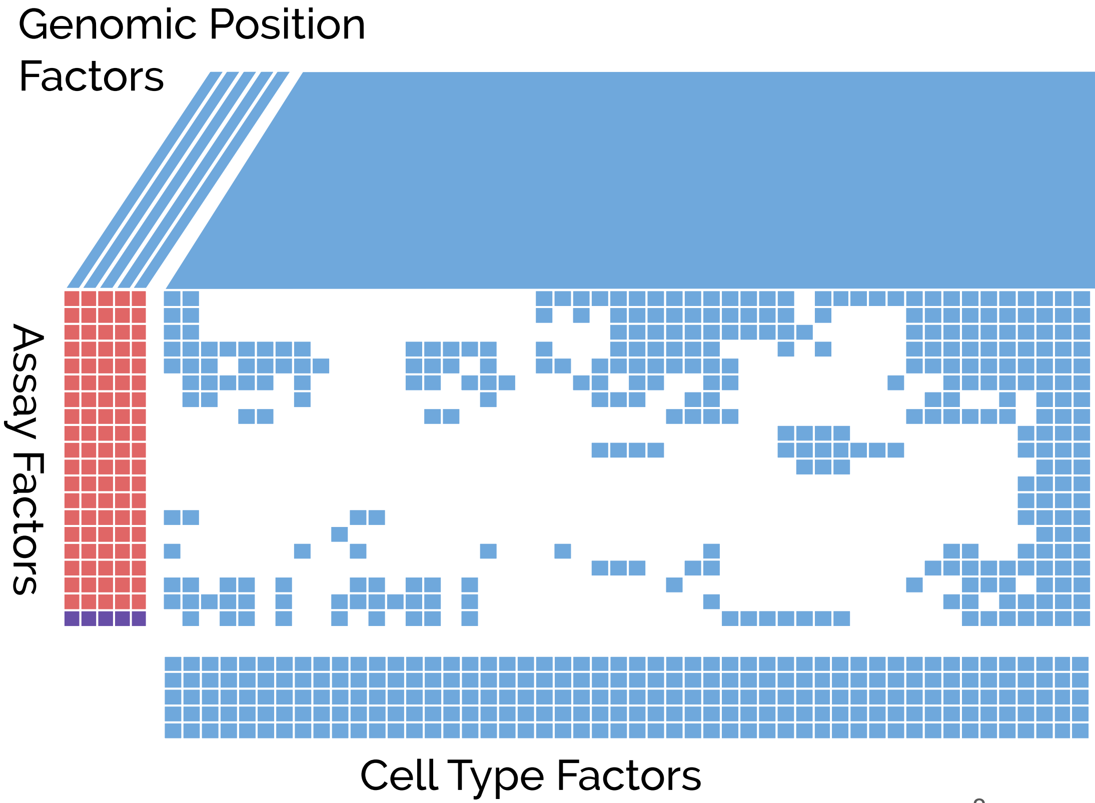
Chromatin Accessibility



The ENCODE compendium has more biosamples...

# Cell Types: from 127 to 400

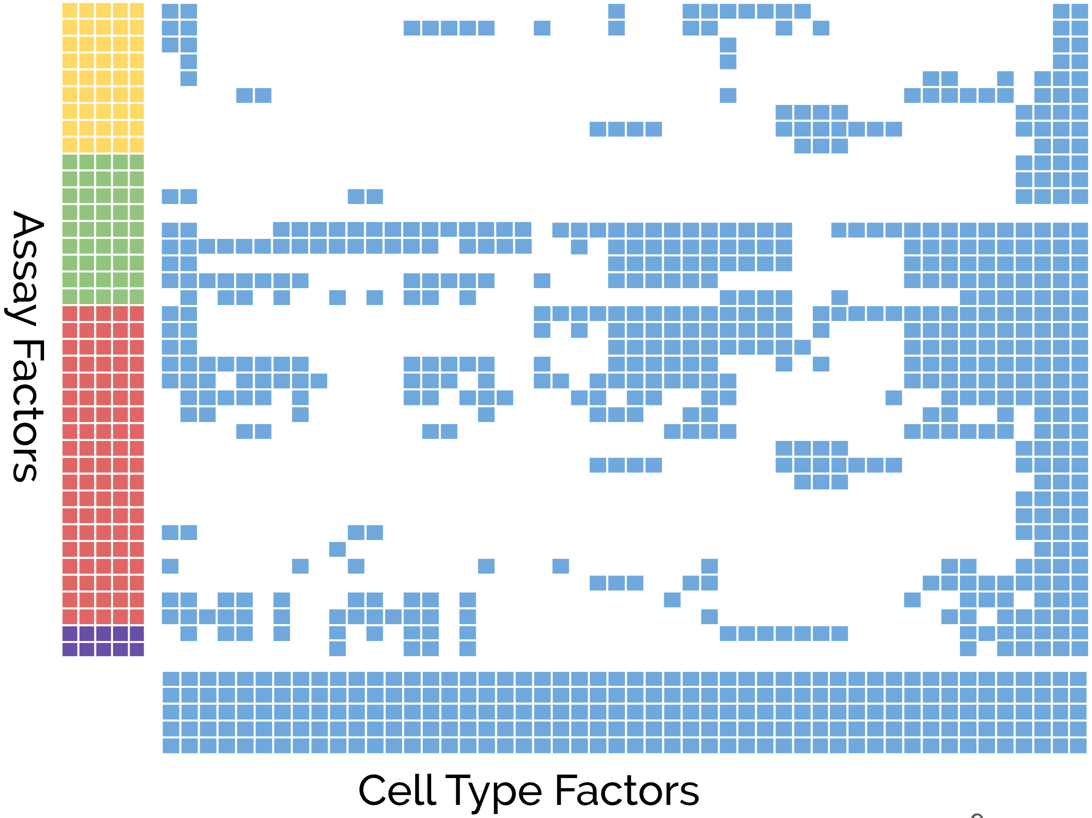
Histone Modification ChIP-seq
Chromatin Accessibility



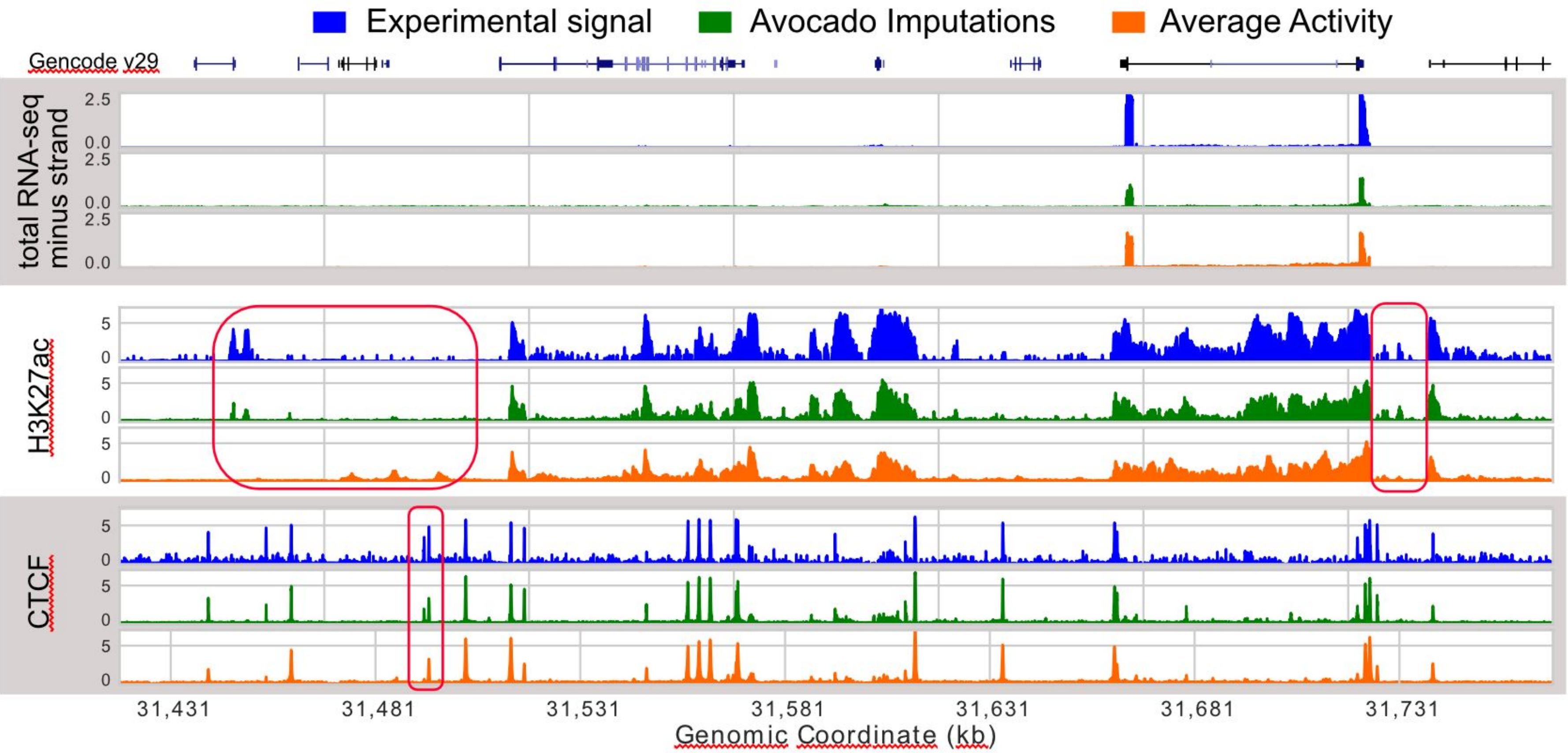
… and more assays

# Cell Types: from 127 to 400
# Assays: from 24 to 76

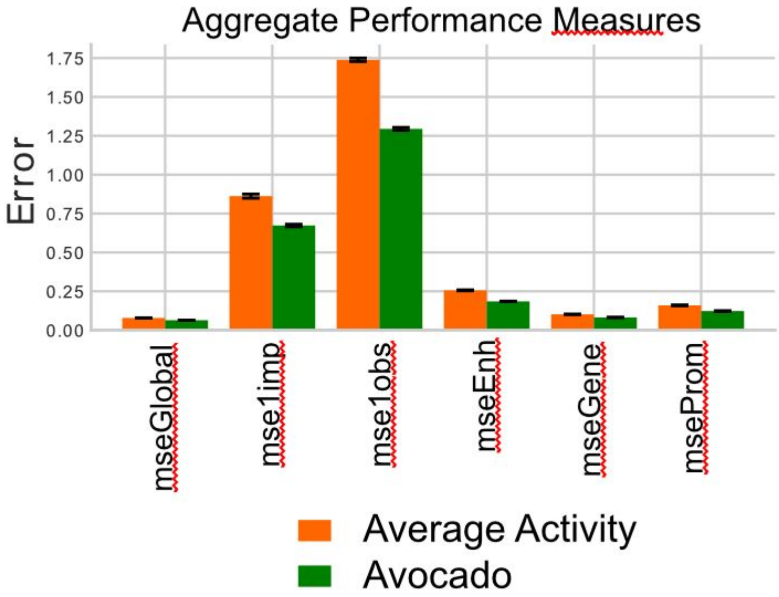
Histone Modification ChIP-seq
Chromatin Accessibility
Gene Transcription
Transcription Factor ChIP-seq

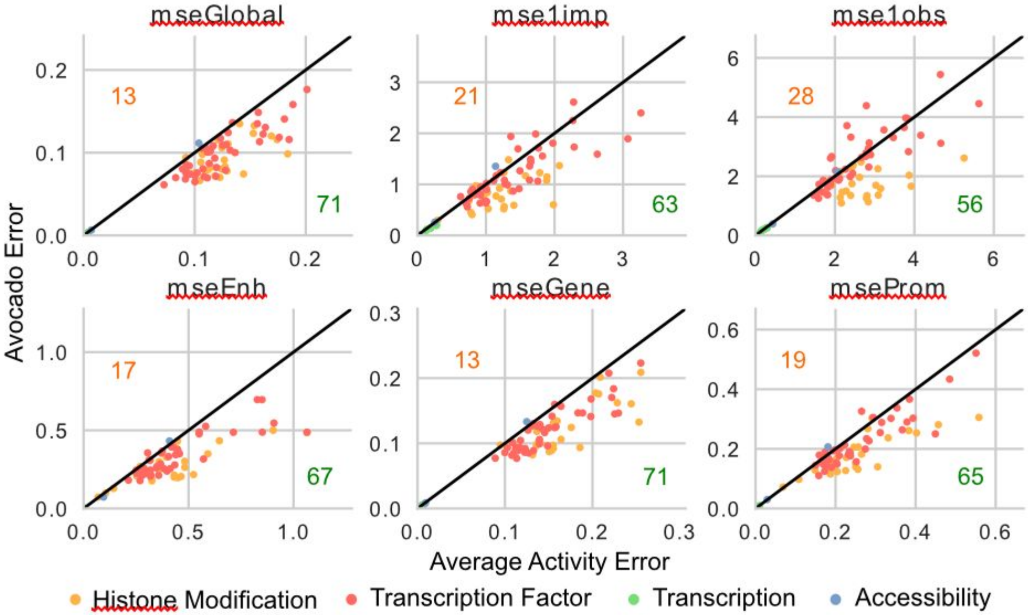


Avocado can jointly model many forms of activity

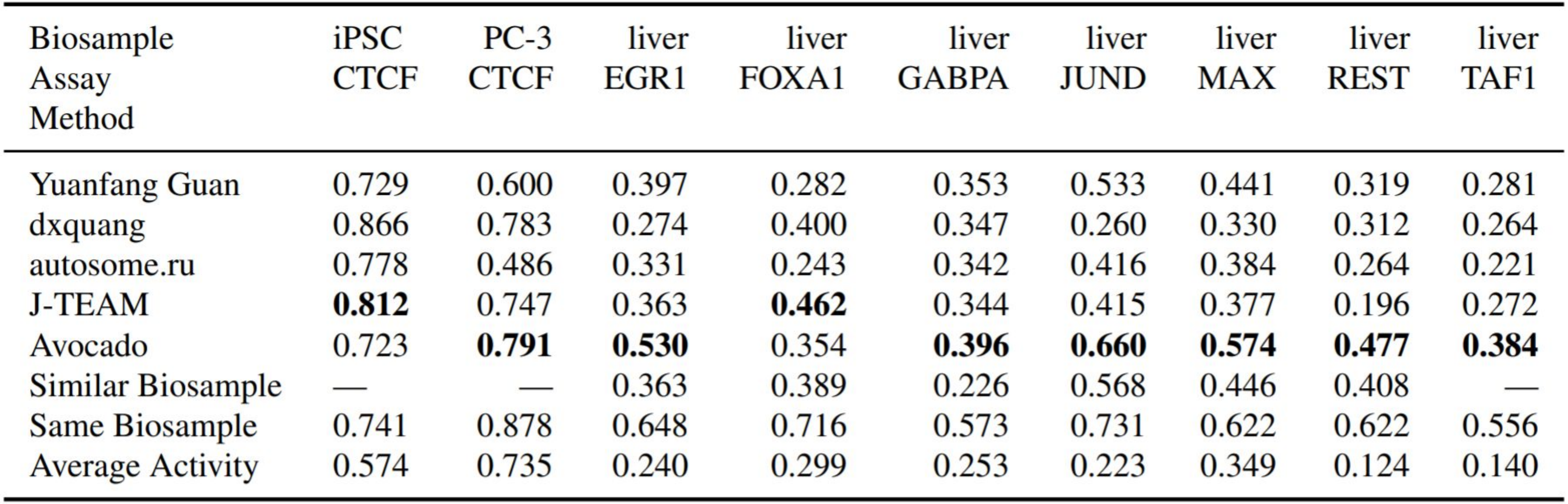


Avocado’s imputations are of high accuracy





Avocado imputes TF binding better than the participants in the ENCODE-DREAM challenge*



Biosample	iPSC	PC-3	liver	liver	liver	liver	liver	liver	liver
Assay	CTCF	CTCF	EGR1	FOXA1	GABPA	JUND	MAX	REST	TAF1
Method									
Yuanfang Guan	0.729	0.600	0.397	0.282	0.353	0.533	0.441	0.319	0.281
dxquang	0.866	0.783	0.274	0.400	0.347	0.260	0.330	0.312	0.264
autosome.ru	0.778	0.486	0.331	0.243	0.342	0.416	0.384	0.264	0.221
J-TEAM	**0.812**	0.747	0.363	**0.462**	0.344	0.415	0.377	0.196	0.272
Avocado	0.723	**0.791**	**0.530**	0.354	**0.396**	**0.660**	**0.574**	**0.477**	**0.384**
Similar Biosample	—	—	0.363	0.389	0.226	0.568	0.446	0.408	—
Same Biosample	0.741	0.878	0.648	0.716	0.573	0.731	0.622	0.622	0.556
Average Activity	0.574	0.735	0.240	0.299	0.253	0.223	0.349	0.124	0.140

Performance metric is auPR (average precision)

* read about the caveats in our preprint

Okay so now have we fully characterized human epigenomics?

No; the ENCODE compendium does not include hundreds of protein binding assays or a number of cell states, diseases, and mutations.

However:

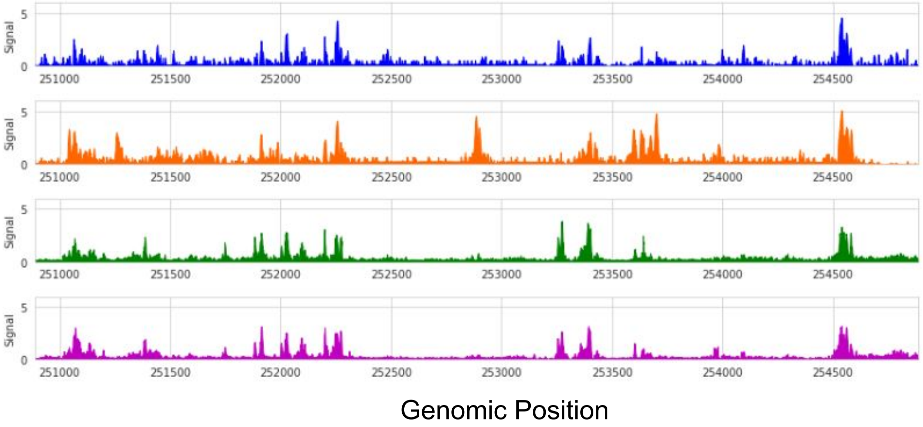
- New biosamples and assays can be added to a pre-trained model with as little as a single experiment

- We are exploring zero-shot imputation approaches that precalculate assay embeddings using protein similarity and interaction networks

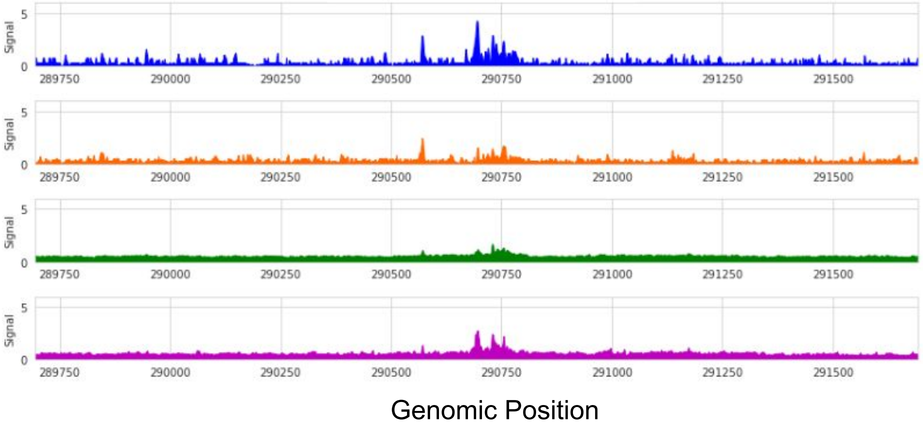
Leveraging the large amount of human data enables zero-shot imputation of TF binding across species

```
Average Activity:                        0.09677
Mouse + 3,814 Human Experiments:  0.09252
Mouse + 6,870 Human Experiments:  0.08570
```

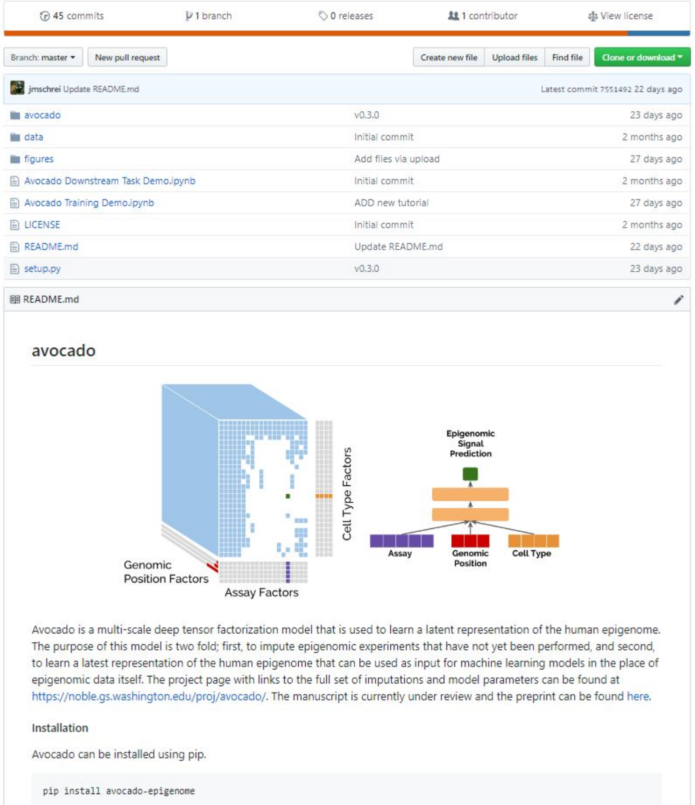
ELF1



MAX



GitHub repo, pretrained models, and preprints online!



jmschrei Update README.md — Latest commit 7551492 22 days ago

avocado	v0.3.0	23 days ago
data	Initial commit	2 months ago
figures	Add files via upload	27 days ago
Avocado Downstream Task Demo.ipynb	Initial commit	2 months ago
Avocado Training Demo.ipynb	ADD new tutorial	27 days ago
LICENSE	Initial commit	2 months ago
README.md	Update README.md	22 days ago
setup.py	v0.3.0	23 days ago

README.md

avocado

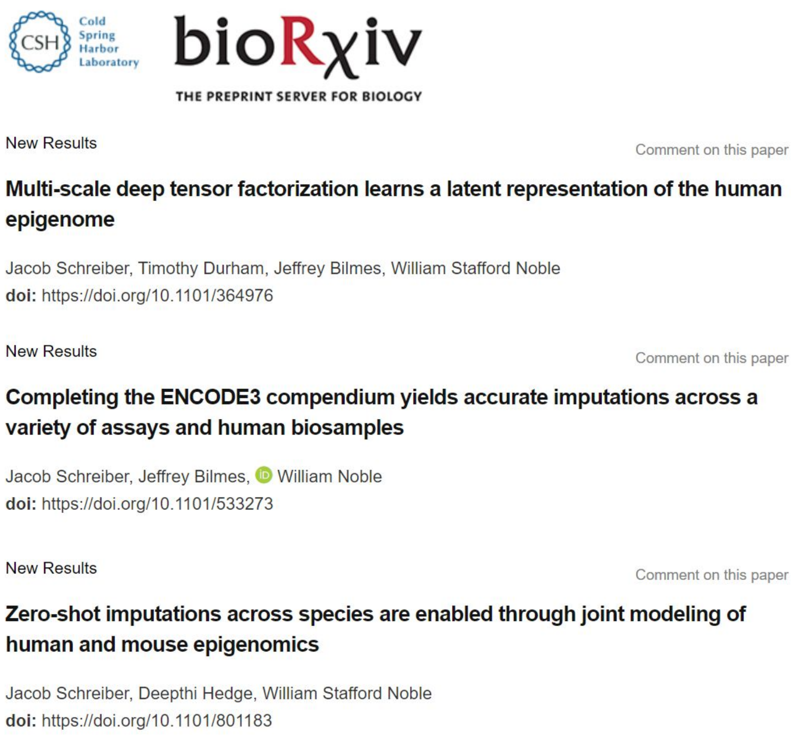
Avocado is a multi-scale deep tensor factorization model that is used to learn a latent representation of the human epigenome. The purpose of this model is two fold; first, to impute epigenomic experiments that have not yet been performed, and second, to learn a latest representation of the human epigenome that can be used as input for machine learning models in the place of epigenomic data itself. The project page with links to the full set of imputations and model parameters can be found at https://noble.gs.washington.edu/proj/avocado/. The manuscript is currently under review and the preprint can be found here.

Installation

Avocado can be installed using pip.

```
pip install avocado-epigenome
```

https://github.com/jmschrei/avocado



New Results

Comment on this paper

Multi-scale deep tensor factorization learns a latent representation of the human epigenome

Jacob Schreiber, Timothy Durham, Jeffrey Bilmes, William Stafford Noble

doi: https://doi.org/10.1101/364976

New Results

Comment on this paper

Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples

Jacob Schreiber, Jeffrey Bilmes, William Noble

doi: https://doi.org/10.1101/533273

New Results

Comment on this paper

Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics

Jacob Schreiber, Deepthi Hedge, William Stafford Noble

doi: https://doi.org/10.1101/801183

Acknowledgements

Timothy Durham

Deepthi Hedge

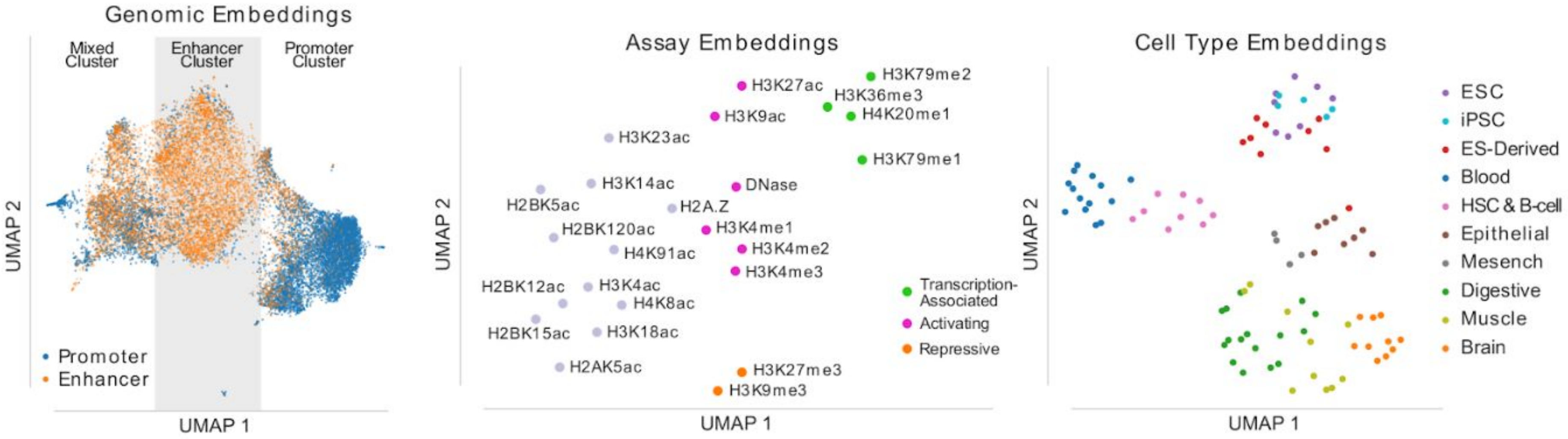
Jeffrey Bilmes

William Noble

eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

The learned latent representations capture known associations


Genomic Embeddings
Mixed Cluster
Enhancer Cluster
Promoter Cluster
Promoter
Enhancer
UMAP 2
UMAP 1
Assay Embeddings
H3K27ac
H3K79me2
H3K36me3
H3K9ac
H4K20me1
H3K23ac
H3K79me1
H3K14ac
DNase
H2BK5ac
H2A.Z
H2BK120ac
H3K4me1
H3K4me2
H4K91ac
H3K4me3
H2BK12ac
H3K4ac
H4K8ac
H2BK15ac
H3K18ac
H2AK5ac
H3K27me3
H3K9me3
Transcription-Associated
Activating
Repressive
UMAP 2
UMAP 1
Cell Type Embeddings
ESC
iPSC
ES-Derived
Blood
HSC & B-cell
Epithelial
Mesench
Digestive
Muscle
Brain
UMAP 2
UMAP 1