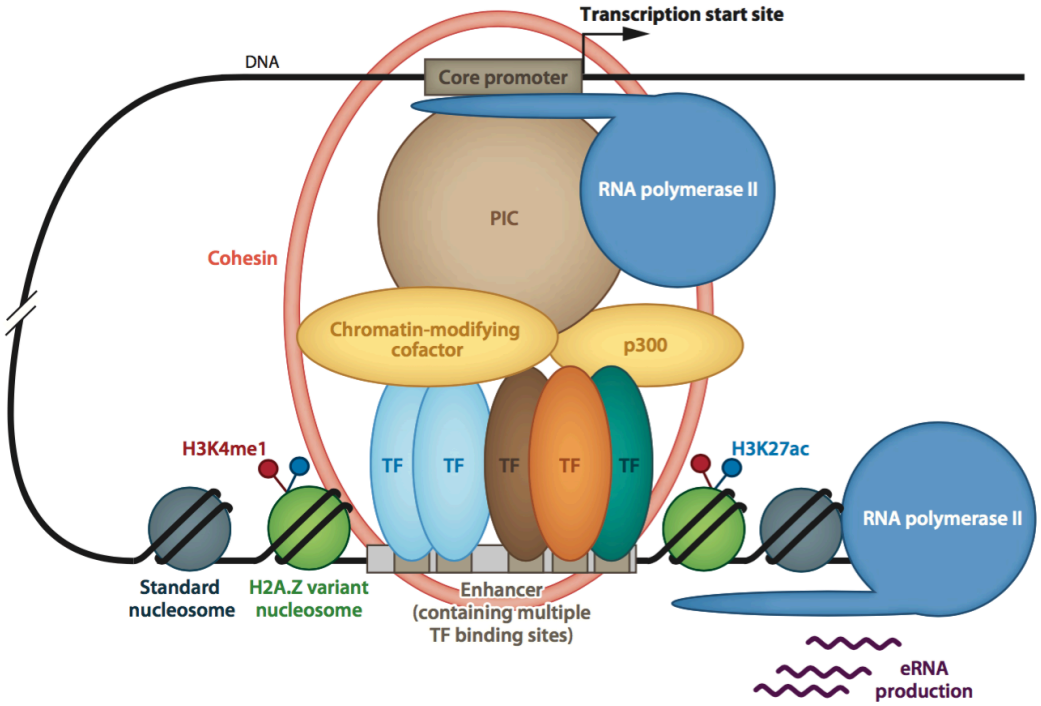
REPTILE: **R**egulatory **E**lement **P**rediction based on **TI**ssue-specific **L**ocal **E**pigenetic marks

Yupeng He
Ecker lab
Salk Institute for Biological Studies

ENCODE AWG
Feb-26-2016

Enhancer Prediction

- Sequence
- p300 binding
- Chromatin modifications
- Open chromatin
- eRNA
- Physical interaction
- Evolutionary conservation



Maston, Glenn A., et al., Annual review of genomics and human genetics 13 (2012): 29-57.

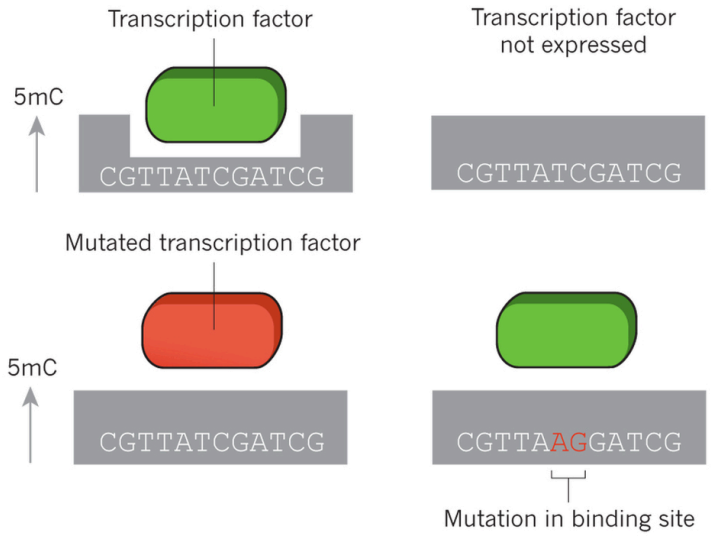
Kleftogiannis et al., Briefings in bioinformatics (2015): bbv101.

Incorporating DNA methylation data into enhancer identification

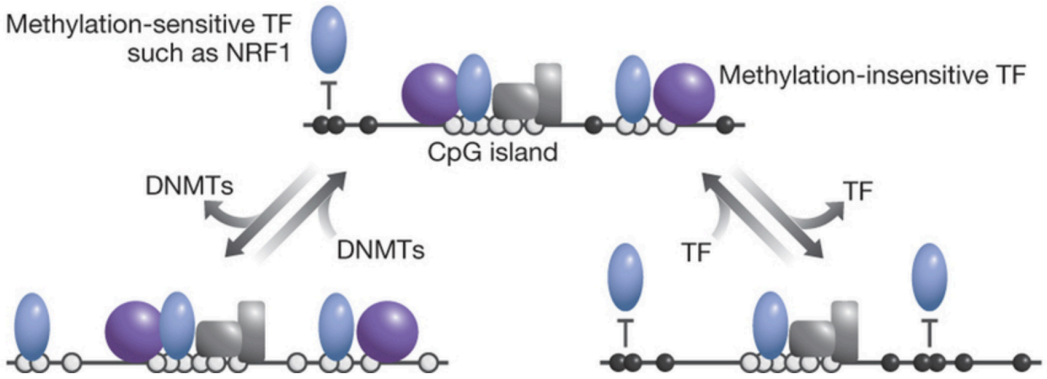
- DNA methylation provides hints about the locations and boundary of enhancers
- Tissue-specific DNA methylation is predictive of enhancers
- DNA methylation has not been effectively incorporated in predicting enhancers

TF binding and DNA methylation

- Information content of DNA methylation

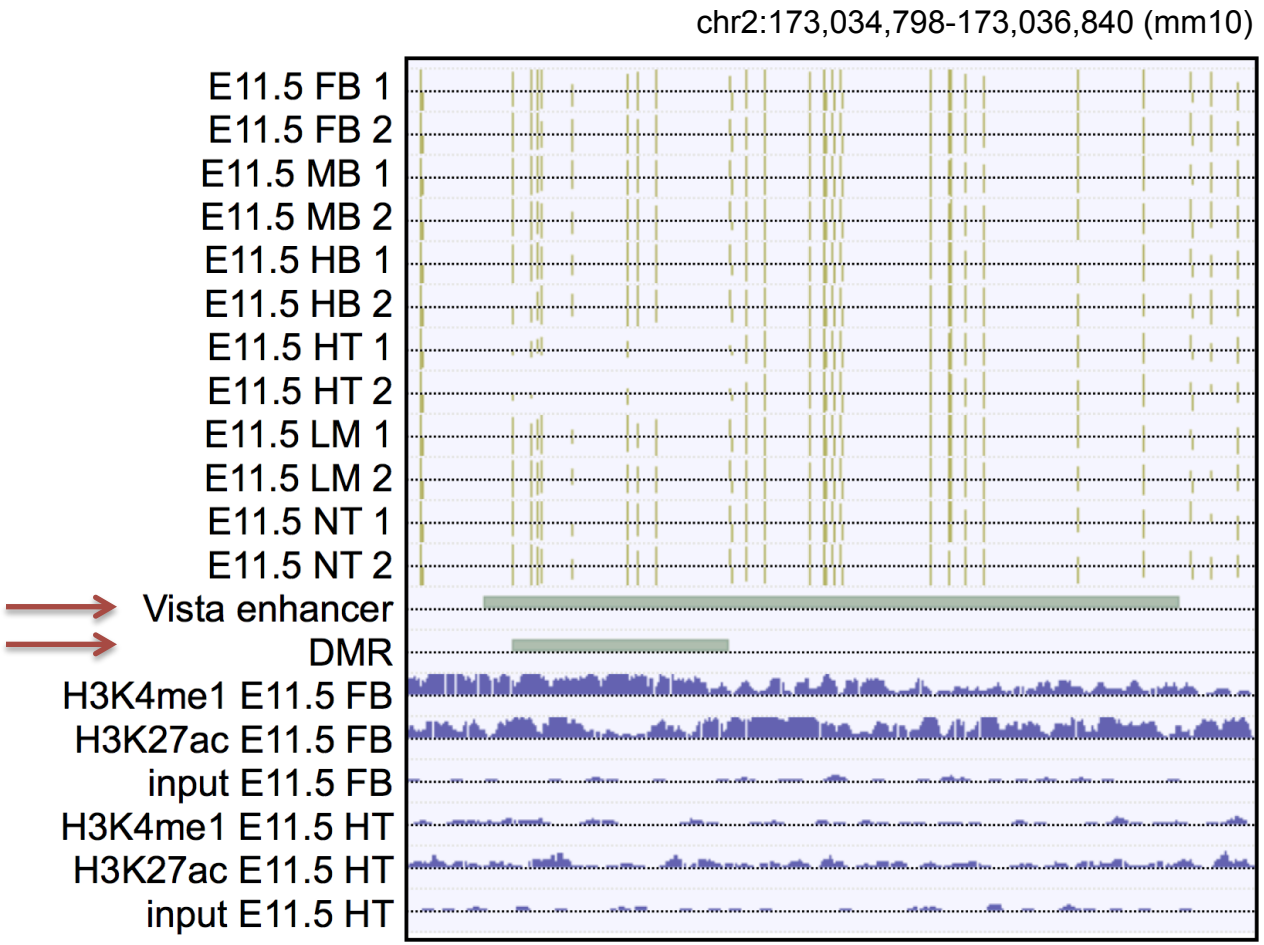


Domcke, Silvia, et al., Nature (2015). APA



Schübeler, Dirk., Nature 517.7534 (2015): 321-326.

Differentially **M**ethylated **R**egions

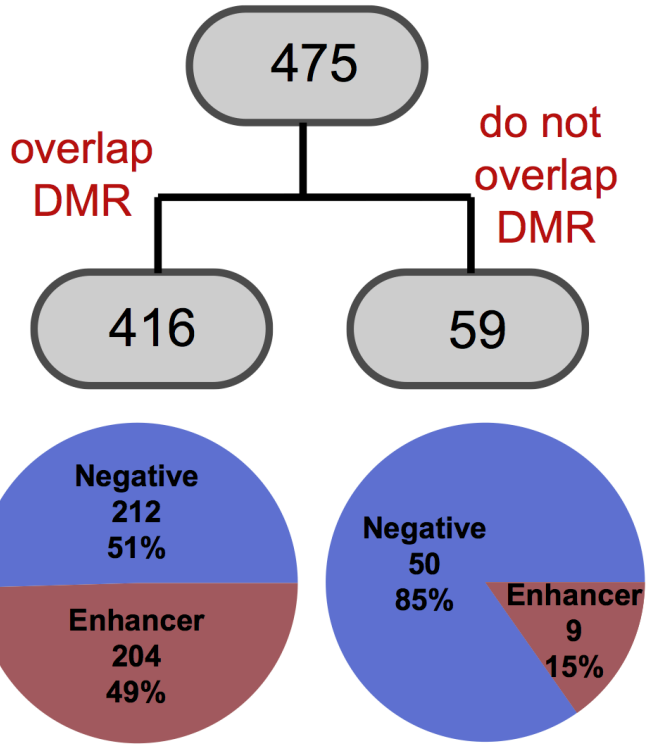


Heart 5/11

FB – forebrain
MB – midbrain
HB – hindbrain
HT – heart
LM – limb
NT – neural tube

Enhancers and DMRs are strongly overlapped

- 475 tested sequences/regions
  - from vista enhancer browser (Oct 12 2015) *
- 416 (87%) overlap DMRs
  - Covering 96% of validated enhancers



* Excluding the 70 test regions in ENCODE enhancer prediction competition

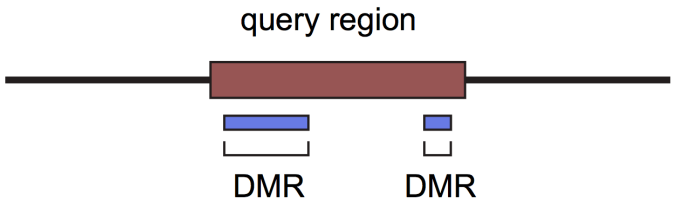
REPTILE

PREDICTING ENHANCERS BASED ON DNA METHLYATION AND CHROMATIN DATA

REPTILE: Regulatory **E**lement **P**rediction based on **TI**ssue-specific **L**ocal **E**pigenetic marks

- Idea
  - Capture local DNA methylation signatures which are washed out in the whole region
  - Improve prediction by looking at the tissue-specificity of DNA methylation

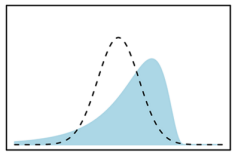
- Input
  - Input regions
  - DMRs
  - Epigenetic data of targeted sample and additional “reference” samples

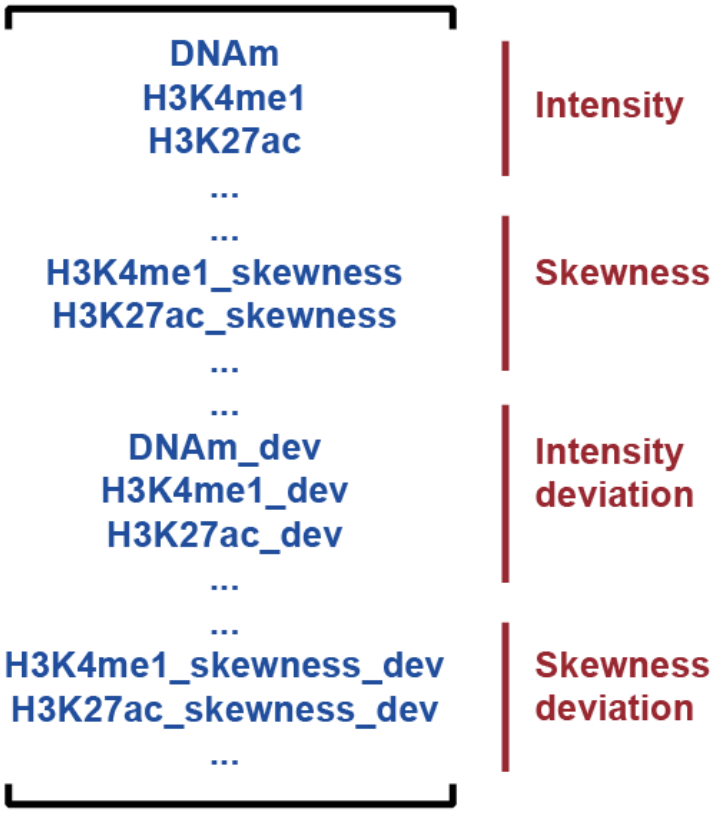


- Output
  - Enhancer activity score for each input region in targeted sample

REPTILE workflow

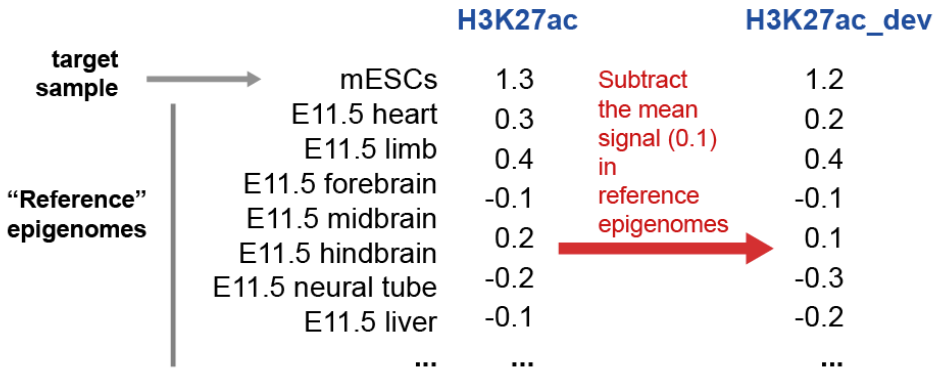
- Each DMR or query region is represented as single high-dimensional feature vector
  - Intensity
    - DNAm, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac, H3K27me3
  - Skewness
    - Chromatin marks
    - Lu, Yiming, et al., PloS one 10.6 (2015): e0130622.



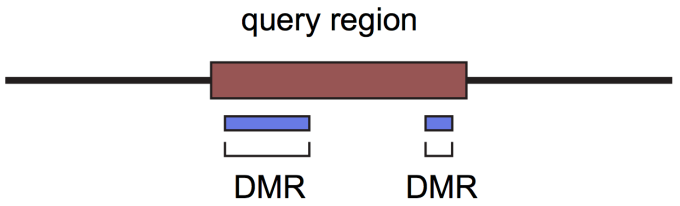


REPTILE workflow (cont.)

- Deviation from “reference epigenomes”
  - To capture the tissue-specific nature of epigenetic marks
  - Calculated for intensity and skewness of chromatin marks



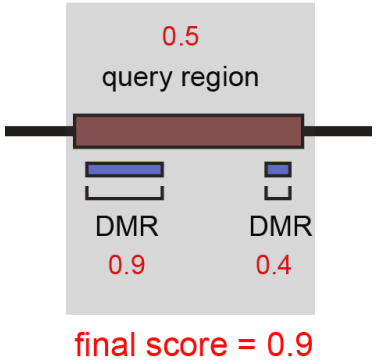
- Random forest classifiers: one for DMRs and one for query regions



REPTILE workflow (cont.)

- Training
  - Supervised
  - Tag DMRs as positives if they overlap validated enhancers and tag the remaining as negatives.
    - Assume only a small fraction will be mislabeled since most DMRs are negative in specific tissue/cell type

- Prediction
  - Predict the enhancer score of both DMRs and query regions
  - For each query region, enhancer score is defined as the **maximum** of score of region itself and scores of DMRs within them



Epigenetic datasets

- 7 types of epigenetic data
  - DNAm
  - H3K4me1
  - H3K4me2
  - H3K4me3
  - H3K27ac
  - H3K9ac
  - H3K27me3
- Mouse embryonic stem cells (mESCs)
- 8 E11.5 mouse tissues
  - Forebrain
  - Midbrain
  - Hindbrain
  - Neural tube
  - Heart
  - Limb
  - Embryonic facial prominence
  - Liver

BENCHMARK NO.1

ENCODE ENHANCER PREDICTION CHALLENGE

ENCODE enhancer prediction challenge

- Predict enhancers on
  - E11.5 Forebrain H3K27ac peaks (n=39; 19 positives)
  - E11.5 Heart H3K27ac peaks (n=31; 8 positives)

- REPTILE
  - **Single enhancer model for both forebrain and heart predictions**
  - Trained on VISTA mouse enhancers (n=363; Jun. 2015 version; same as the version used by many participated methods)
  - Incorporating DNA methylation and chromatin data of all E11.5 tissues

- Evaluation
  - AUROC: how well the classifier separate positives from negatives
  - AUPR: average precision of predictions. Better metric for imbalanced dataset

Enhancer Prediction Challenge

- REPTILE ranks top 2 in both tasks

Forebrain enhancer prediction			
Rank	**Method**	**AUROC**	**AUPR**
1	Beer3	0.708	0.741
2	**REPTILE**	**0.724**	**0.736**
3	Brown1	0.682	0.719
4	Brown5	0.647	0.7
5	Lowe1	0.634	0.694
6	Lowe4	0.605	0.691
7	Weng3	0.657	0.688
8	Weng5	0.657	0.688
9	Brown2	0.666	0.677
10	Beer1	0.737	0.675
11	Lowe2	0.674	0.672

Heart enhancer prediction			
Rank	**Method**	**AUROC**	**AUPR**
1	**REPTILE**	**0.626**	**0.592**
2	Valouev4	0.641	0.489
3	Beer5	0.701	0.445
4	Wang	0.565	0.408
5	Yuan1	0.511	0.382
6	Beer1	0.576	0.38
7	Yuan4	0.549	0.356
8	Yuan3	0.704	0.346
9	Keles9	0.484	0.334
10	Brown3	0.375	0.331
11	Valouev2	0.595	0.329

Comparison with ENSEMBLE method

Forebrain enhancer prediction		
Method	**AUROC**	**AUPR**
Best ENSEMBLE method (PL)	0.72	0.81
Best single method	0.74	0.74
REPTILE	0.72	0.74
Worst single method	0.38	0.43

Heart Enhancer prediction		
Method	**AUROC**	**AUPR**
REPTILE	0.63	0.59
Best ENSEMBLE method (PL)	0.70	0.51
Best single method	0.70	0.49
Worst single method	0.27	0.18

1. REPTILE performs comparably well as ENSEMBLE approach
2. Same set of parameters was used in REPTILE for both tasks
3. REPTILE can be applied to different tissues and cells without re-training (shown in later slides)

Result of ENSEMBLE is from slides by Anurag Sethi

BENCHMARK NO.2

MOUSE ESC ENHANCER PREDICTION

mESC enhancer prediction

- Dataset with validated enhancers
  - 211 regions tested by **luciferase reporter assay** in mESCs
  - 131 positives and 80 negatives

- Training
  - Positives: top 5000 p300 binding sites from ChIP-seq
  - Negatives: 5000 randomly chosen promoters and 30,000 randomly chosen 2kb regions

- DMRs for REPTILE
  - Called by comparing the methylomes of mESCs and 6 E11.5 mouse tissues
  - n=497,934
  - Average size 479bp after 150bp extension from both sides

Yue, Feng, et al., Nature 515.7527 (2014): 355-364.

Against published approaches

- RFECS
  - Random forest
  - Based on the shape and intensity of chromatin ChIP-seq data in 2kb sliding windows across the genome
- DELTA
  - Adaptive Boosting (AdaBoost)
  - Based on the shape (represented as three scores: kurtosis, skewness, bimodality) and intensity of chromatin ChIP-seq data in 2kb sliding windows
- CSIANN
  - Neural network
  - Based on the intensity of chromatin ChIP-seq data in 2kb sliding windows

RFECS - Rajagopal, Nisha, et al., PLoS Comput Biol 9.3 (2013): e1002968.
DELTA - Lu, Yiming, et al., PloS one 10.6 (2015): e0130622.
CSIANN - Firpi, Hiram A., Duygu Ucar, and Kai Tan., Bioinformatics 26.13 (2010): 1579-1586.

Results

- Predicting enhancer activity of the 211 regions

Model	AUROC	AUPR	Positives in Top5	Pos in Top10	Pos in Top20
REPTILE	**0.727**	**0.820**	5	10	20
RFECS	0.720	0.785	5	8	15
DELTA	0.719	0.816	5	10	20
CSIANN	0.712	0.795	5	9	19

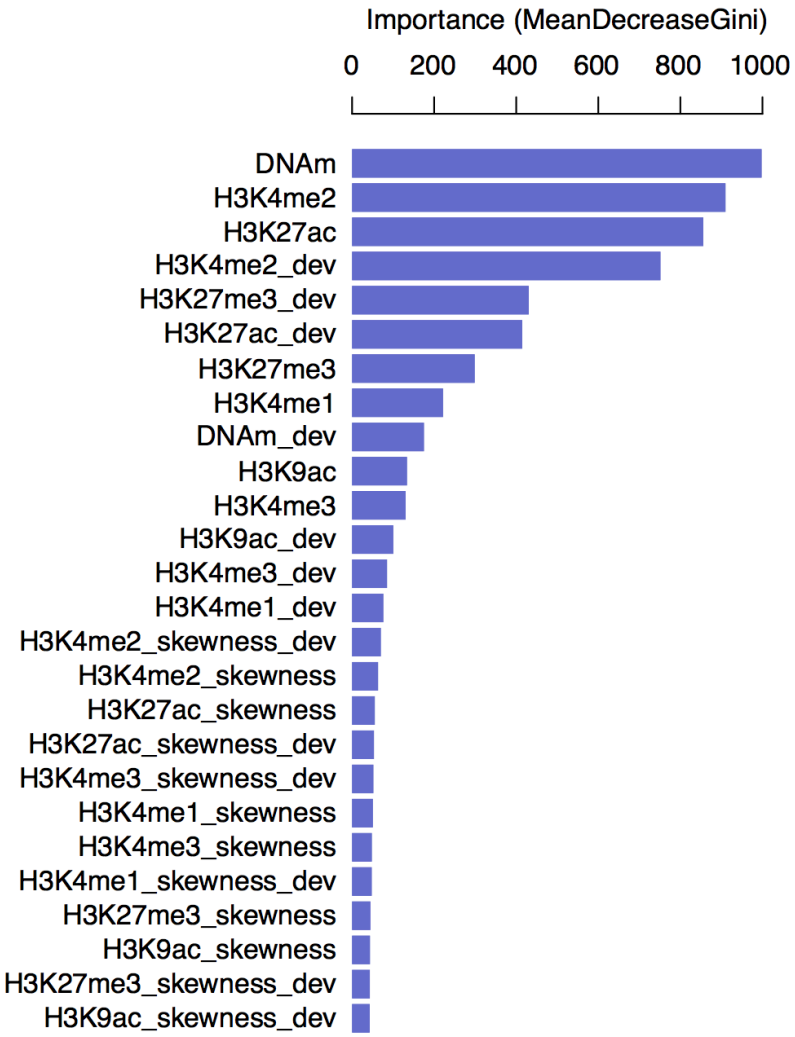
- Genome-wide enhancer predictions overlap open chromatin

Method	% of putative enhancers overlapped DHSs (narrow peaks)	% of putative enhancers overlapped DHSs (broad peaks)	% of putative enhancers overlapped no DHSs	% putative enhancer base pairs in DHSs (narrow peaks)	% putative enhancer base pairs in DHSs (broad peaks)
REPTILE	**77.2%**	**95.4%**	**4.6%**	**15.1%**	**60.0%**
RFECS	75.2%	94.0%	6.0%	14.7%	58.3%
DELTA	68.1%	91.2%	8.8%	13.7%	57.7%
CSIANN	50.0%	86.3%	13.7%	8.5%	40.9%

Importance of variables

- Top3: DNAm, H3K4me2 and H3K27ac

- Deviation of H3K4me2, H3K27me3 and H3K27ac are also very predictive



BENCHMARK NO.3

CAN MODEL TRAINED ON ONE SAMPLE BE USED FOR DIFFERENT SAMPLES?

Benchmark setup

- **Datasets with experimentally validated elements (by transgenic mouse reporter assays)**

Tissues	Source	Total	Positives	Positive%
E11.5 heart	VISTA enhancer browser	545	110	20%
E11.5 limb	VISTA enhancer browser	545	72	13%
E11.5 forebrain	VISTA enhancer browser	545	70	13%
E11.5 midbrain	VISTA enhancer browser	545	59	11%
E11.5 hindbrain	VISTA enhancer browser	545	40	7%
E11.5 neural tube	VISTA enhancer browser	545	30	6%
embryonic heart	Narlikar et al.	36	14	39%

Visel, Axel, et al., Nucleic acids research 35.suppl 1 (2007): D88-D92.

Narlikar, Leelavati, et al., Genome research 20.3 (2010): 381-392.

- **Predicting enhancers using models trained on data of mESCs**

Results

- REPTILE, trained on data of one cell type, can accurately predict enhancers on samples of **different cell types and tissues types**

Method	Data	AUPR						
		E11.5 Heart	E11.5 Limb	E11.5 Forebrain	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural Tube	Heart (Narlikar et al.)
REPTILE	Chromatin + DNAm	**0.59**	**0.44**	**0.47**	**0.36**	**0.37**	**0.18**	**0.74**
RFECS	Chromatin	0.55	0.36	0.42	0.33	0.29	0.16	0.55
DELTA	Chromatin	0.56	0.32	0.37	0.32	0.30	0.17	0.38
CSIANN	Chromatin	0.48	0.25	0.30	0.24	0.23	0.13	0.45

Method	Data	AUROC						
		E11.5 Heart	E11.5 Limb	E11.5 Forebrain	E11.5 Midbrain	E11.5 Hindbrain	E11.5 Neural Tube	Heart (Narlikar et al.)
REPTILE	Chromatin + DNAm	**0.85**	**0.84**	**0.85**	**0.81**	**0.81**	**0.77**	**0.73**
RFECS	Chromatin	0.84	0.79	0.83	0.80	0.79	0.72	0.63
DELTA	Chromatin	0.81	0.76	0.77	0.73	0.78	0.70	0.51
CSIANN	Chromatin	0.80	0.72	0.78	0.72	0.76	0.72	0.61

Summary

1. REPTILE outperforms other methods in predicting enhancer activity

2. REPTILE, trained on data of one cell type, can accurately predict enhancers on samples of different cell types and tissues types.

3. mESC enhancer predictions from REPTILE are supported by open chromatin data

4. By incorporating base-resolution DNA methylation, we are able to improve the accuracy and resolution of enhancer predictions

Acknowledgements

Ecker Lab
- Joe Ecker
- Chongyuan Luo
- Manoj Hariharan
- Joe Nery
- Rosa Castanon
- Huaming Chen
- Mark Urich

UC San Diego
Ren Lab
- Bing Ren
- David Gorkin

Wang Lab
- Wei Wang
- Andre Wildberg
- Tung Nguyen

LBNL
- Len Pennacchio
- Axel Visel

Yue lab
- Feng Yue

U54 HG006997