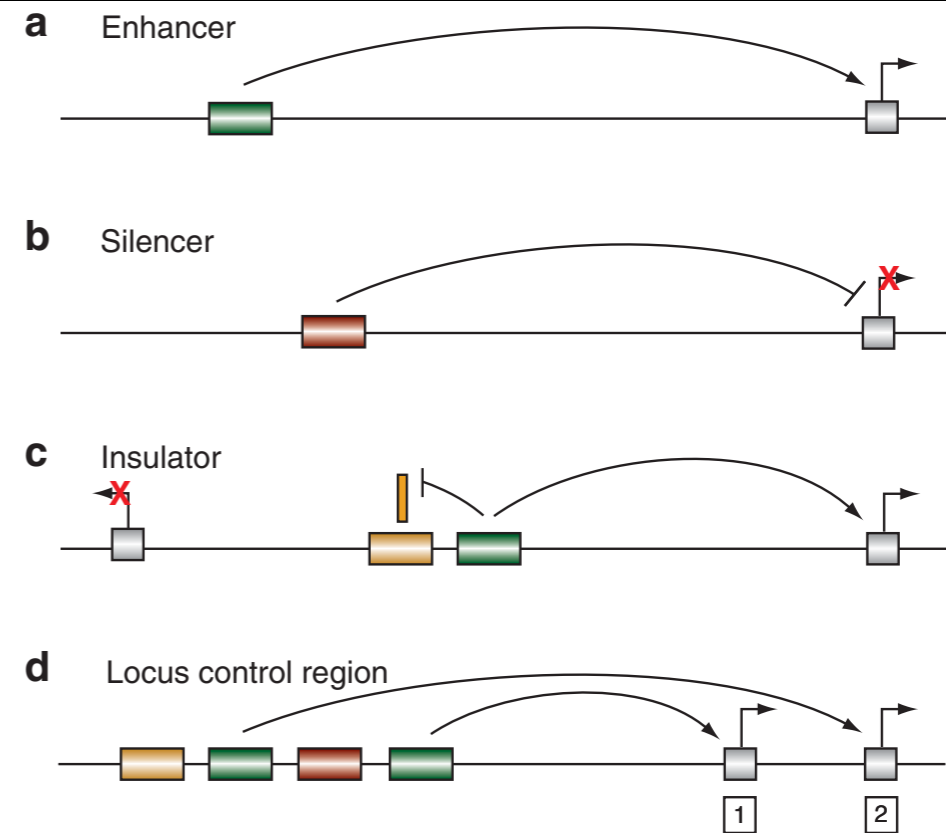
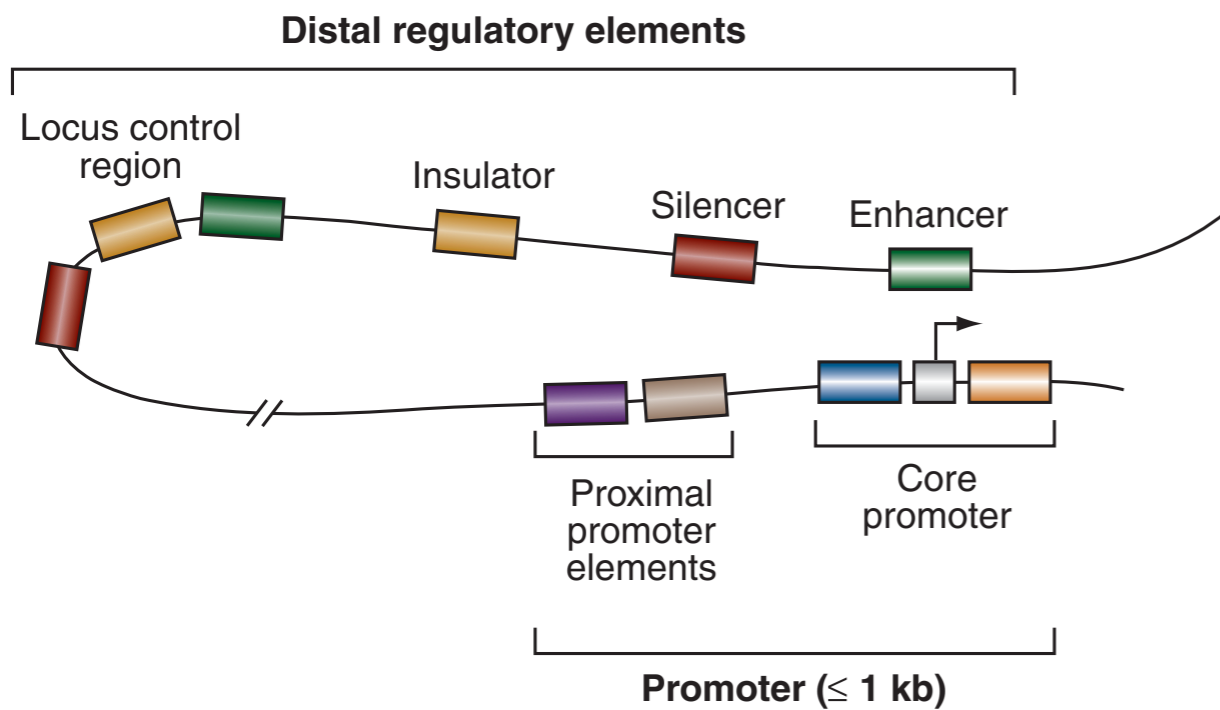


The big bad messy world of enhancers

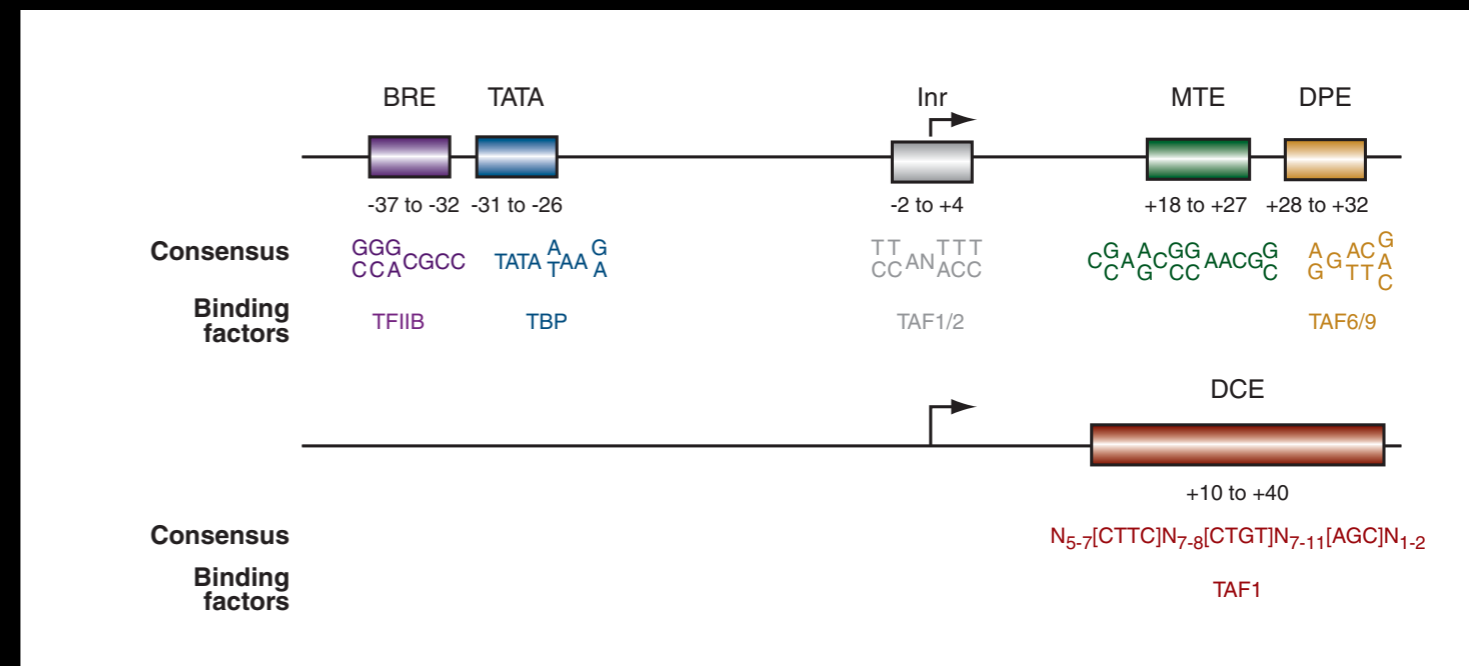
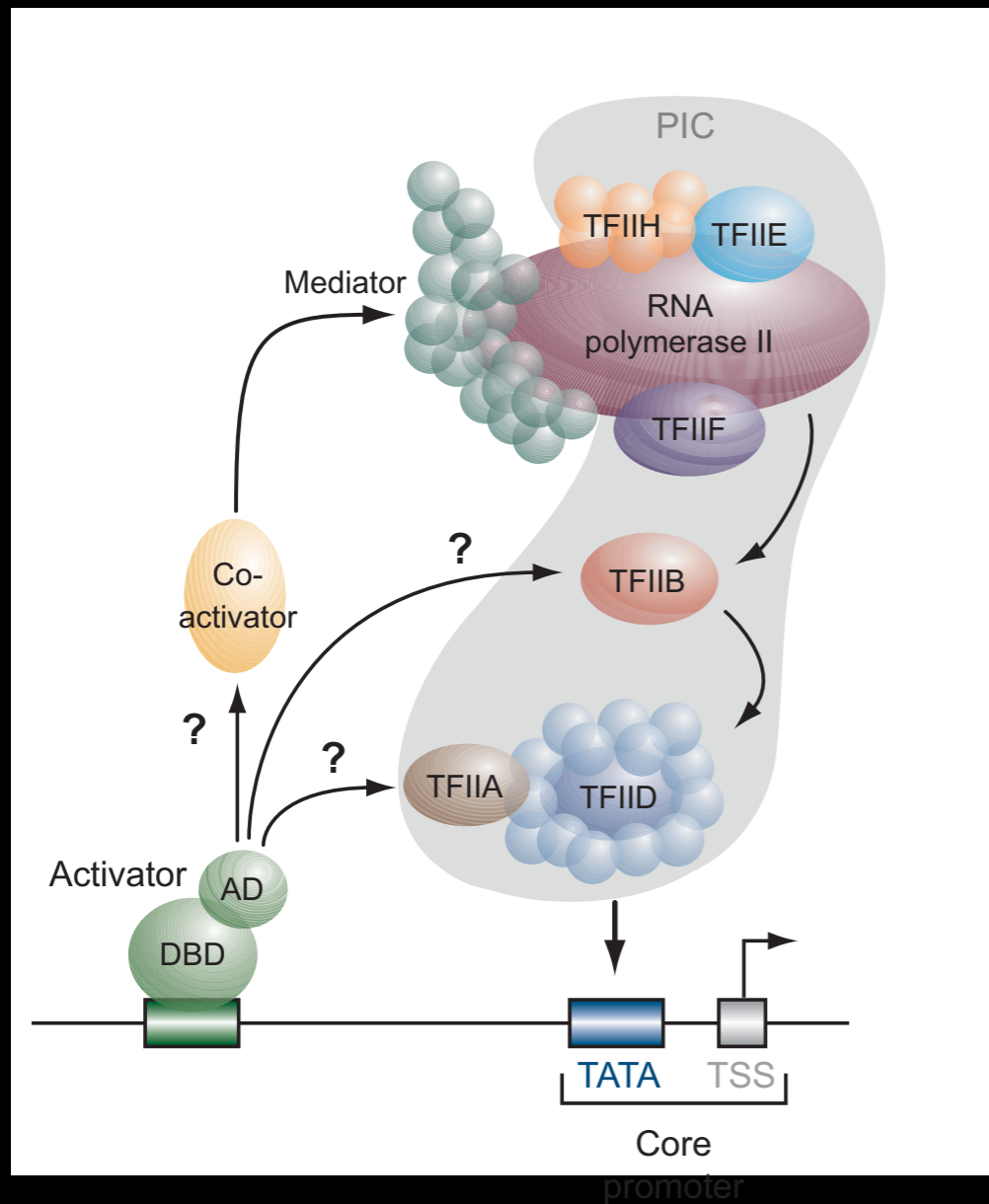
gpmtg 2016
Anurag Sethi

What is a transcriptional enhancer?



There are various categories of noncoding regulatory elements.

The core promoter is required for basal transcription



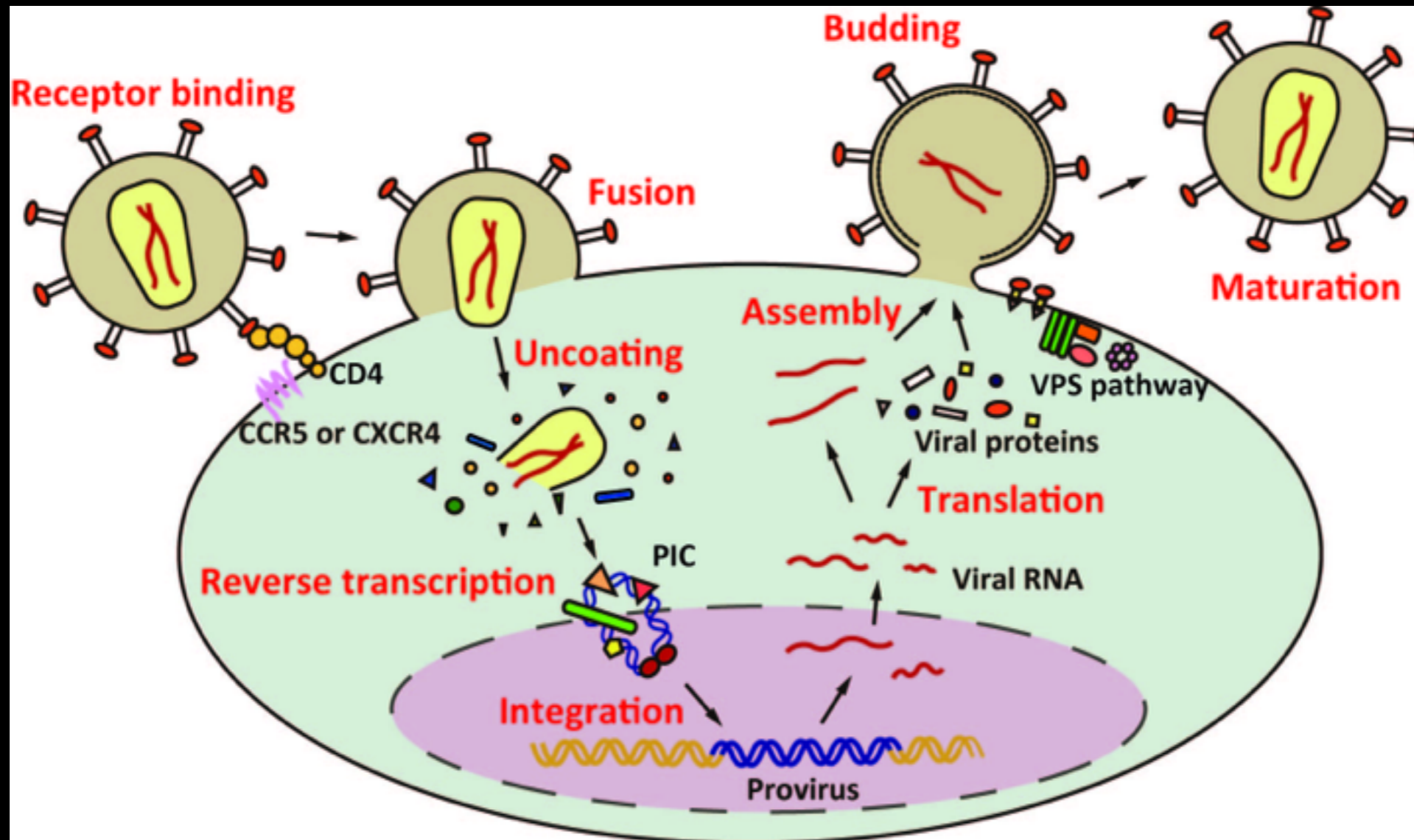
The pre-initiation complex (PIC) forms on the core promoter. There are various different kinds of core promoters (and not all of them are known/characterized).

How do you test the activity of an enhancer? - easy assay



Plasmids with region to be tested for activity and luciferase/GFP gene can be used to test regions for enhancer activity

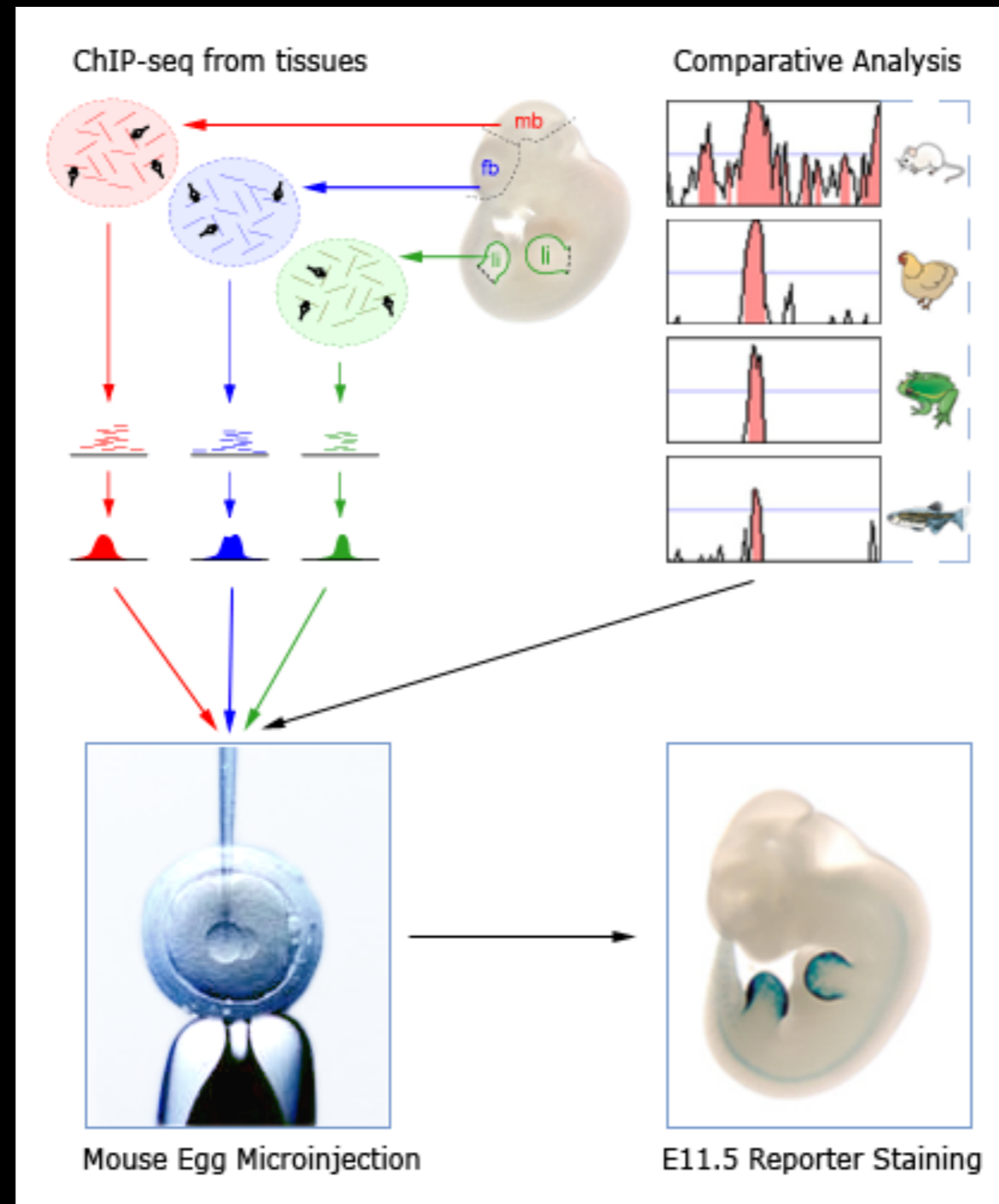
How do you test the activity of an enhancer? - harder assay



Integrate the gene + probable enhancer into the genome and transcription has to take place in genomic environment before it is tested.

Note that this can be done either *in vivo* or in culture.

How do you test the activity of an enhancer? - Transgenic assay



In Len's assay the tested region is inserted into host DNA within the mouse egg to ensure that the mouse embryo has this region as it grows - you get a single assay to test for all tissues at a particular timepoint.

Artifacts from assays

	Transfection (plasmid)	Transduction (integration)	Transgenic (Len)
Native Context	No	No	No
Chromatin Context	No	Typically present	Typically present
Copy Number Artifact	Yes	Can be avoided	Can be avoided
<i>In vivo</i>	Depends	Depends	Yes
Integration bias	No	Yes	Yes
Core promoter dependence	Yes	Probably	Probably



Difficulty of assay increases
Confidence of assay increases

A number of massively parallel assays have been developed in the last 5 years for testing enhancer activity

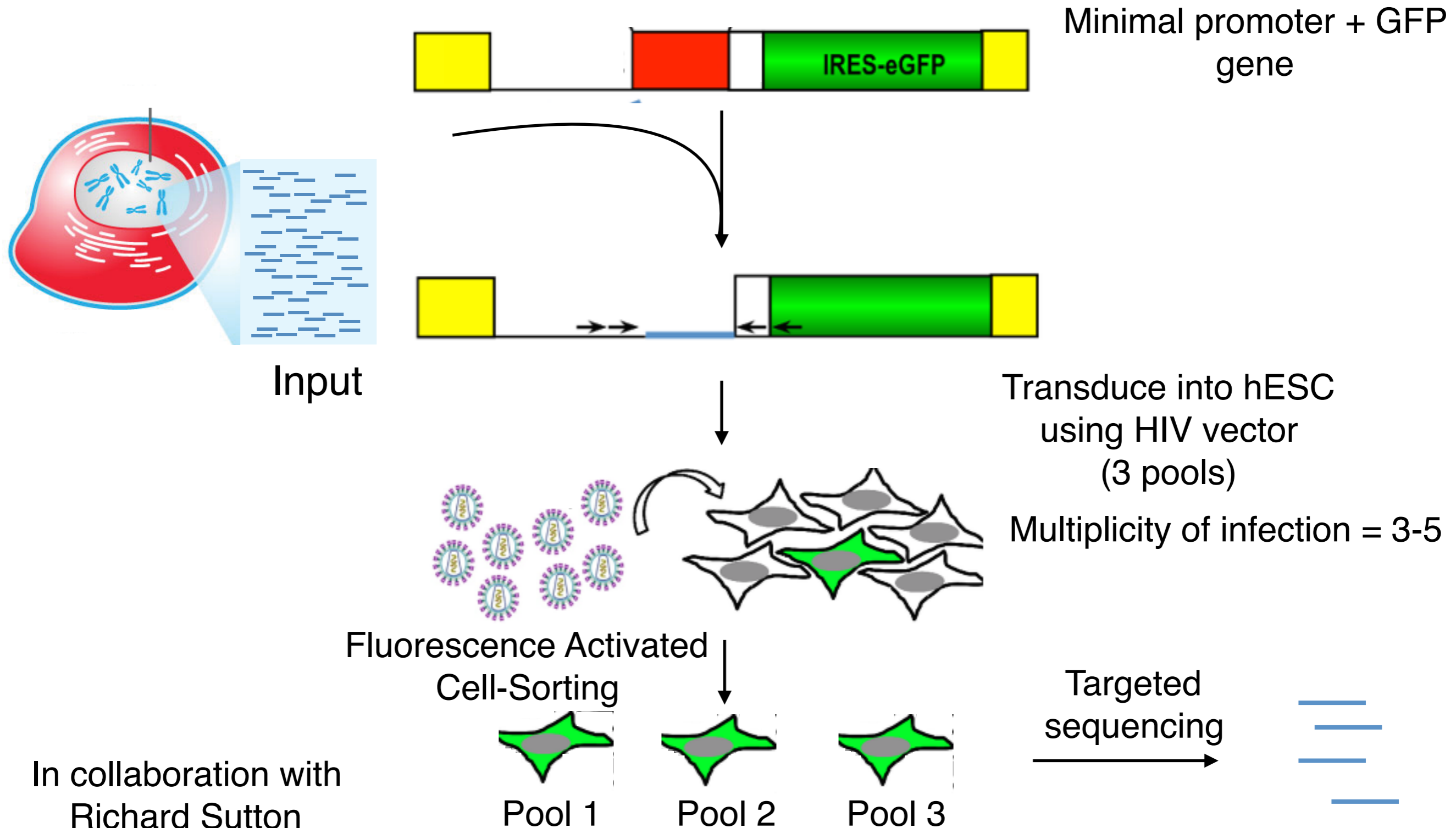
Technique	Plasmid/Chromatin	Length of element tested	Elements
<i>In-vitro</i> transcription (Shendure, Nat. Biotech, 2009)	In-vitro (100K)	200 bp / 3-4 promoters	Effect of variants
MPRA (Tarjei, Nature Biotech, 2012)	Transfection/ human cells (40K) - RNAseq	87 bp / 2 enhancers	Effect of variants (indels/subs)
MPFD (Shendure, Nat. Methods, 2012)	Transfection/mouse (100K)	1kb /3-4 enhancers	Effect of variants
eFS (Bulyk, Nat. Methods, 2013)	Transduction/fly 1 clone per cell (500)	1 kb/ ChIP-seq of TF	Finding enhancers
STARR-Seq (Stark, Science, 2013)	Transfection/fly (whole genome - fly)	600 bp/whole genome	Finding enhancers
CRE-Seq (Cohen, Genome Res, 2014)	Transfection/human (2000 x 2)	132 bp/chromHMM and Segway	Accuracy of predictions
FIREWACH (Dailey, Nature Methods, 2014)	Transduction/mouse 1 clone per cell (80K)	100-300bp/DNase	Finding enhancers
SIF-Seq (Pennacchio, Nature Methods, 2014)	Transduction/mouse 1 clone per cell (500kb)	1-2 kb/specific regions of genome	Finding enhancers

MPRAs

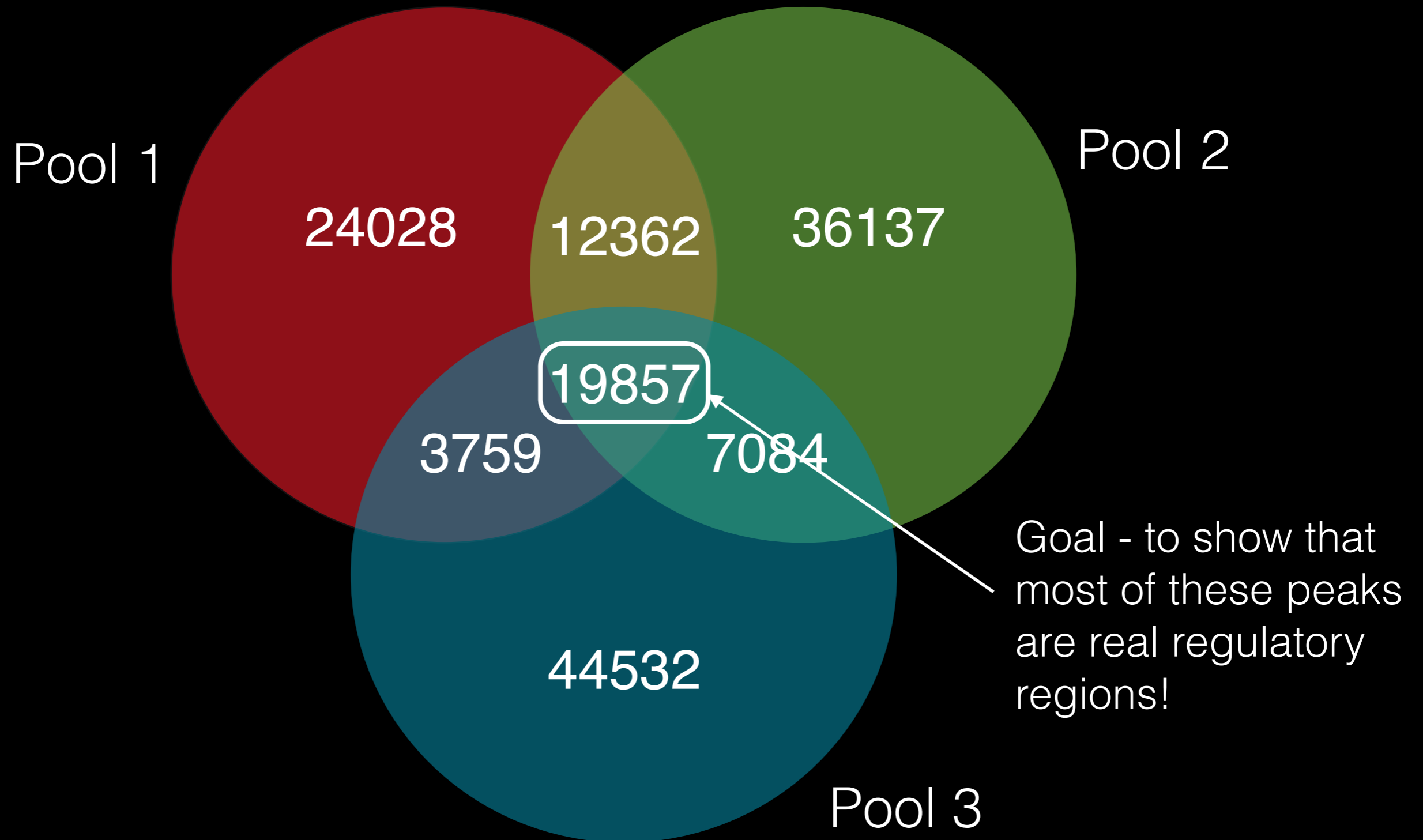
None of these assays have worked on the whole genome for mammalian cells so far (published studies).

Transduction > Transfection (chromatin context)

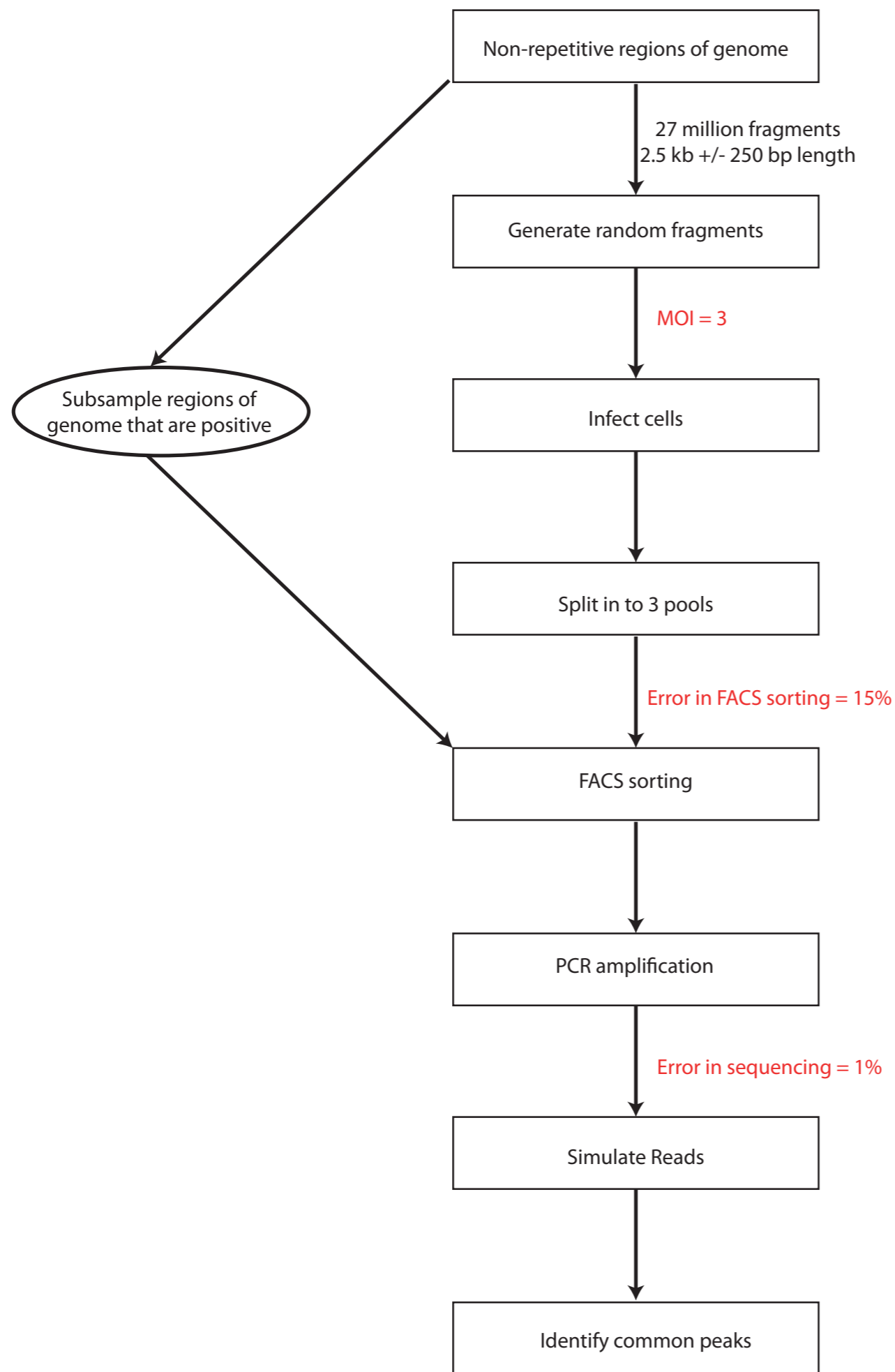
ImPACT-Seq - Immensely Parallel Assay using Cellular Transduction - A new massively parallel assay for identifying regulatory regions in the genome.



ImPACT-Seq - identifying regulatory regions in the genome

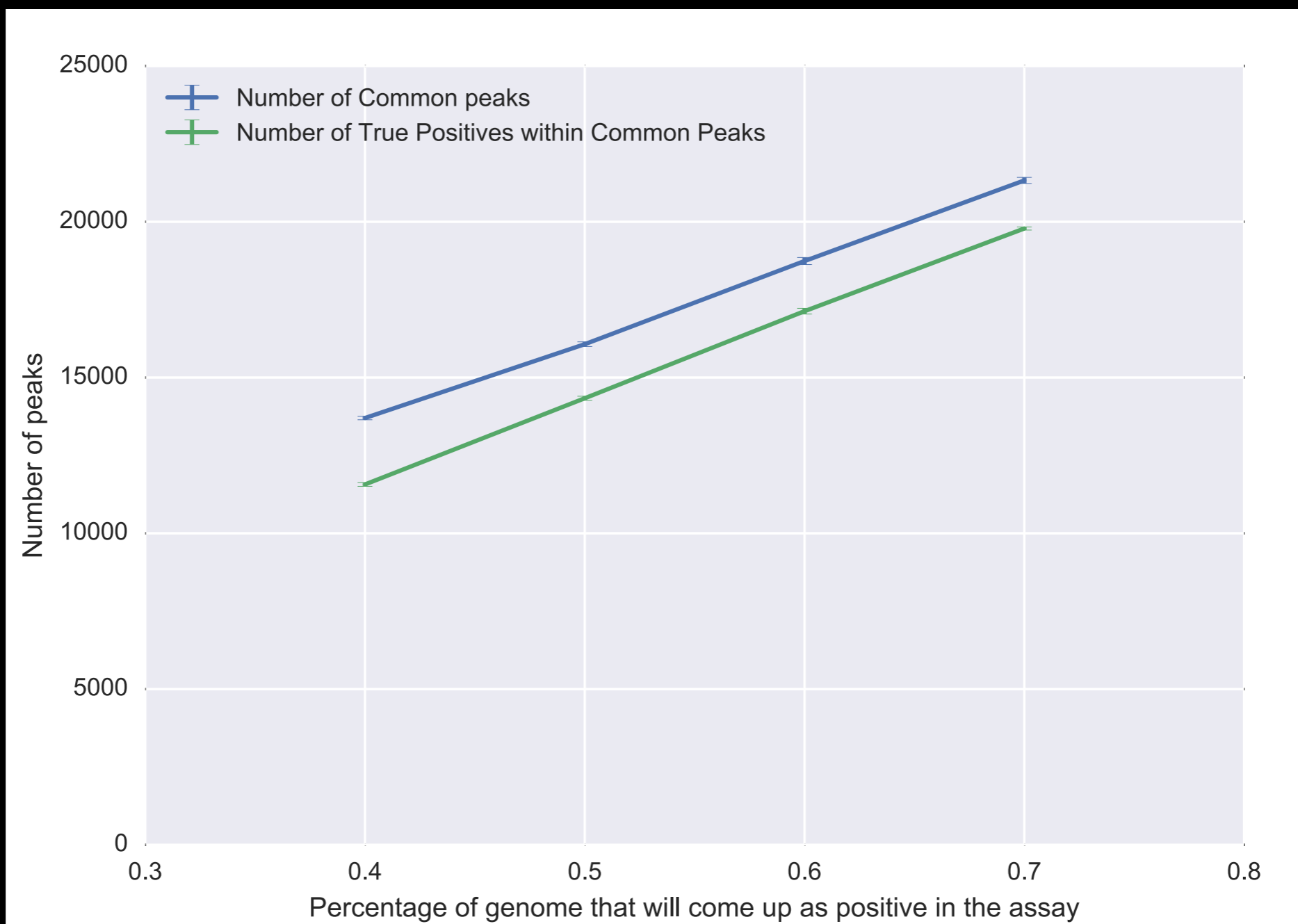


25-33% of peaks within a pool are consistent with peaks across all three pools.



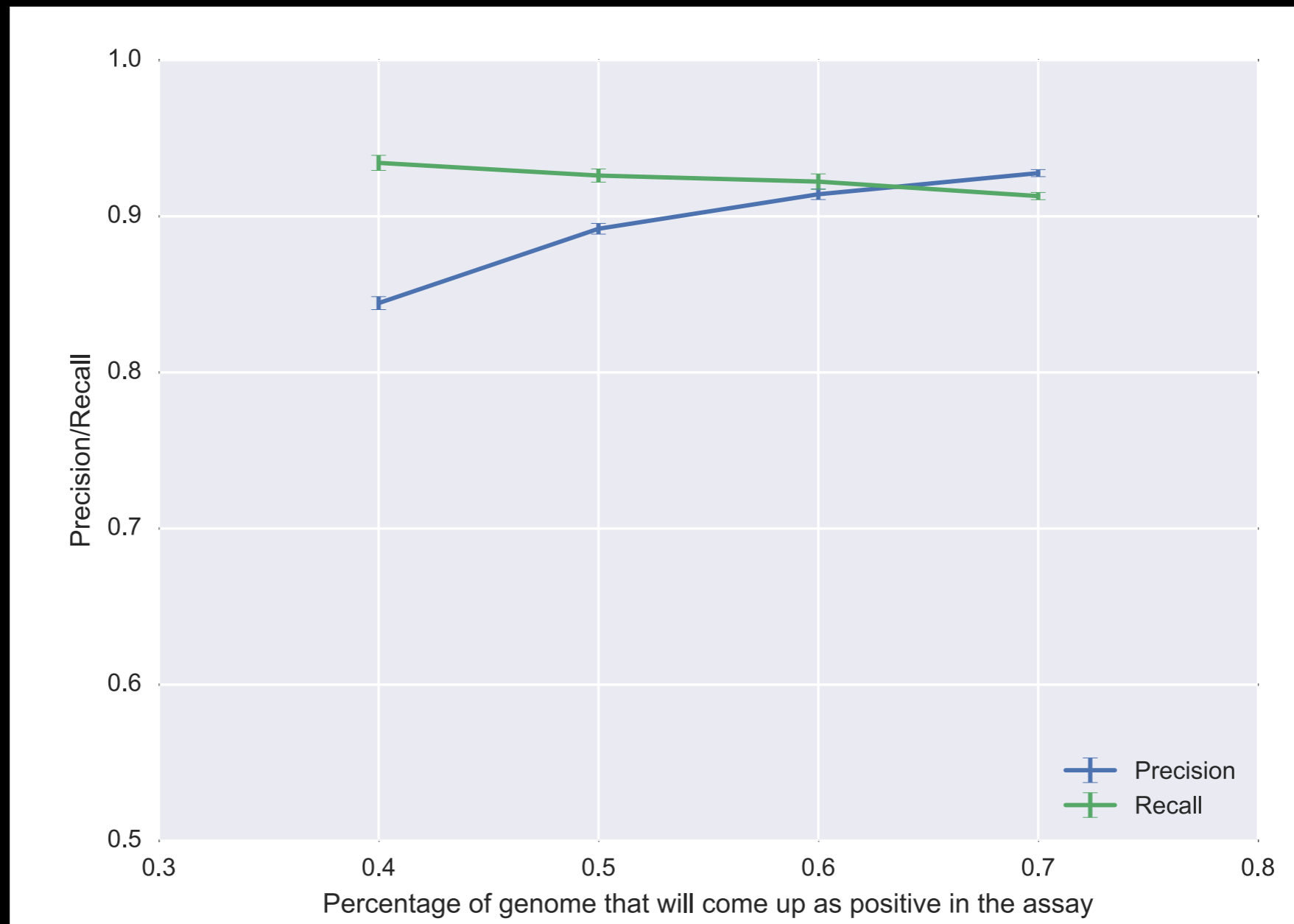
Simulate what we think is happening in the experiment

ImPACT-Seq simulations can be used to identify what genome we expect to be positive in assay



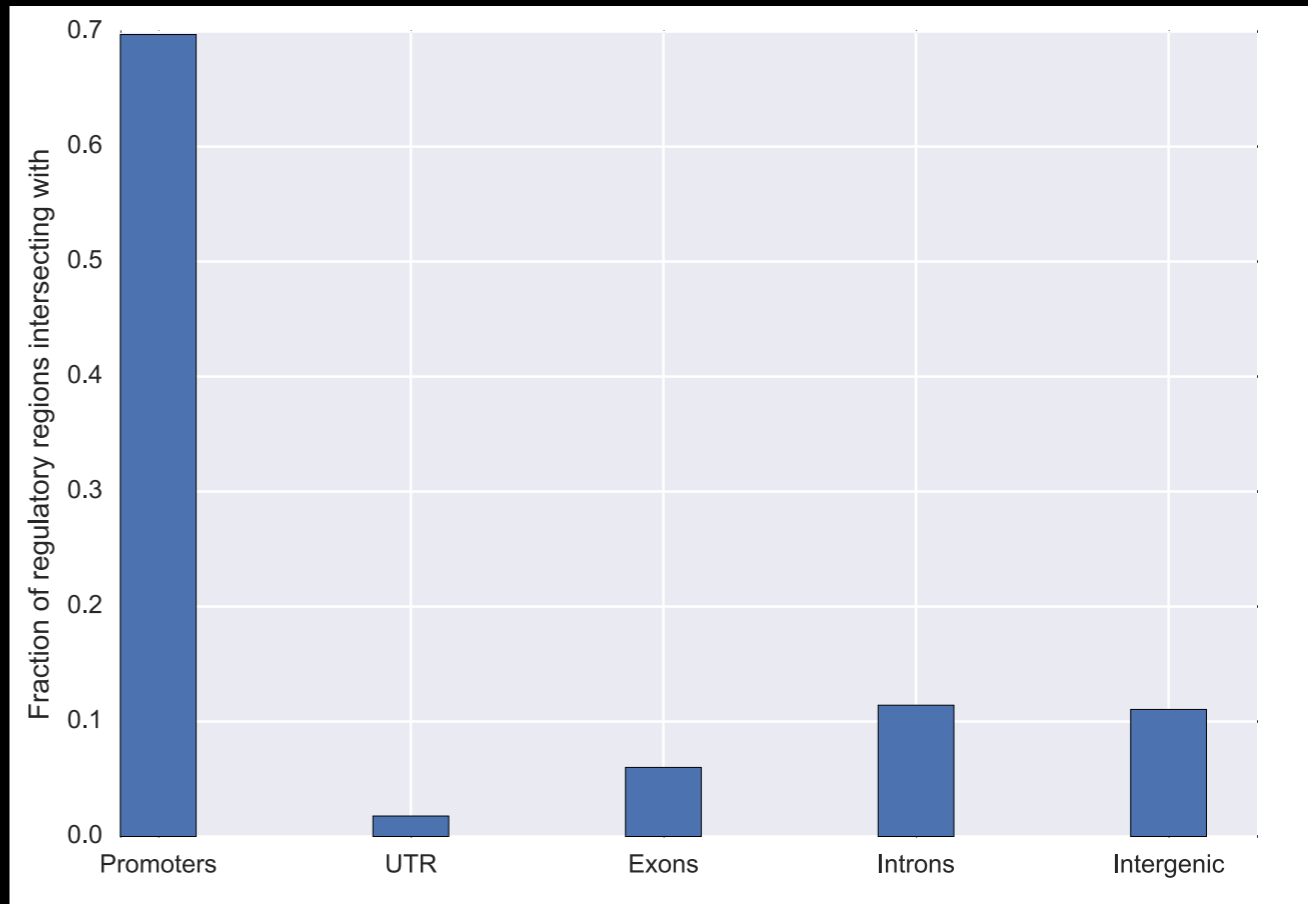
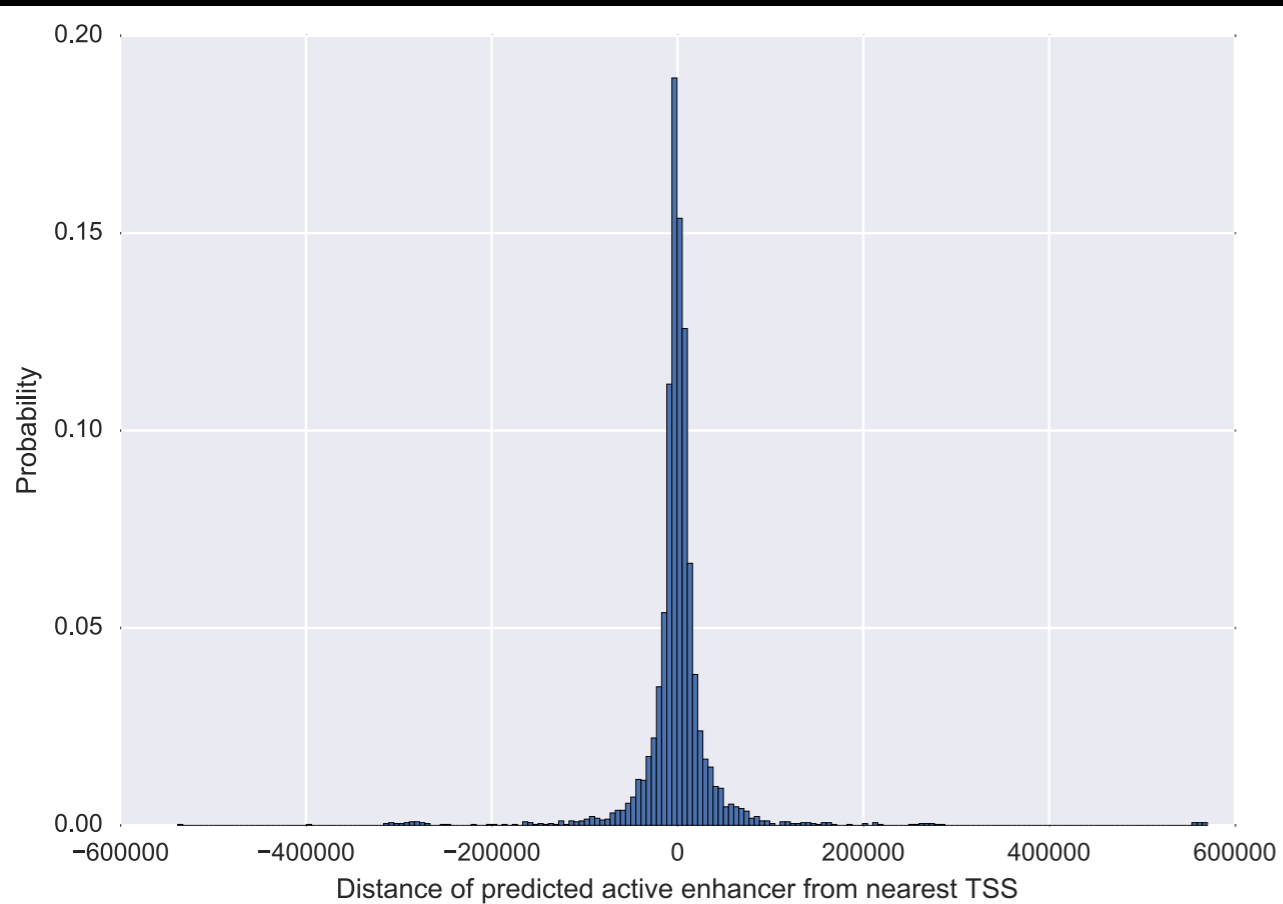
About 0.6-0.7% of genome could be expected to come as positive in the assay.

ImPACT-Seq simulations can also be used to calculate the expected precision and recall in the assay



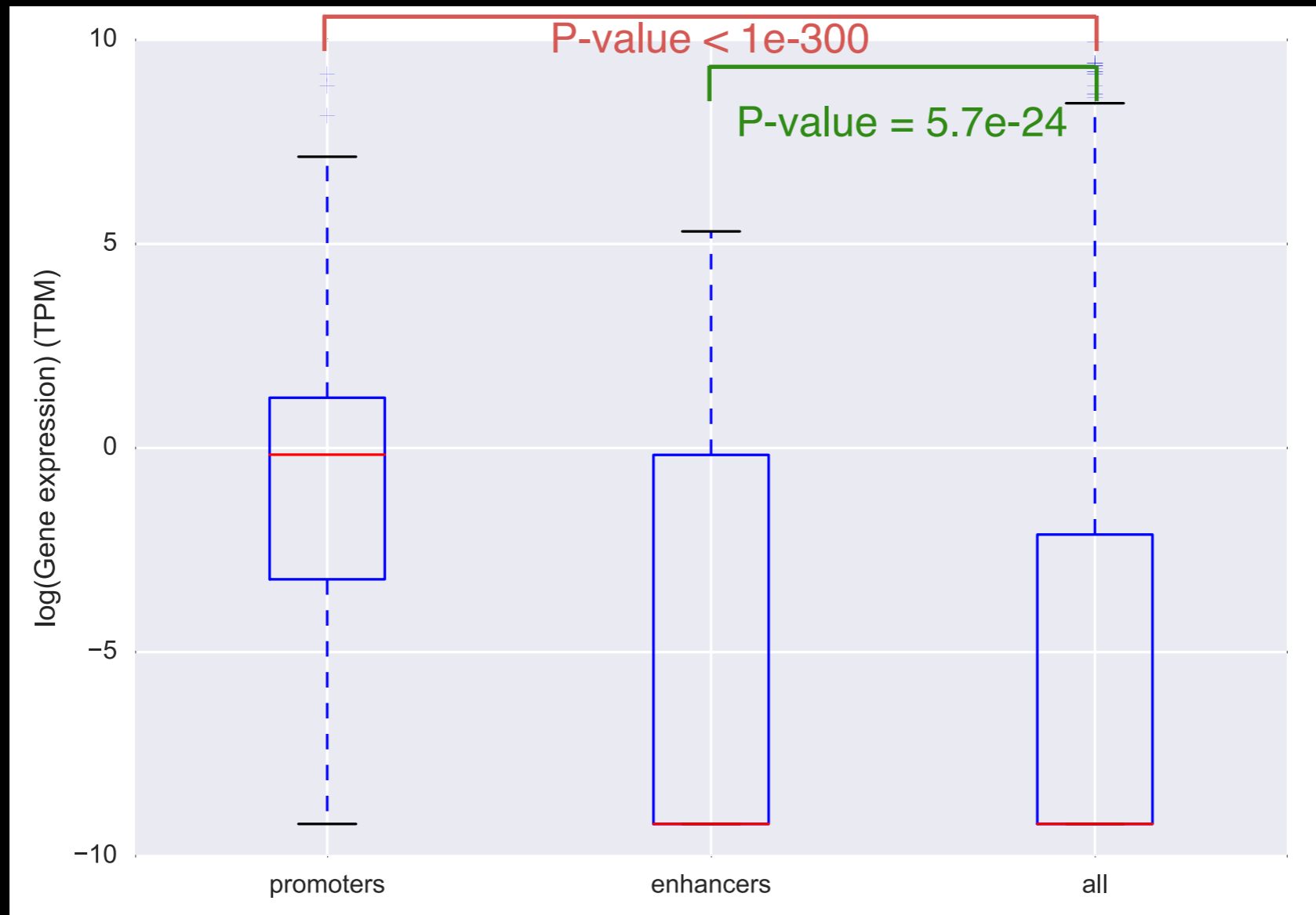
The FDR is expected to be <10% according to these simulations.

ImPACT-Seq identifies regulatory elements that are close to a number of active genes









- Fewer enhancers come out as positives in this assay might be due to :
- Enhancers are weak promoters and may be more prone to be mistaken as negatives in FACS sorting.
 - All enhancers might not function as promoters during this testing.

ImPACT-Seq promoters and enhancers are closer to active genes



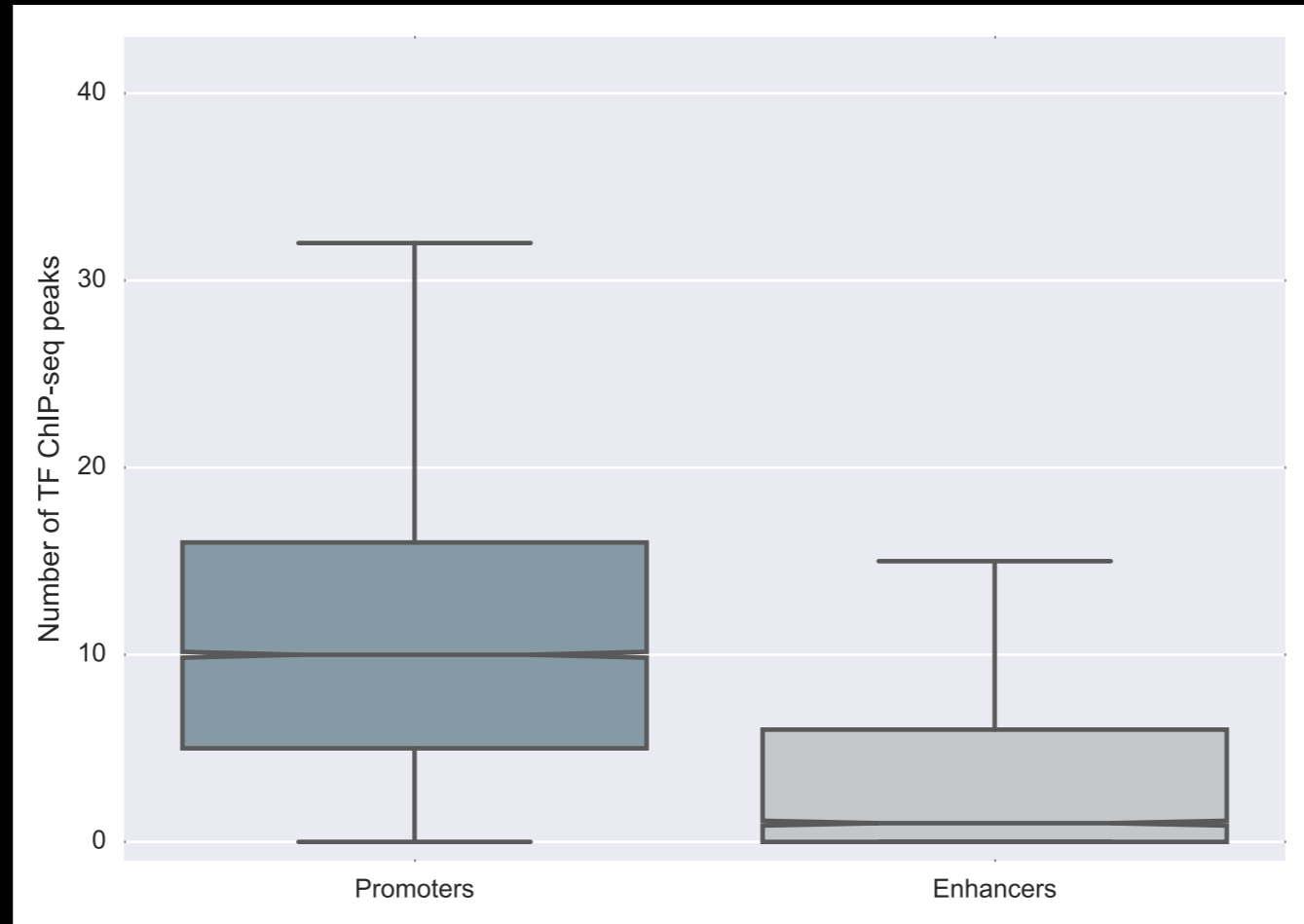
Enhancers may not be regulating closest gene in a majority of cases.

Sequence determinants for regulatory regions

De novo Motifs	E-Value (DREME)	Motif Matches
	1.4e-197	SP/KLF family (SP1, SP2, KLF4, KLF5) E2F family (E2F1, E2F3, E2F4, E2F6) EGR family (EGR1, EGR2) GC box
	3.9e-111	SP/KLF family (SP1, SP2) ZNF263/MZF-1 family
	1.9e-107	NRF1
	5.3e-84	E2F family (E2F1, E2F3, E2F4, E2F6) SP/KLF family (KLF4, KLF5) STAT (STAT1, STAT3) ETS family (ELK1, ELK4, SPIB, GABPA)
	1.3e-24	AP2 family (TFAP2A, TFAP2C) EBF1 Zfx
	3.0e-21	Helix-loop-helix family (TCF3, TCF12, Myod1, Myog, Atoh1, NHLH1) RFX5 CTCF

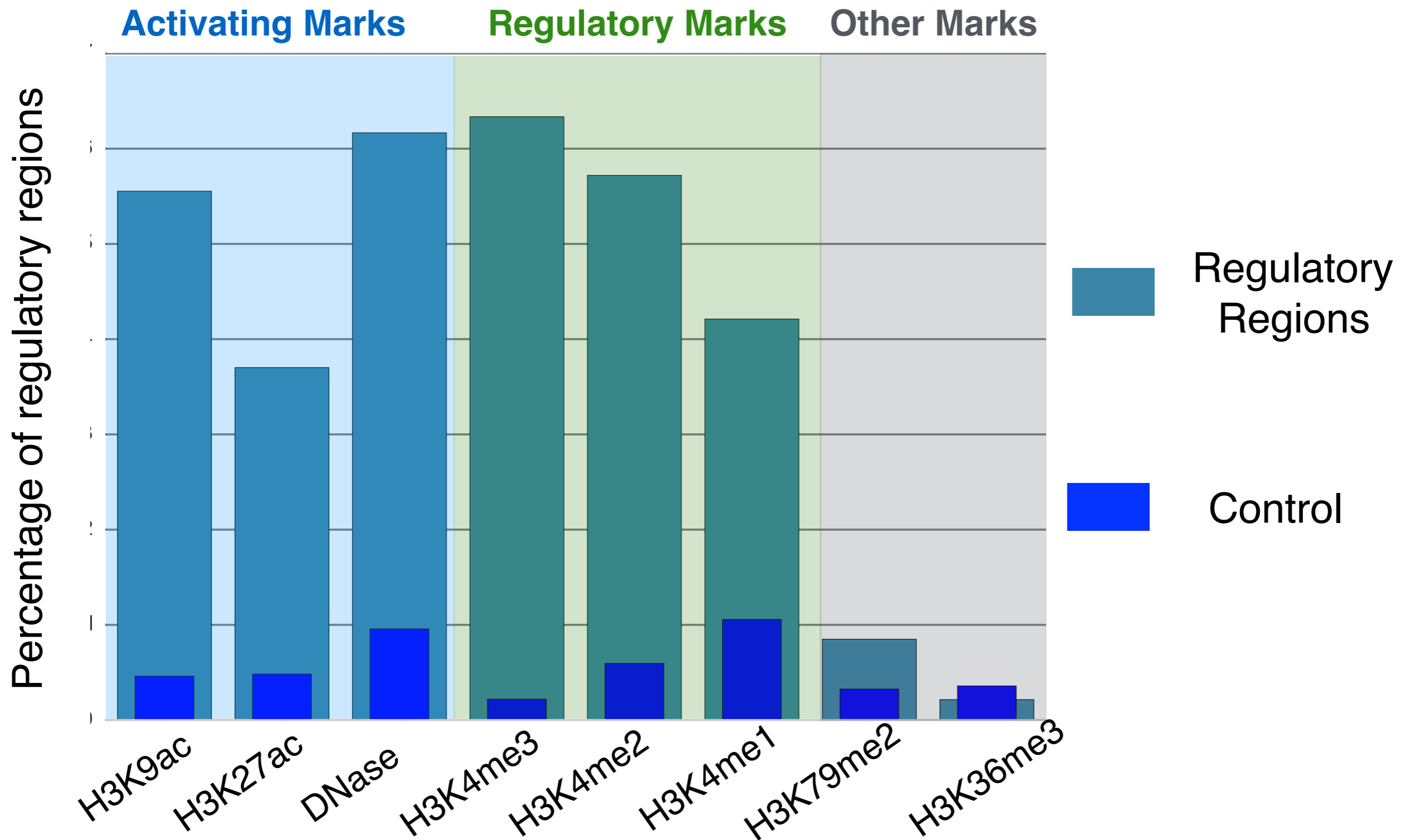
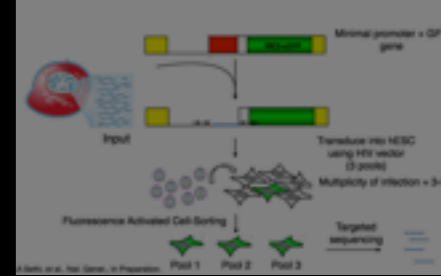
Known **transcription factor binding motifs** are enriched in the identified regulatory regions.

A lot more ENCODE2 ChIP-Seq peaks occur on promoters as compared to enhancers

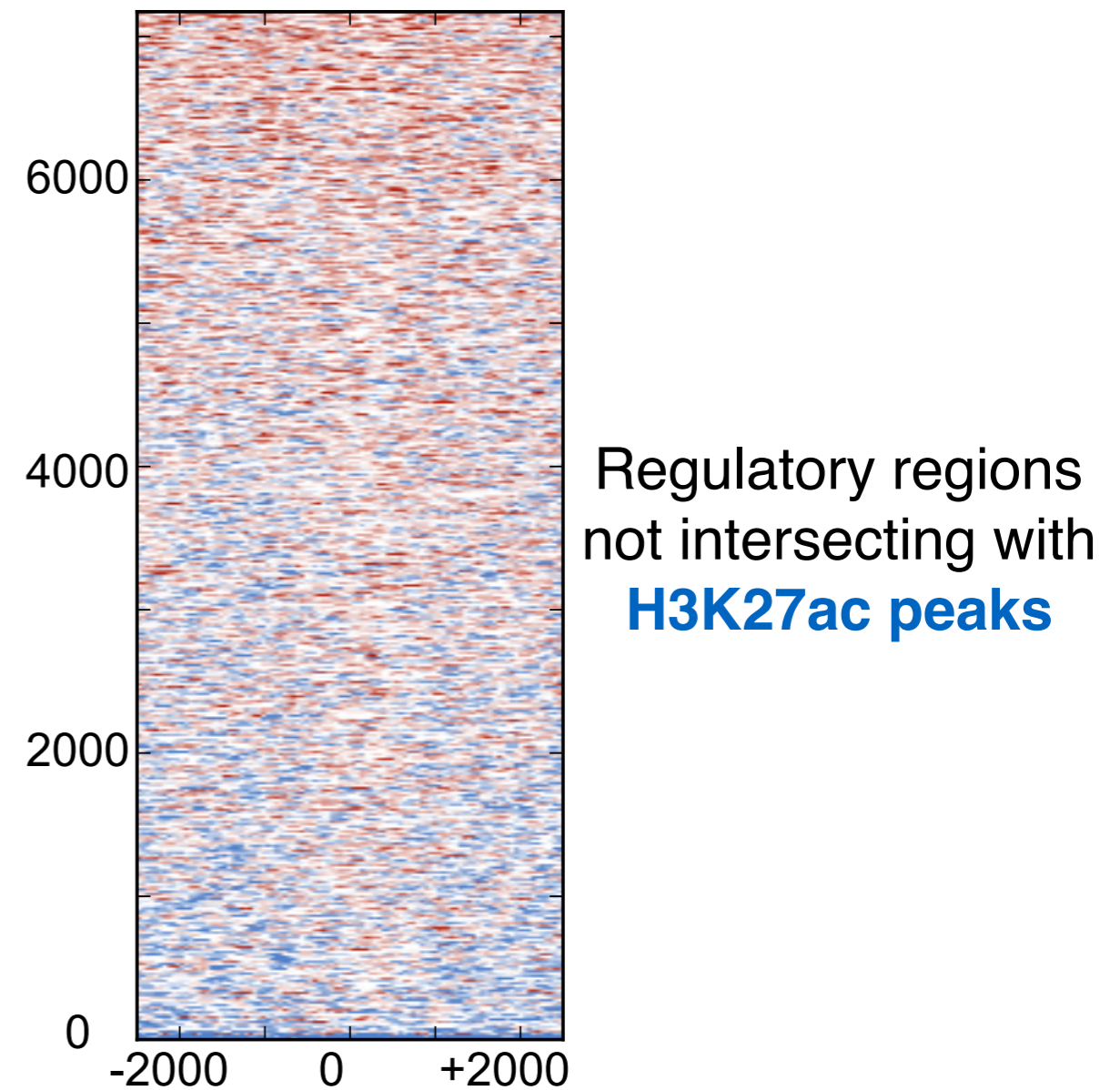
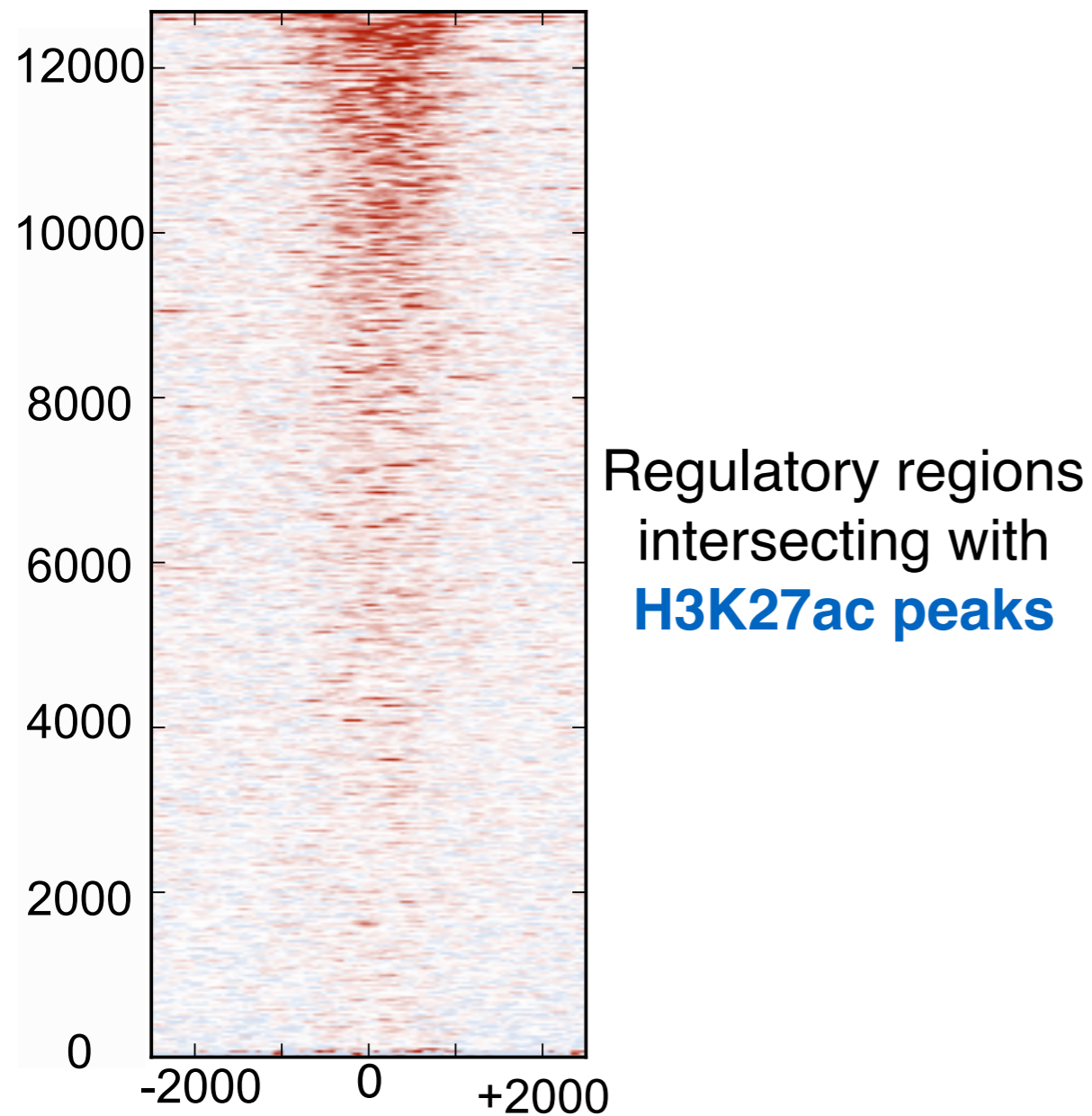


Enhancers may be more heterogeneous and ENCODE2 ChIP-Seq dataset might have focused on the TRFs associated with promoters.

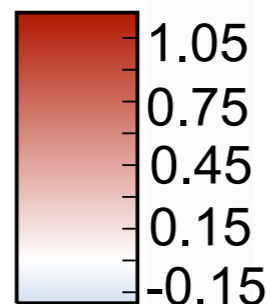
The regulatory regions display significant overlap with histone peaks



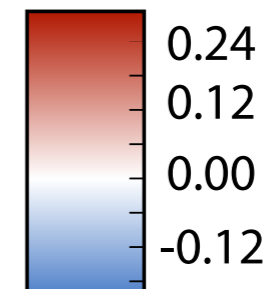
No single histone mark is able to identify regulatory regions in the genome



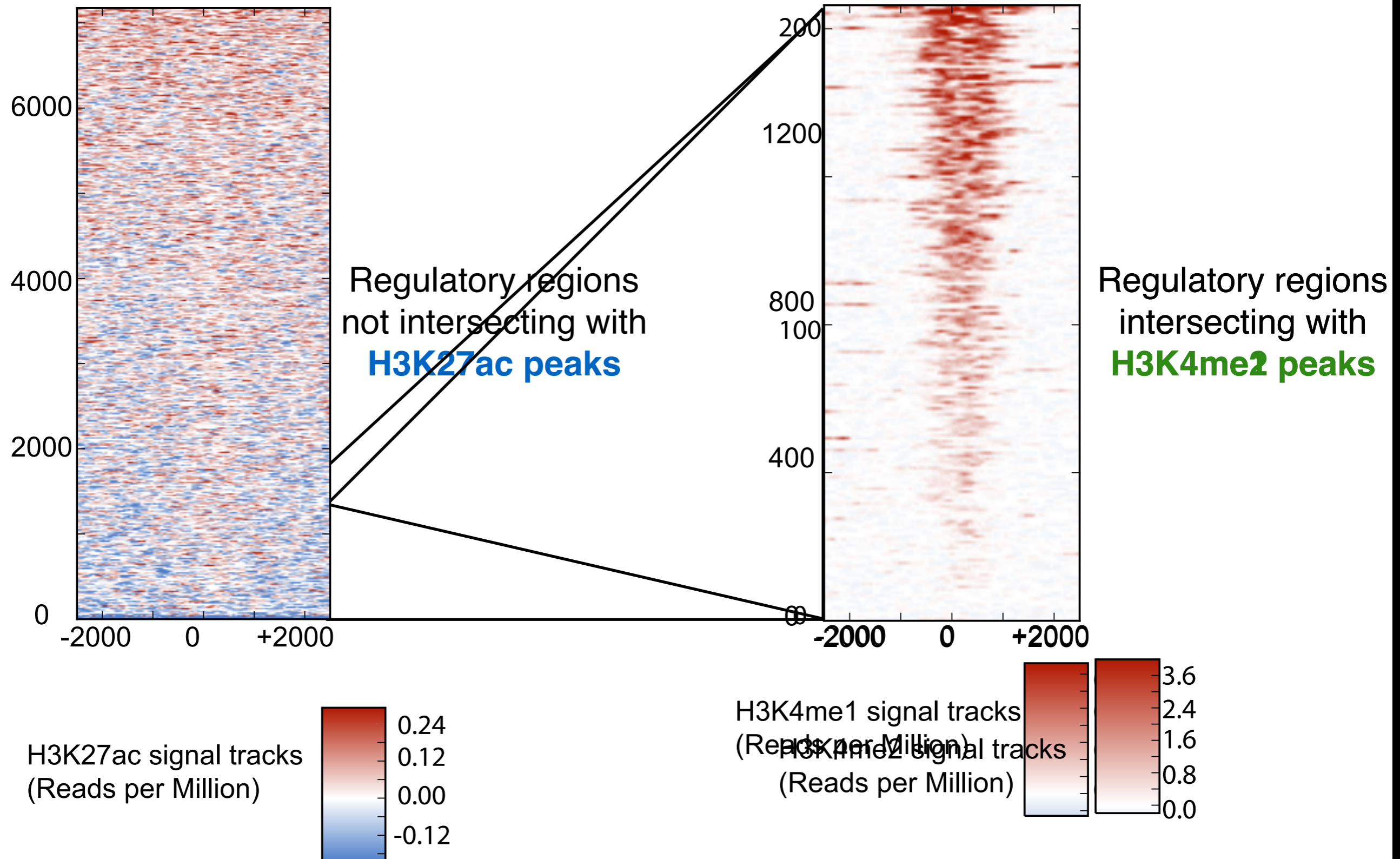
H3K27ac signal tracks
(Reads per Million)



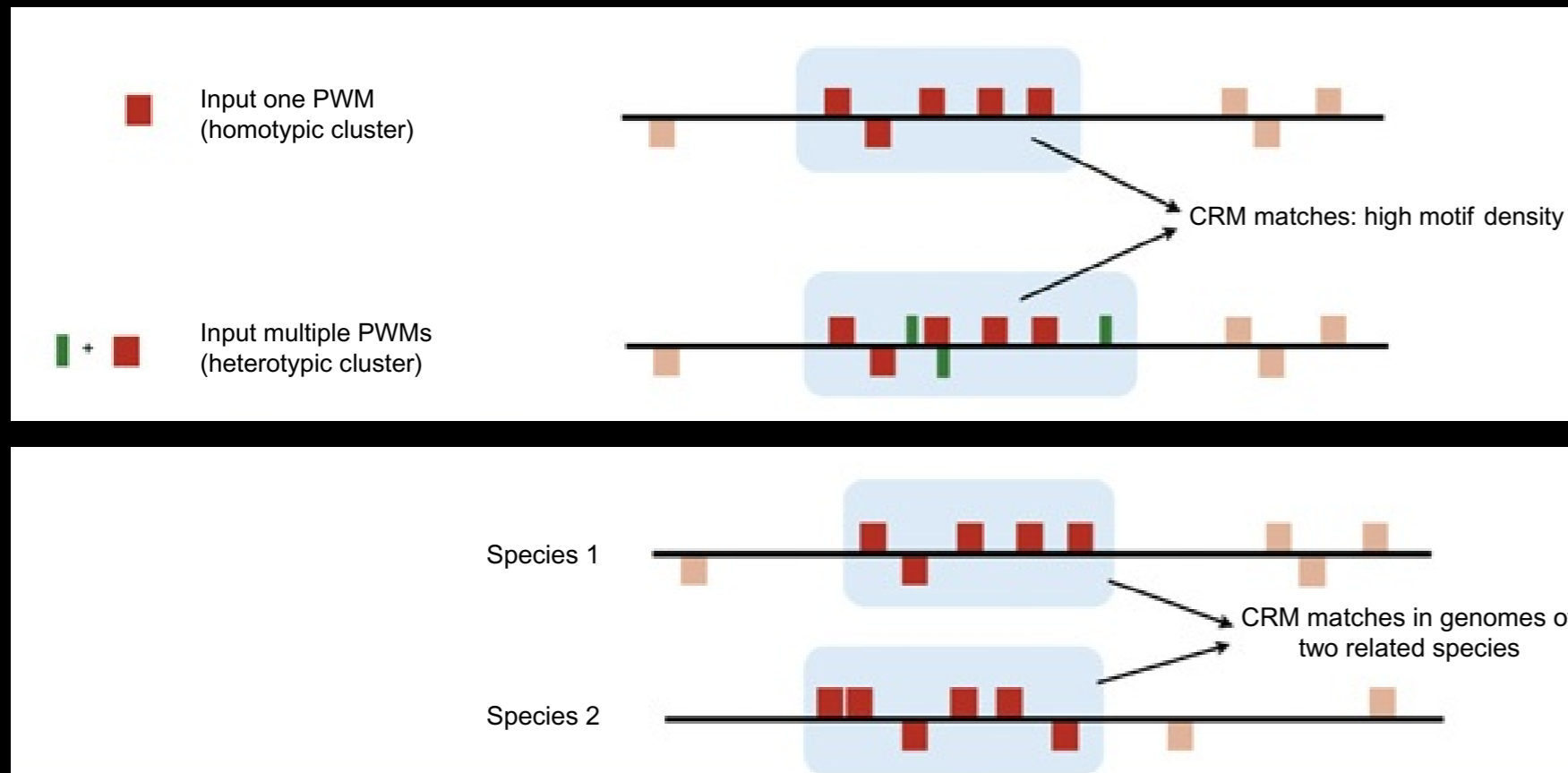
H3K27ac signal tracks
(Reads per Million)



No single histone mark is able to identify regulatory regions in the genome



Data types used to predict active enhancers - Evolutionary Constraints and Motif Content



High density of TF binding motifs

Evolutionary conservation of TF binding motifs, Ultraconserved sites

Approximately 30-40% of noncoding sites under high evolutionary pressure tested positive for enhancer activity.

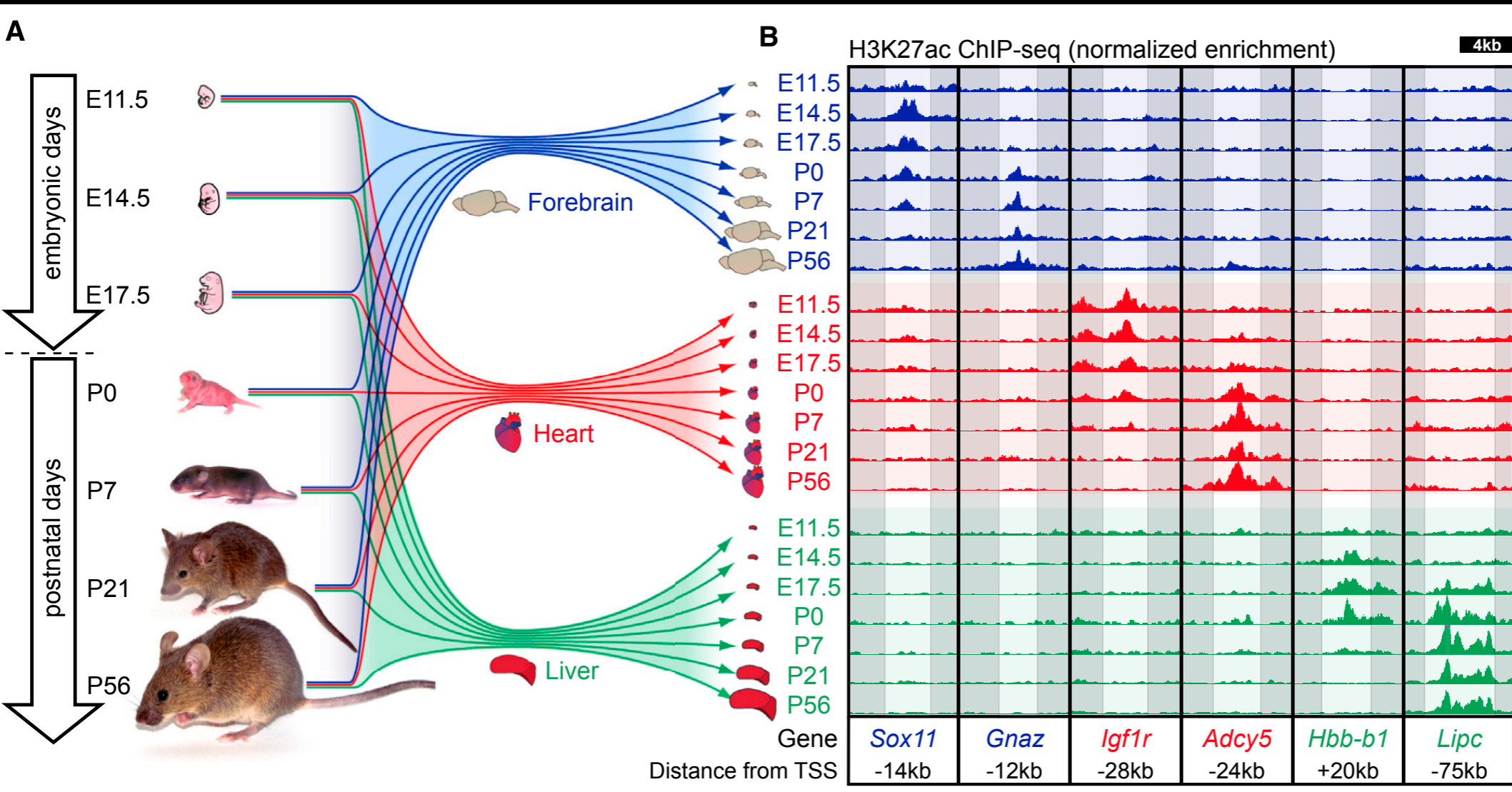
Problems with this Approach:

Cannot predict tissue specificity.

A number of conserved elements serve alternate functions.

All enhancers may not be under very high evolutionary pressure.

Data types used to predict active enhancers - Epigenetic datasets (Open chromatin and Histone Modification)



Associated with active enhancers:
H3K27ac or
H3K79me3 +
H3K4me1

Associated with inactive enhancers:
H3K27me3 +
H3K4me1

Approximately 67% of “dynamic” H3K27ac peaks tested positive for enhancer activity.

Problems with this Approach:

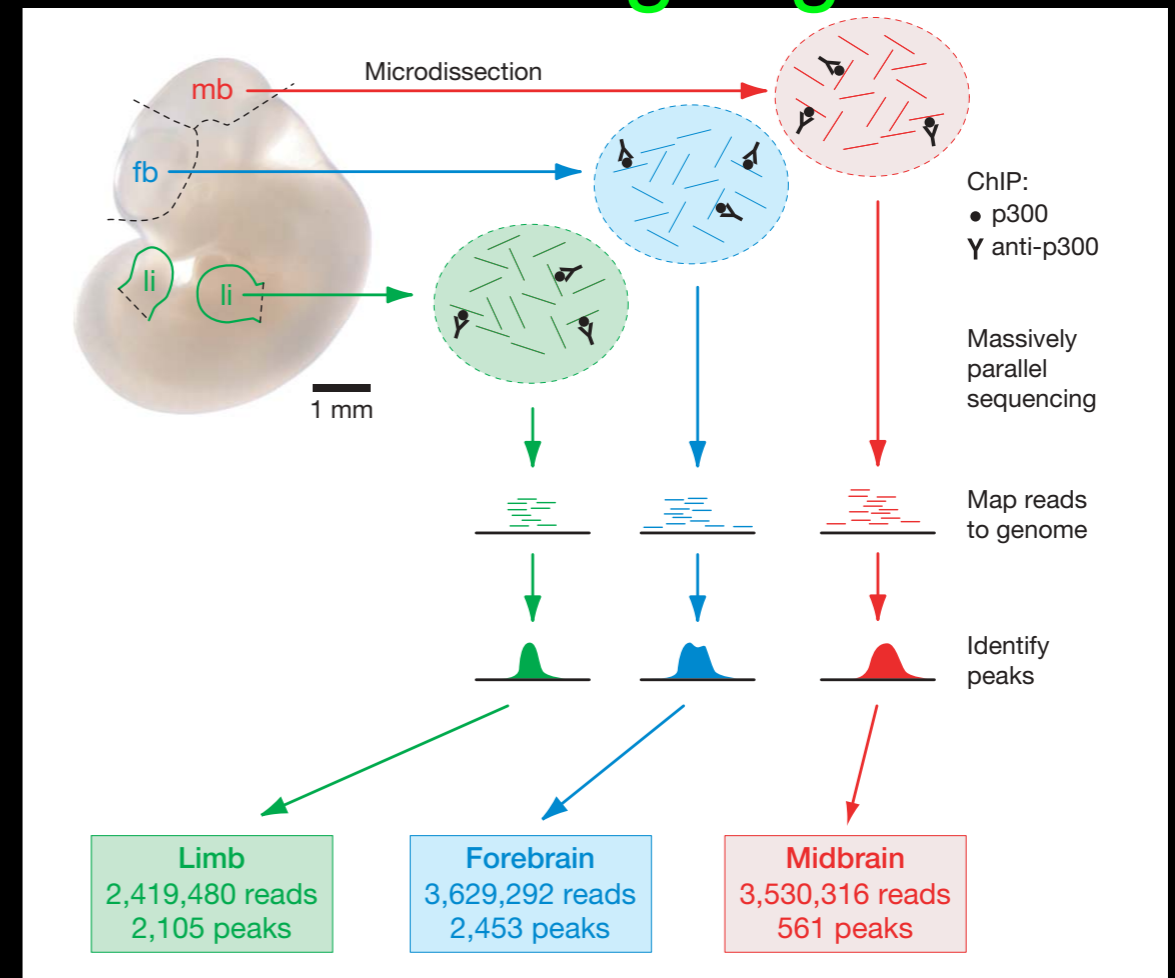
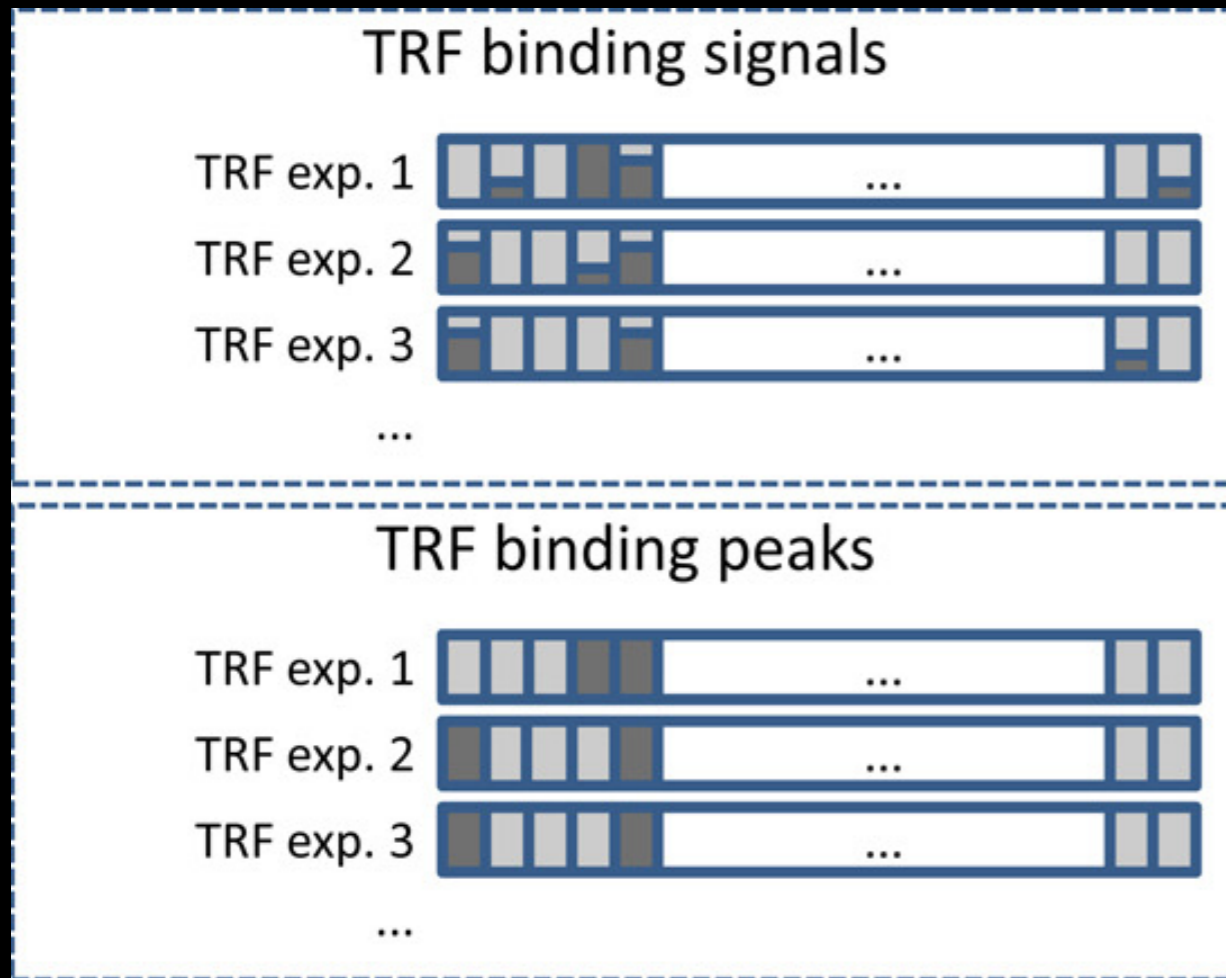
No single histone modification or combination of histone modifications have been associated with all enhancers.

Nord, et al., Cell, 2013.

Rajagopal, et al., PLoS CB, 2012.

Ernst and Kellis, Nature Methods, 2012.

Data types used to predict active enhancers - Models based on transcription factor binding regions



Large cluster of TF binding (ChIP-Seq)

Related approach: p300/Cbp binding peaks

Approximately 58-82% of p300 peaks tested positive for enhancer activity.

Problems with this Approach:

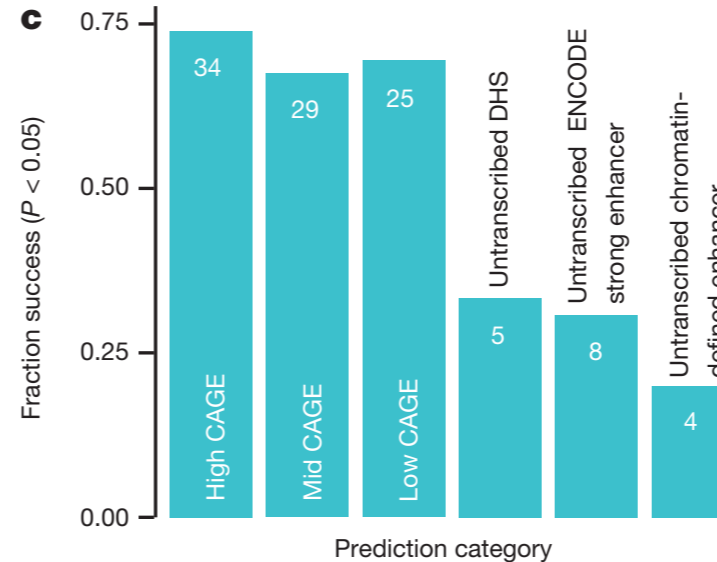
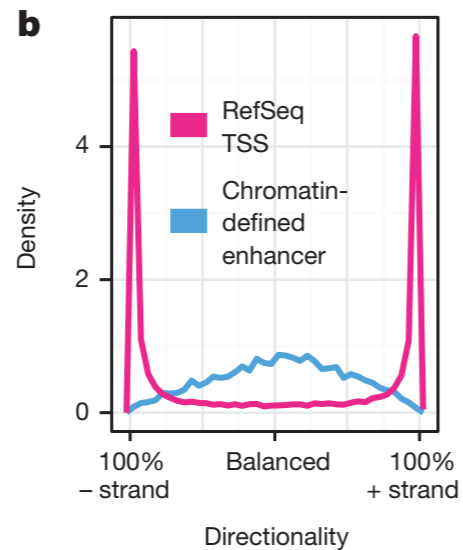
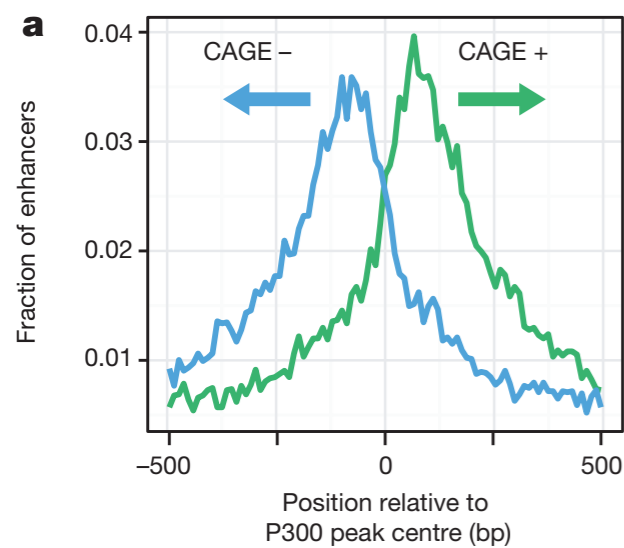
A large number of experiments required.

Difficult to distinguish between functional and passive TF binding.

Yip, et al., Genome Biology, 2012.

Visel, et al., Nature, 2009.

Data types used to predict active enhancers - Bidirectional nonpolyadenylated CAGE peaks (eRNA)



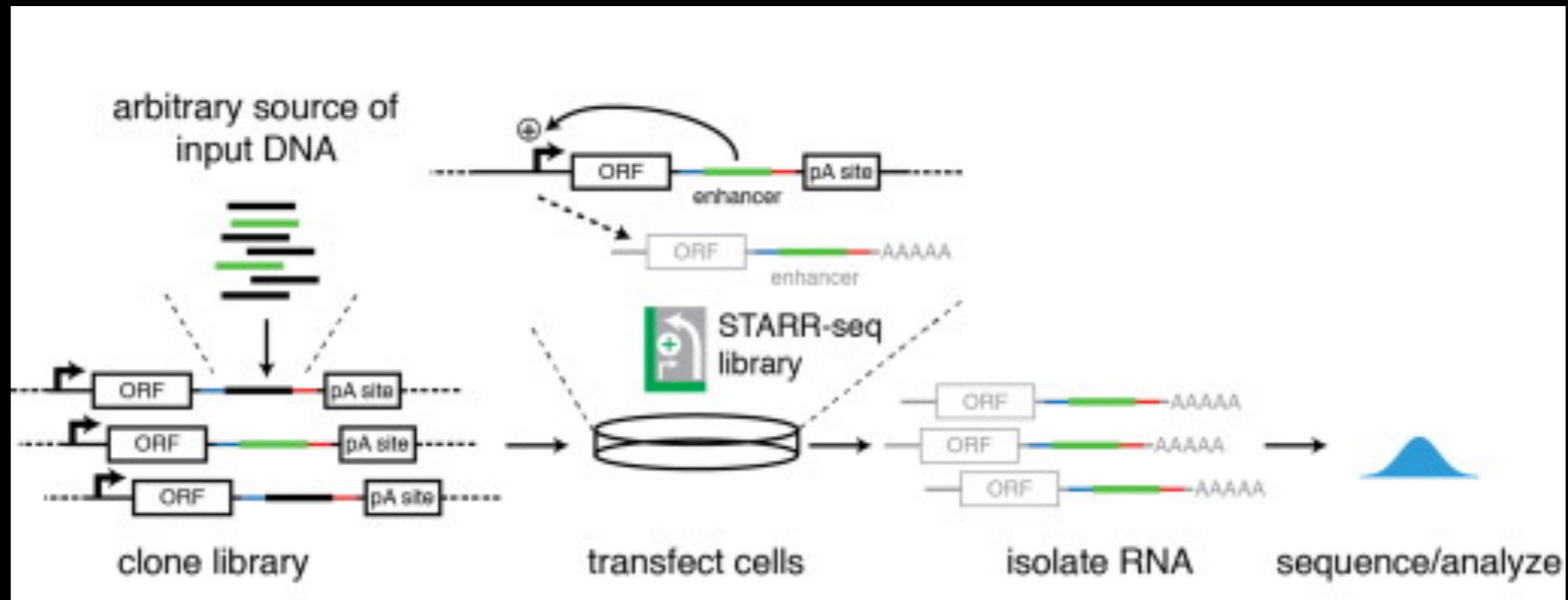
bi-directional CAGE peaks are present near some enhancers

Approximately 75% of predicted enhancers in the vicinity of bidirectional CAGE peaks tested positive for enhancer activity.

Problems with this Approach:

Not sure if all active enhancers can be found using this approach.
May not be able to distinguish from random transcription.

STARR-seq is a massively parallel NGS assay that utilizes transduction to identify enhancers on a genome-wide scale



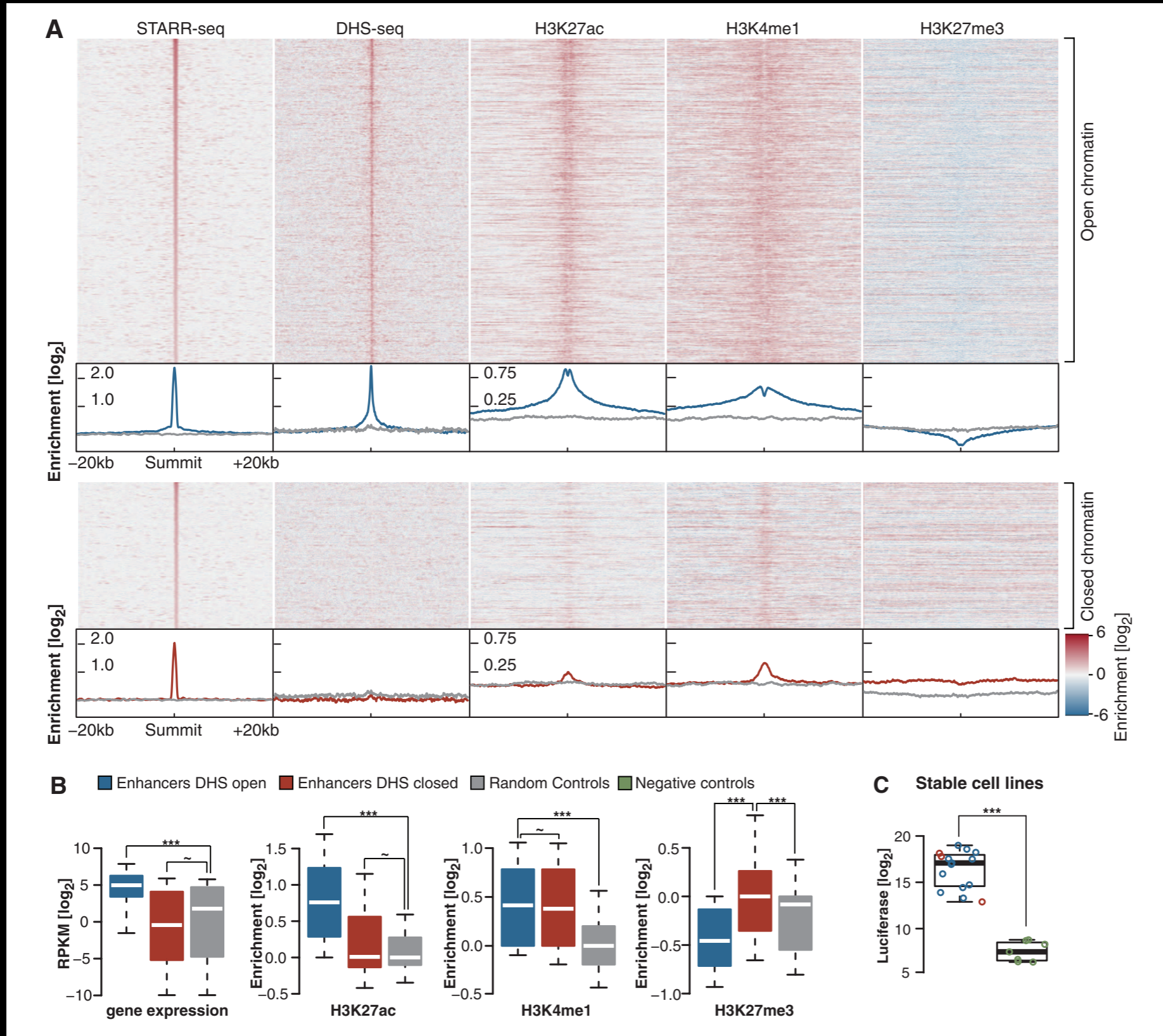
STARR-seq was developed in flies and translating it to whole genome for mammalian cells has been challenging.

Both active and poised enhancers can come positive in this assay

Active enhancers

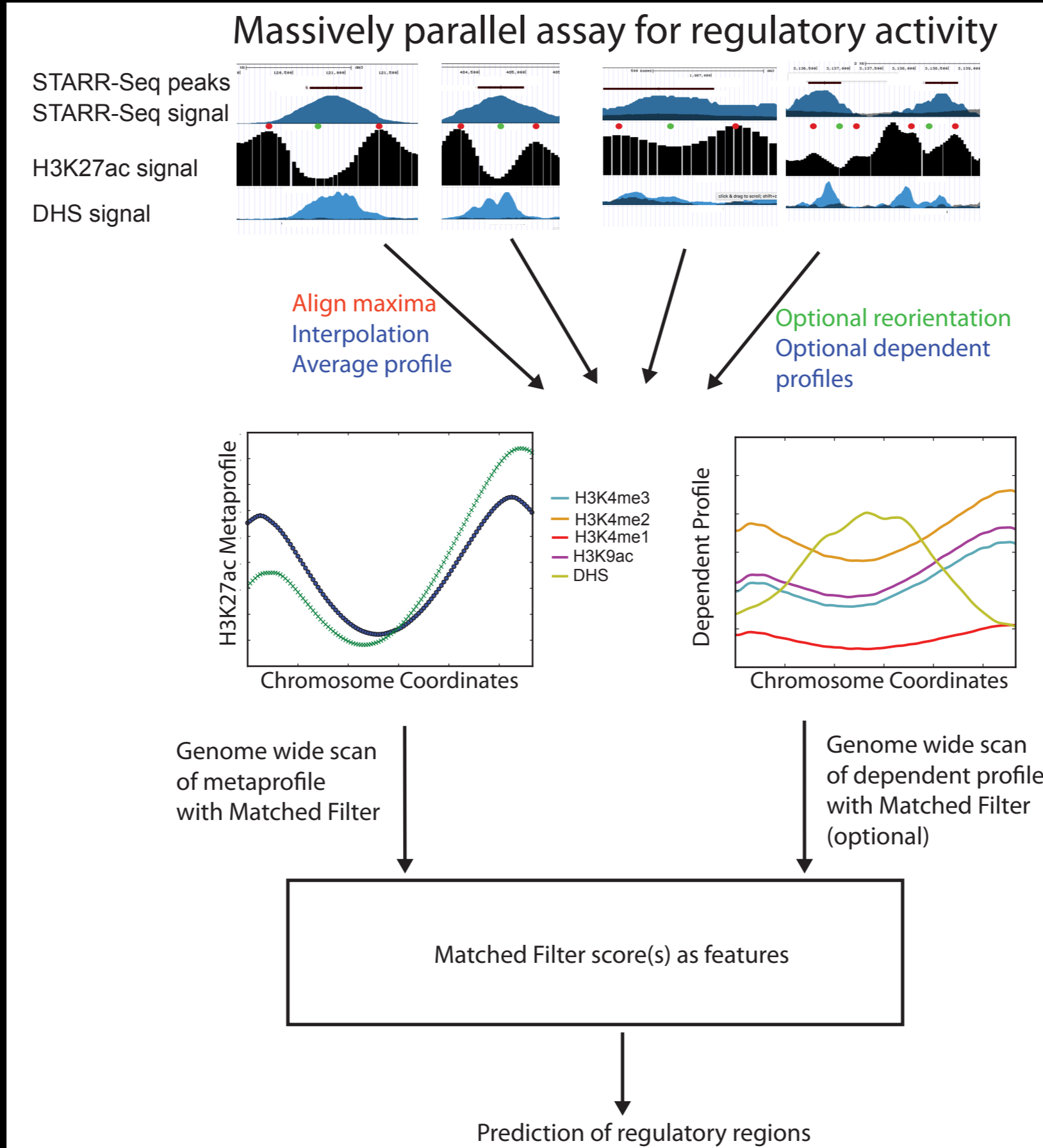
Poised enhancers

Gene expression of flanking genes

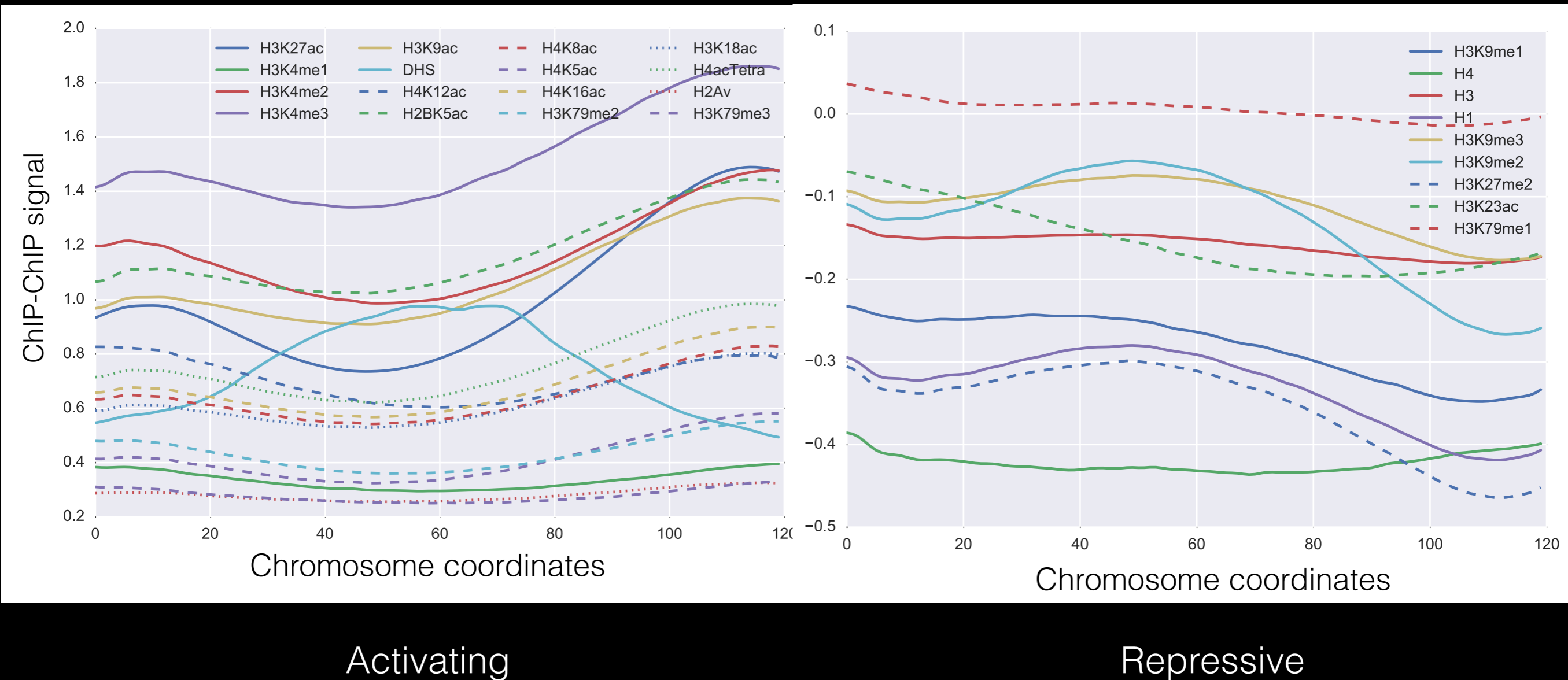


Can we use these epigenetic underpinnings to predict enhancers in a tissue-specific manner?

MPRAs can be used to learn the signal shape in different epigenetic marks at active regulatory regions of the genome

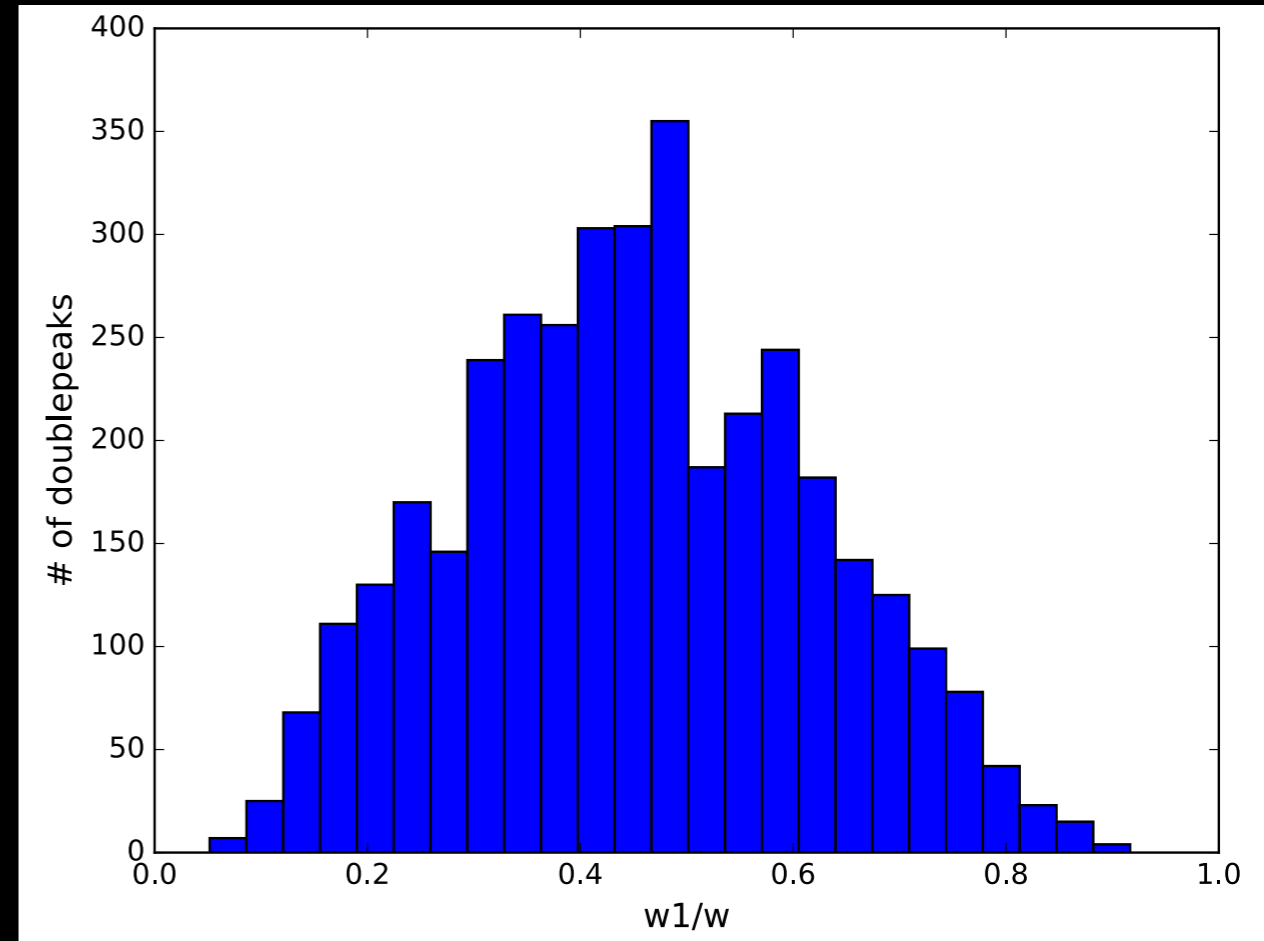
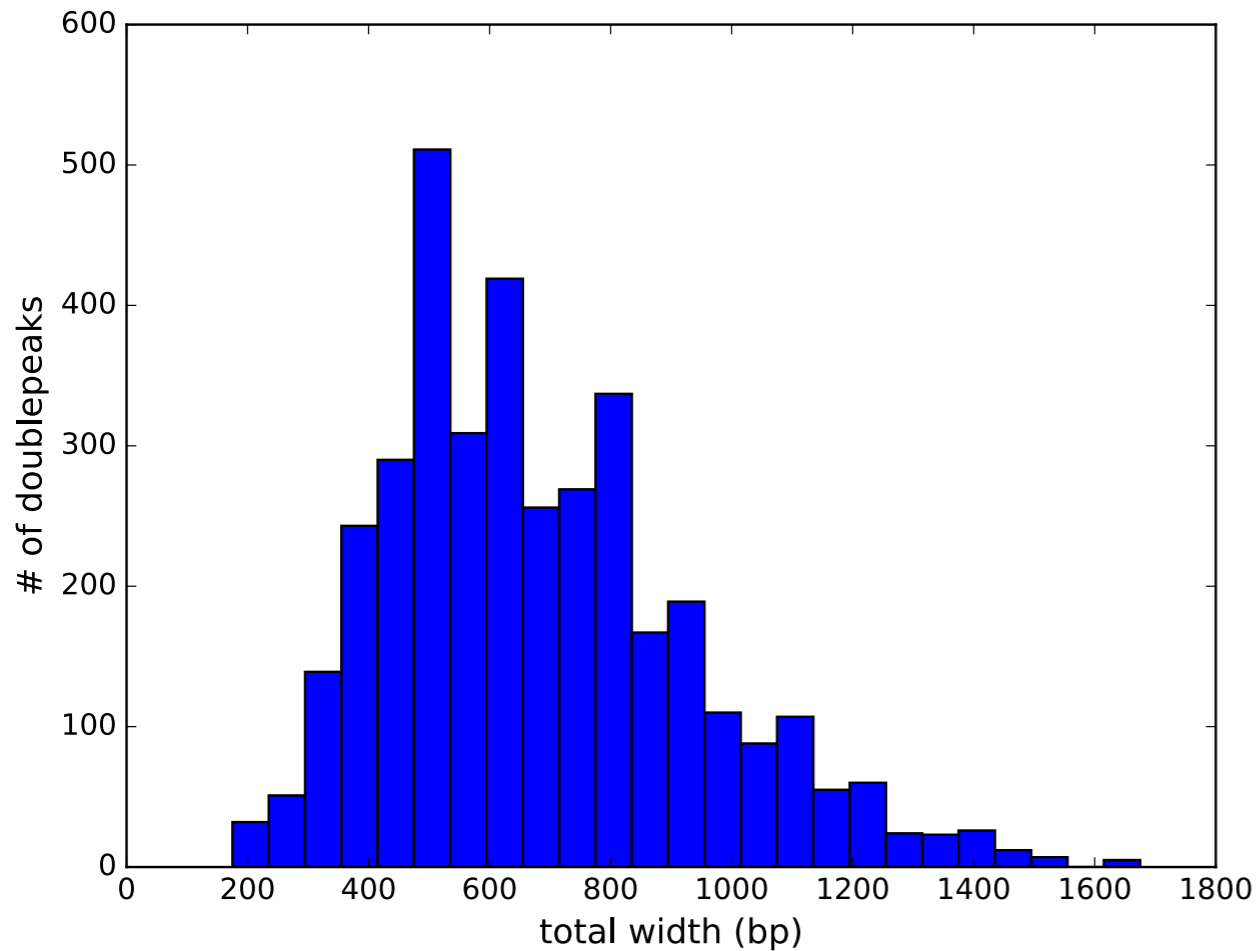


We can roughly split the chromatin marks by their metaprofiles



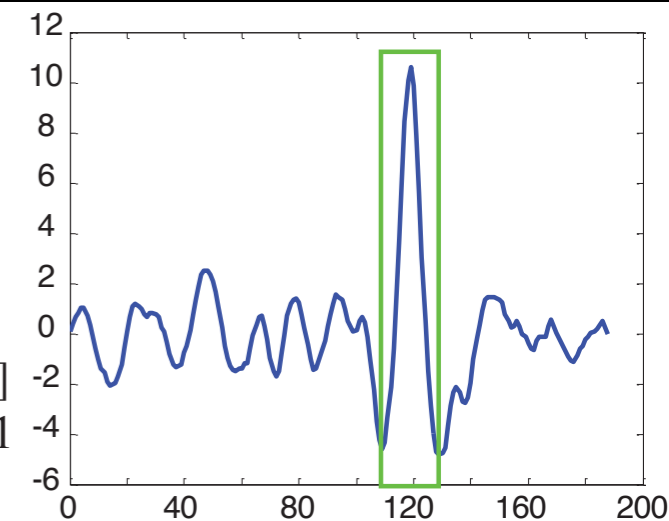
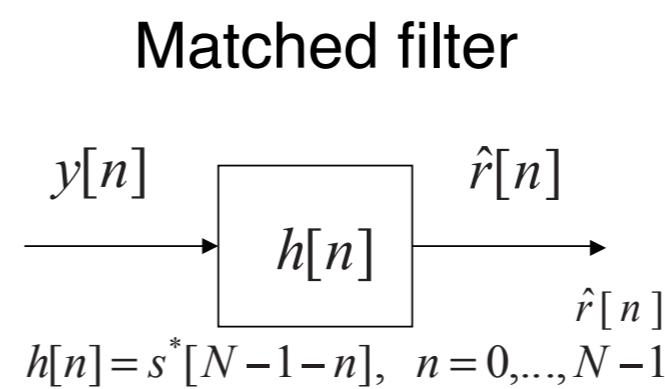
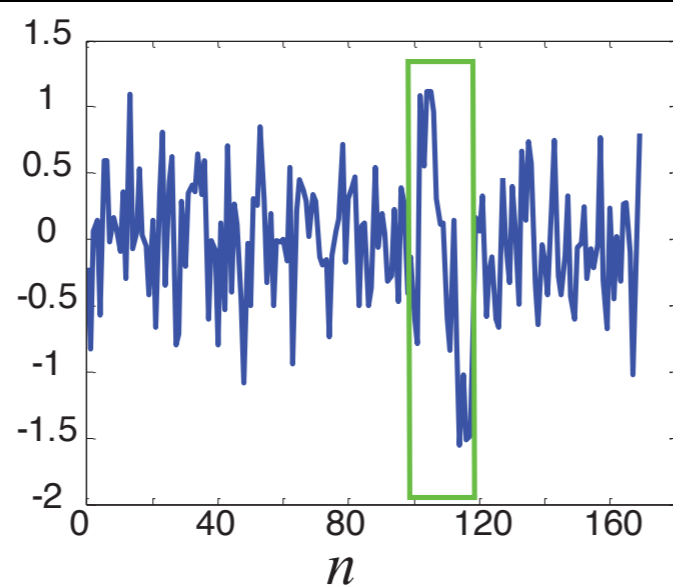
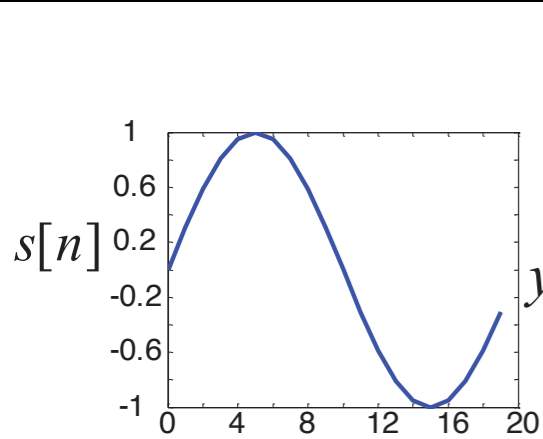
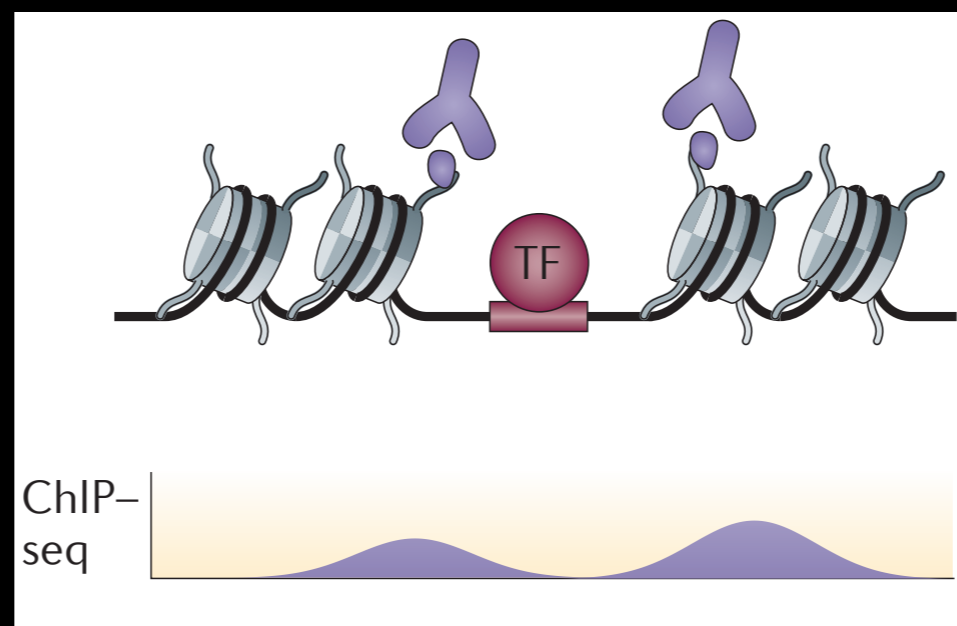
The peaks on either side represent the peak position of modified nucleosomes.

There is variability in the histone modification profiles

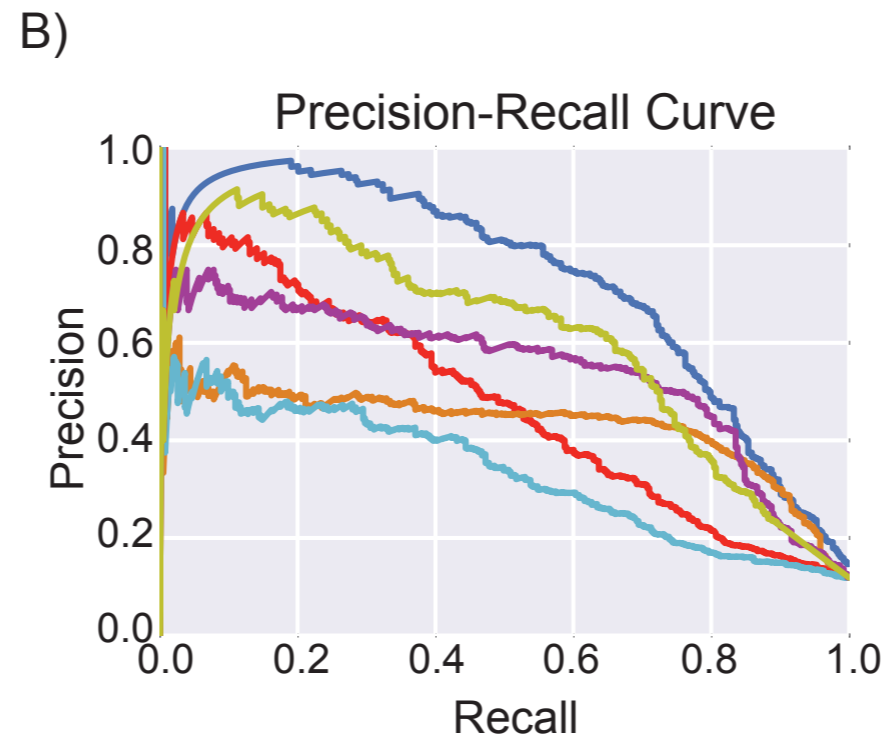
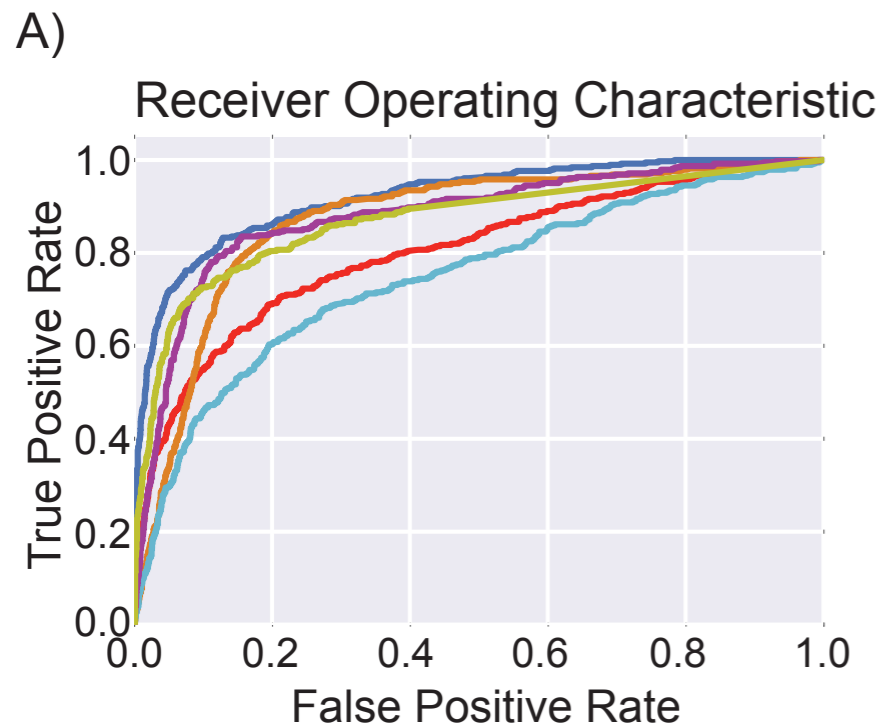


The distance between the two maxima can vary between 300-1100 bp.
On average, the two profiles are pretty symmetric.

Signal processing approach for predicting active regulatory regions



The histone marks can be used to predict occurrence of regulatory regions

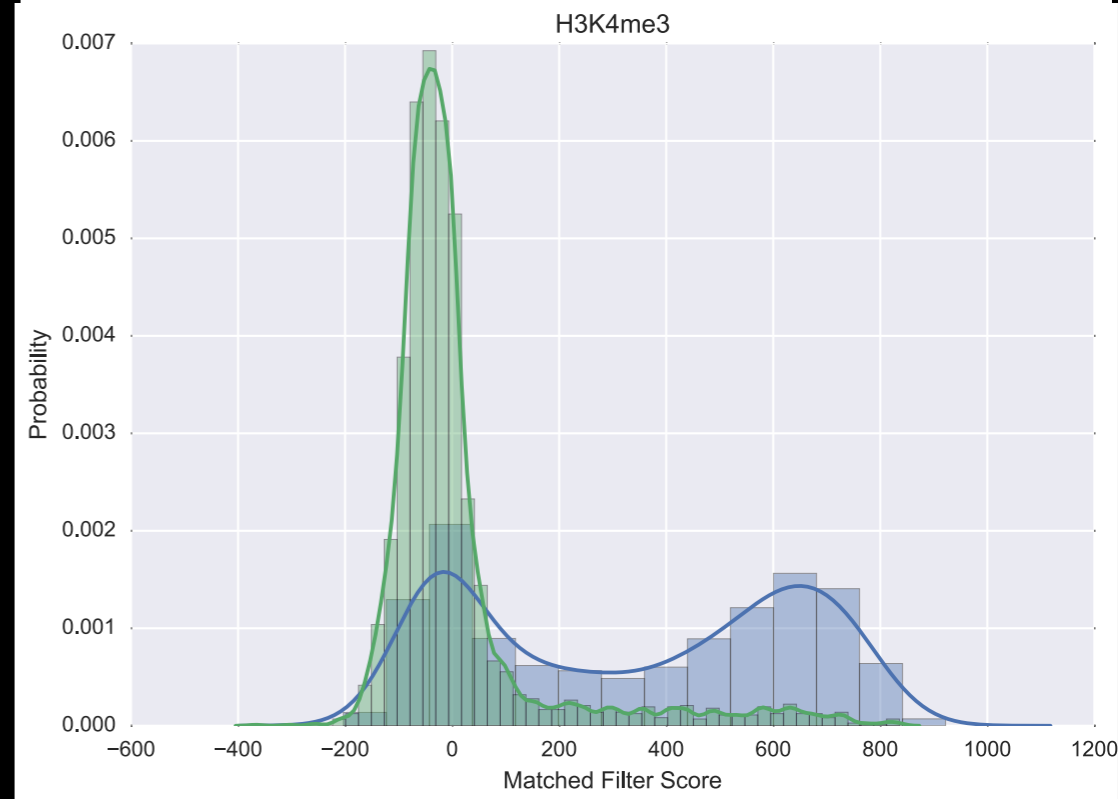
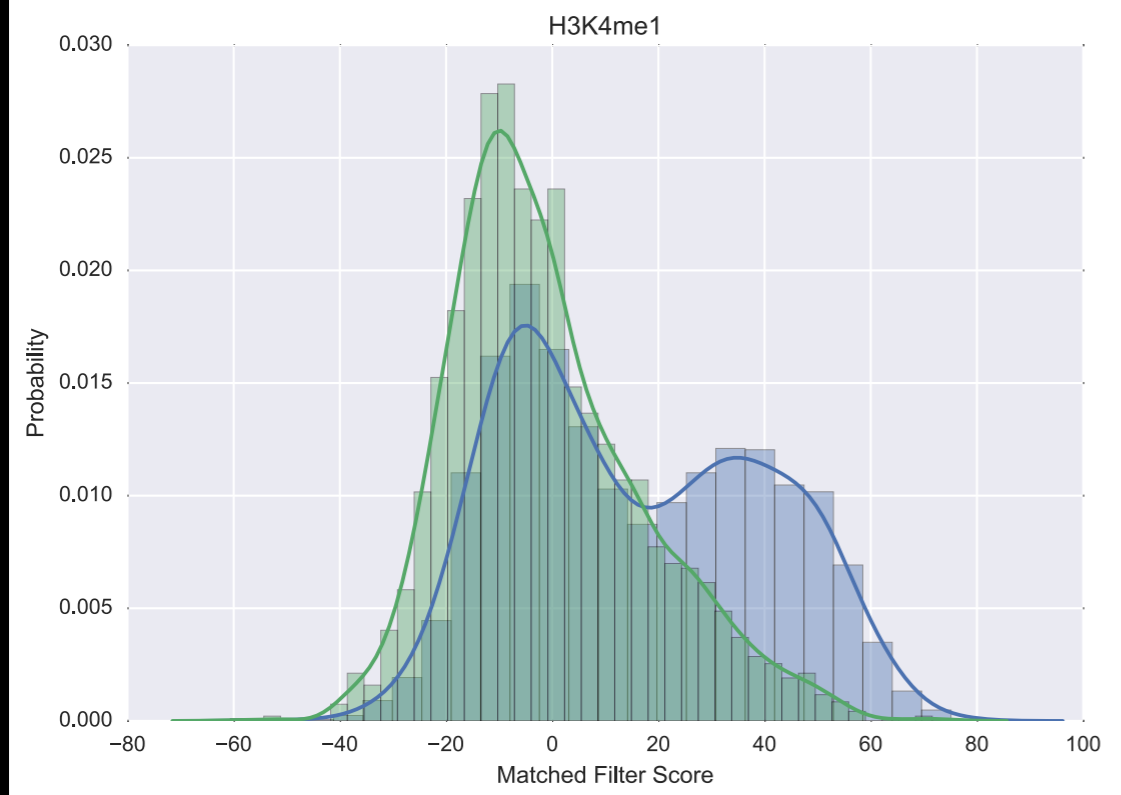
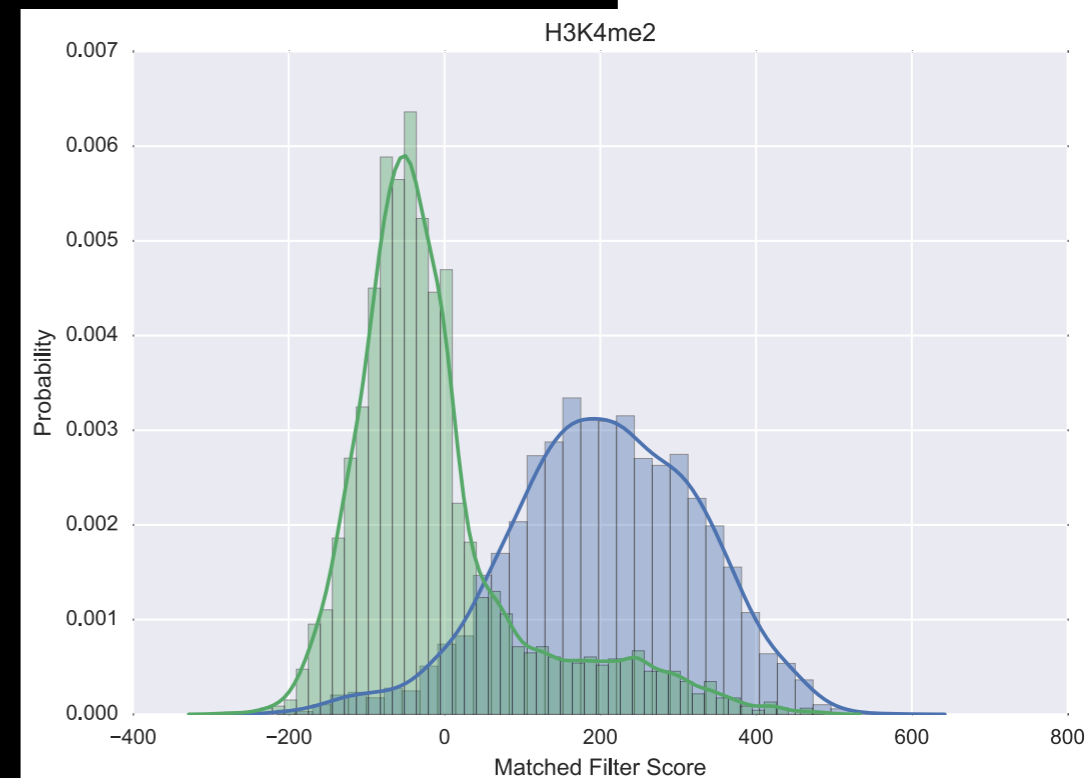
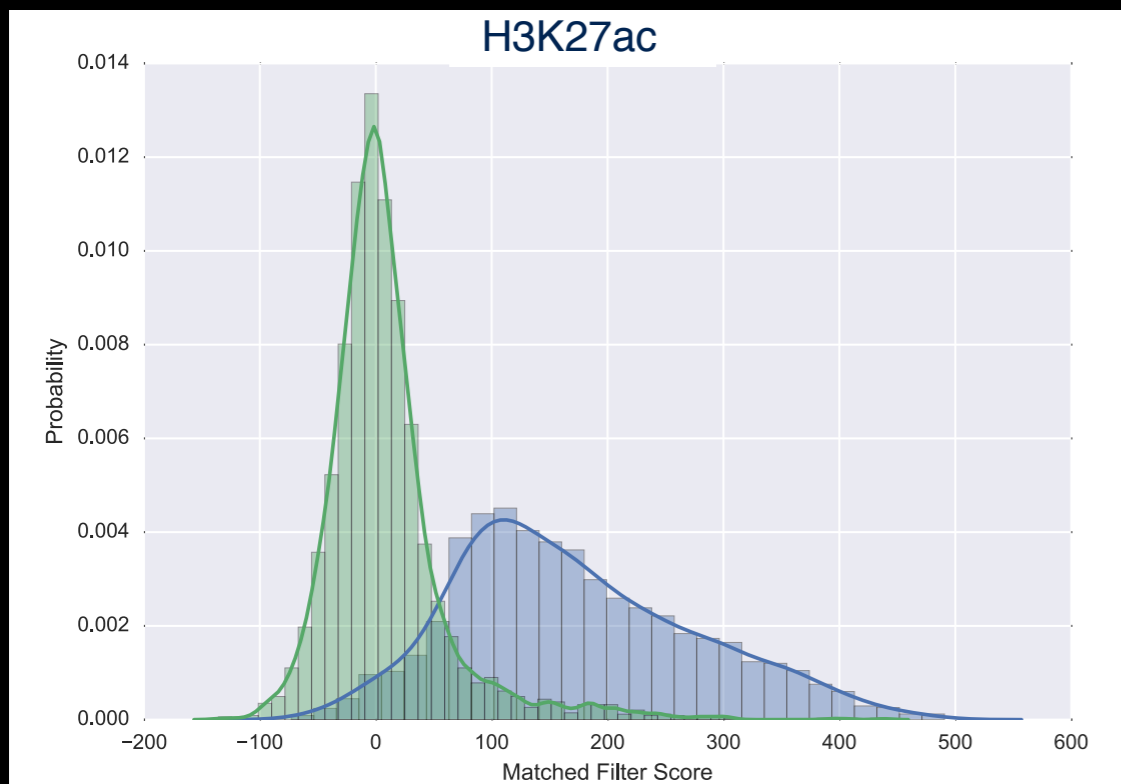
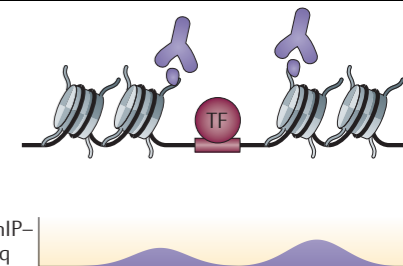


C)

Feature	AUROC	AUPR
H3K27ac	0.92	0.72
H3K4me1	0.80	0.46
H3K4me2	0.87	0.41
H3K4me3	0.73	0.32
H3K9ac	0.89	0.52
DHS	0.86	0.58

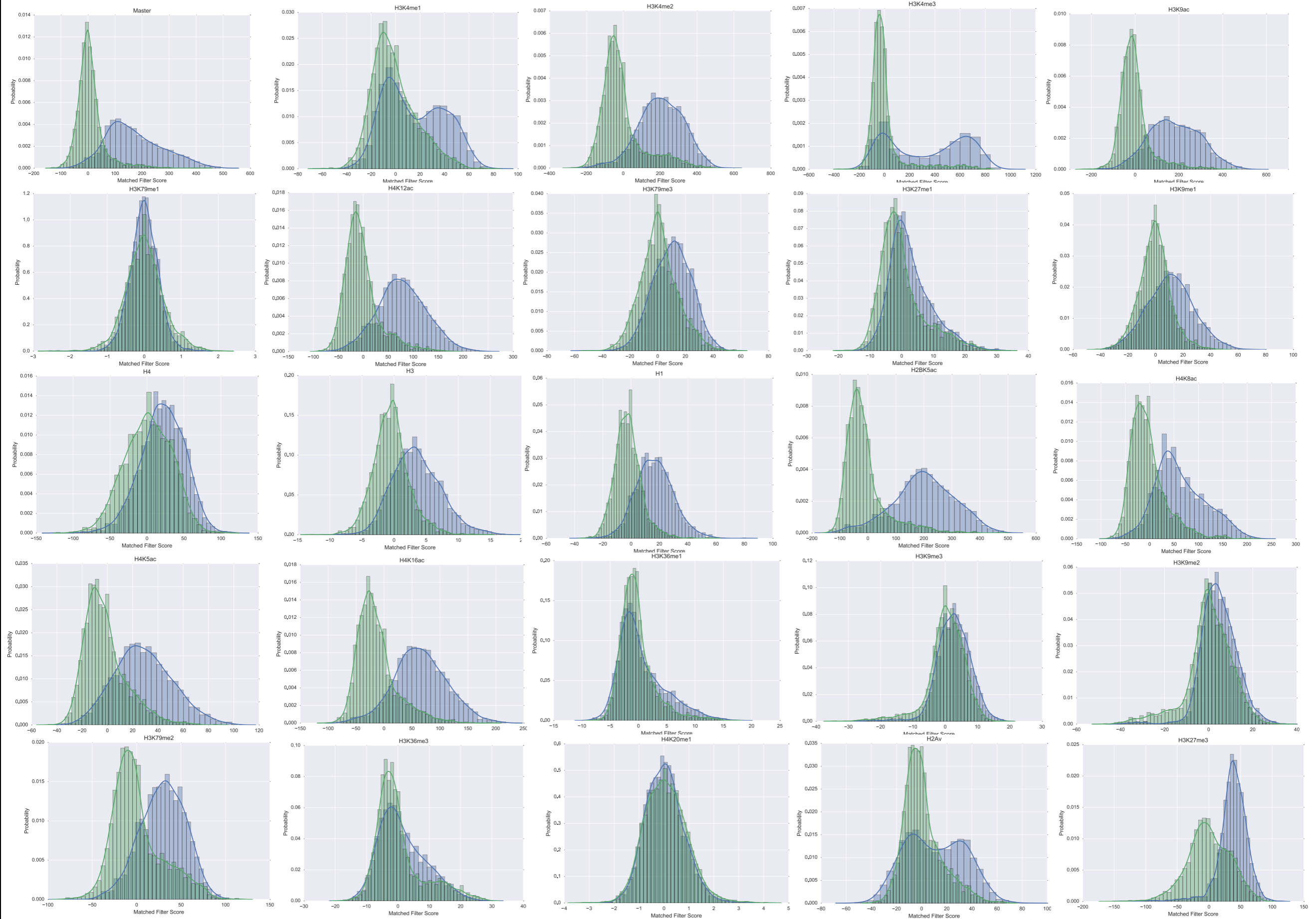
The Matched filter for each histone mark lead to accurate prediction of enhancer regions.

Good separation for each feature

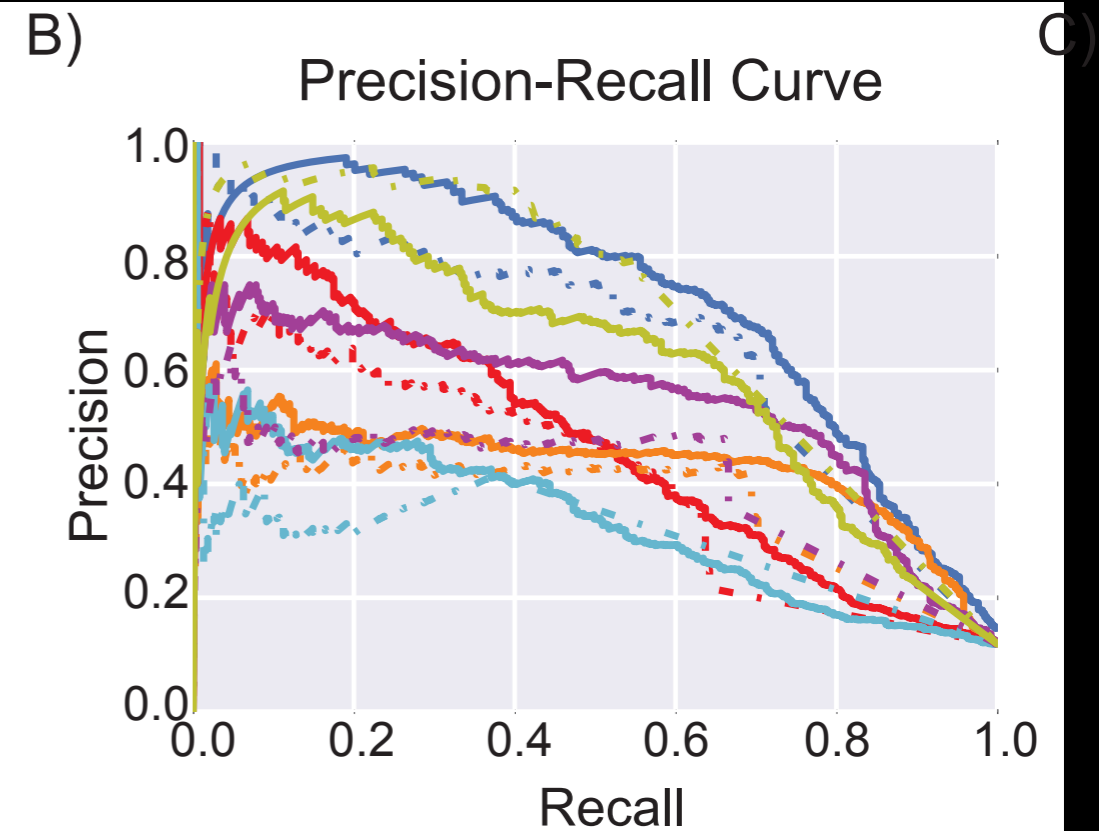
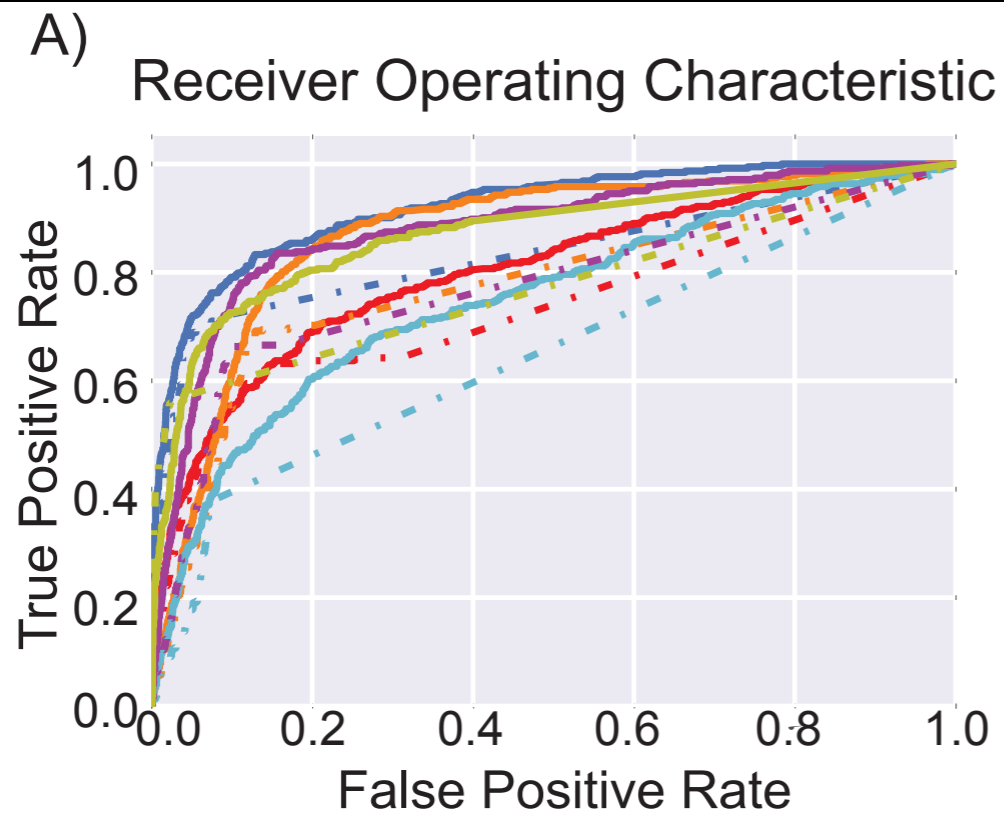


H3K4me1 and H3K4me3 alone displays two Gaussians among positives

Gaussians can fit most matched filter scores for most features



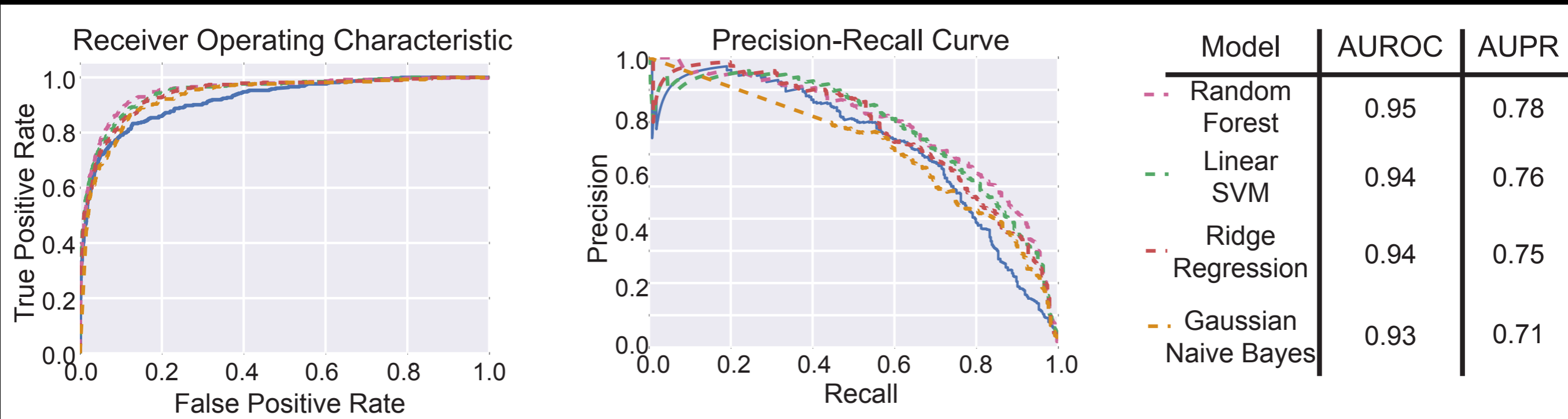
AUROC/AUPR - Comparison to peaks



Feature	AUROC	AUPR
H3K27ac	0.92 (0.83)	0.72 (0.63)
H3K4me1	0.80 (0.72)	0.46 (0.39)
H3K4me2	0.87 (0.75)	0.41 (0.34)
H3K4me3	0.73 (0.64)	0.32 (0.28)
H3K9ac	0.89 (0.77)	0.52 (0.39)
DHS	0.86 (0.77)	0.58 (0.67)

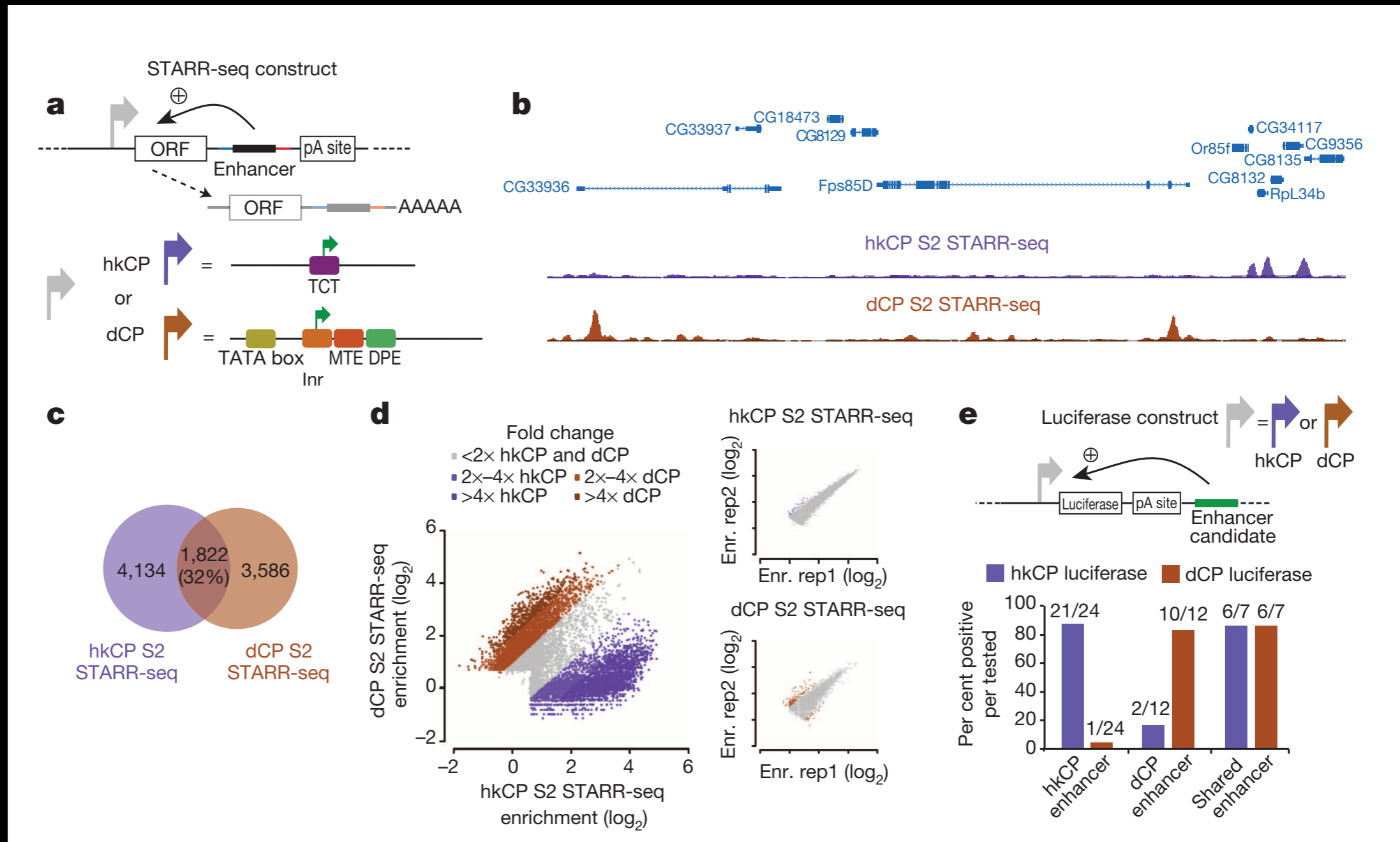
The matched filters do better than individual peaks (except for DHS).

The matched filter scores can be combined to make even more accurate models.

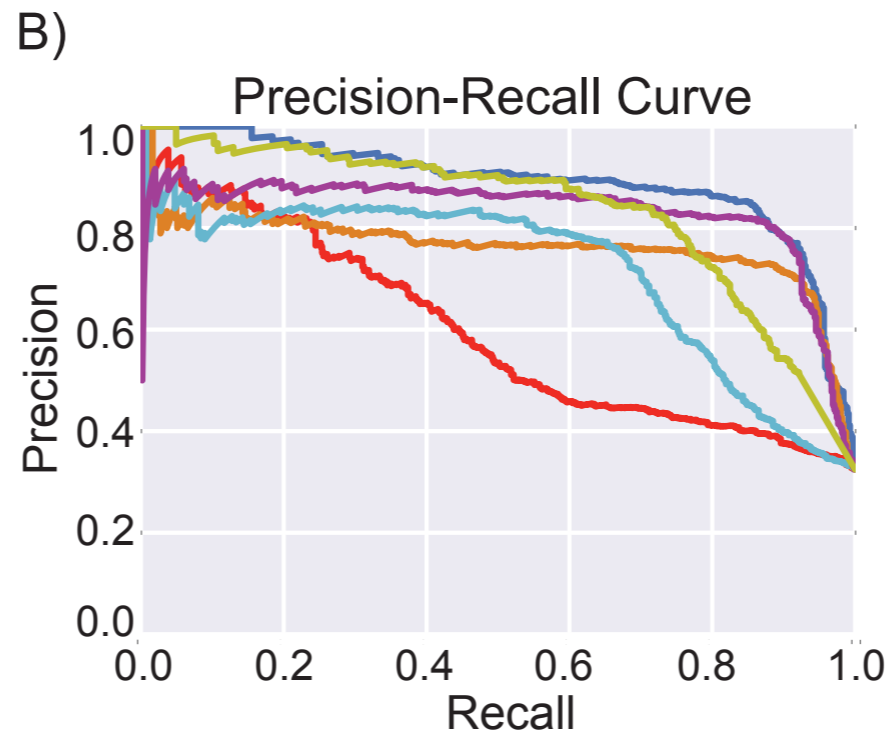
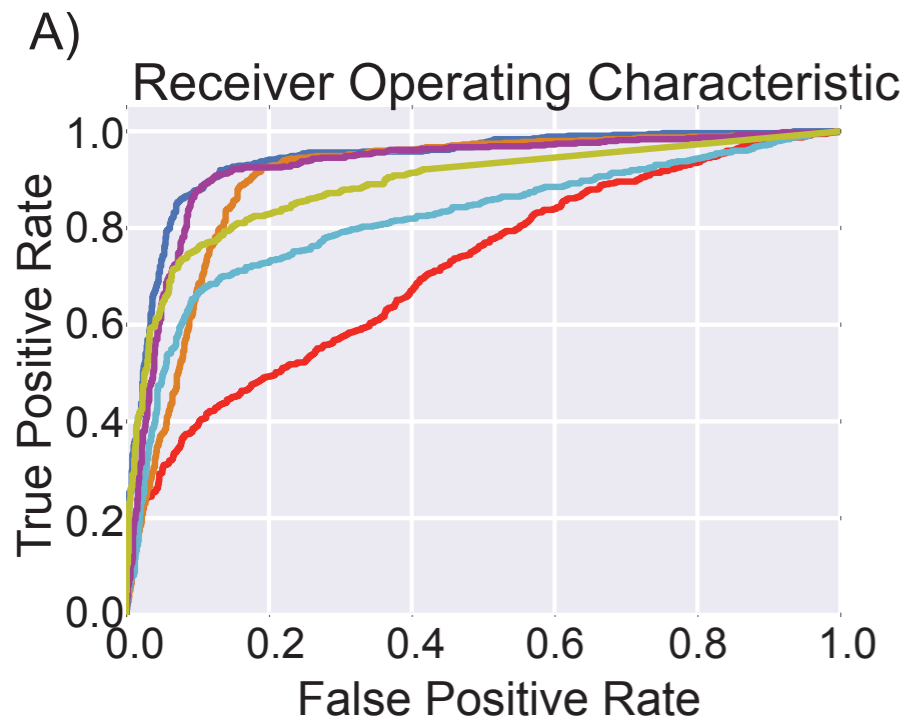


Naive Bayes does not work that well.

The core promoter used in the assay can influence the enhancers that come up active in the assay.

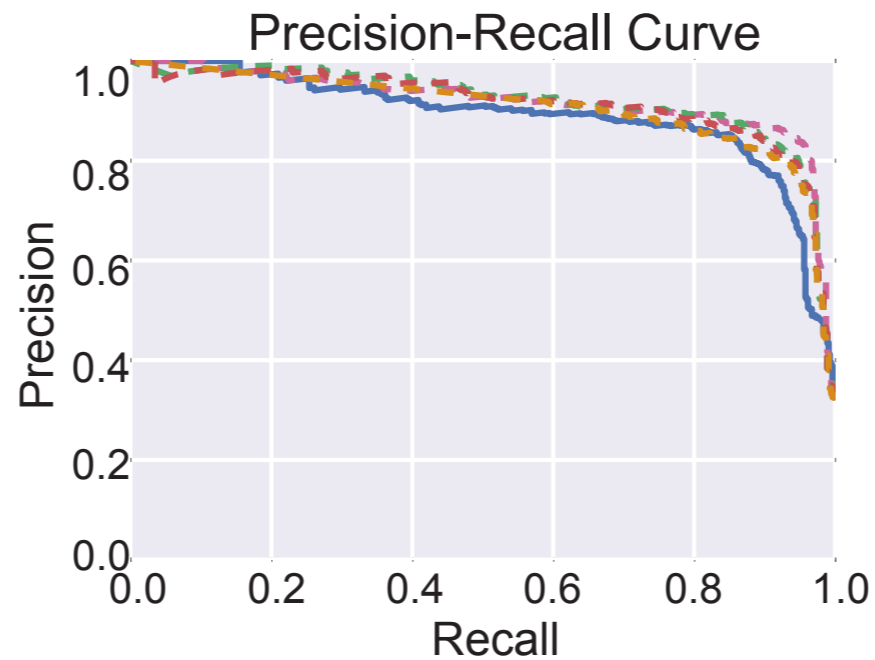
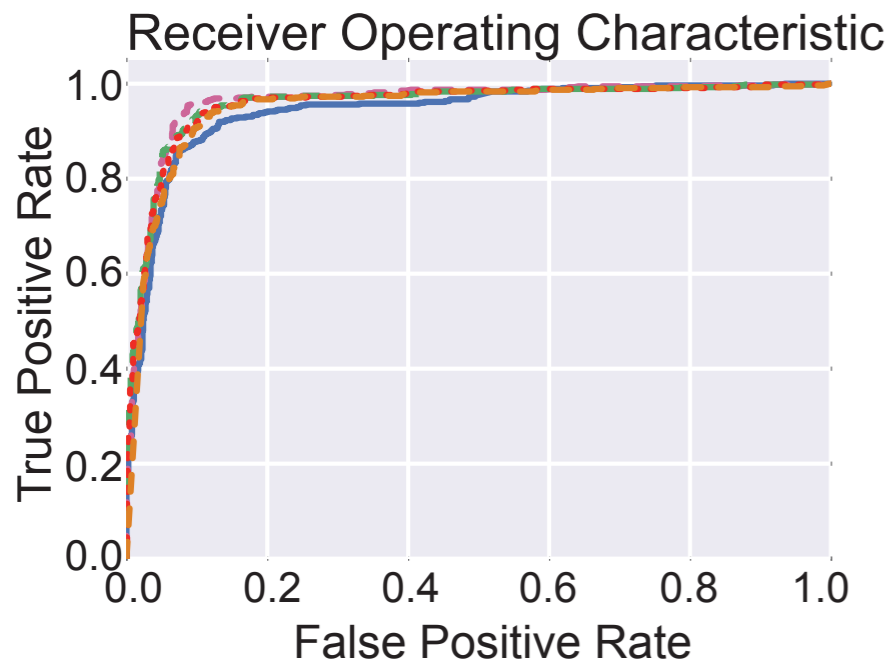


Combining across different core promoters



C)

Feature	AUROC	AUPR
H3K27ac	0.95	0.89
H3K4me1	0.70	0.56
H3K4me2	0.90	0.73
H3K4me3	0.82	0.71
H3K9ac	0.92	0.82
DHS	0.88	0.79



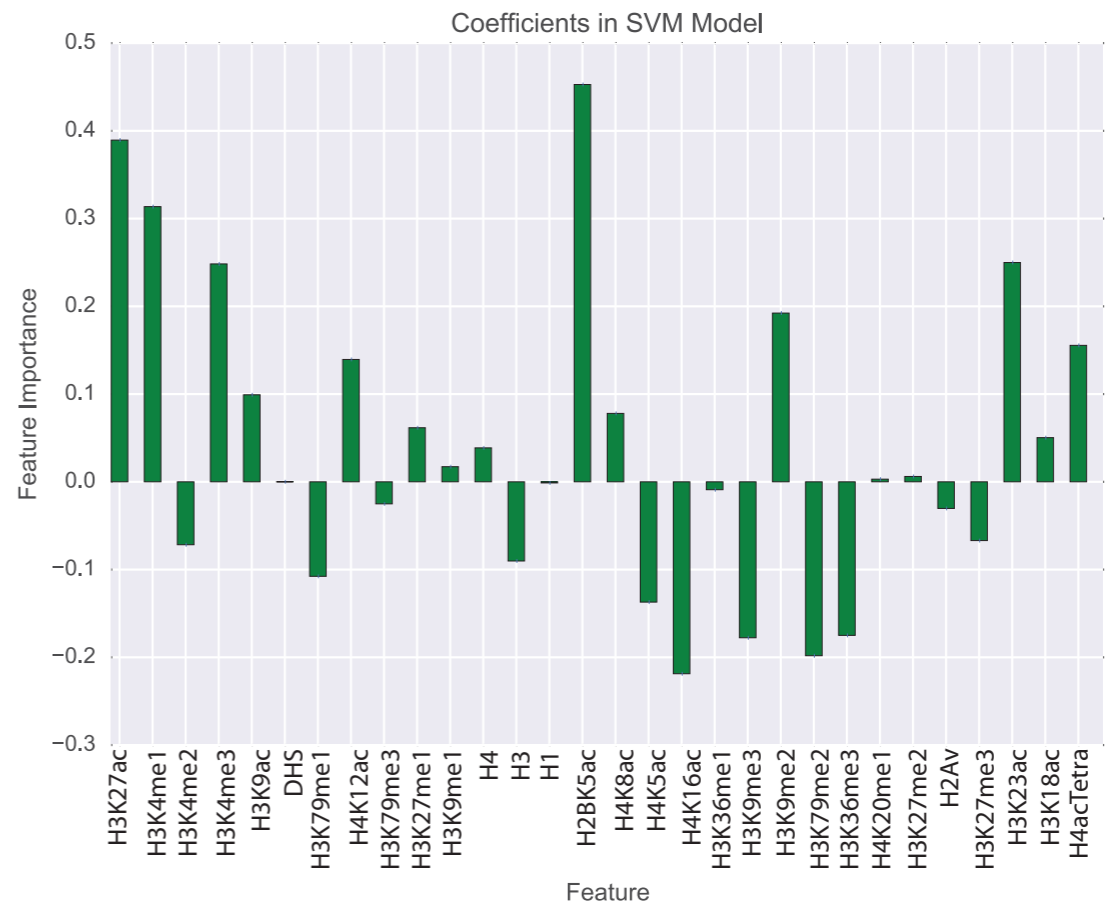
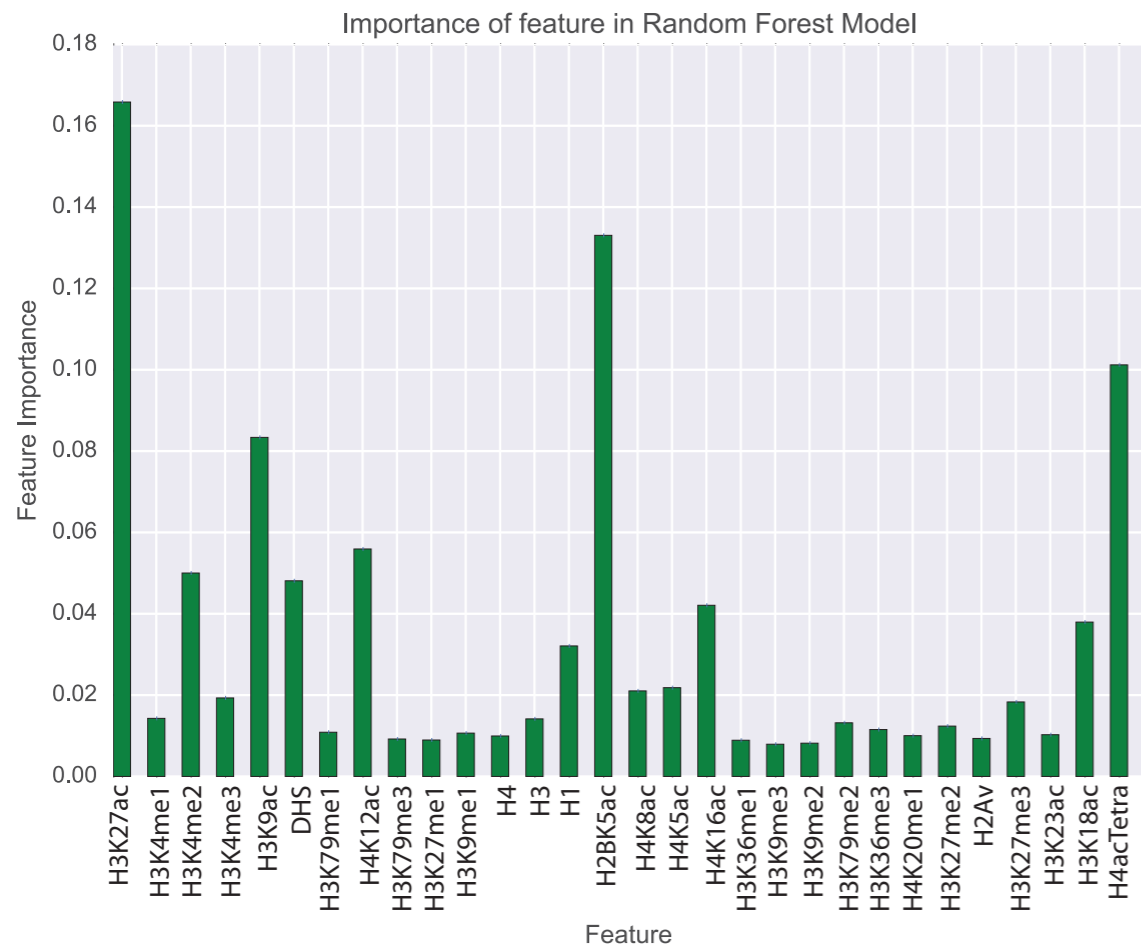
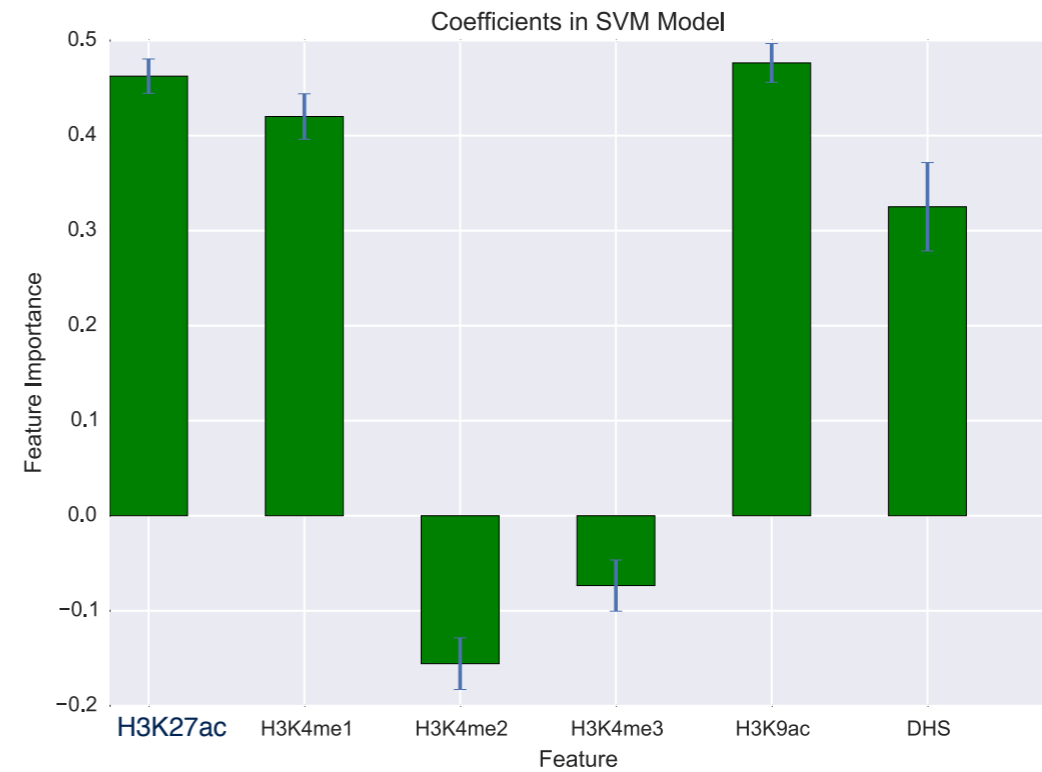
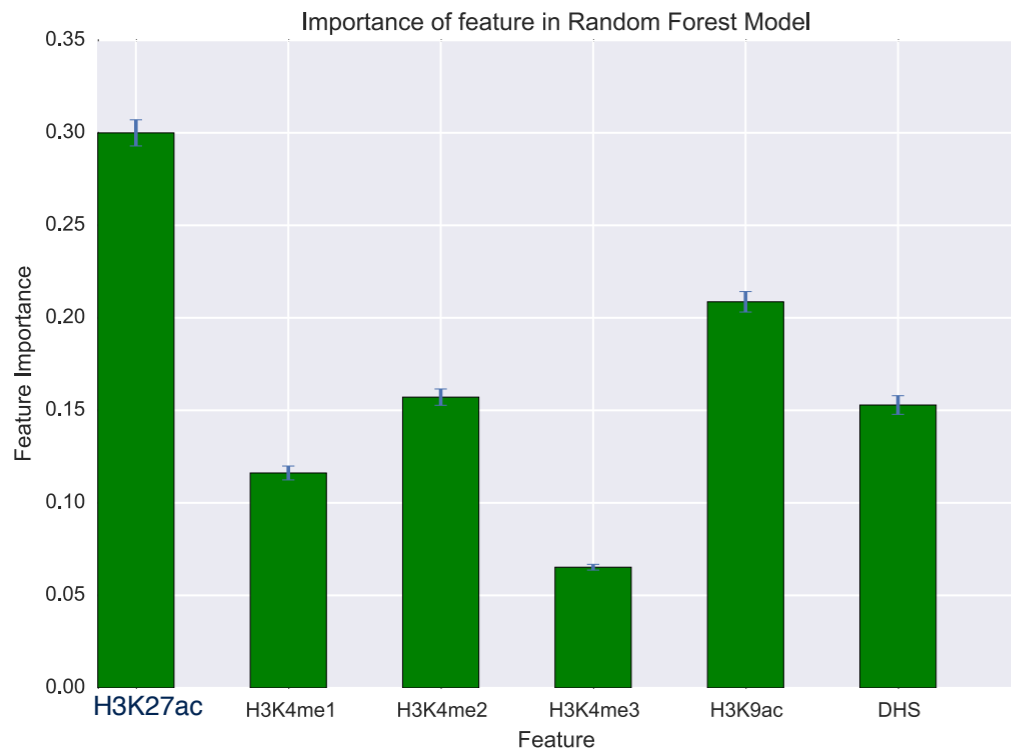
Model	AUROC	AUPR
Random Forest	0.96	0.91
Linear SVM	0.96	0.91
Ridge Regression	0.95	0.90
Gaussian Naive Bayes	0.95	0.89

Feature	AUROC	AUPR
H3K27ac	0.95	0.90
H3K4me1	0.70	0.59
H3K4me2	0.91	0.79
H3K4me3	0.84	0.76
H3K9ac	0.92	0.85
H4K12ac	0.92	0.86
H3	0.80	0.70
H1	0.88	0.81
H2BK5ac	0.94	0.90
H4K8ac	0.88	0.79
H4K5ac	0.87	0.79
H4K16ac	0.89	0.72
H3K18ac	0.90	0.84
H3K9me1	0.71	0.61
H3K79me2	0.79	0.58
H4K27me2	0.81	0.68
H2Av	0.66	0.57
H3K27me3	0.83	0.64
H3K23ac	0.66	0.46
H3K79me3	0.70	0.51
H3K27me1	0.64	0.43
H4	0.67	0.49
H3K36me1	0.54	0.41
H3K9me3	0.59	0.42
H3K9me2	0.60	0.41
H3K36me3	0.57	0.38
H4K20me1	0.47	0.31
H3K79me1	0.47	0.30

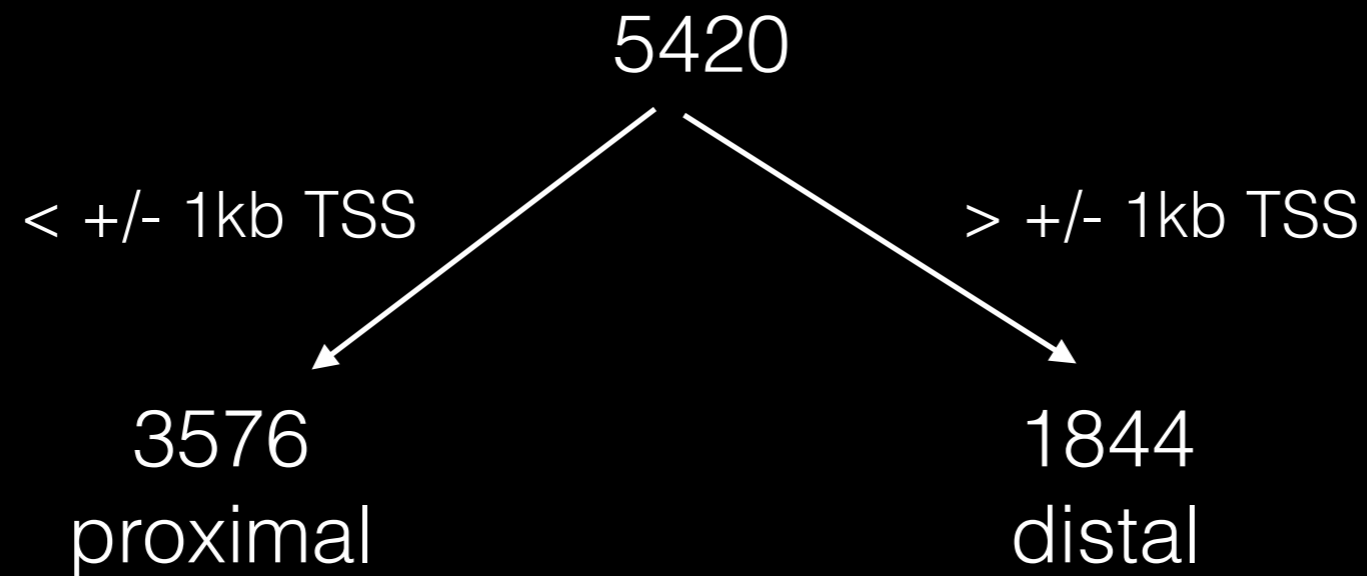
Acetylations tend to be the strongest marks for active regulatory regions

Combining all the marks can lead to slightly higher accuracy (AUROC=0.97 and AUPR=0.93)

There is consistency in the importance of features in the different machine learning models

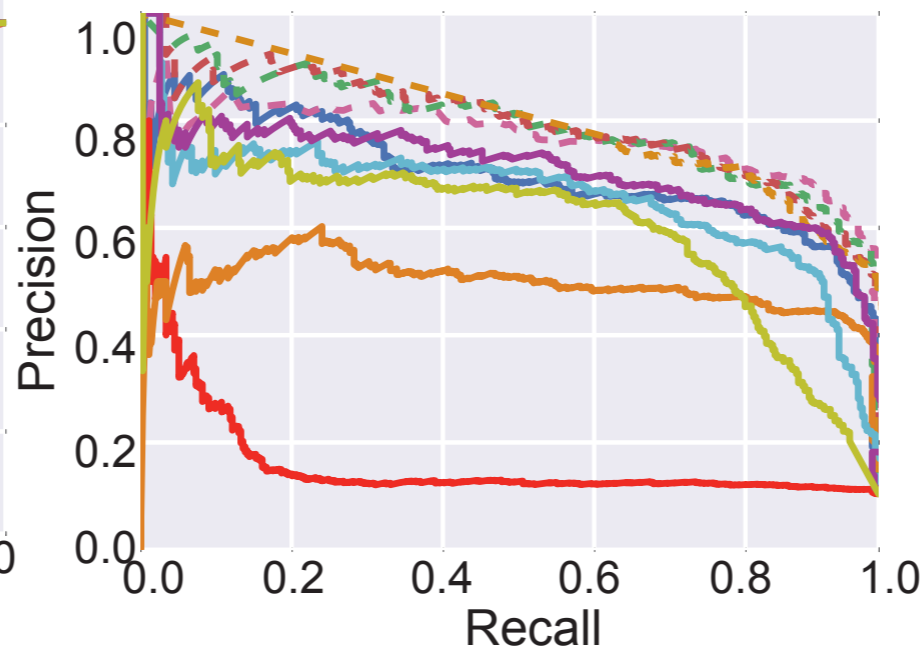
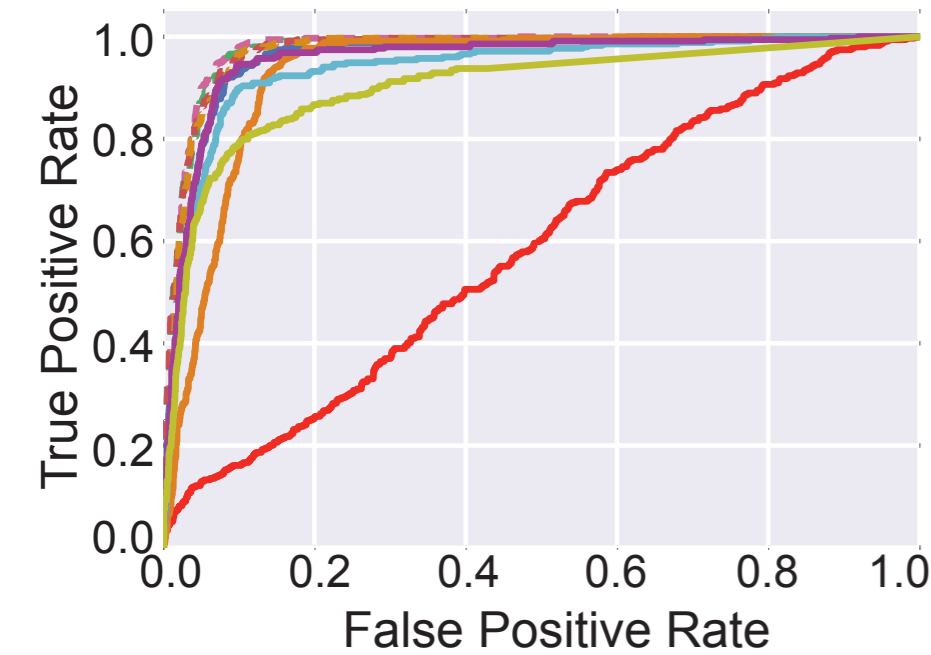


Are there real differences between proximal and distal regulatory elements?



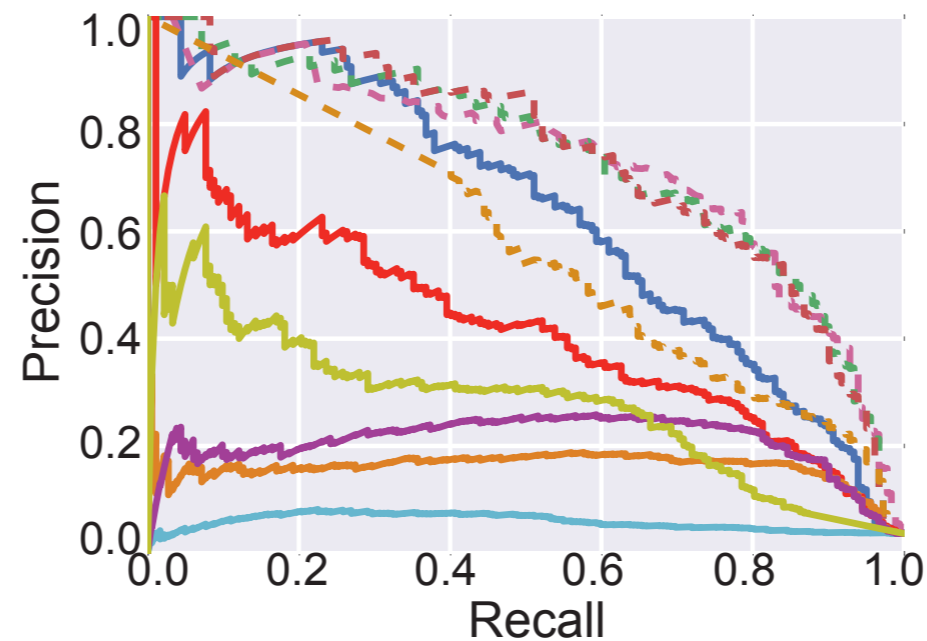
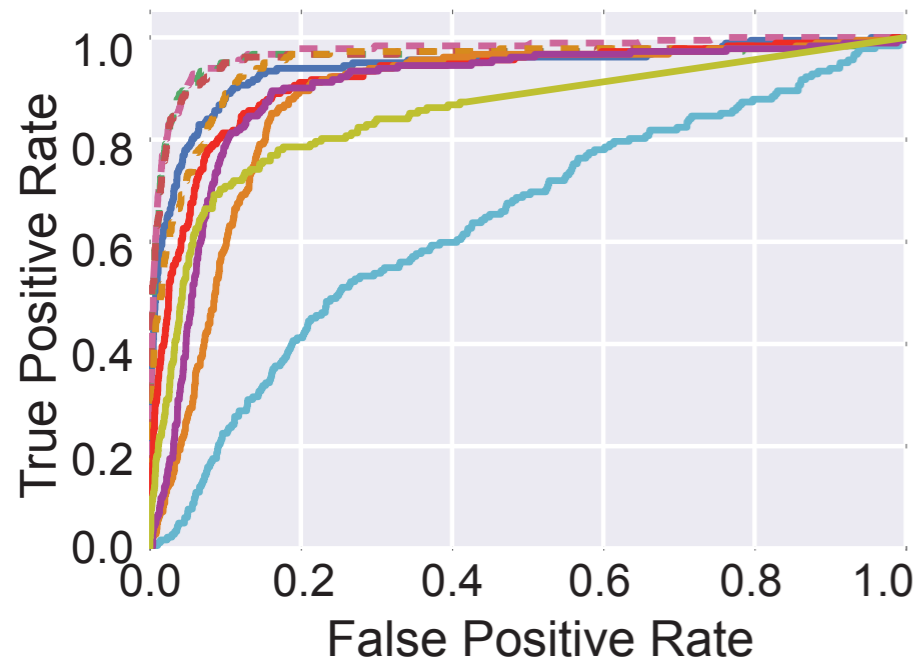
The chromatin marks are better for proximal regulatory elements

Proximal Regulatory Regions



Feature/Model	AUROC	AUPR
H3K27ac	0.96	0.71
H3K4me1	0.59	0.16
H3K4me2	0.92	0.49
H3K4me3	0.93	0.64
H3K9ac	0.95	0.69
DHS	0.89	0.59
Random Forest	0.97	0.78
Linear SVM	0.97	0.78
Ridge Regression	0.97	0.78
Naive Bayes	0.97	0.79

Distal Regulatory Regions



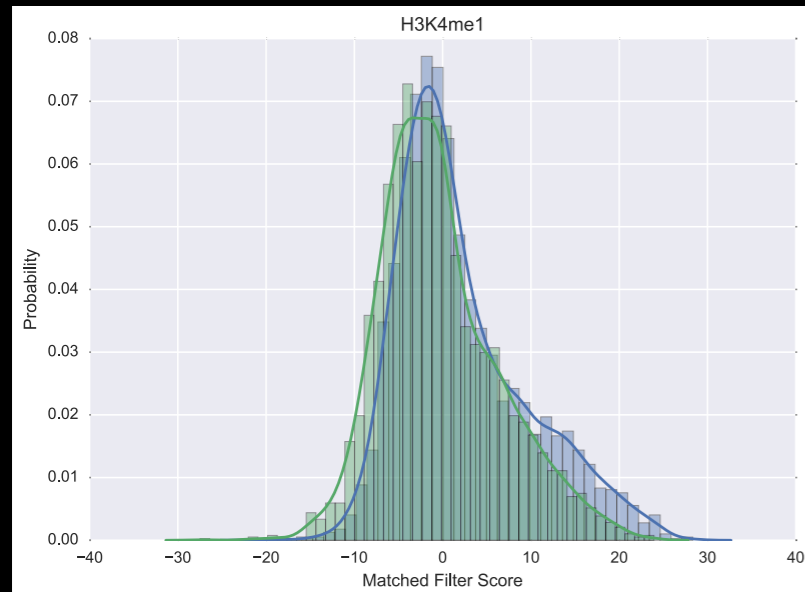
Feature/Model	AUROC	AUPR
H3K27ac	0.92	0.55
H3K4me1	0.90	0.36
H3K4me2	0.85	0.15
H3K4me3	0.63	0.06
H3K9ac	0.87	0.19
DHS	0.83	0.28
Random Forest	0.95	0.66
Linear SVM	0.94	0.66
Ridge Regression	0.94	0.65
Naive Bayes	0.92	0.54

Same machine learning models with all features - AUPR goes to 0.73

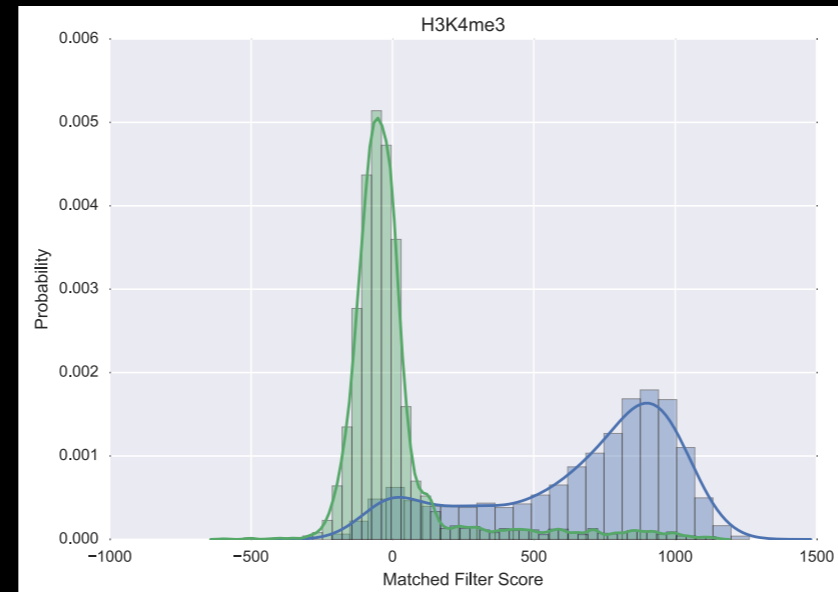
Marks and differences between promoters and enhancers

Promoters

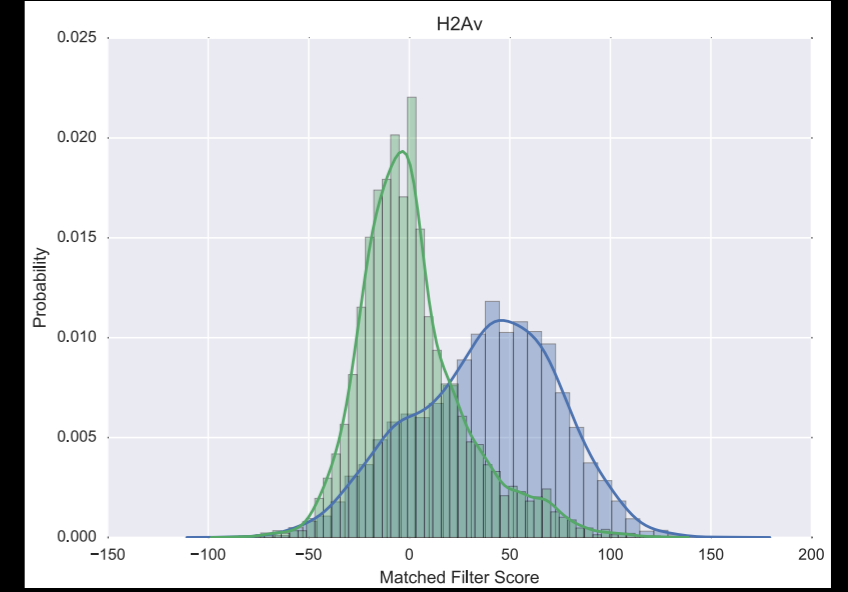
H3K4me1



H3K4me3

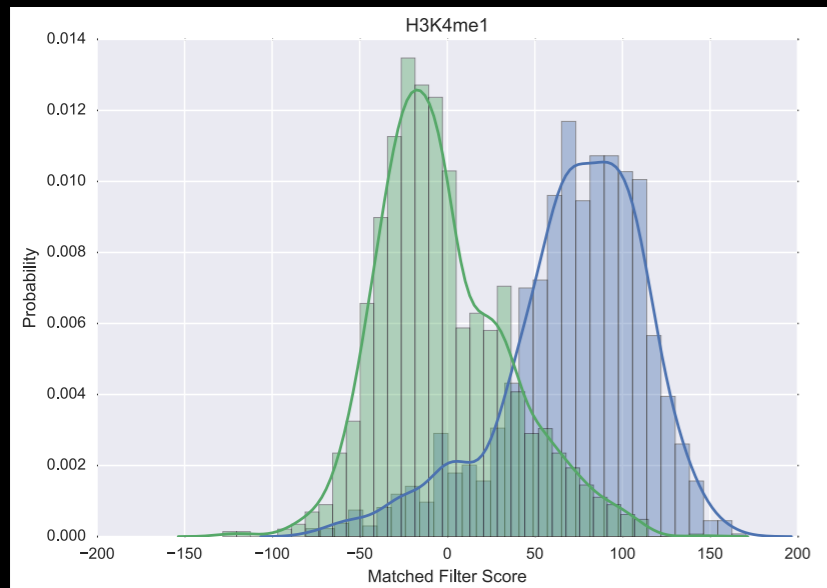


H2Av

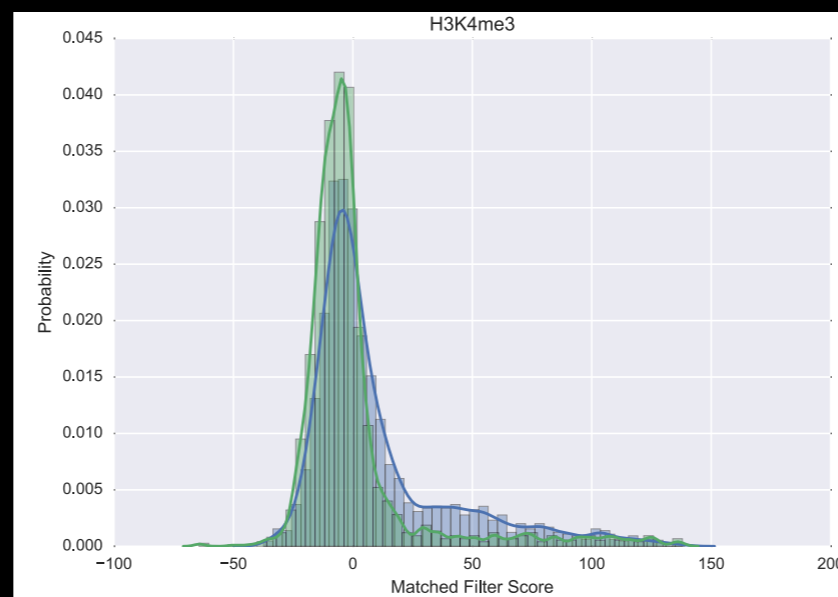


Enhancers

H3K4me1



H3K4me3



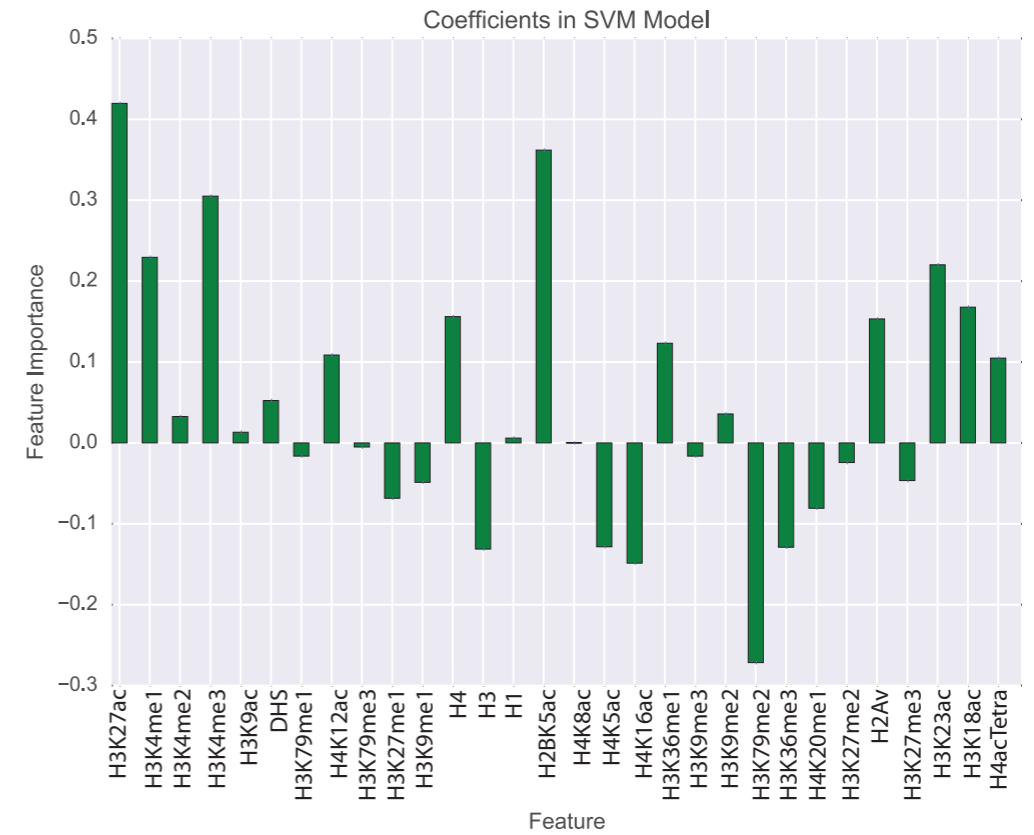
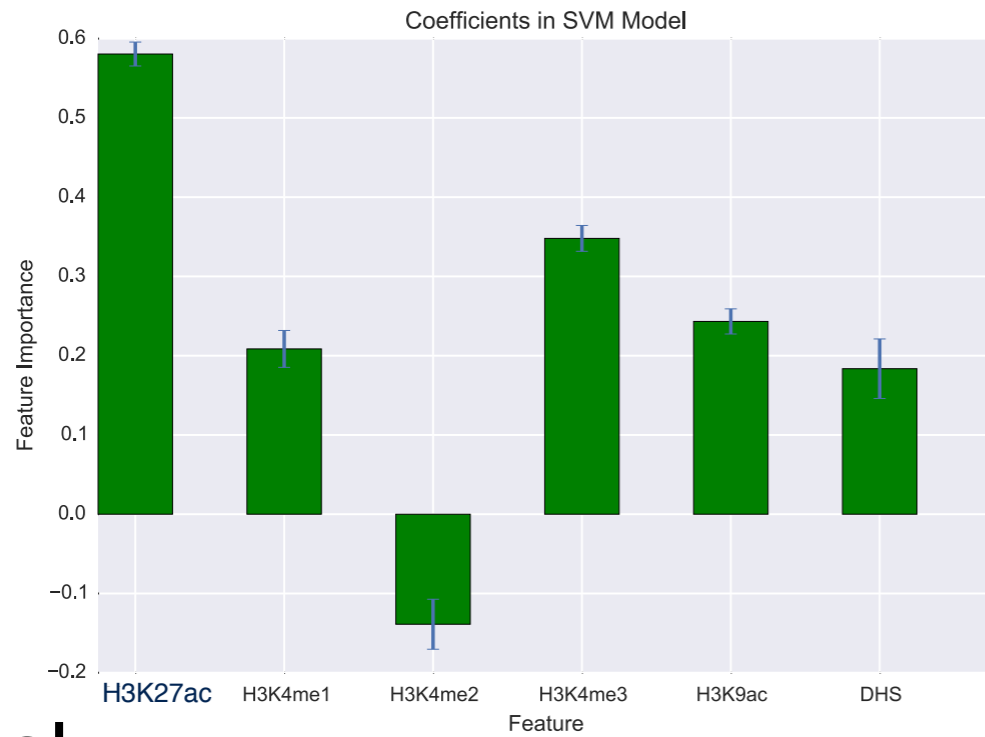
H2Av



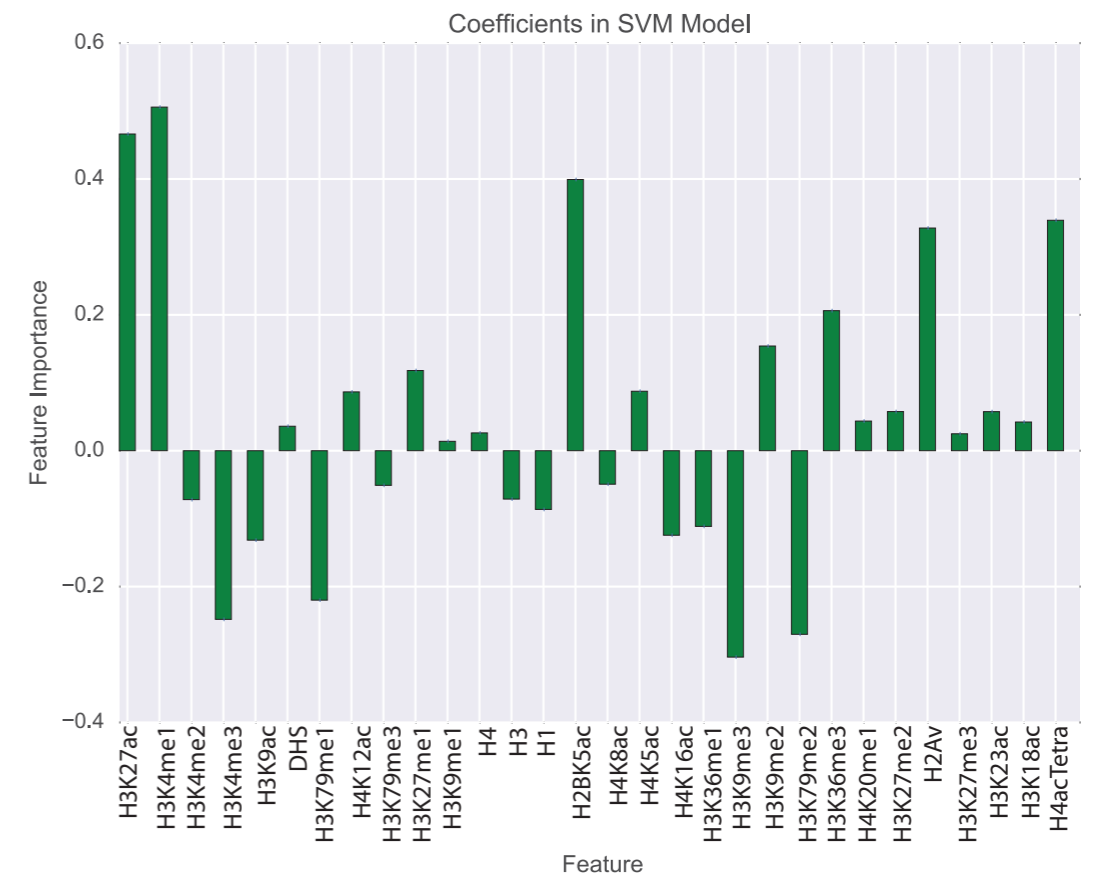
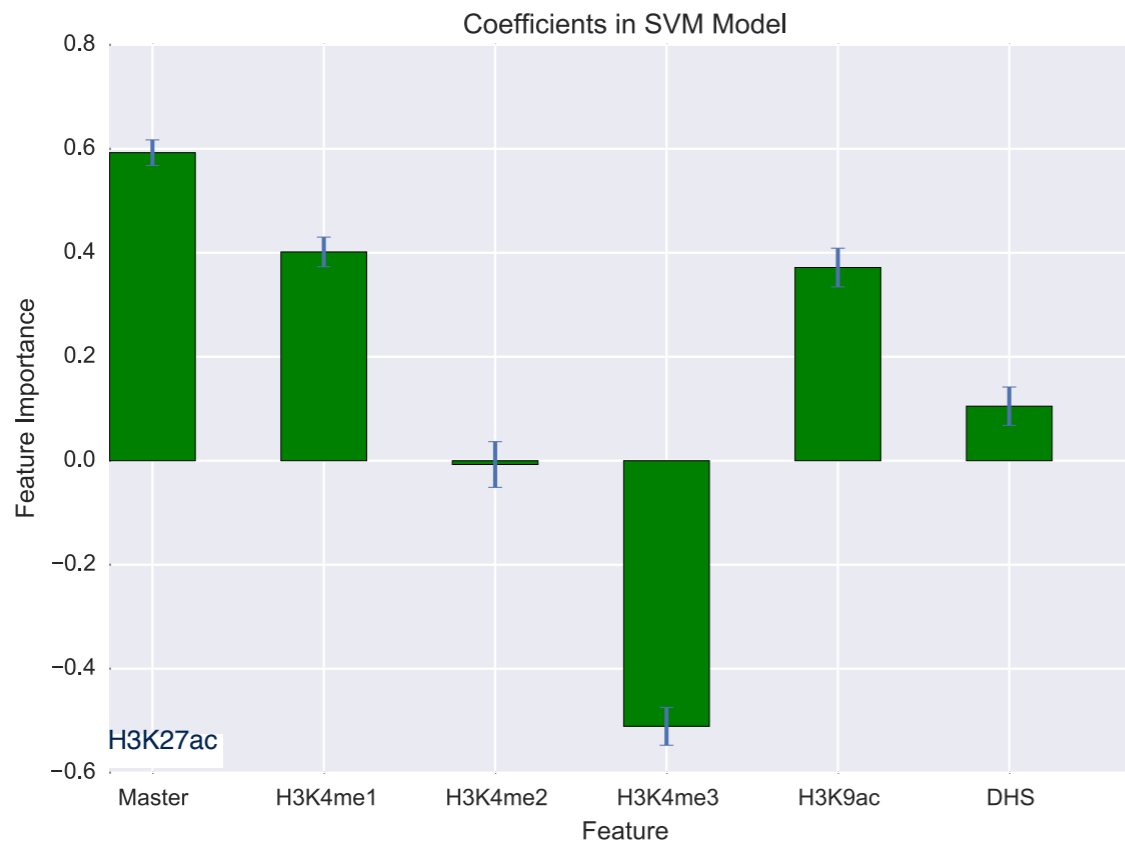
Enhancers contain unimodal distributions on H3K4me1, H3K4me3, and H2Av. Promoters might still contain a few elements that are more “enhancer-like”.

Change in importance of features for distal and proximal predictions based on these histone marks

Proximal



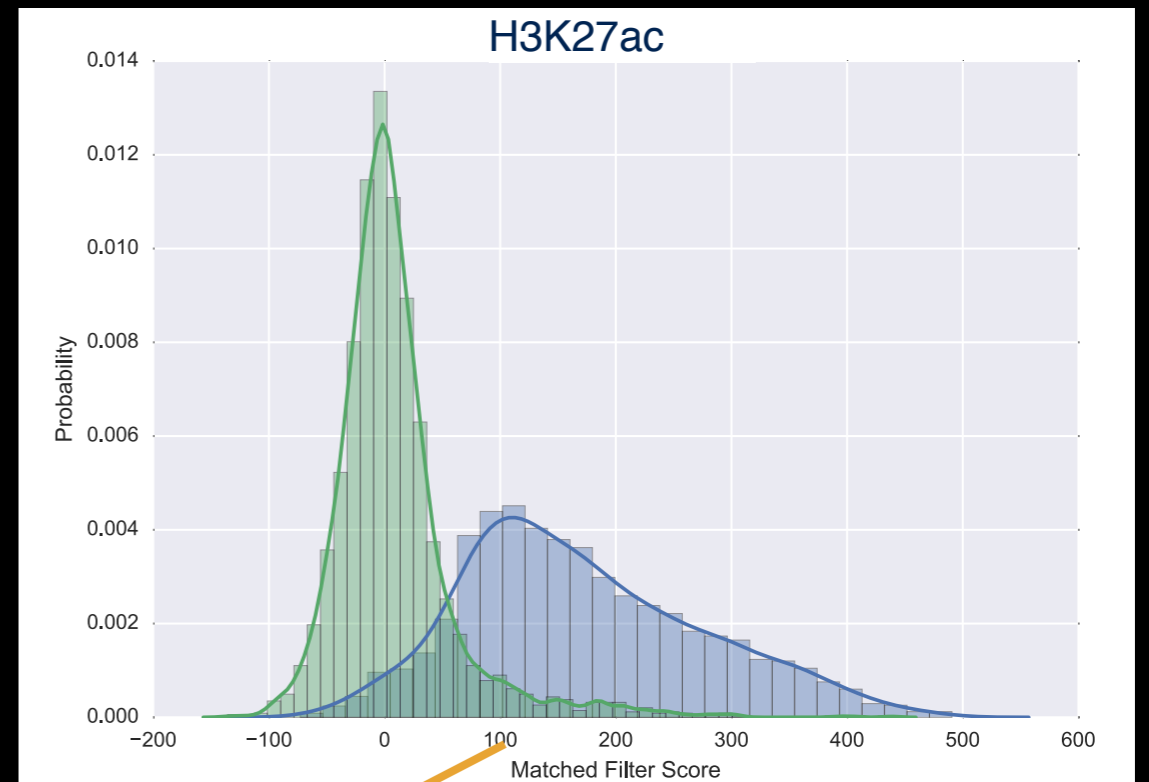
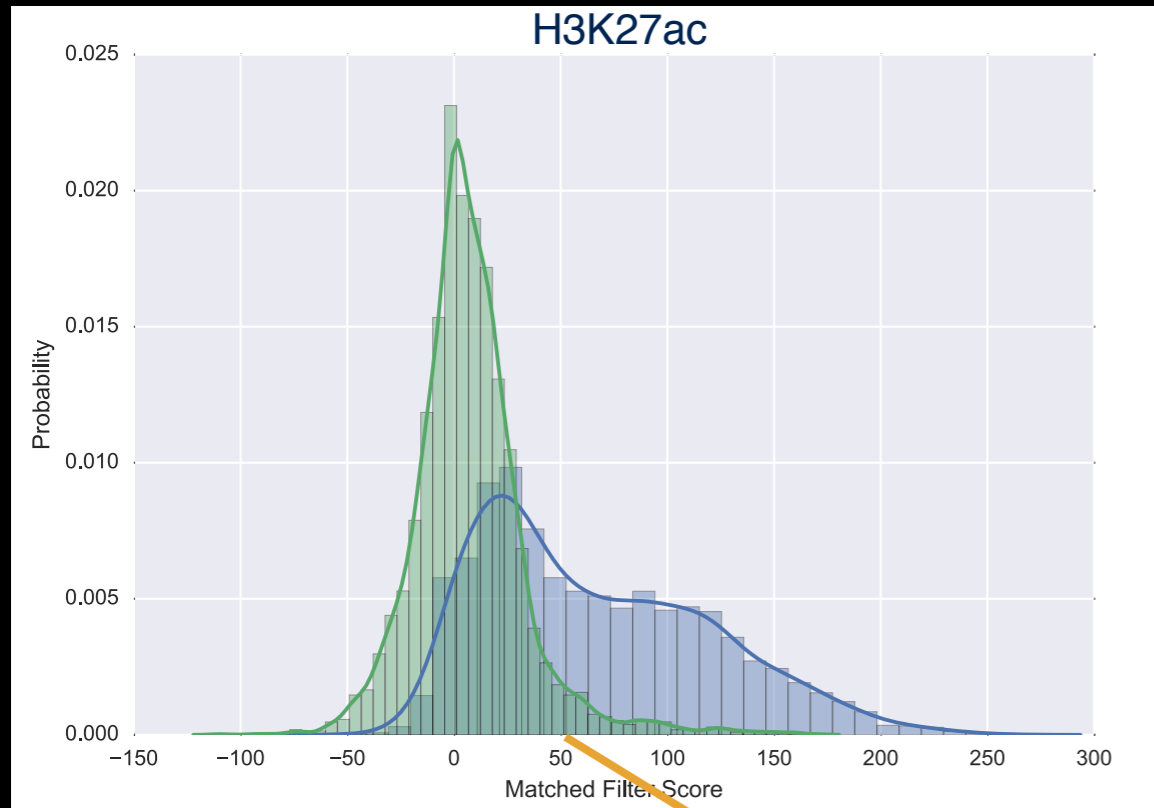
Distal



How conserved are these metaprofiles?

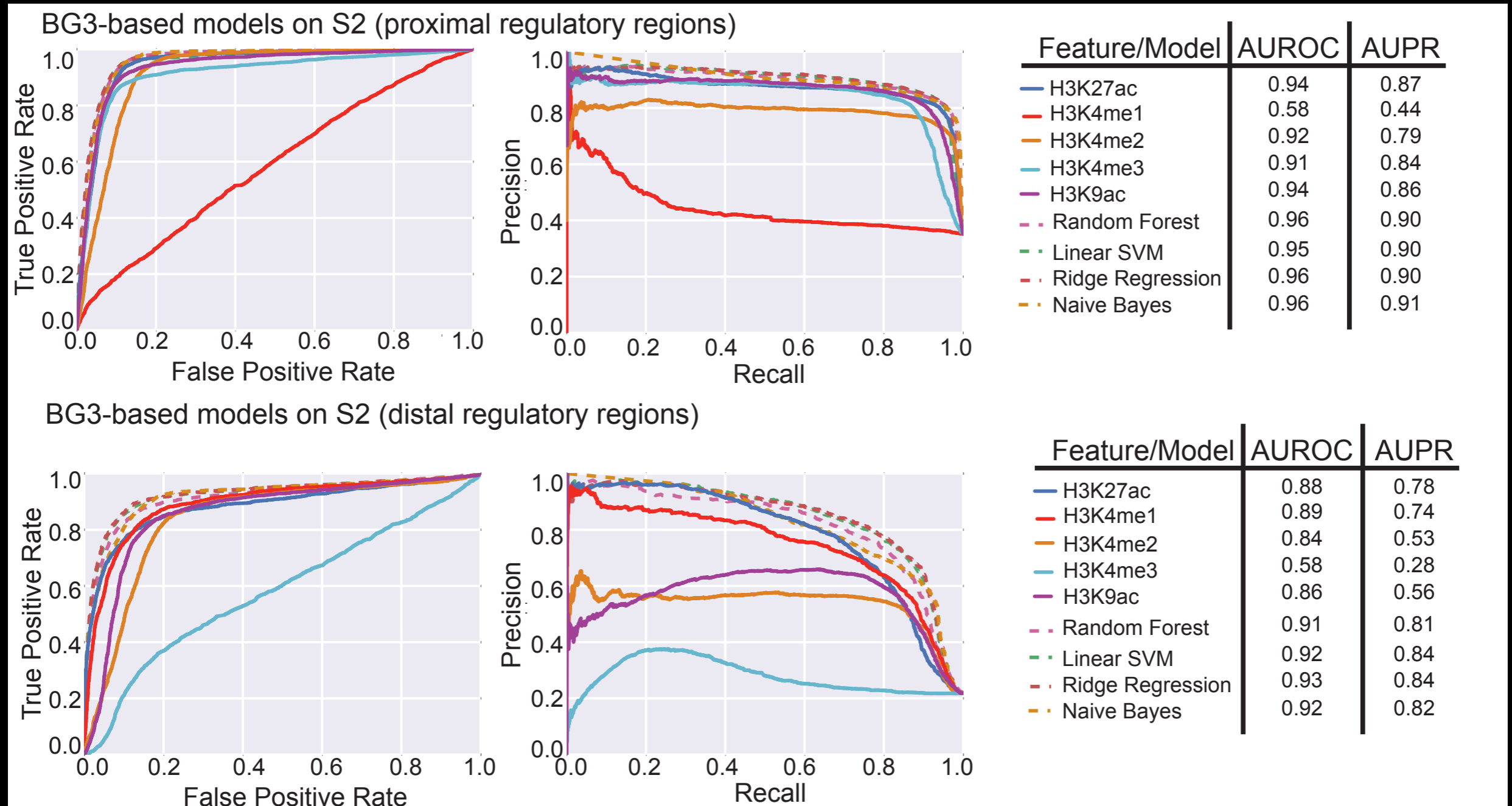
Can we use these machine learning models across tissues
and species?

Different cell-lines features histograms



Notice the difference in distribution of scores

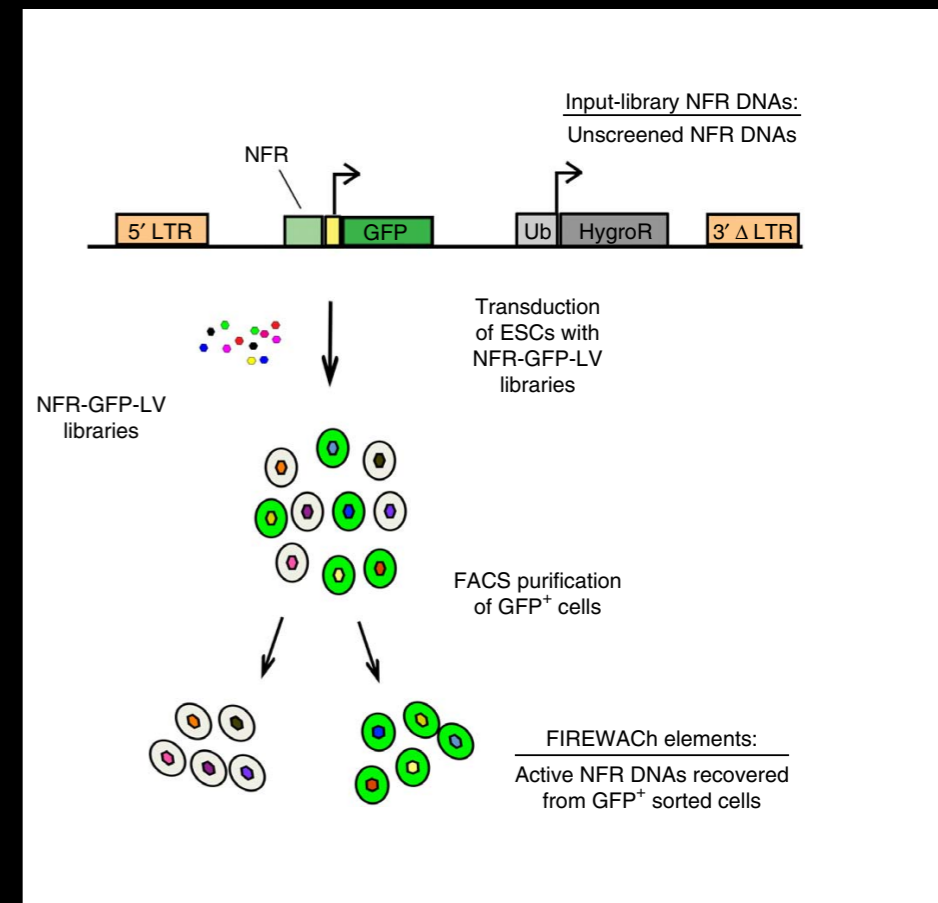
Matched filters can be used to predict enhancers across tissues and cell-lines



Histone enrichment profiles are conserved across regulatory elements from different cell-lines and tissues.

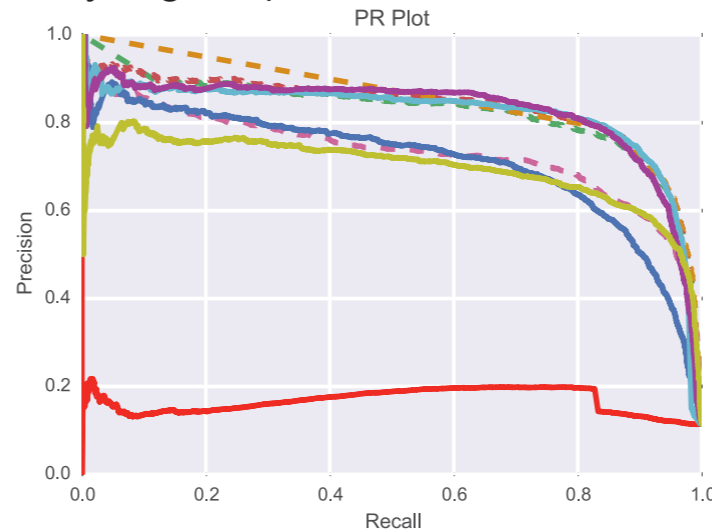
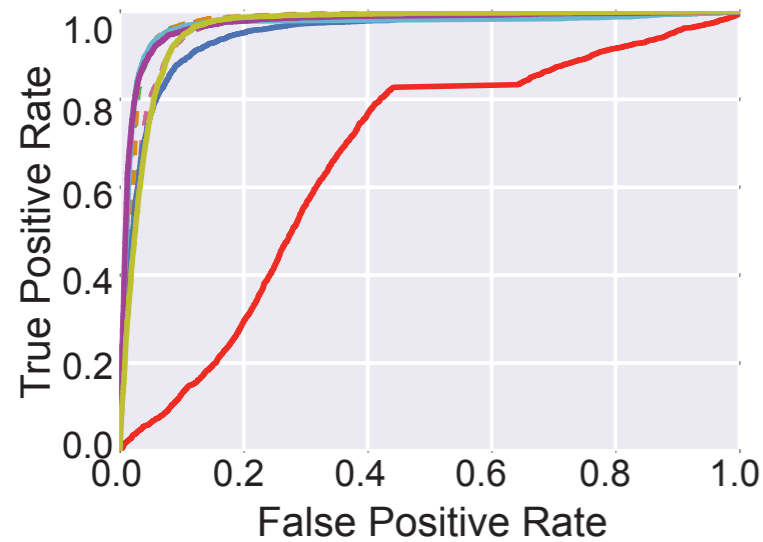
FIREWACH study design - massively parallel enhancer assay

- Enhancer candidates chosen based on open DNA in cell-line (murine ESC).
- Integrated into virus particles close to a minimal promoter and GFP.
- **Integrated into genome randomly with 1 clone per cell (H1-hESC).**
- **One potential enhancer of length 100-300 bp per cell.**
- FACS to sort cells expressing GFP.
- Small population of cells show positive enhancer activity.
- Amplified positive enhancer sequences with PCR using primers recognizing the flanking sequences.
- Tested enhancer activity using traditional assays.



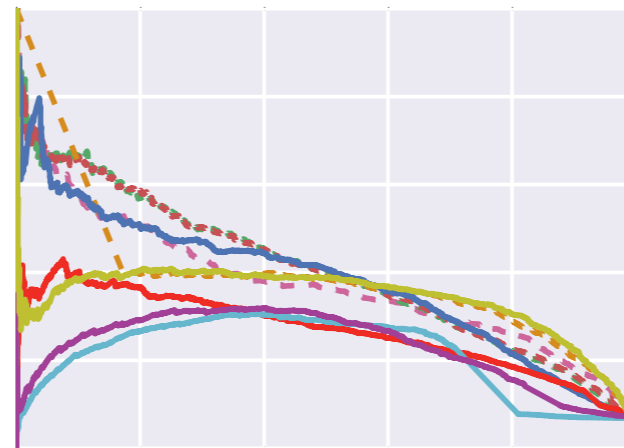
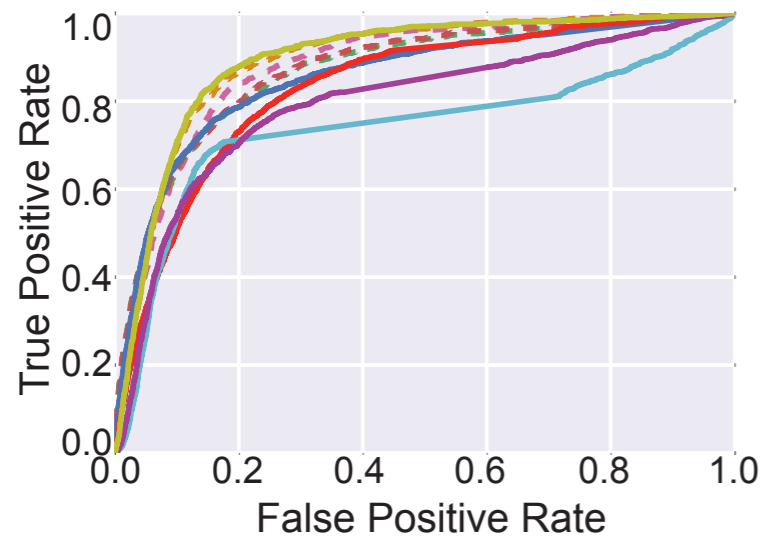
Conservation of histone modification profile across species

S2-based models on mESC14 (proximal regulatory regions)



Feature/Model	AUROC	AUPR
H3K27ac	0.95	0.71
H3K4me1	0.66	0.17
H3K4me3	0.97	0.82
H3K9ac	0.97	0.83
DHS	0.96	0.70
Random Forest	0.96	0.73
Linear SVM	0.98	0.83
Ridge Regression	0.97	0.83
Naive Bayes	0.97	0.86

S2-based models on mESC14 (distal regulatory regions)



Feature/Model	AUROC	AUPR
H3K27ac	0.86	0.38
H3K4me1	0.83	0.27
H3K4me3	0.74	0.21
H3K9ac	0.80	0.23
DHS	0.90	0.34
Random Forest	0.88	0.38
Linear SVM	0.88	0.40
Ridge Regression	0.87	0.40
Naive Bayes	0.89	0.39

The models from fly work in mouse though there is reduction in AUPR - especially for enhancers!

Why does mouse have lower AUPR than fly with these models?

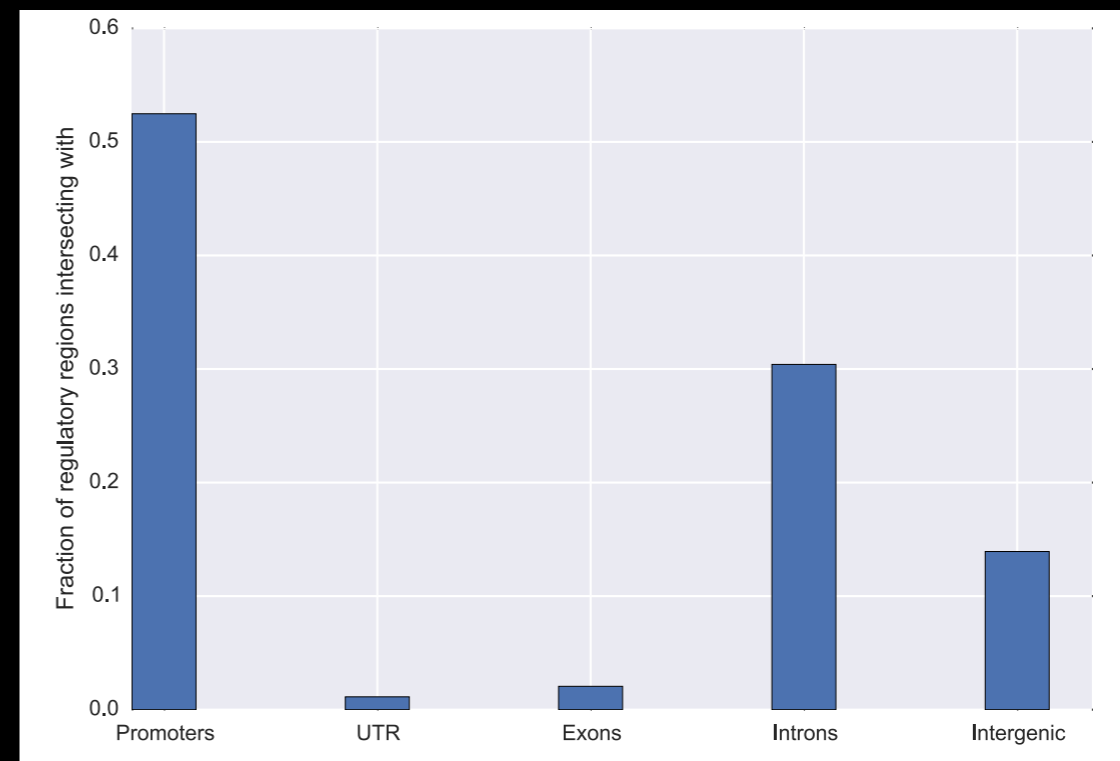
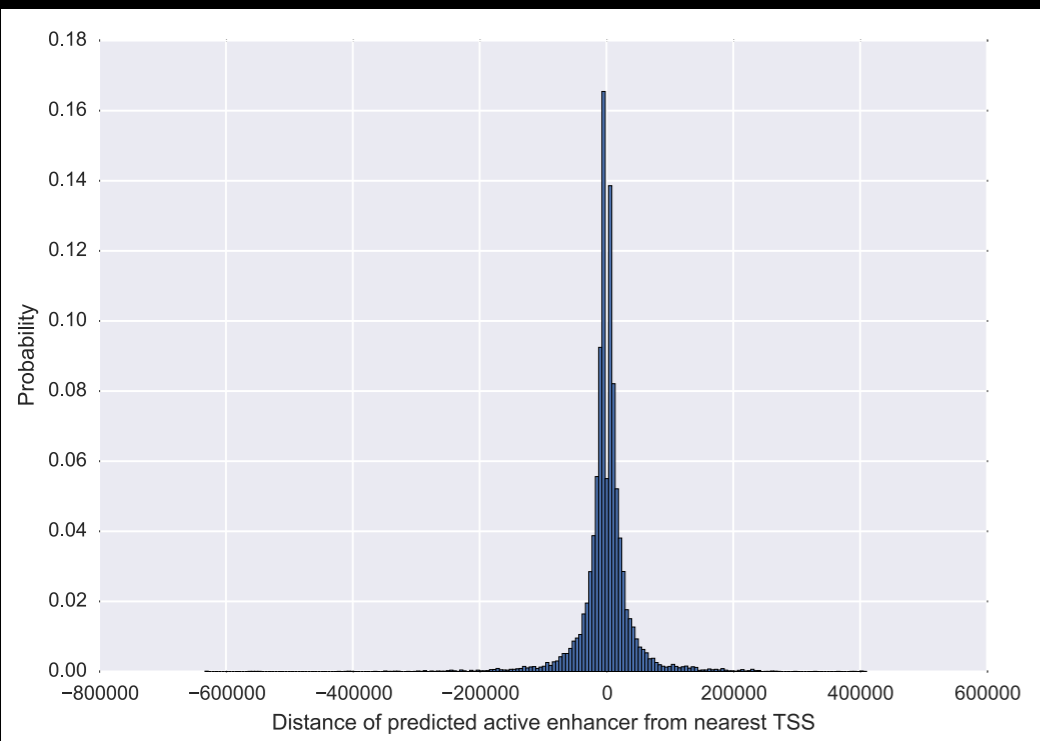
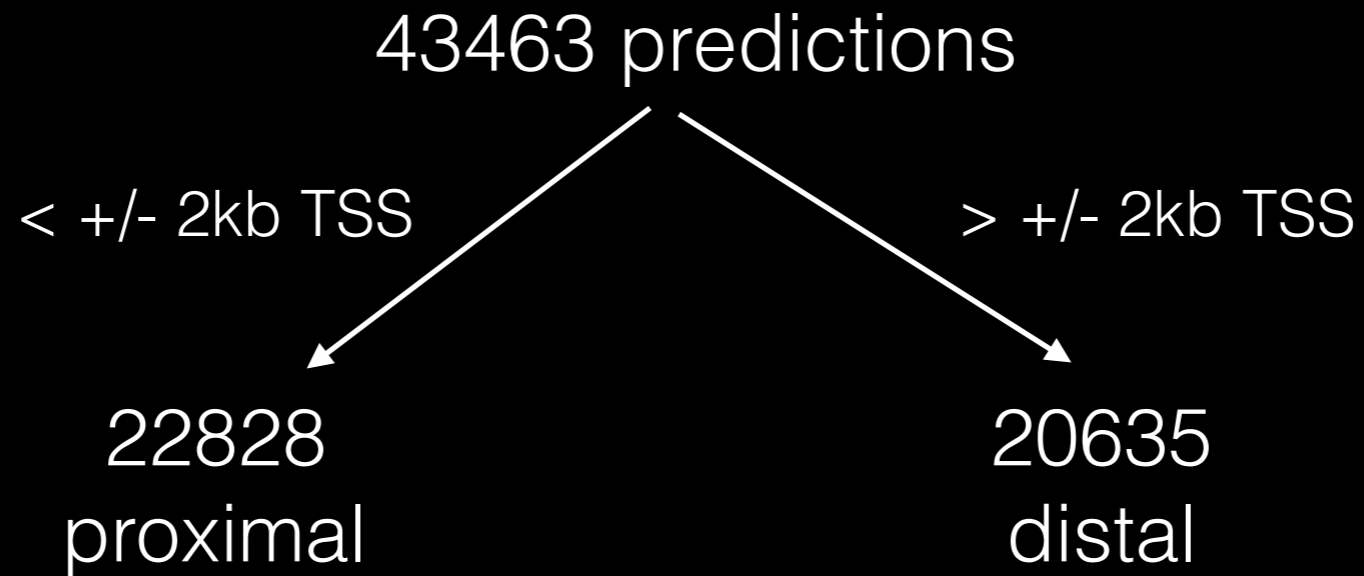
Mammalian genomes may have more types (or more complex regulation) of core promoters than fly (we know from fly that AUPR goes up when you include enhancers from multiple core promoters).

These assays have low dynamic range and many enhancers are labeled as false positives even though they may be true positives (bigger problem in mammals which have larger genomes and more negatives).

A larger proportion of enhancers in mammals might not display functionality in non-native context.

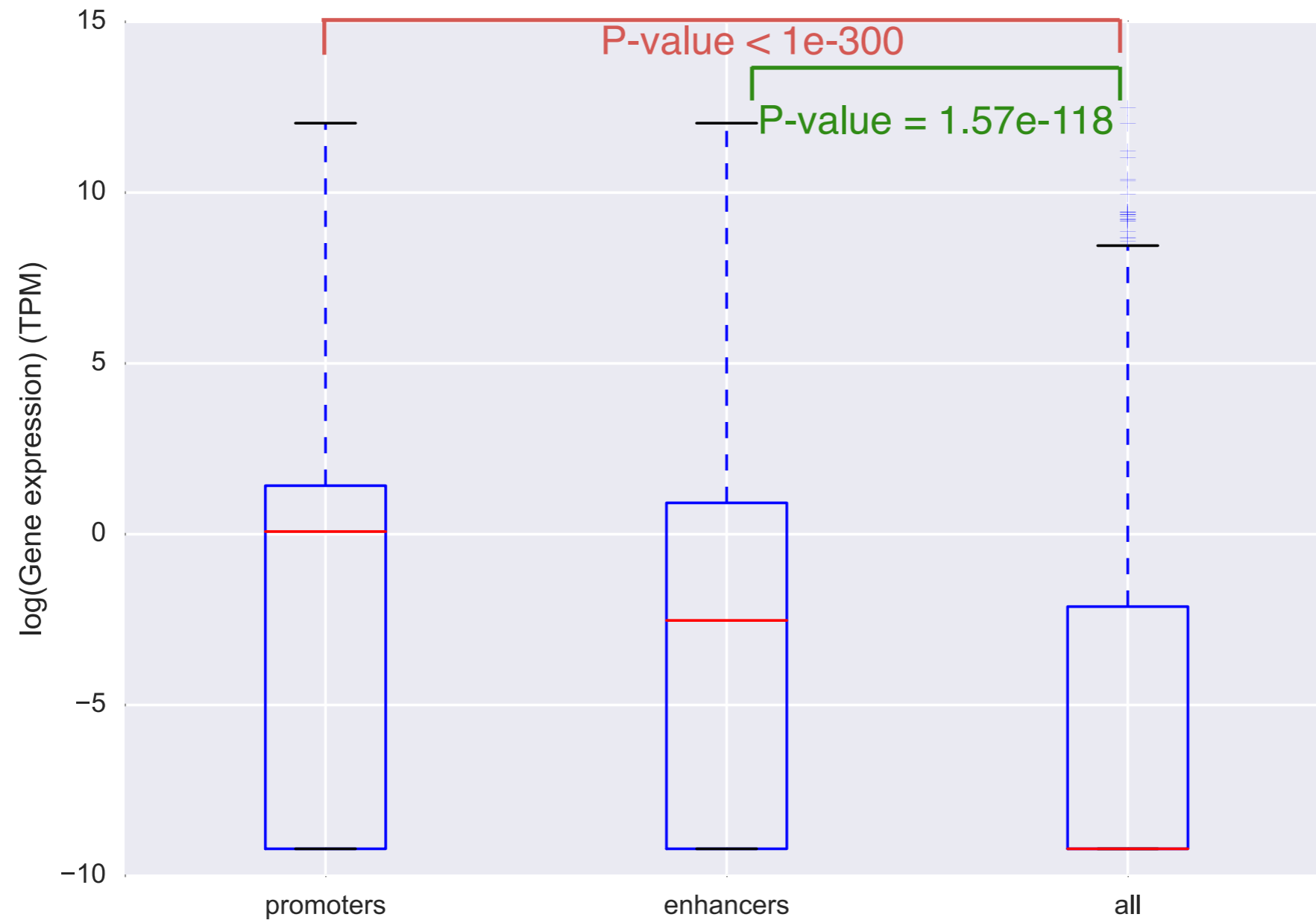
These histone marks may have evolved new functional roles in mammals.

Applying matched filter for whole genome prediction



Ideally, I would follow this up with some experimental validation.

Gene expression - regulatory regions



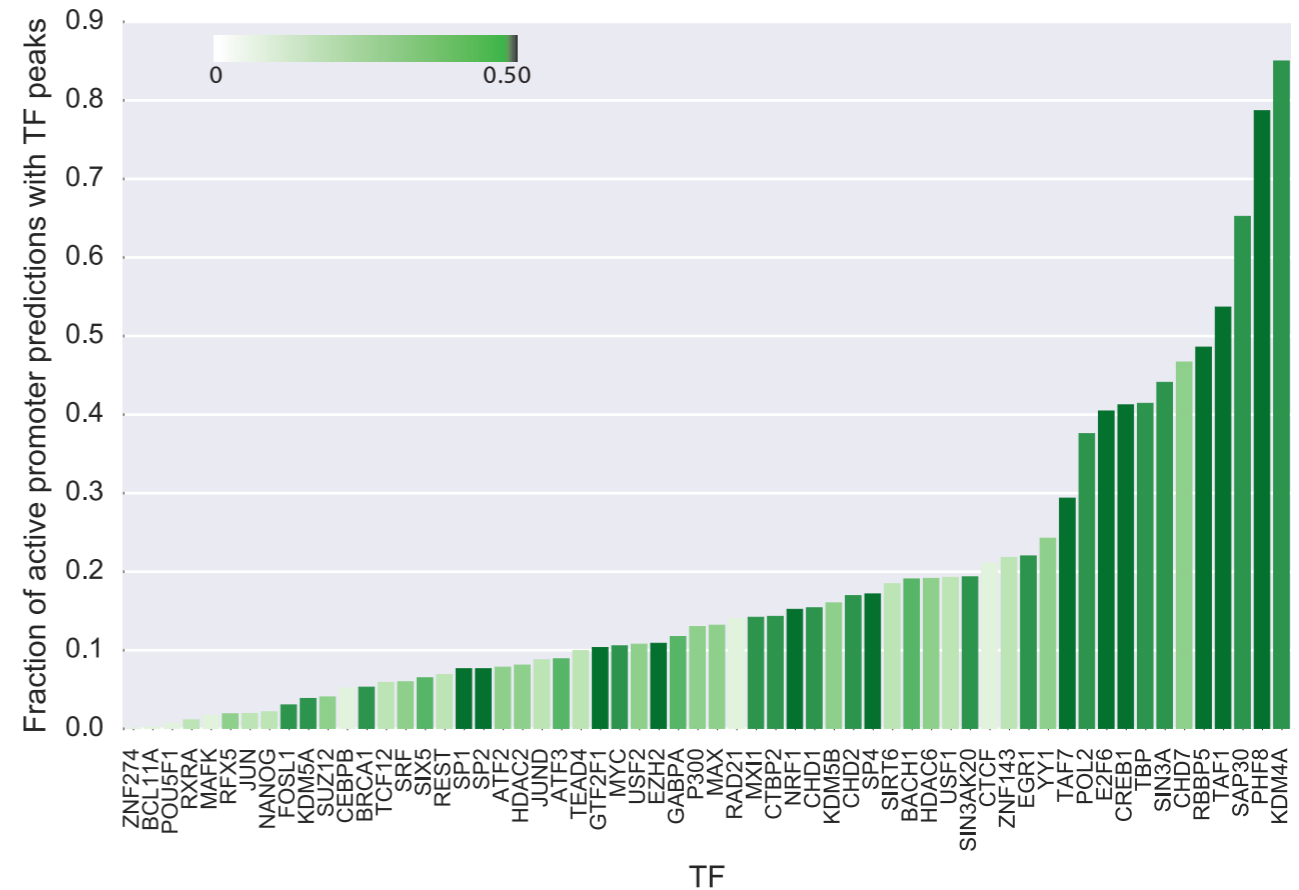
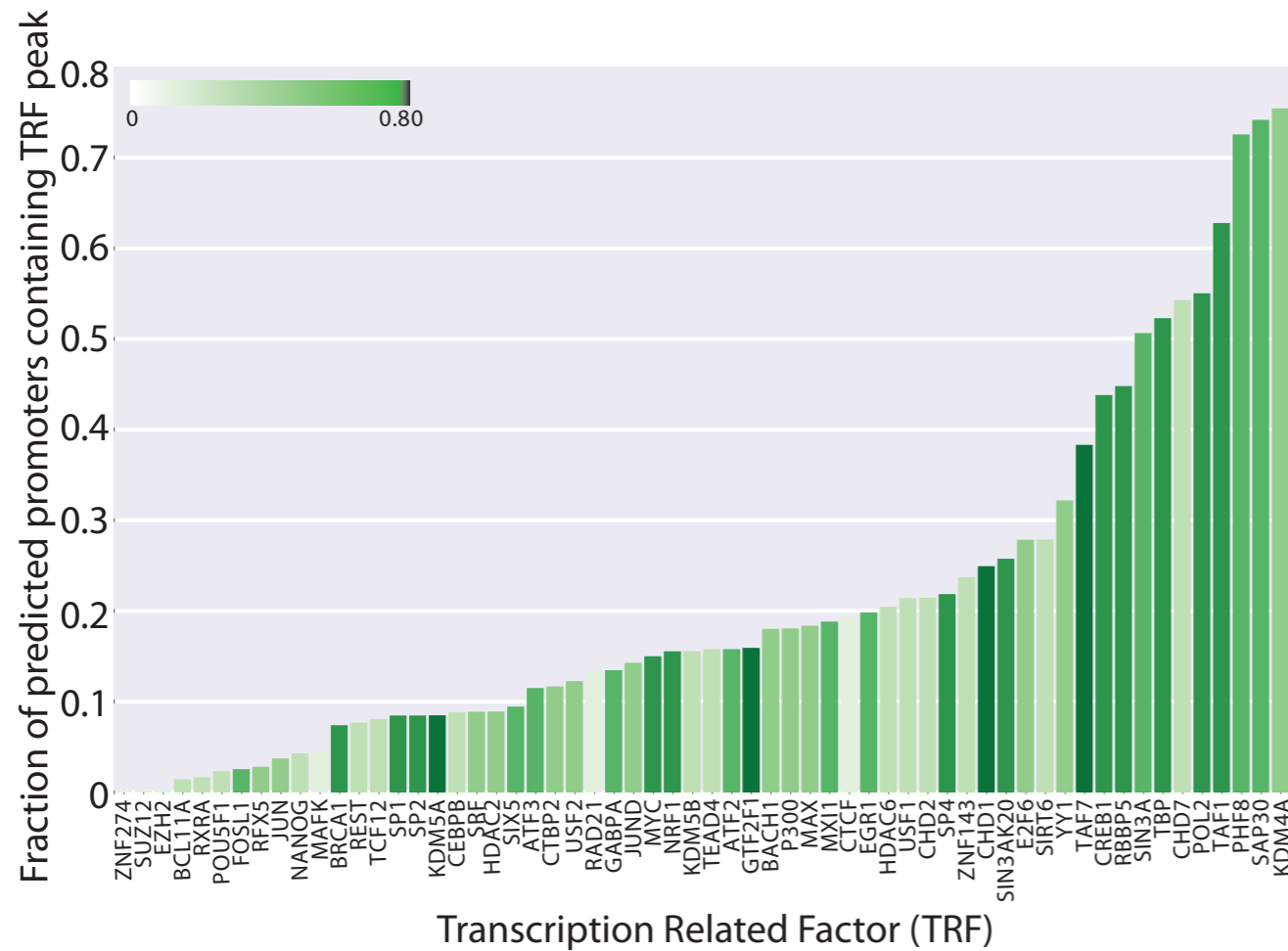


Matched Filter v SuttonSeq

Matched Filter v ImPACT-seq - The battle of the proximal regulatory regions

Matched Filter Predictions

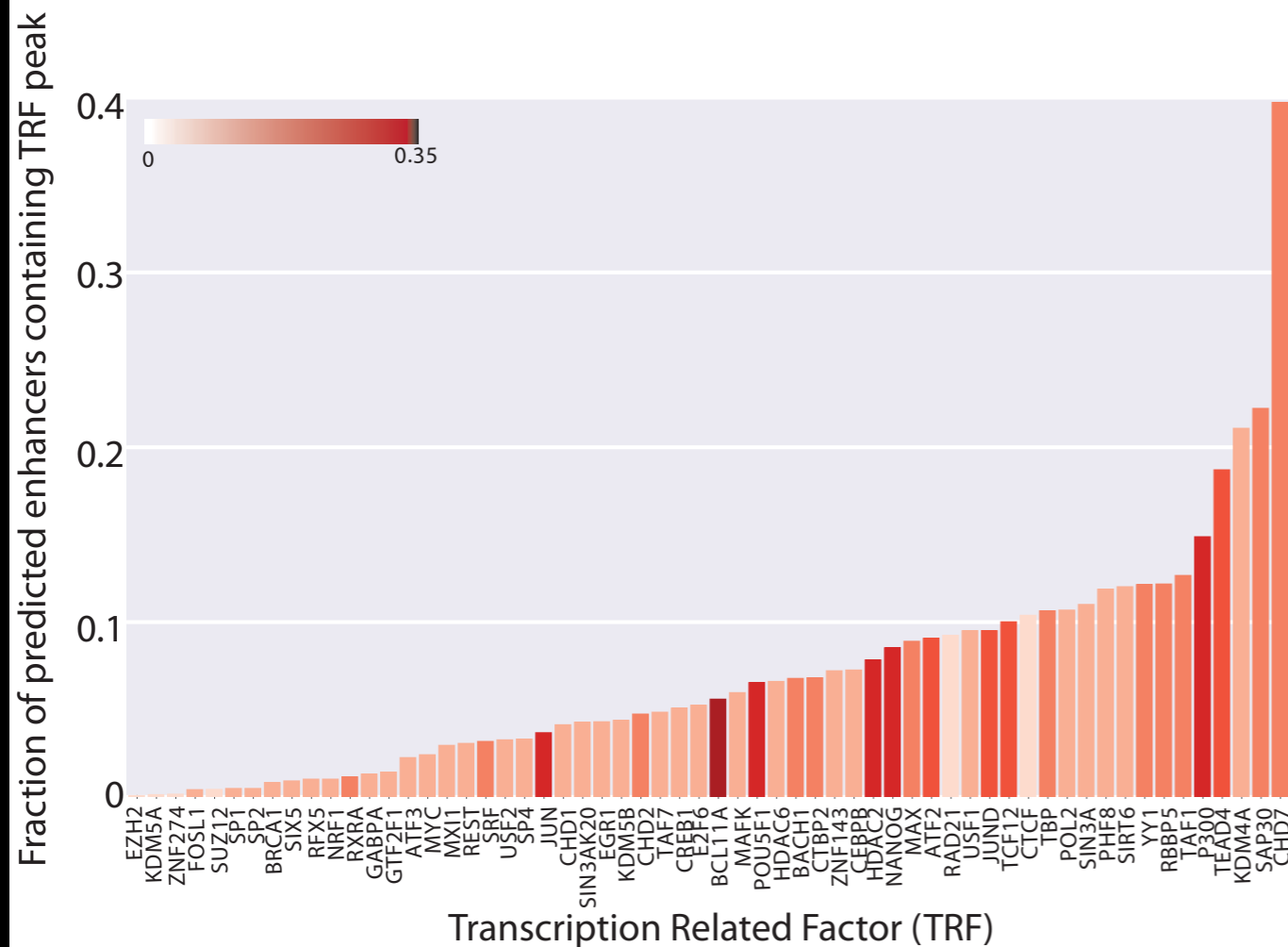
ImPACT-seq



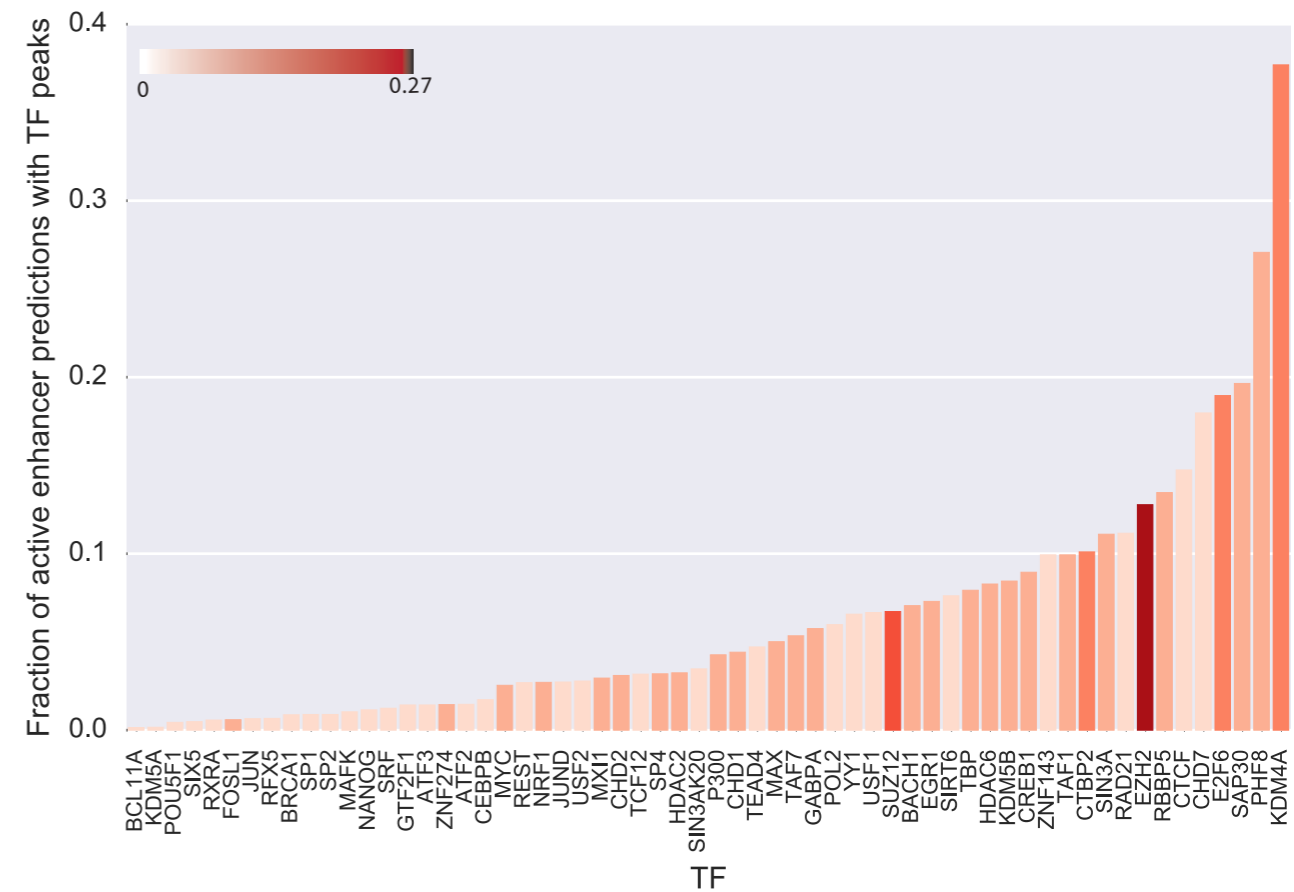
The same TFBSs are enriched in our predictions

Matched Filter v ImPACT-seq - The battle of the distal regulatory regions

Matched Filter Predictions

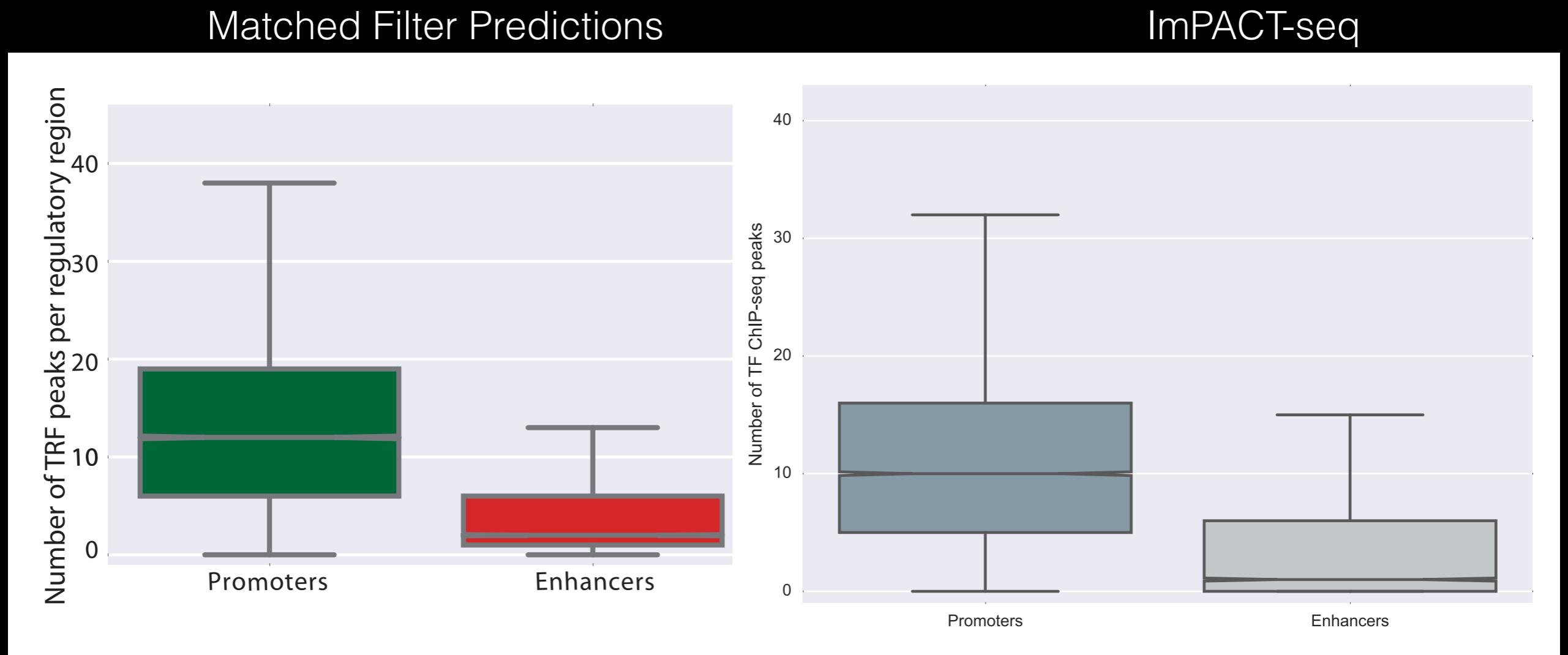


ImPACT-seq

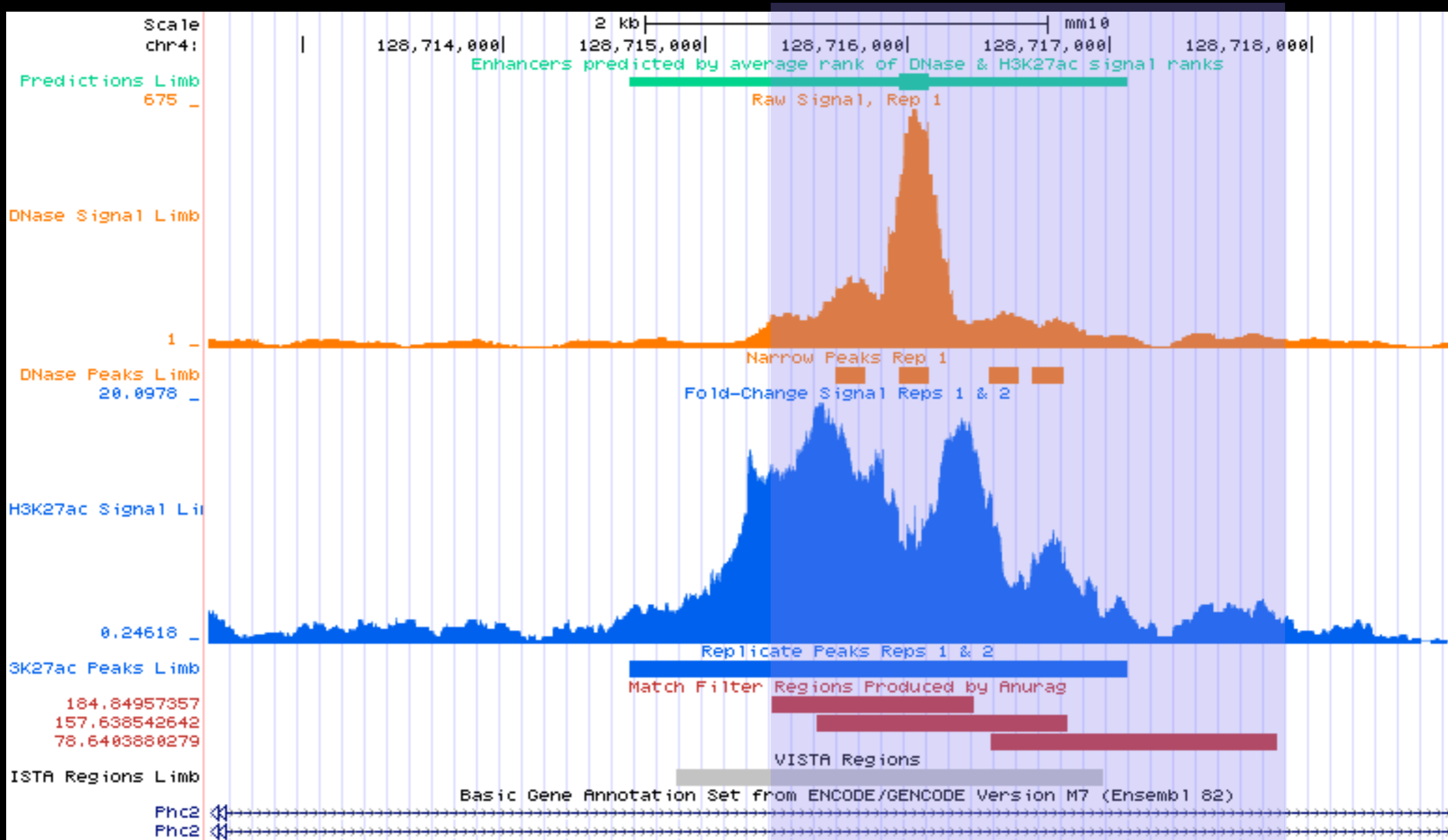


Enhancers found using ImPACT-seq are also diverse in terms of TRF binding. EZH2 and SUZ12 are repressors that are found in a number of ImPACT-seq peaks because it also finds poised enhancers.

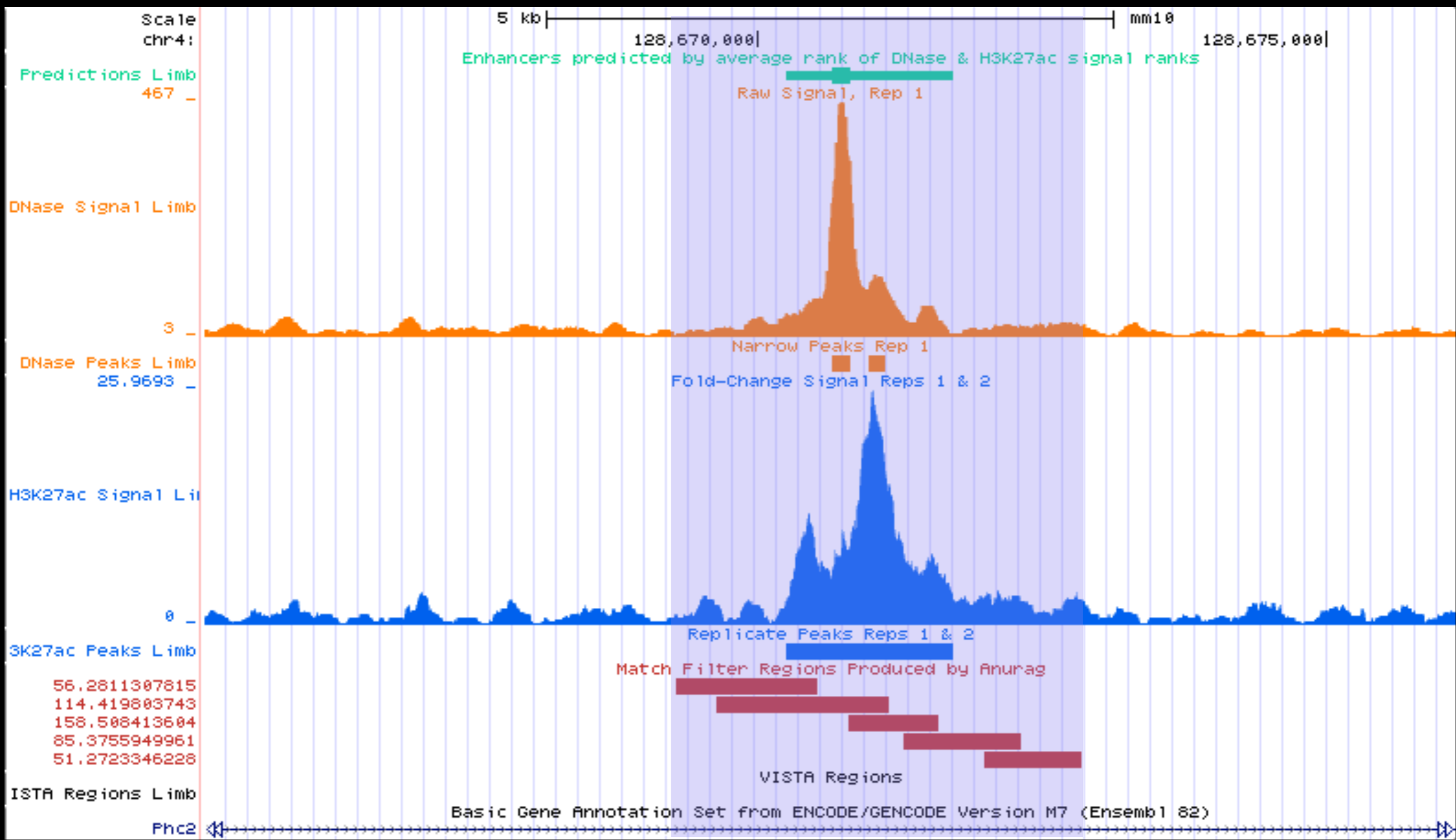
Matched Filter v ImPACT-seq - The battle of the ChIP-Seq peaks



Overall, the number of ChIP-Seq peaks per prediction is slightly higher than that identified in the experiment.



Marking positions of maxima among the whole MF region



Marking positions of maxima among the whole MF region

Conclusions

Our collaborators developed a new method to identify regulatory regions.

We showed that these regions could function as regulatory regions based on their properties.

We developed a new method to predict regulatory activators that utilizes information in the shape of chromatin data.

The enhancers coincided with TFBS and we were able to identify a few promoter-associated TFs.

The enhancers tended to occur closer to active genes (maybe add 3d context to this sentence).

Acknowledgements

Mark

Enhancer prediction: Kevin, Joel, Landon, Xue

ImPACT-seq: Sutton, Joel

ENCODE Enhancer: Zhiping, Jill, Jing, Sushant, and many many more

ENCODE Cancer: Yunsi, Jing, Donghoon

tsSIN: Declan, Rob, Koon-Kiu

COSB: Declan, Jieming, Sushant

Allostery: Declan, Sushant, Shantao

Gerstein Lab