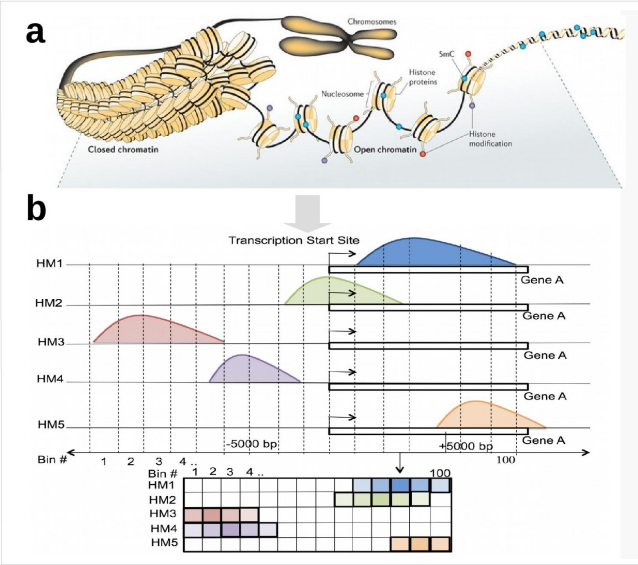
ReChrome: Recursive Convolutional Neural Networks for Epigenomics

Aikaterini Symeonidi, Nicolaou Anguelos, Frank Johannes, Vincent Christlein

Histone modifications and Gene Expression

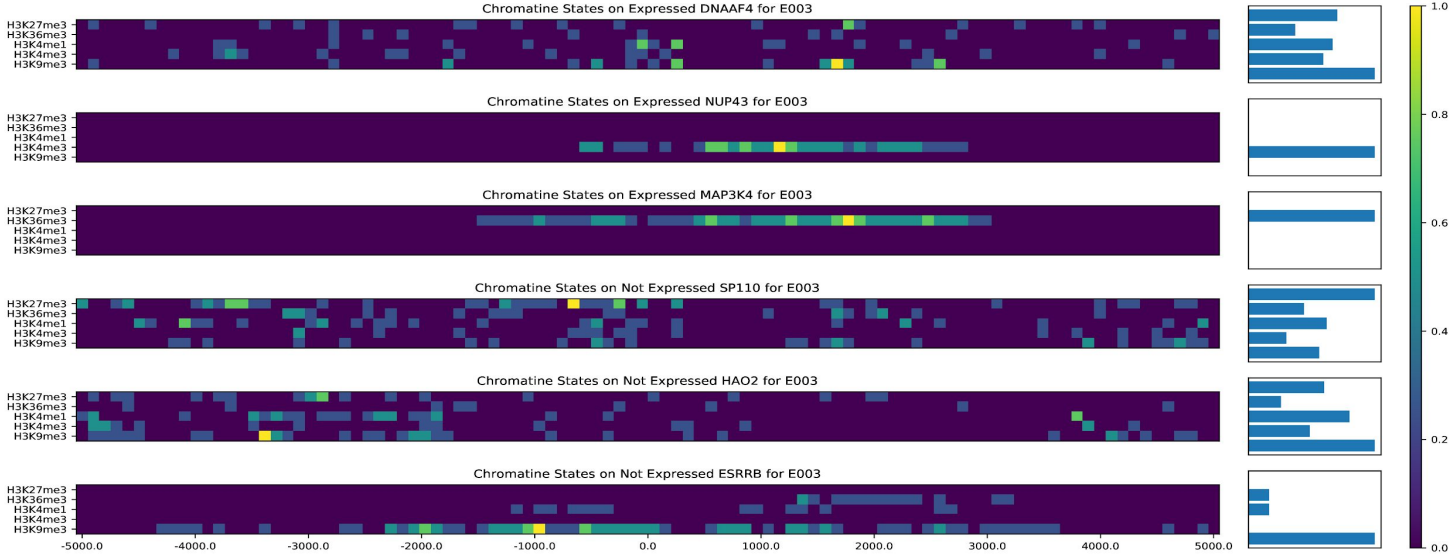
- Histone modifications affect DNA coiling
- DNA coiling affects Gene expression
- Histone modifications are sequenced and mapped to the genome



Modified from Taudt et al. and Sing et al.

Dataset Structure

- Each Gene TSS is a sample
- 5 modifications -> 5x1D signals
- Map modification signals to Expressed/Non-expressed



If you torture the data long enough,
it will confess to anything.

Ronald Coase (1910-2013)

Cardinalities

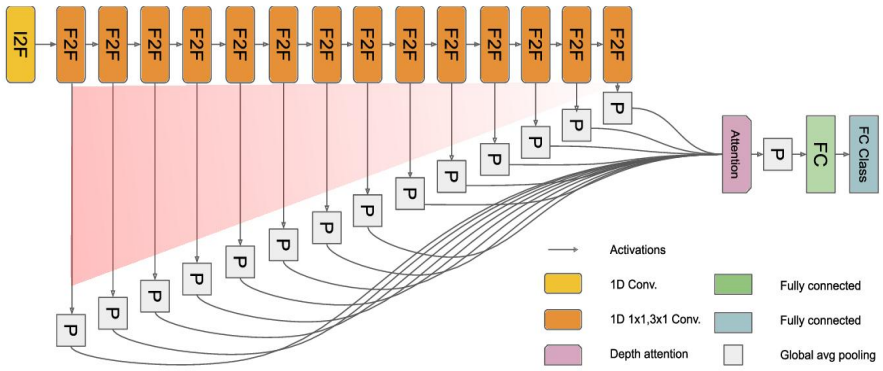
- Binary classifier
- 56 Datasets (human tissues)
- ~20,000 samples per dataset (genes)
- Input: 5 channels x (10,000 - 30,000 bases) per sample

Motivation

- Restrict capacity to force generalisation
  - Weight sharing
- Explore sampling frequencies and sample sizes
  - Multiscale
  - Variable depth

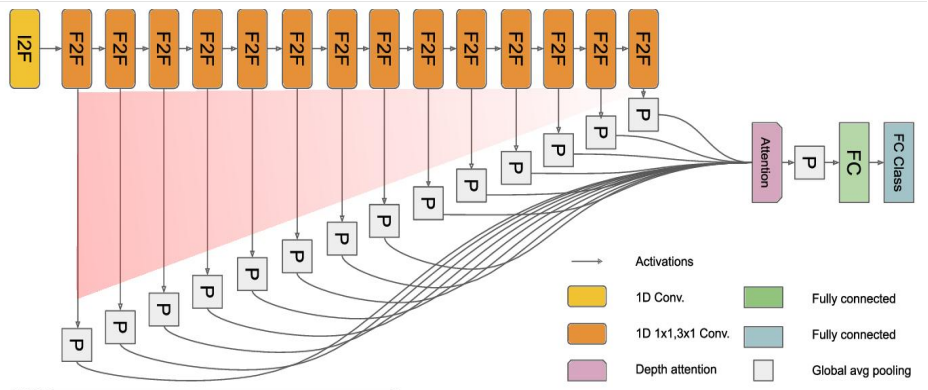
ReChrome Architecture

- Global average pooling
- Multiple routes for information
- Channels, C is for capacity:
  - I2F: 5xT -> CxT
  - F2F: CxT -> Cx((T/2)-2)
  - FC: C -> C
  - FC Class: C -> 2



ReChrome Architecture: Variants

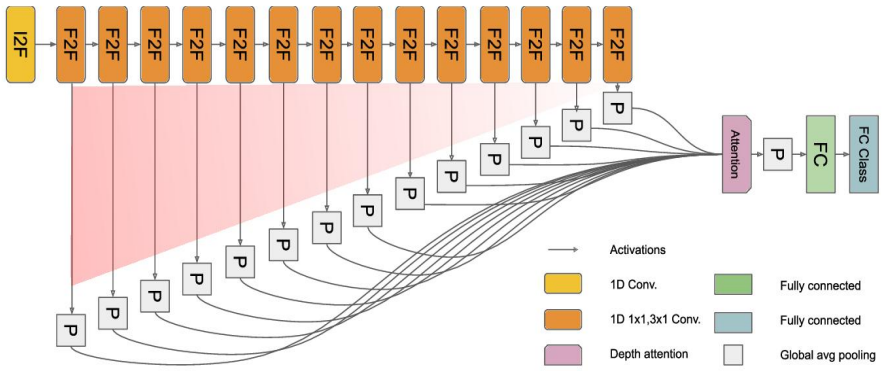
C is for capacity!



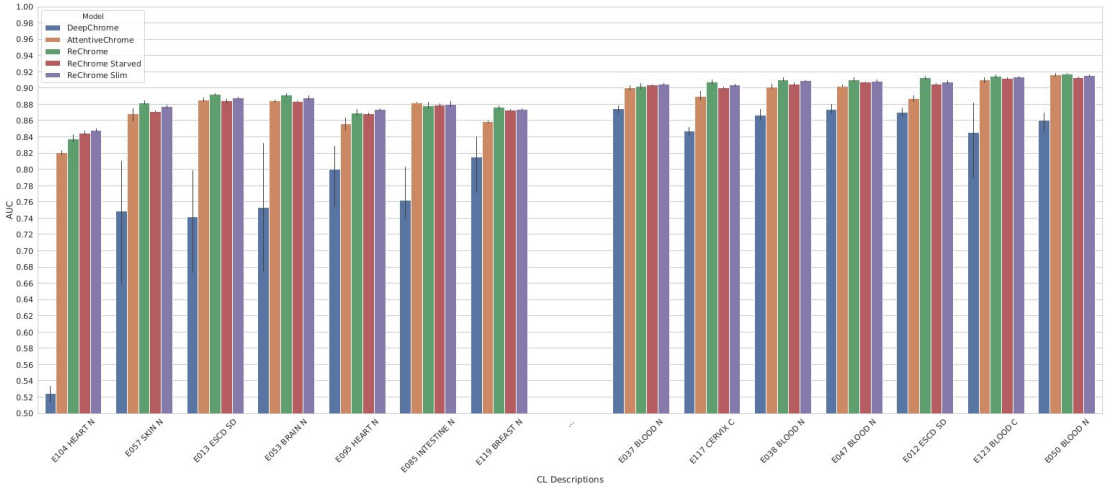
Rechrome Variant	**C**hannels	Parameters
Starved	10	416
Slim	30	3076
Normal	100	31016

ReChrome Architecture: F2F

- Inception inspired
- 1x1 convolution
- 1x3 convolution
- Batch norm



Experiments

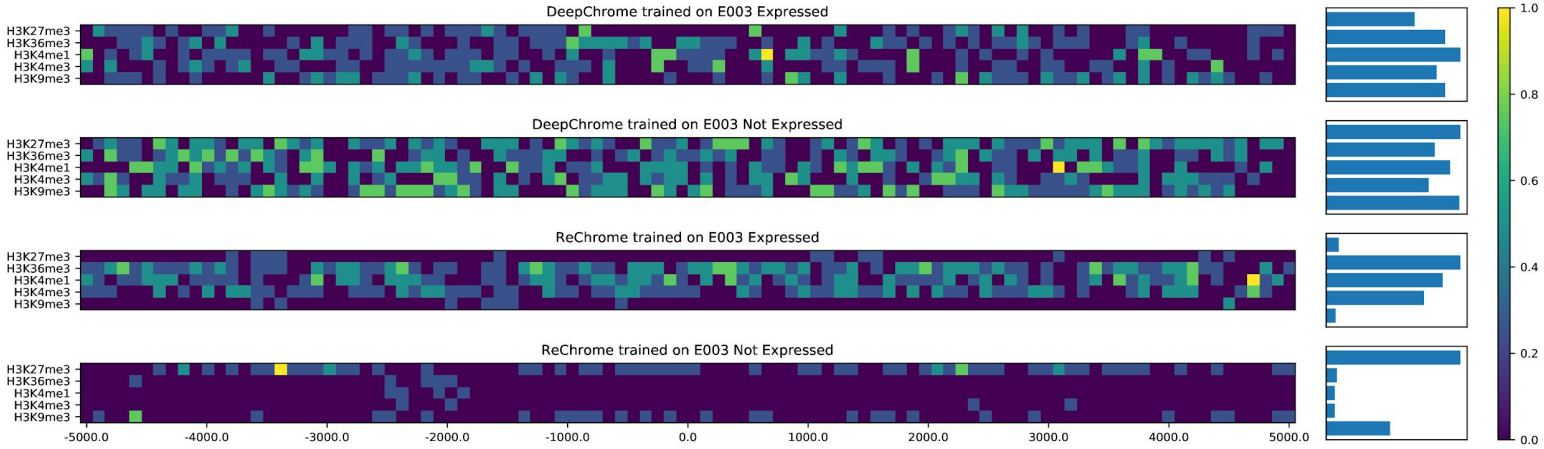


Model	Parameters	Val. set	Test set	Cross test set
DeepChrome	644 177	87.36	81.43	79.78
AttentiveChrome	55 681	86.36	86.80	NA
ReChrome	31 016	**87.54**	87.73	86.13
ReChrome Slim	3076	87.06	**87.75**	**86.56**
ReChrome Starved	416	86.55	86.45	86.45

Context and Sampling

Model	bin size	bin count	TSS context (bases)	AUC [%]
ReChrome	1	30 000	$\pm 15,000$	85.35
ReChrome	100	100	± 5000	87.54
ReChrome	150	66	± 4950	87.64
ReChrome	150	200	$\pm 15\,000$	87.63
ReChrome	300	34	$\pm 10\,200$	88.05
ReChrome	300	100	$\pm 15\,000$	88.07
ReChrome	400	26	± 5200	**88.14**
ReChrome	400	76	$\pm 15\,200$	88.09
ReChrome	15 000	1	$\pm 15\,000$	87.35

What Does ReChrome Dreams of?



Conclusions and Discussion

- SotA methods and variants perform between 85% and 90%
- Histone Modifications are not a full predictor for Gene Expression
- Performance Factors:
  - Cell type
  - Model Capacity
  - Sample Size and Sampling rate
- Capacity Restriction can eliminate overfitting