Enhancer Predictions - nonlocal features

Anurag Sethi, Jing Zhang, and Sushant Kumar
P2 - TECH

June 2014

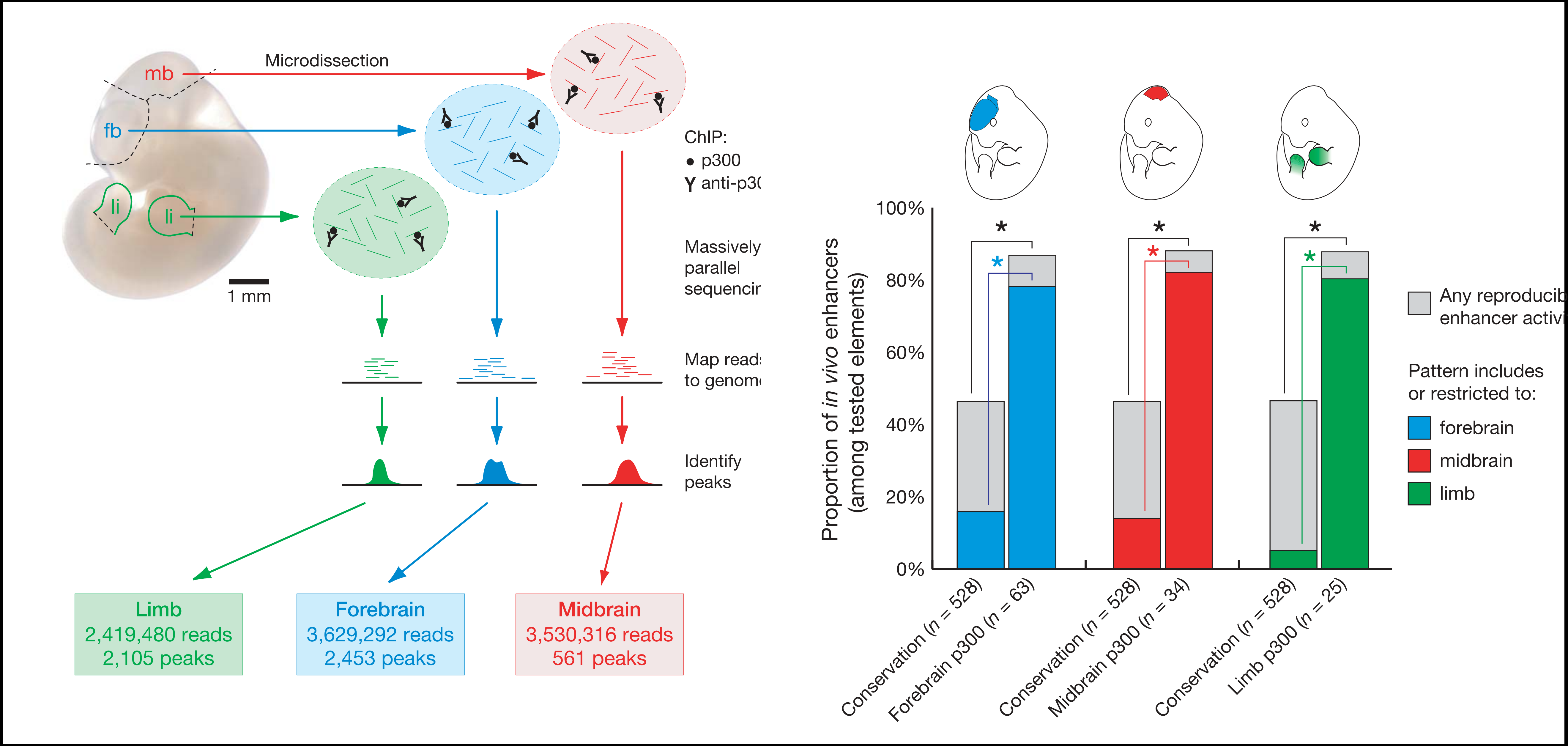
Outline of talk:

Mouse Enhancer Predictions - unsupervised predictions.

Generating a believable set of conditions to define true enhancers.

Models learned using local features versus those learned using features from flanking regions.

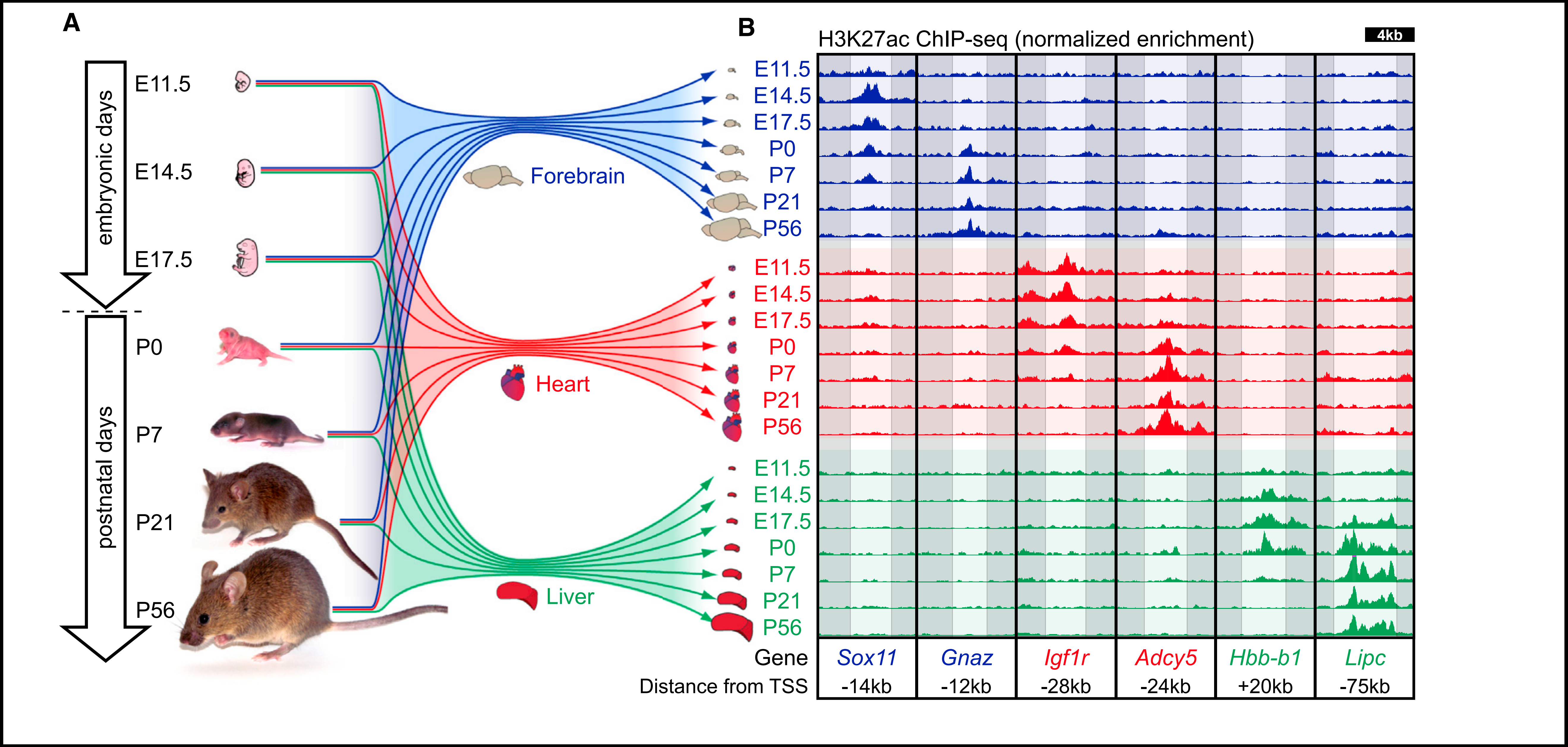
CBP/p300 is a cofactor that is important for enhancer function



p300 peaks were shown to be good predictors of enhancer activity.

Visel, et al., Nature 2009

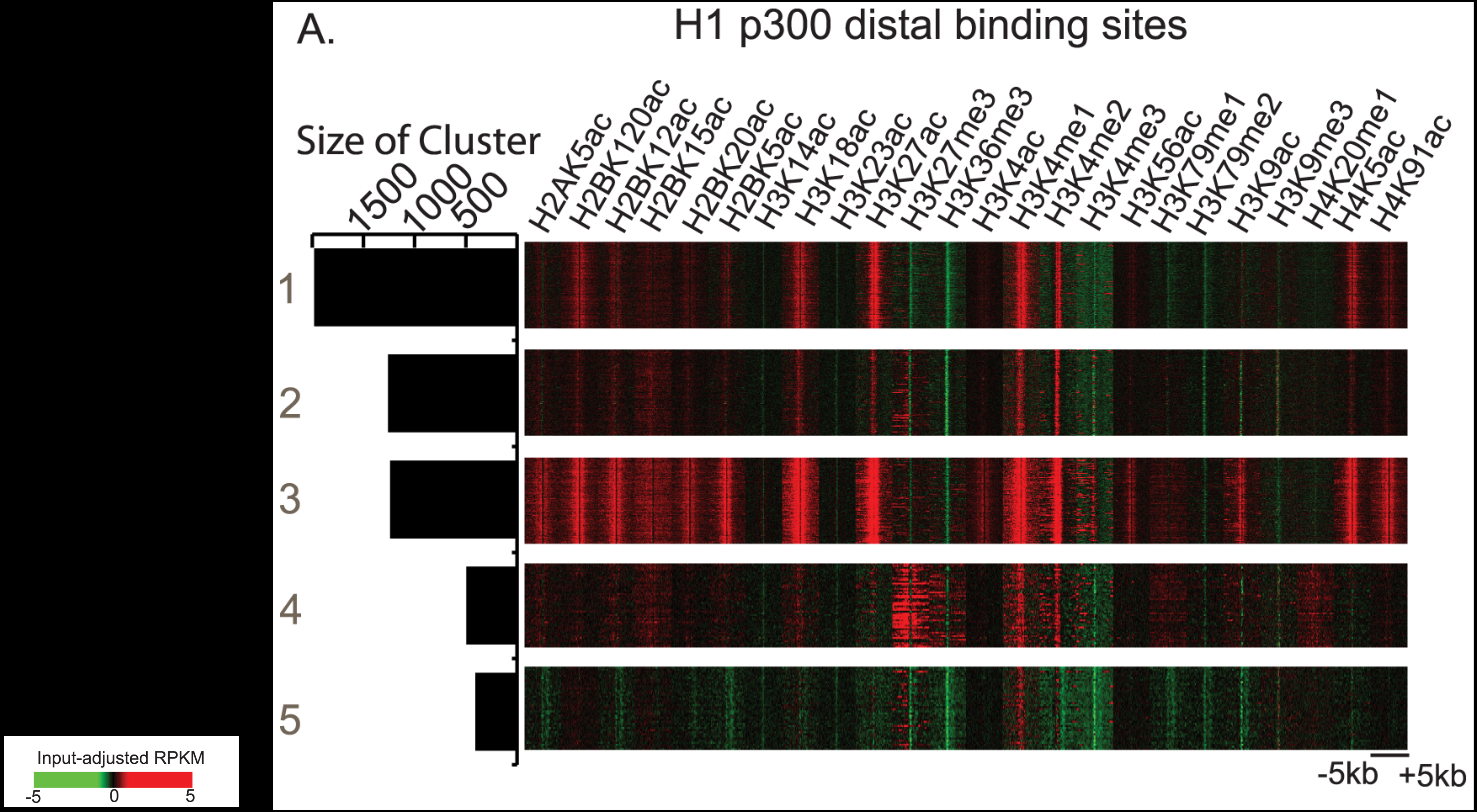
H3K27ac is an important mechanism to regulate the activity of enhancers in different developmental stages



Epigenetically, H3K27ac marks are present near active enhancers.

Nord, et al., Cell, 2013

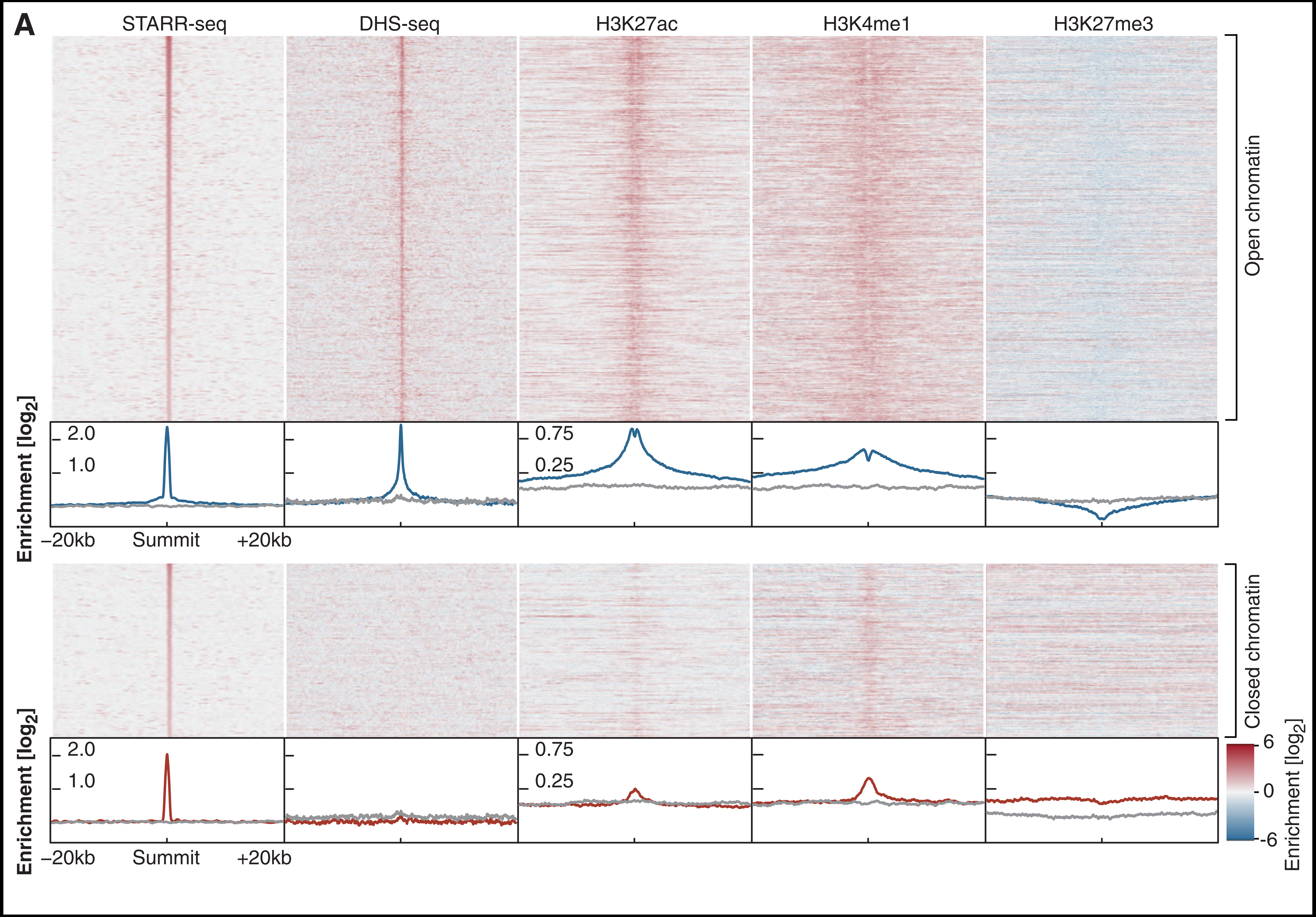
Epigenetic marks associated with putative enhancer regions



Putative enhancer sites are associated with enrichment of H3K4me1 signal and depletion of H3K4me3 signal. Different clusters of putative enhancers have varying levels of H3K27ac (activating) and H3K27me3 (repressive) signals.

Rajagopal, et al., PLoS CB, 2013

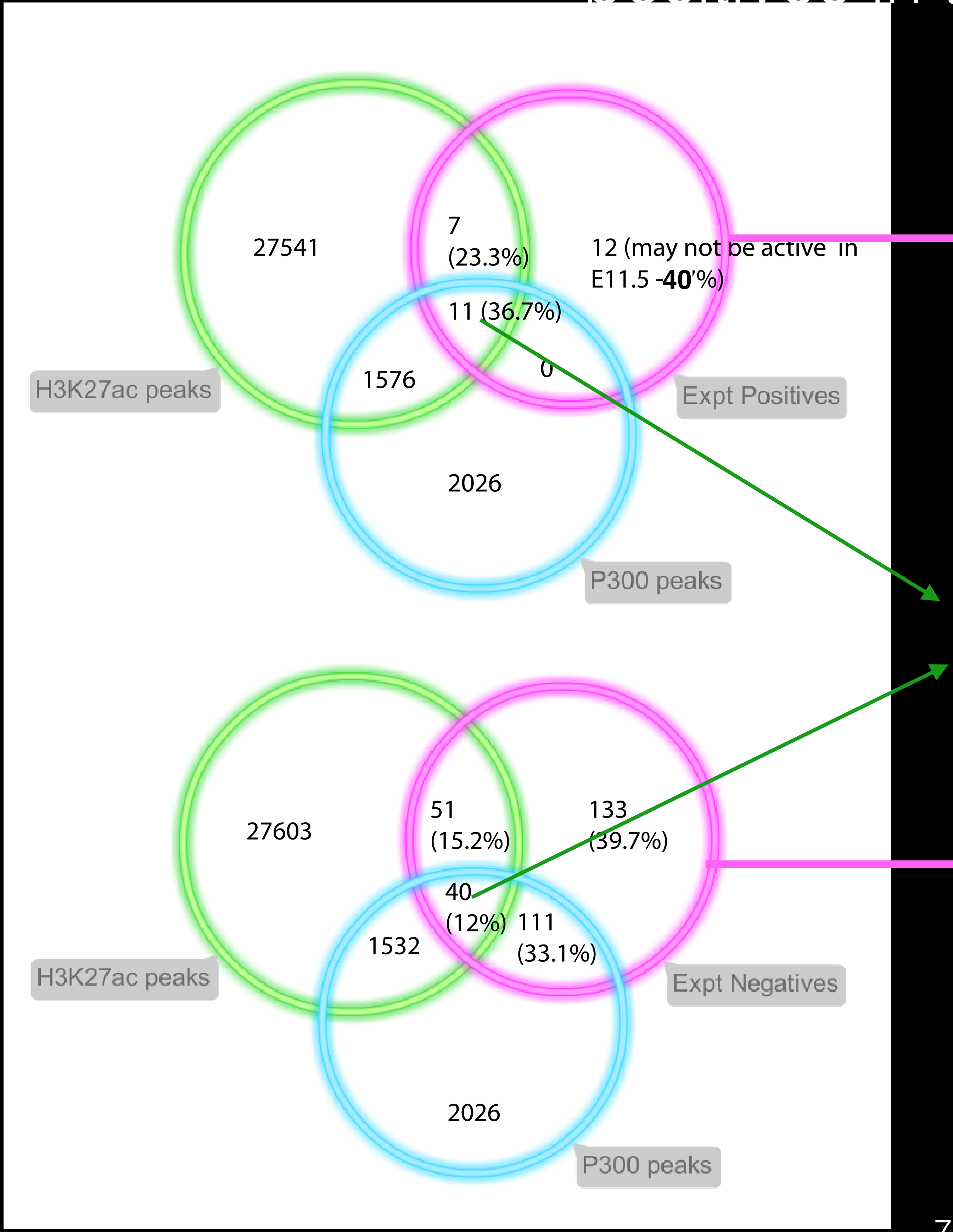
Epigenetic Properties of Enhancers from STARR-Seq assay



Enhancers intersect with DNase hypersensitive sites and are within regions with high H3K27ac/H3K4me1 signals (with higher signals in flanking regions) and H3K27me3 depleted regions.

Arnold, et al., STARR-seq paper, Science, 2013

The H3K27ac and p300 peaks are not sufficient to choose positives in the genome

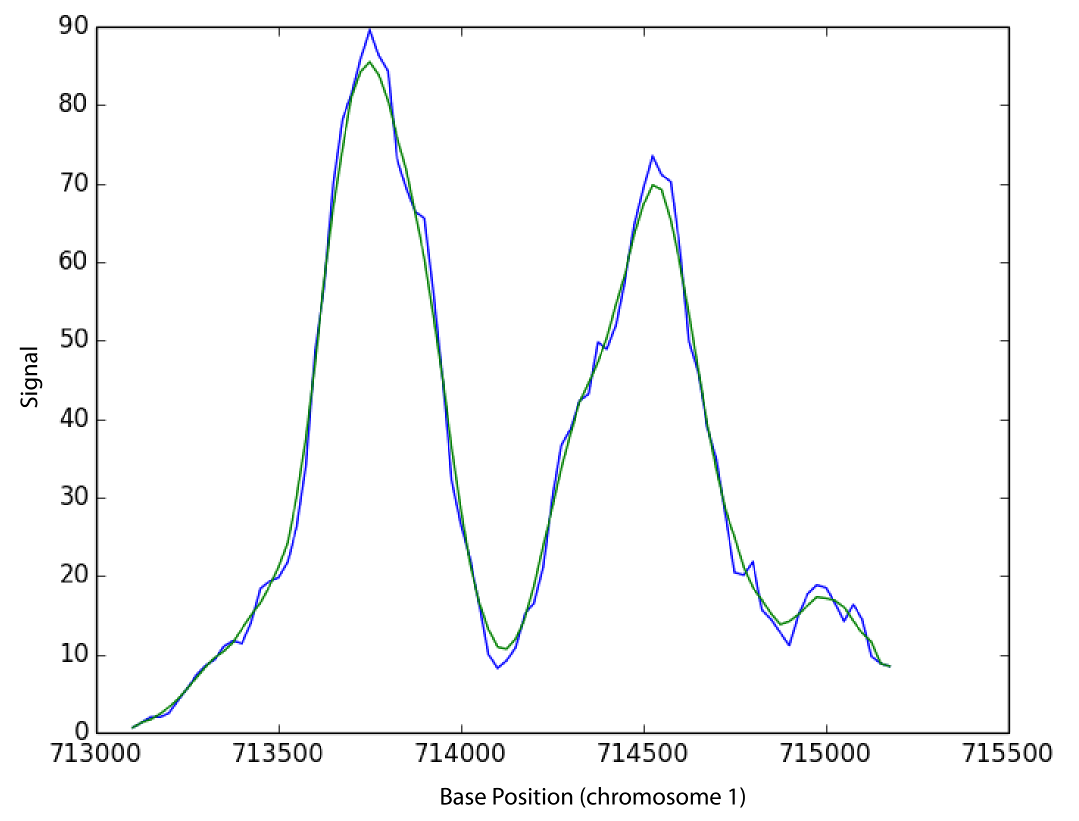


True Positives

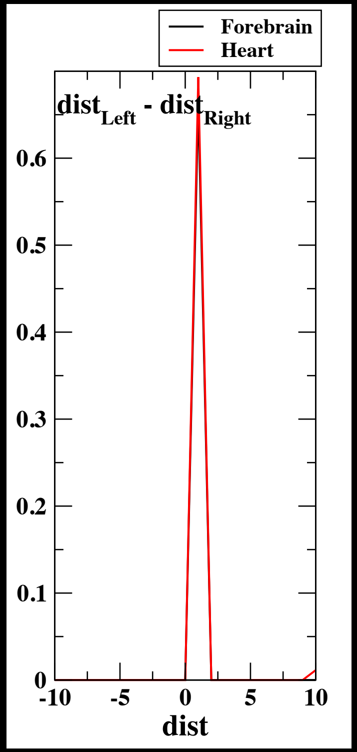
p300 peaks may miss out a number of active enhancers.
p300 and H3K27ac peaks alone can also have a number of false positives

True Negatives

Having a double peak requirement reduces the false positive rate

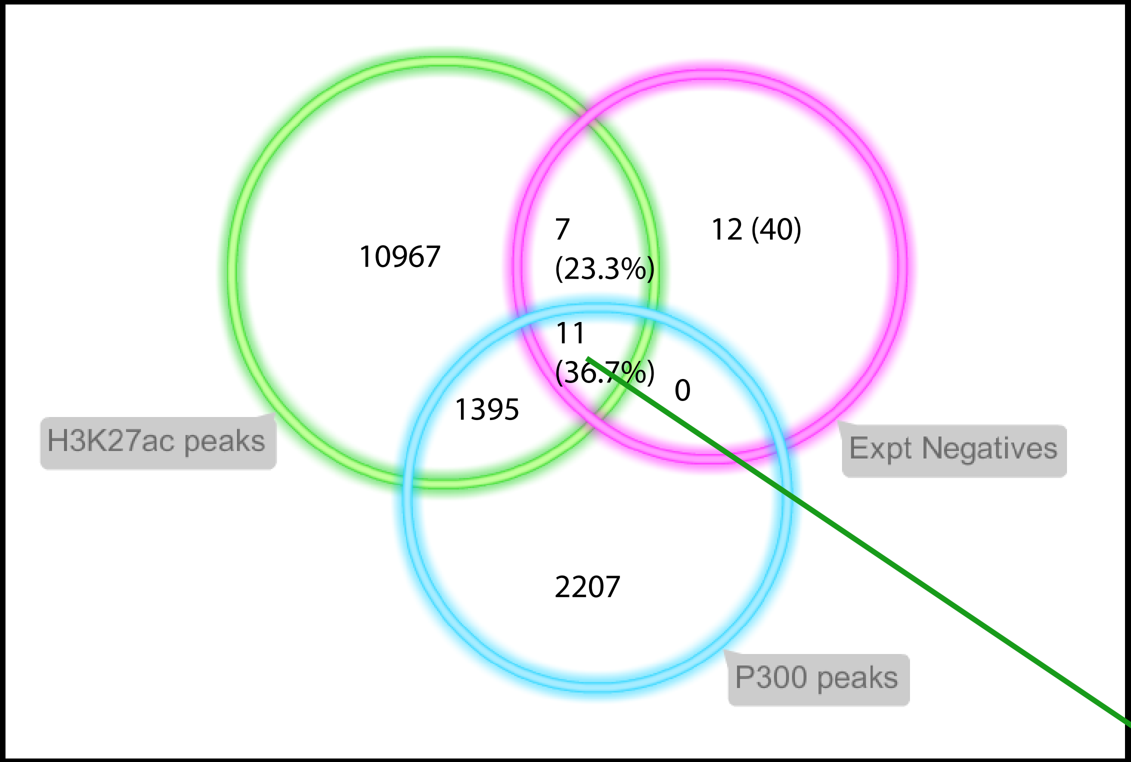


H3K27ac

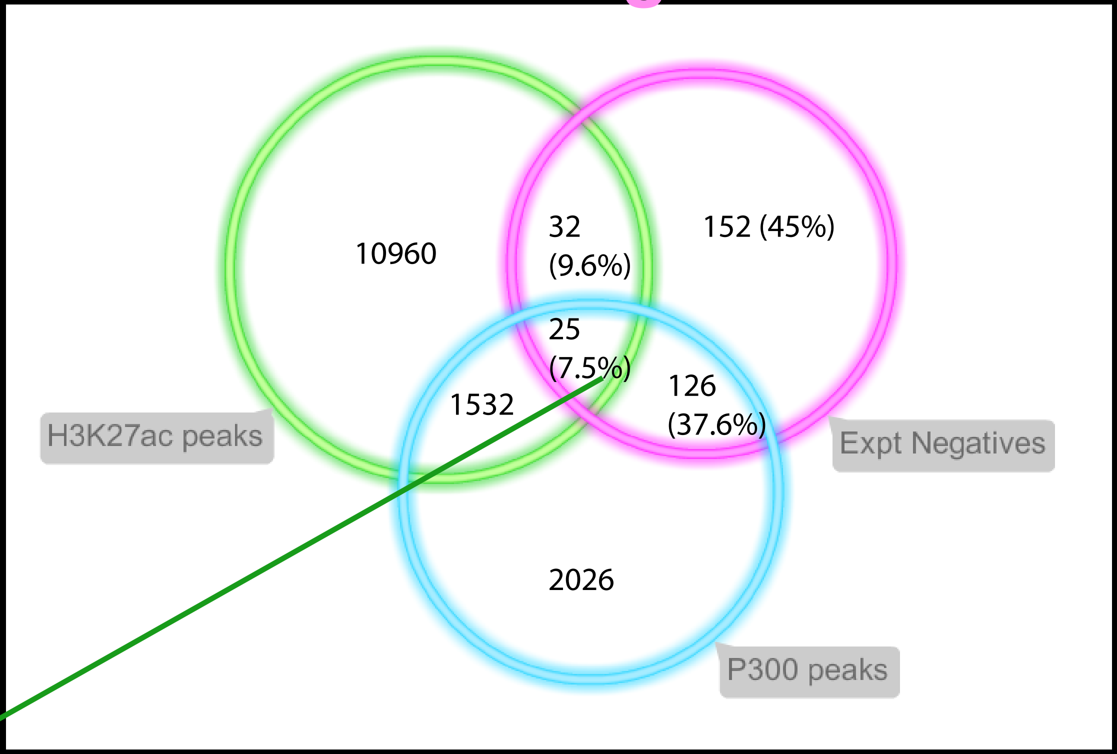


Distance between peaks within 1 kb

True Positives



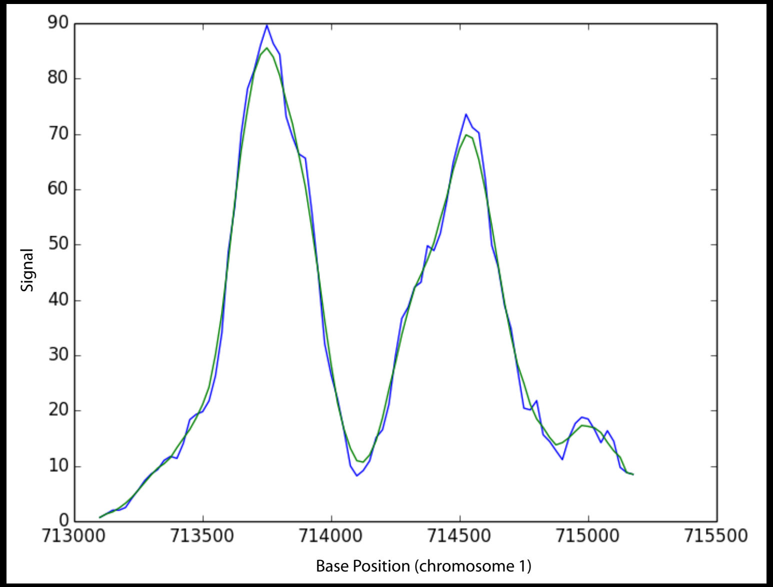
True Negatives



Reduction in false positive rate without affecting true positive rate

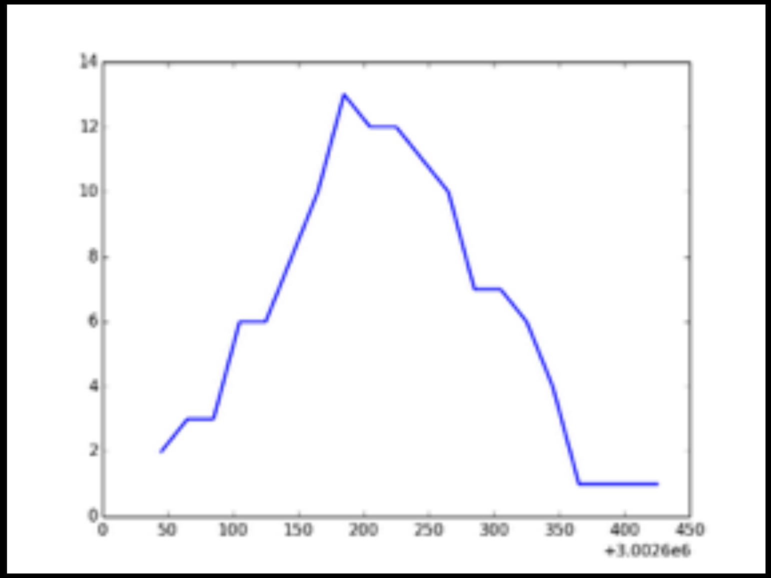
Training Data

Positive Definition



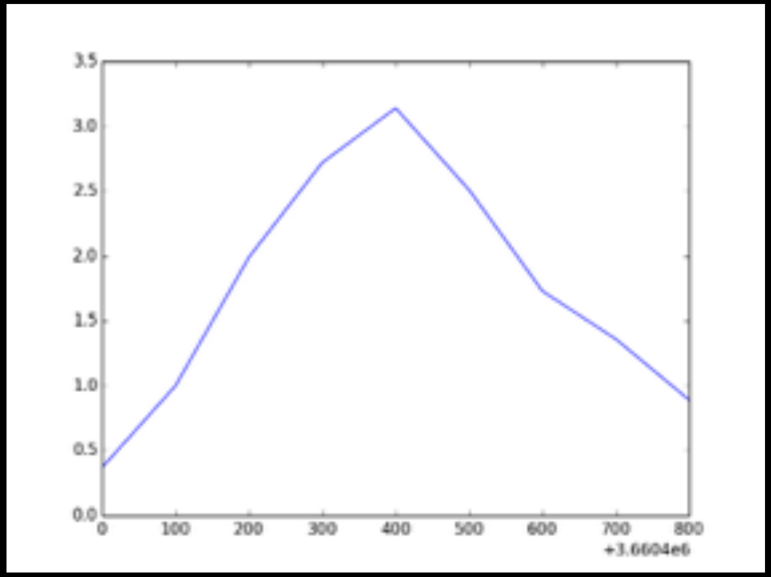
H3K27ac double peaks

∩



DNase HS peaks

-



CTCF peaks

Typically 22000 100 bp bins for mouse

Negative Categories

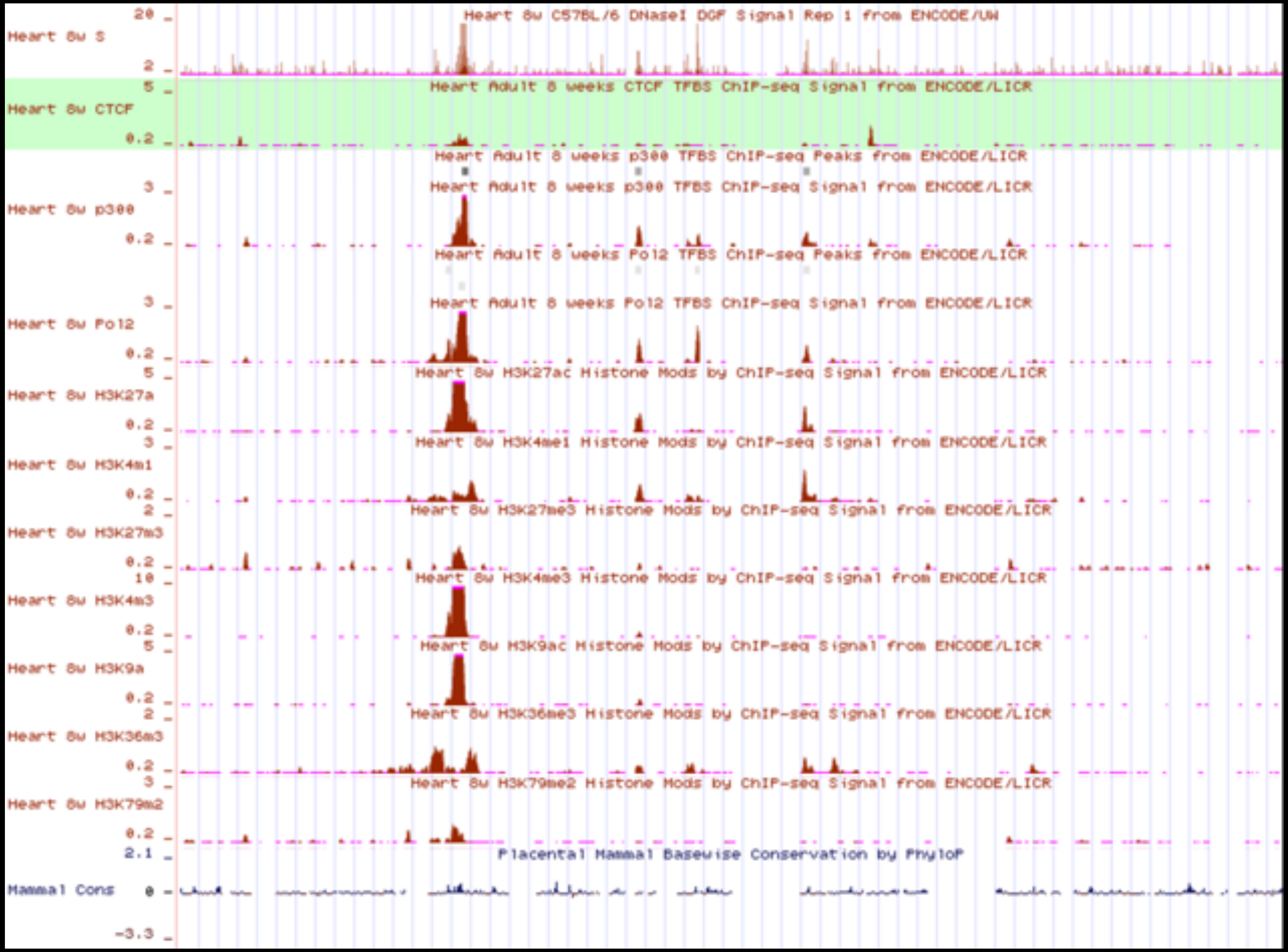
Promoters
Random intergenic regions
H3K27me3 peaks

1.5 x bins randomly sampled
of each category

Features Used for Predicting Enhancers:

Local Model

Histone ChIP-Seq Signal (H3K27ac, H3K4me1, H3K4me3, H3K27me3 necessary)
Transcription Factor ChIP-Seq Signal (p300, CTCF, maybe pol2)
RNA-Seq Signal
Conservation (placental, vertebrate)
GC content

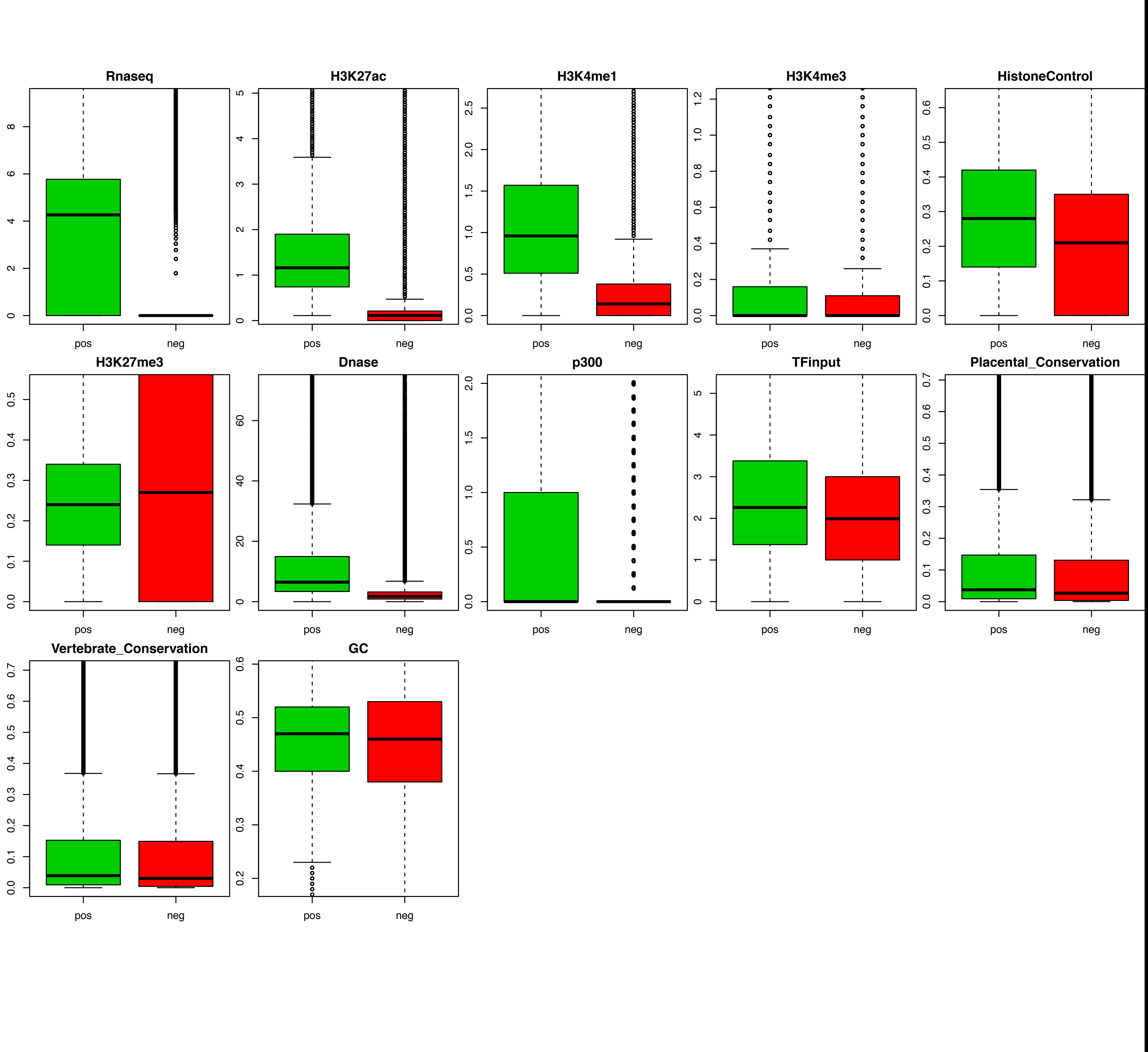

Heart 8w S
Heart 8w C57BL/6 DNaseI DGF Signal Rep 1 from ENCODE/UW
Heart 8w CTCF
Heart Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR
Heart Adult 8 weeks p300 TFBS ChIP-seq Peaks from ENCODE/LICR
Heart 8w p300
Heart Adult 8 weeks p300 TFBS ChIP-seq Signal from ENCODE/LICR
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR
Heart 8w Pol2
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K27a
Heart 8w H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K4m1
Heart 8w H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K27m3
Heart 8w H3K27me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K4m3
Heart 8w H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K9a
Heart 8w H3K9ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K36m3
Heart 8w H3K36me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K79m2
Heart 8w H3K79me2 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
20
2
5
0.2
3
0.2
3
0.2
5
0.2
3
0.2
2
0.2
10
0.2
5
0.2
2
0.2
3
0.2
2.1
0
-3.3

Postprocessing steps

Remove bins overlapping with promoters, gencode exons, and blacklist regions.

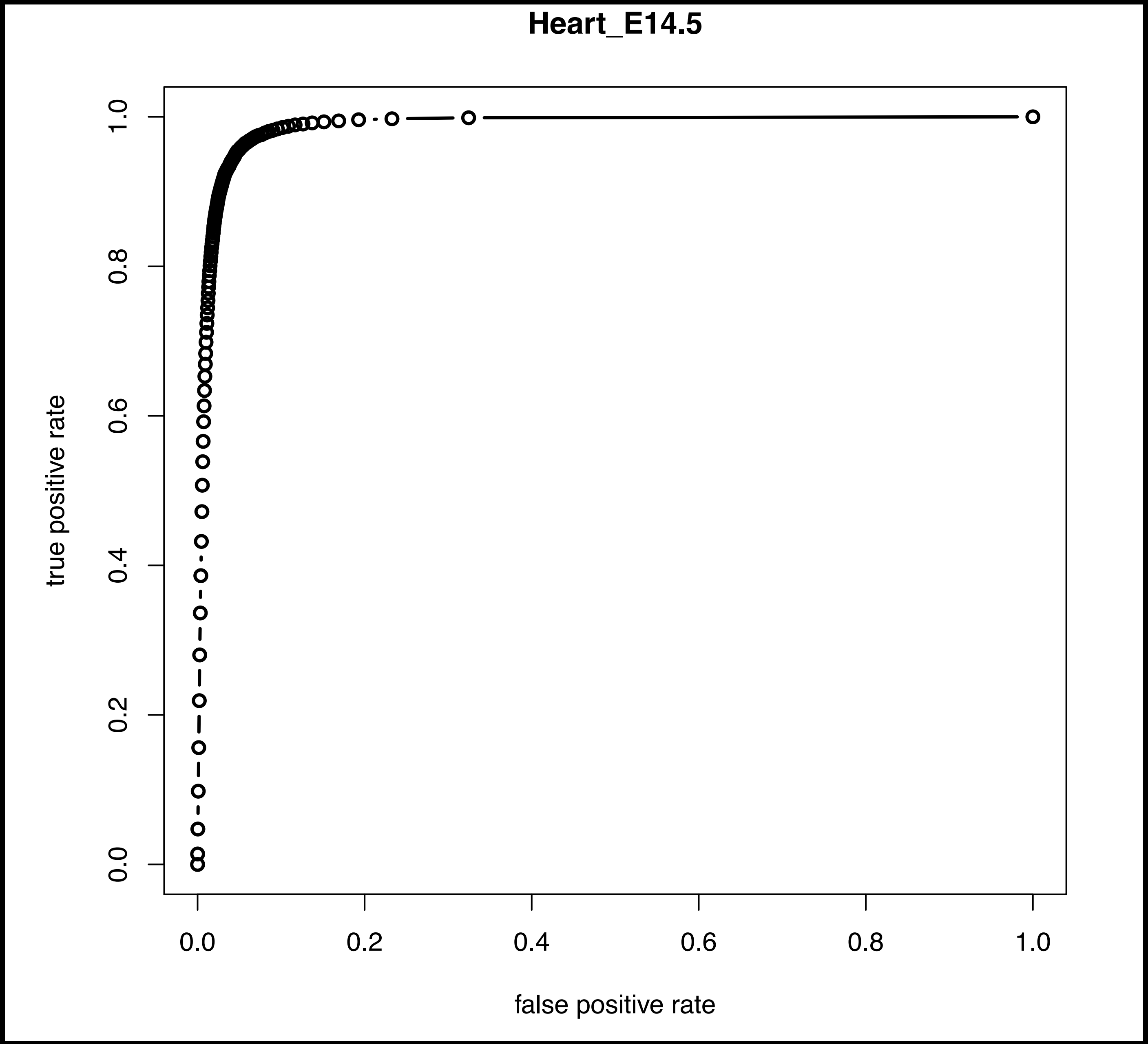
Merge bins from highest scoring bins - peaked regions 1000bp predictions - working on this step.

Local Model - Signal differences in positives and negatives



H3K27ac and H3K4me1 signals have big differences.

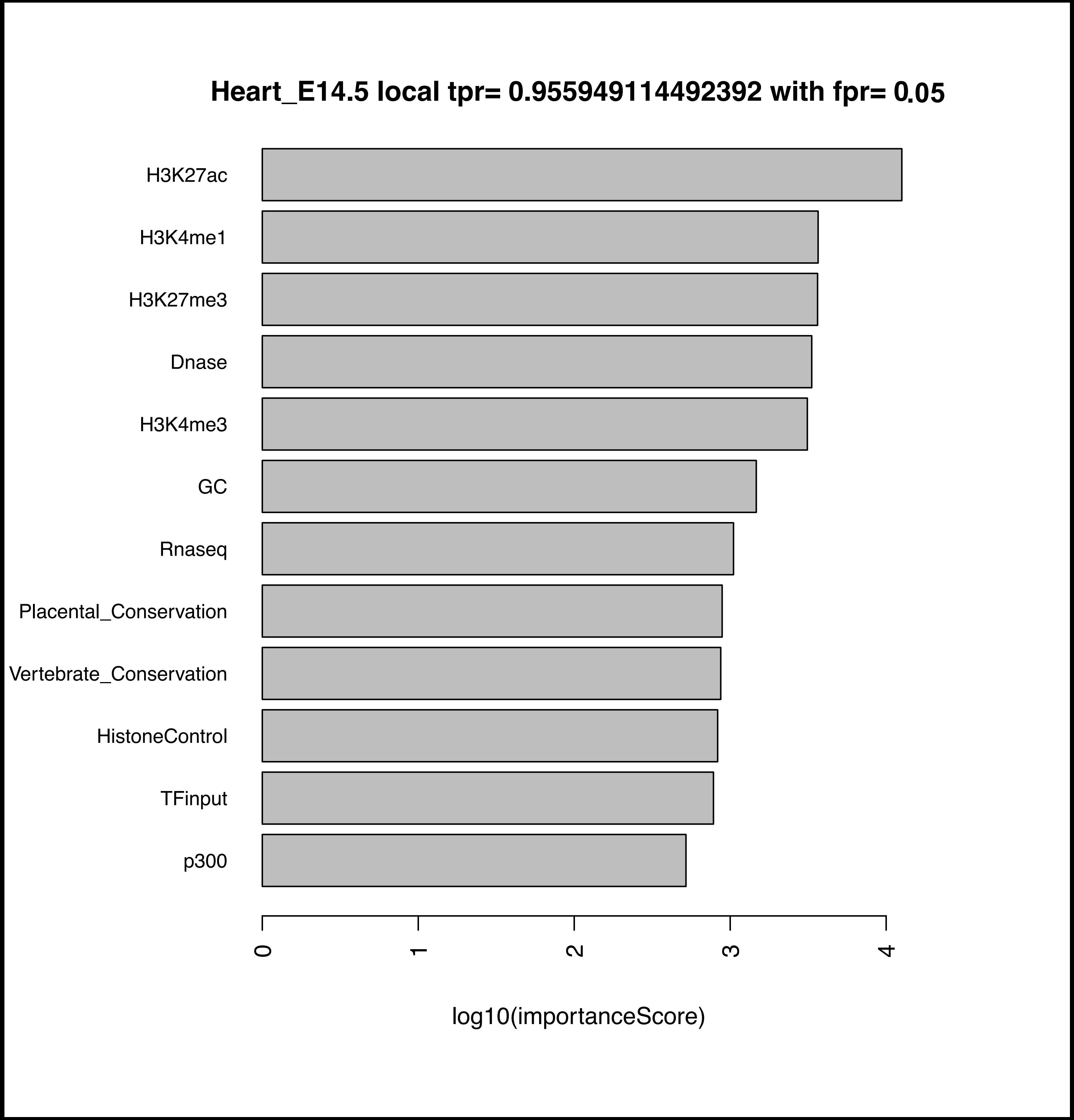
Local Model - results - AUC plot



AUC = 0.94

Local features do reasonably well in model

Local Model - results - importance of features

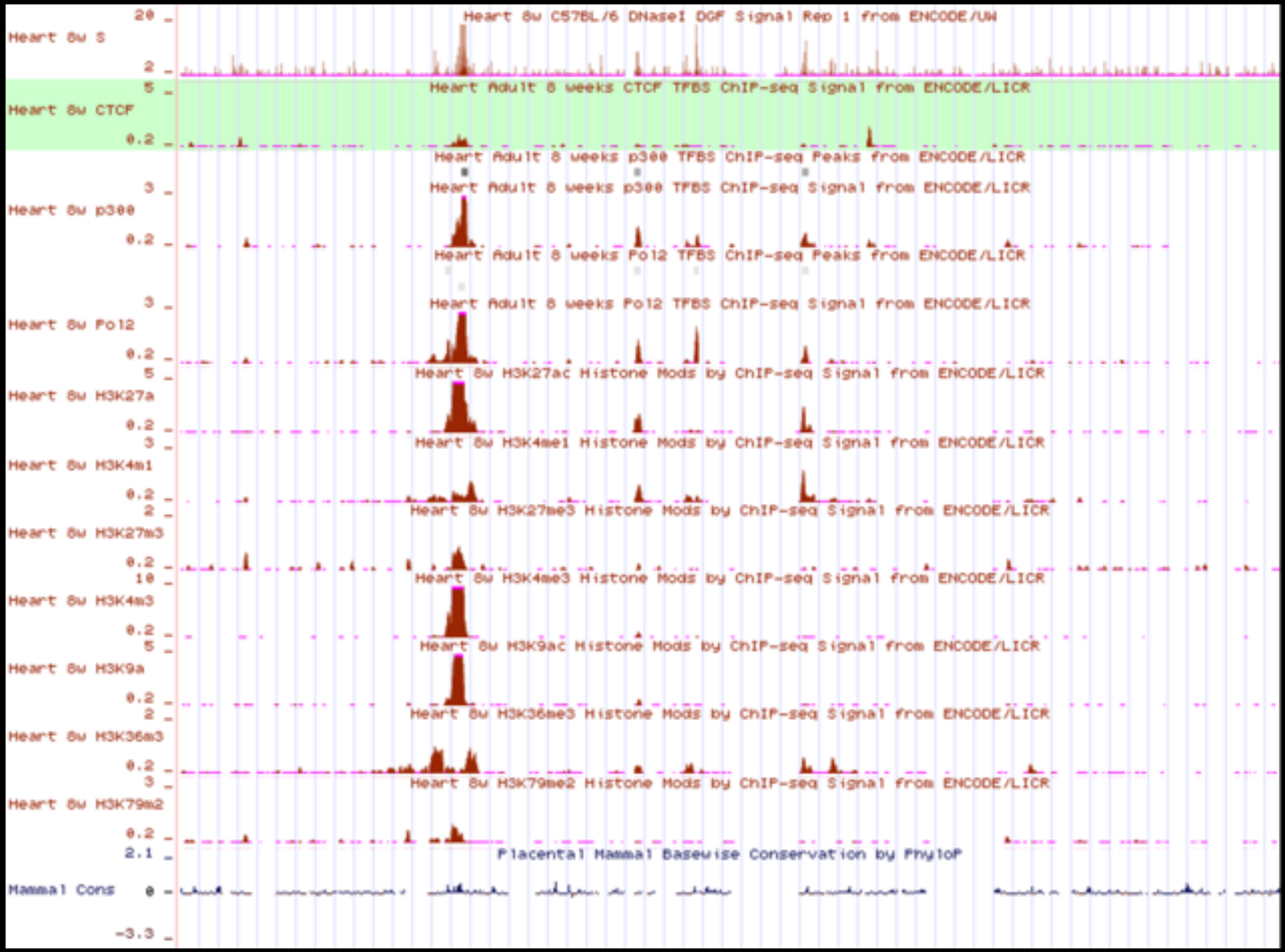


The usual suspects are important - H3K27ac, H3K4me1, H3K4me3, DNase, H3K27me3

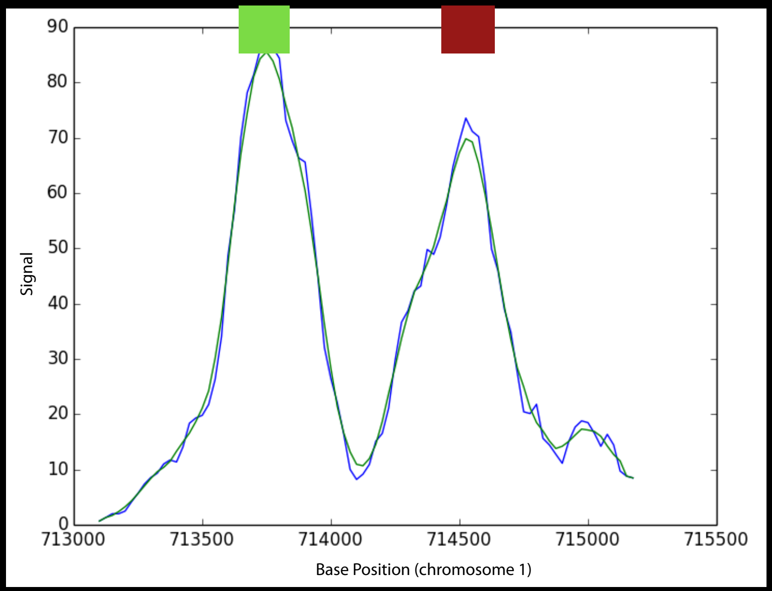
Features Used for Predicting Enhancers:

Nonlocal Model

Histone ChIP-Seq Signal (H3K27ac, H3K4me1, H3K4me3, H3K27me3 necessary)
Transcription Factor ChIP-Seq Signal (p300, CTCF, maybe pol2)
RNA-Seq Signal
Conservation (placental, vertebrate)
GC content

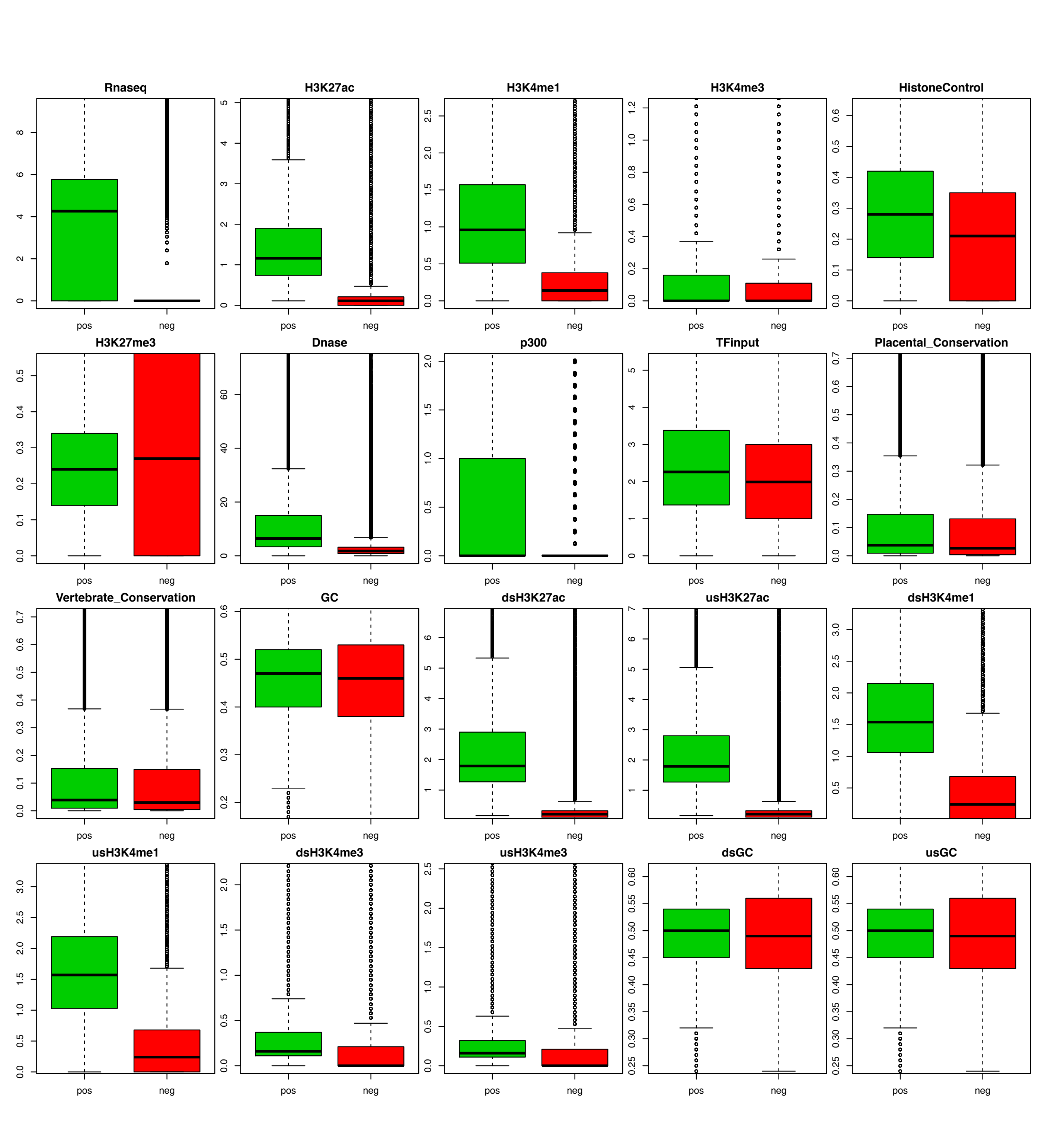

Heart 8w C57BL/6 DNaseI DGF Signal Rep 1 from ENCODE/UW
20
Heart 8w S
2
5
Heart Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR
Heart 8w CTCF
0.2
Heart Adult 8 weeks p300 TFBS ChIP-seq Peaks from ENCODE/LICR
3
Heart Adult 8 weeks p300 TFBS ChIP-seq Signal from ENCODE/LICR
Heart 8w p300
0.2
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR
3
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
Heart 8w Pol2
0.2
5
Heart 8w H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K27a
0.2
3
Heart 8w H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K4m1
0.2
2
Heart 8w H3K27me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K27m3
0.2
10
Heart 8w H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K4m3
0.2
5
Heart 8w H3K9ac Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K9a
0.2
2
Heart 8w H3K36me3 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K36m3
0.2
3
Heart 8w H3K79me2 Histone Mods by ChIP-seq Signal from ENCODE/LICR
Heart 8w H3K79m2
0.2
2.1
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0
-3.3

Upstream and downstream features



H3K27ac
H3K4me1
H3K4me3
GC content

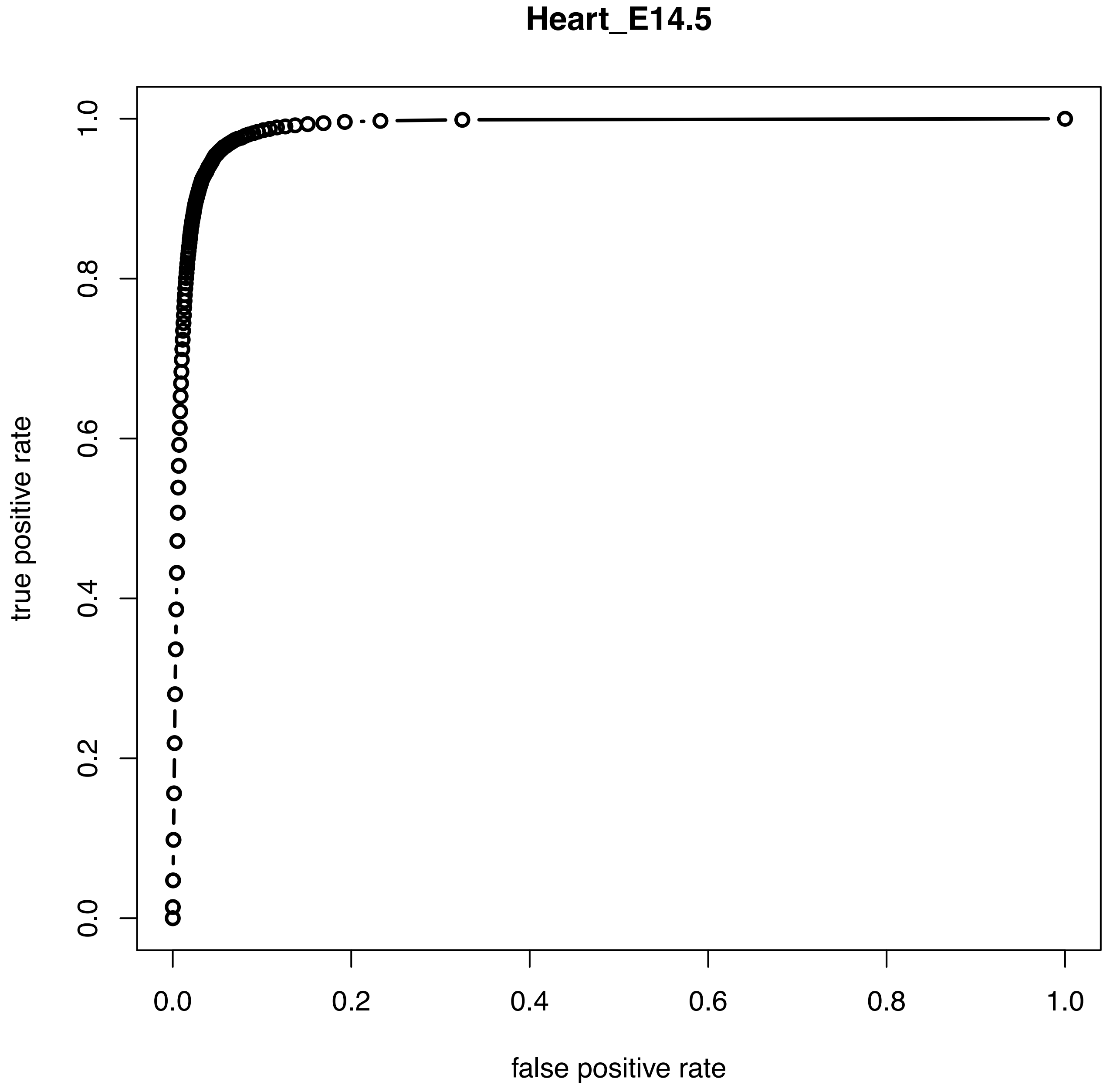
Nonlocal Model - Differences between positives and negatives



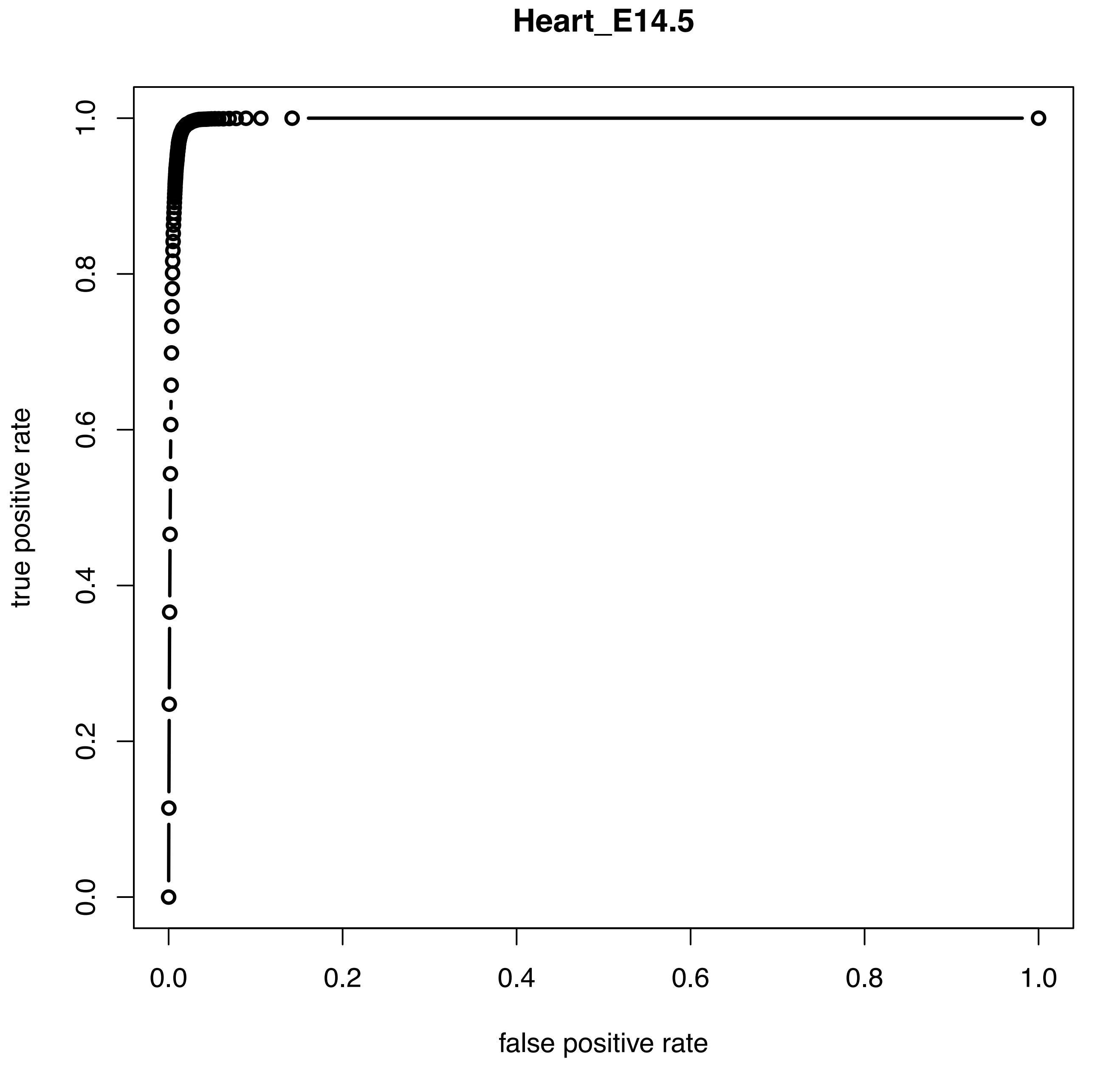
upstream and downstream H3K27ac and H3K4me1 signals have big differences.

Nonlocal Model - AUC plot

AUC = 0.94

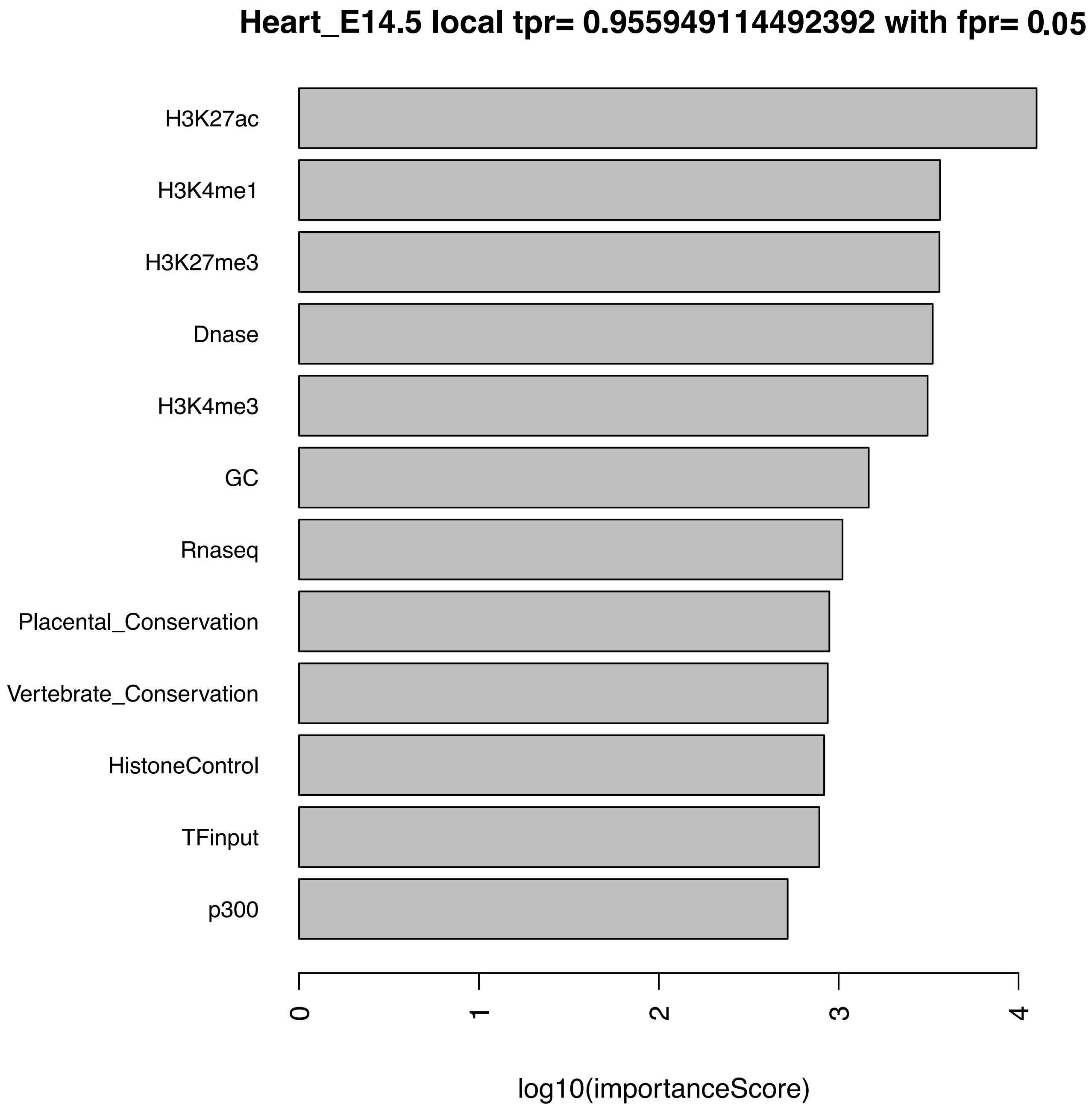


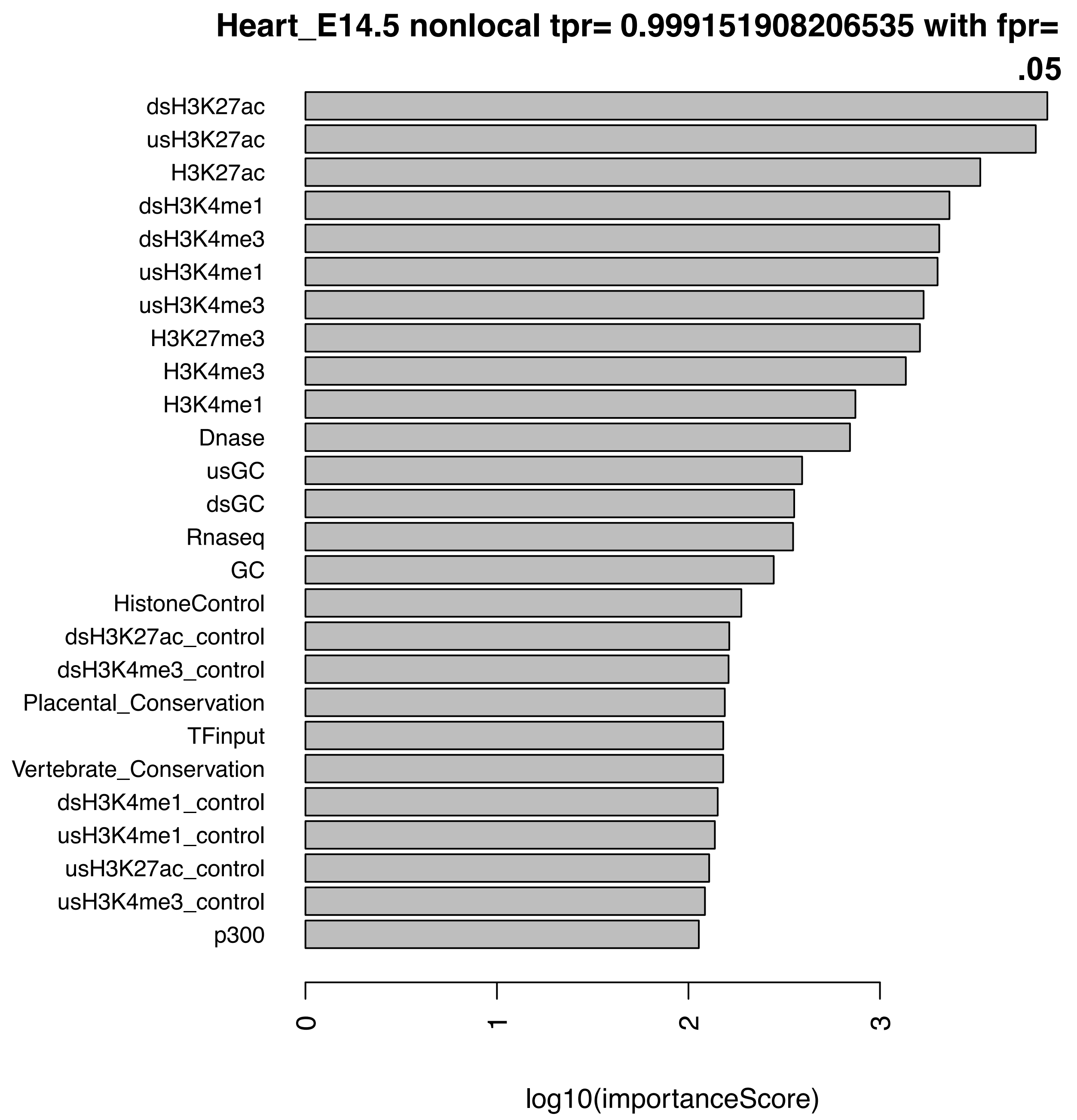
AUC = 0.999



Including features from flanking regions improves the accuracy of model

Nonlocal Model - results - importance of features


Heart_E14.5 local tpr= 0.955949114492392 with fpr= 0.05


Heart_E14.5 nonlocal tpr= 0.999151908206535 with fpr= .05

Conclusions

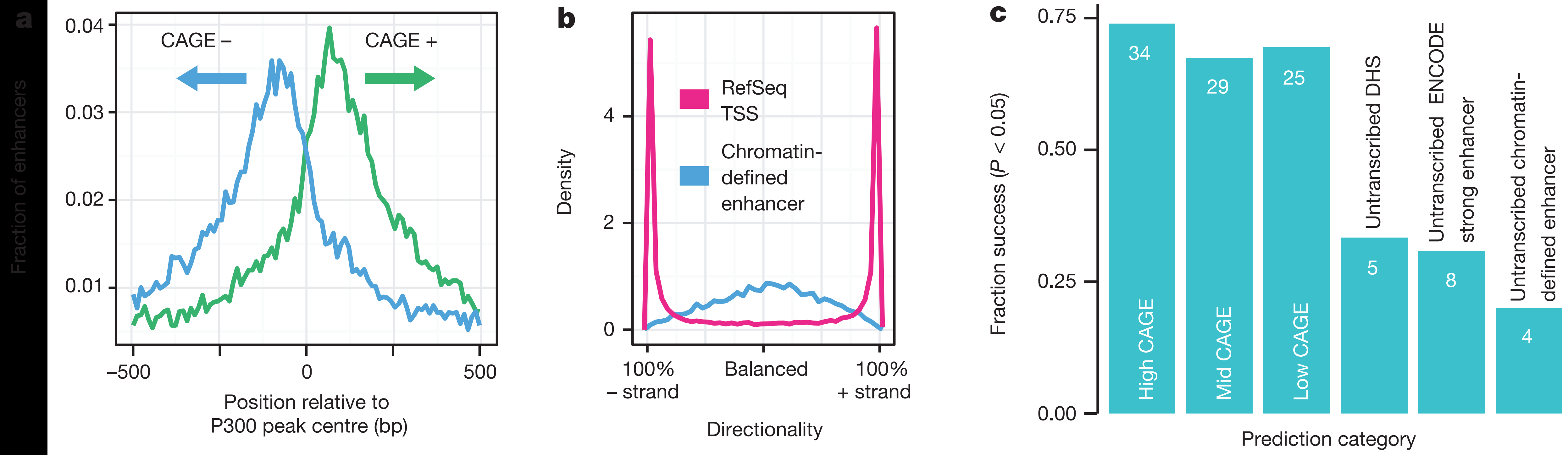
Nonlocal features rock!

Future Work

Add new datasets from mouse ENCODE and redo the models.
Analyze overlap of different predictions with our predictions (similar analysis with VISTA positives and negatives).
Use new method on human data.
Move on to developing methods for target predictions.

Extra Slides

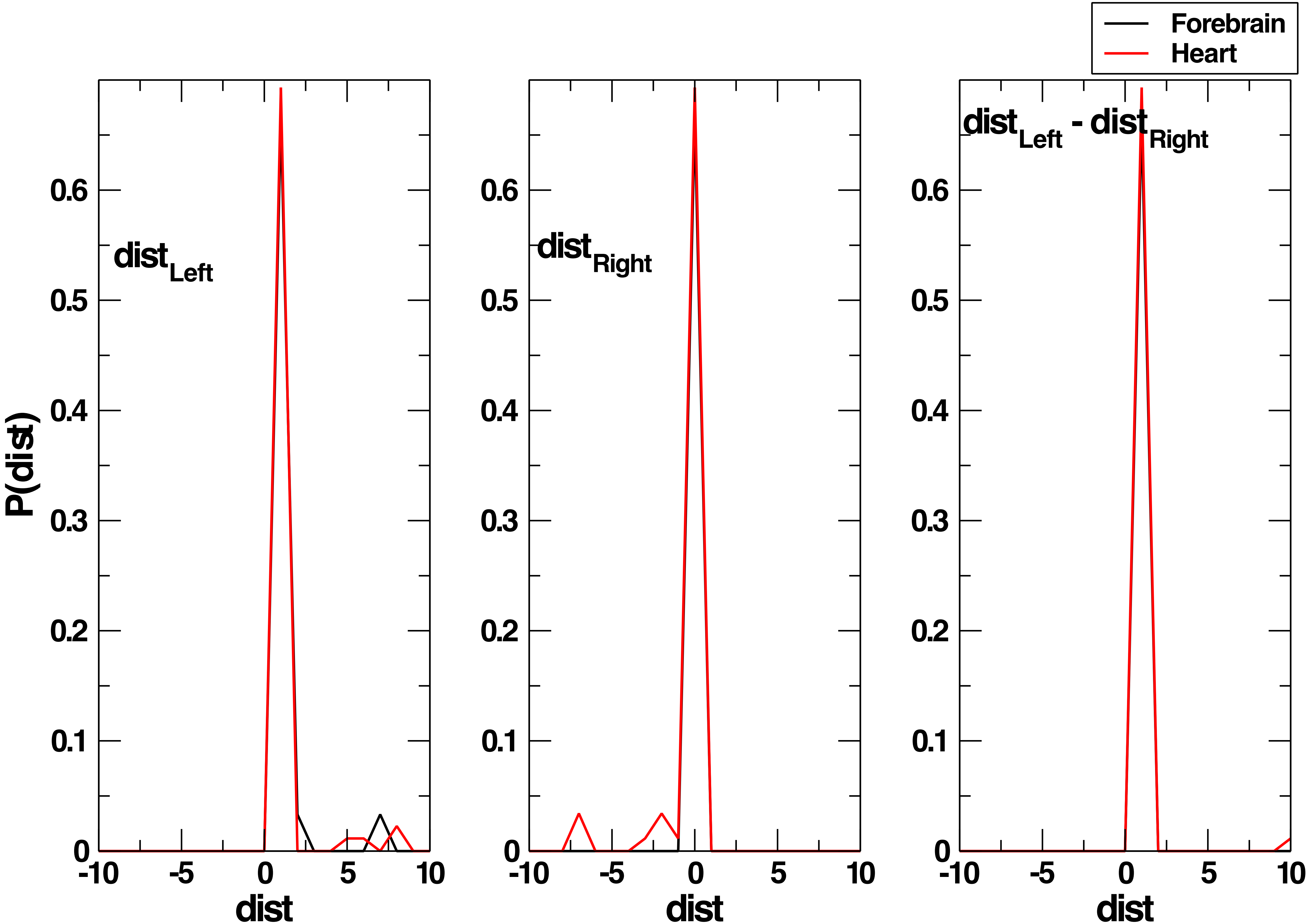
A Trend that Didn’t Work Out



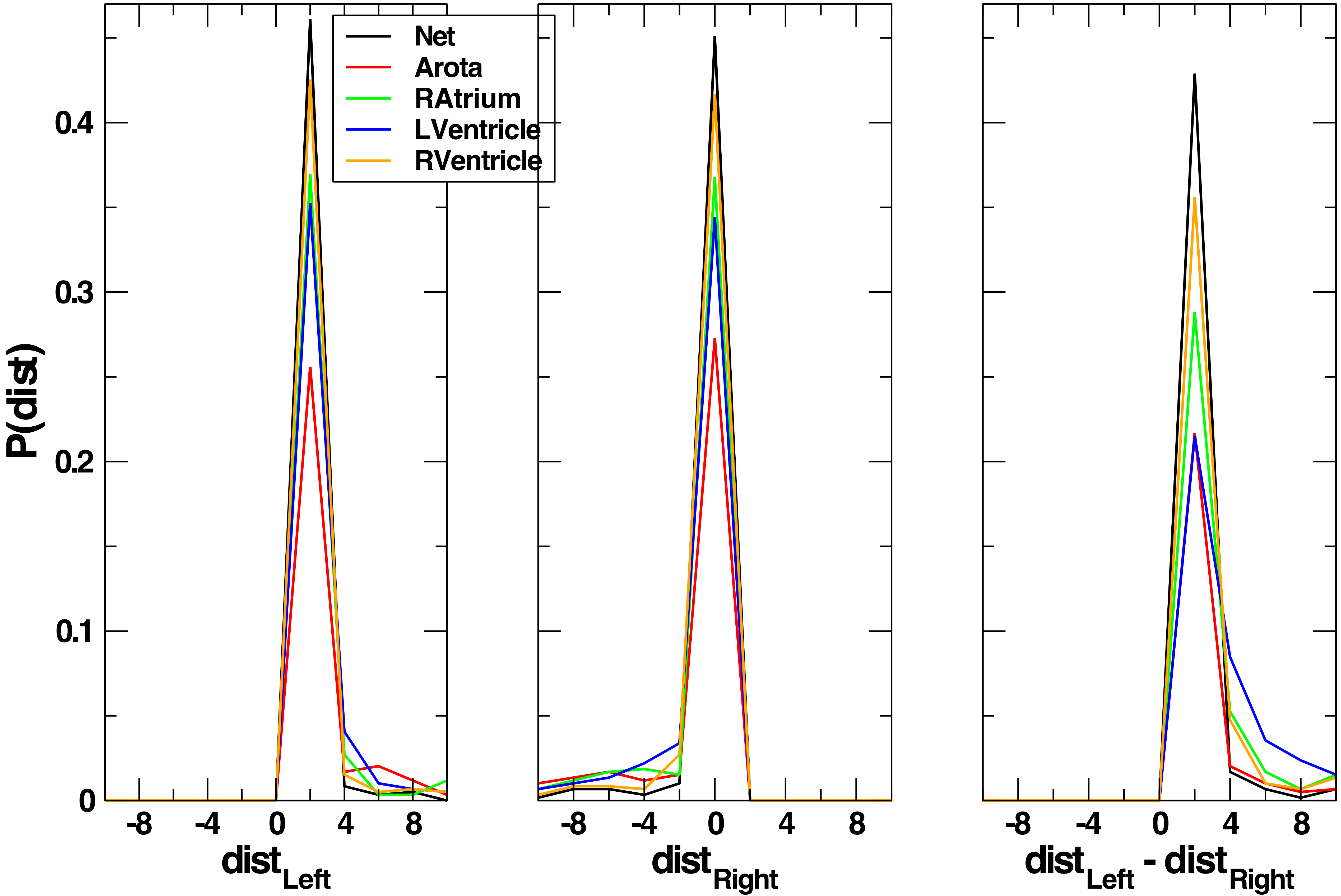
This pattern of CAGE peaks was only present in 1 out of experimentally verified active enhancers in mouse E11.5

FANTOM consortium paper, Nature, 2014

Vista list based analysis



FANTOM based analysis


Net
Arota
RAtrium
LVentricle
RVentricle
P(dist)
$\mathrm{dist}_{\mathrm{Left}}$
$\mathrm{dist}_{\mathrm{Right}}$
$\mathrm{dist}_{\mathrm{Left}} - \mathrm{dist}_{\mathrm{Right}}$