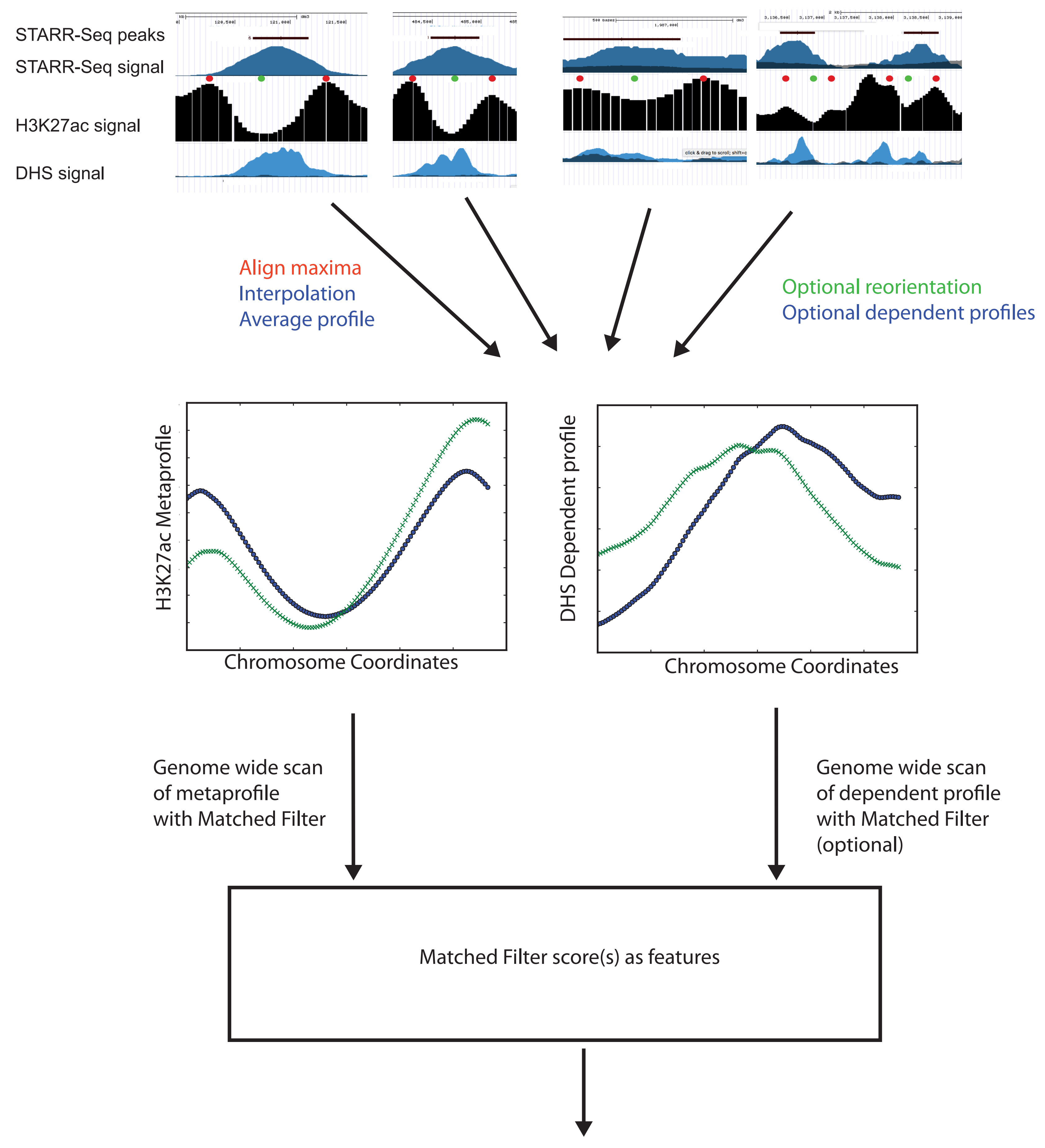
Multisignal matched filter for enhancer prediction

Anurag Sethi
TECH

Massively parallel assay for regulatory activity



Align maxima
Interpolation
Average profile

Optional reorientation
Optional dependent profiles

Genome wide scan
of metaprofile
with Matched Filter

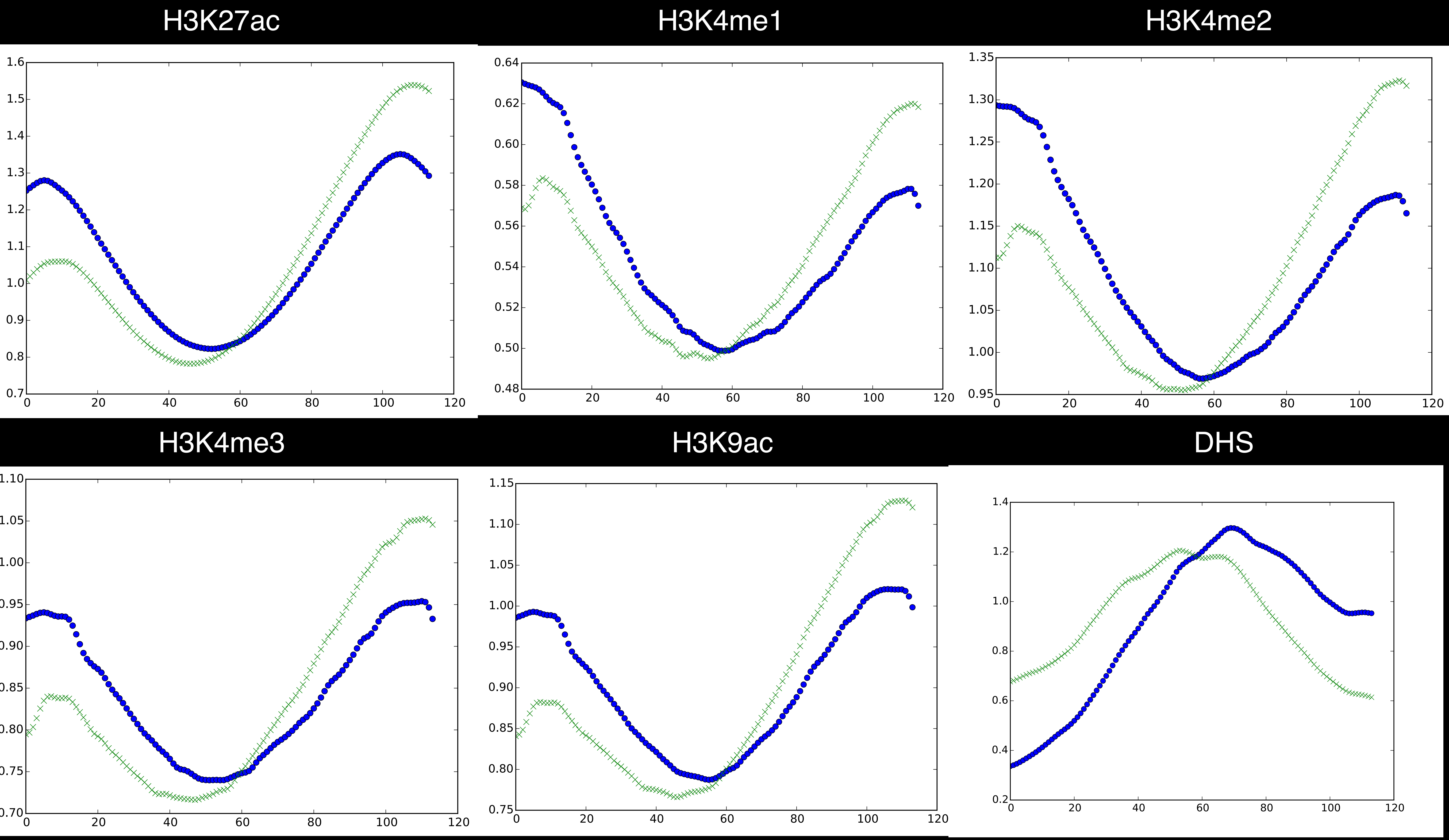
Genome wide scan
of dependent profile
with Matched Filter
(optional)

Matched Filter score(s) as features

Prediction of regulatory regions

Initial analyses performed based on single STARR-seq experiment

H3K27ac = Master Signal for active regulatory regions


H3K27ac

H3K4me1

H3K4me2

H3K4me3

H3K9ac

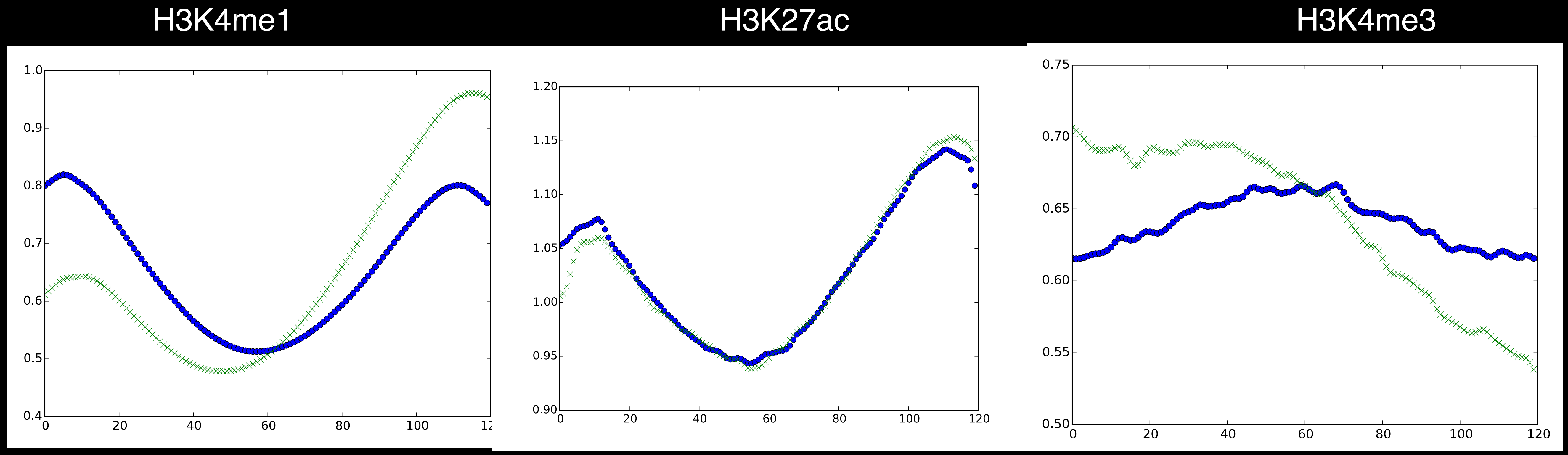
DHS

Without reorienting

After reorienting based on signal on both H3K27ac peaks

Interestingly, the double peak is visible in all the regulatory histone marks and they are correlated (higher maxima are oriented towards each other).

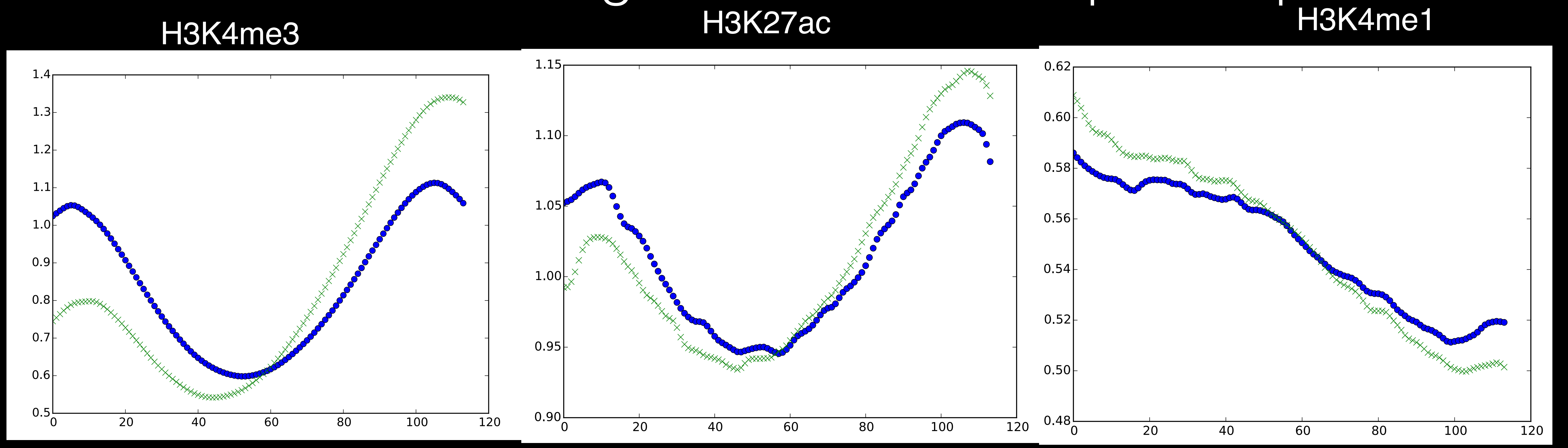
H3K4me1 = Master Signal for active and poised enhancers


H3K4me1

H3K27ac

H3K4me3

H3K4me3 = Master Signal for active and poised promoters

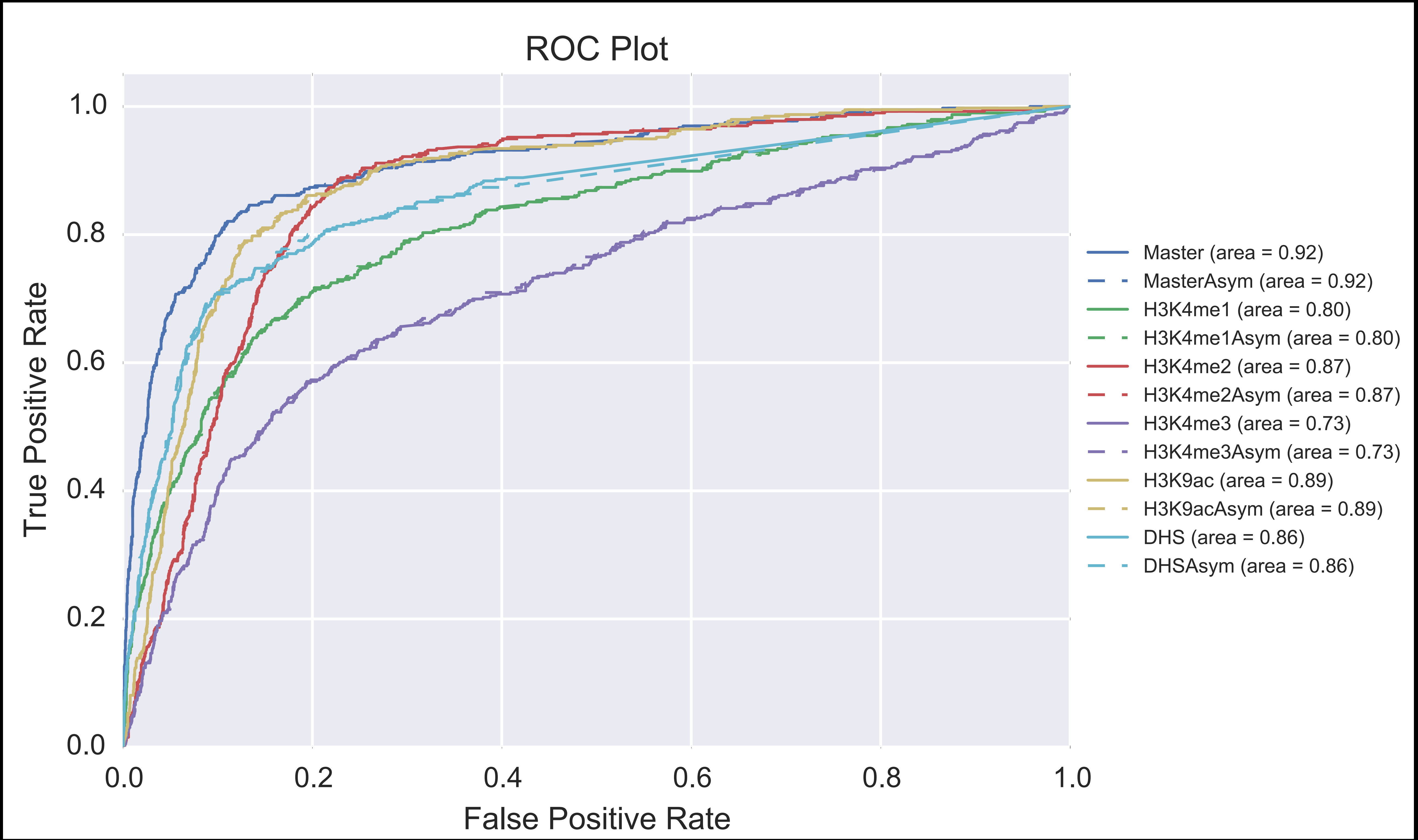

H3K4me3

H3K27ac

H3K4me1

Promoters (enhancers) may have enriched H3K4me1 (H3K4me3) signal but the double peak pattern may be present only in H3K4me3 (H3K4me1).

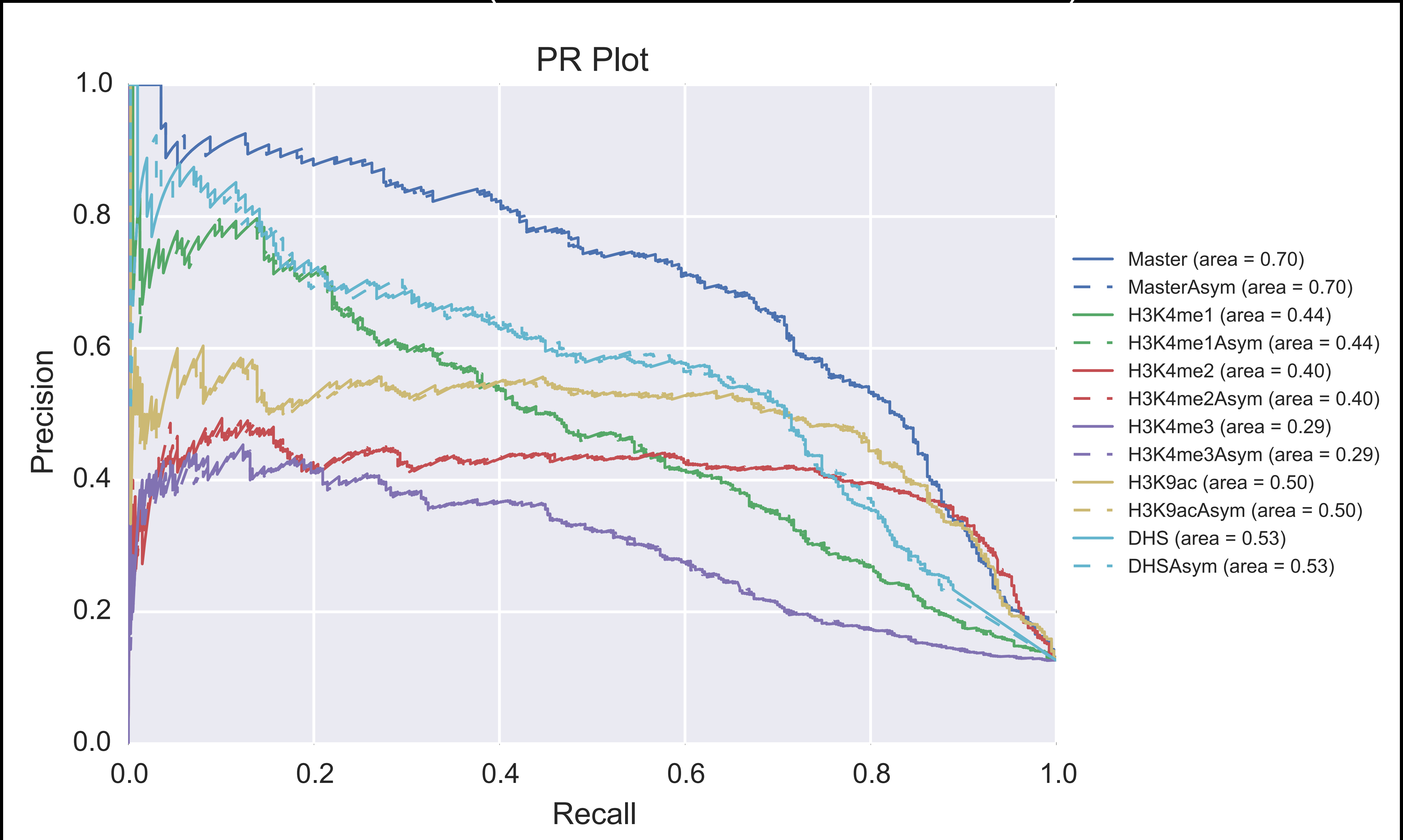
Comparison of performance of averaged and asymmetric marks (10-fold cross validation)



H3K27ac > H3K9ac > H3K4me2=DHS > H3K4me1 > H3K4me3

Asymmetric profiles have similar AUROC/AUPR as symmetric profiles

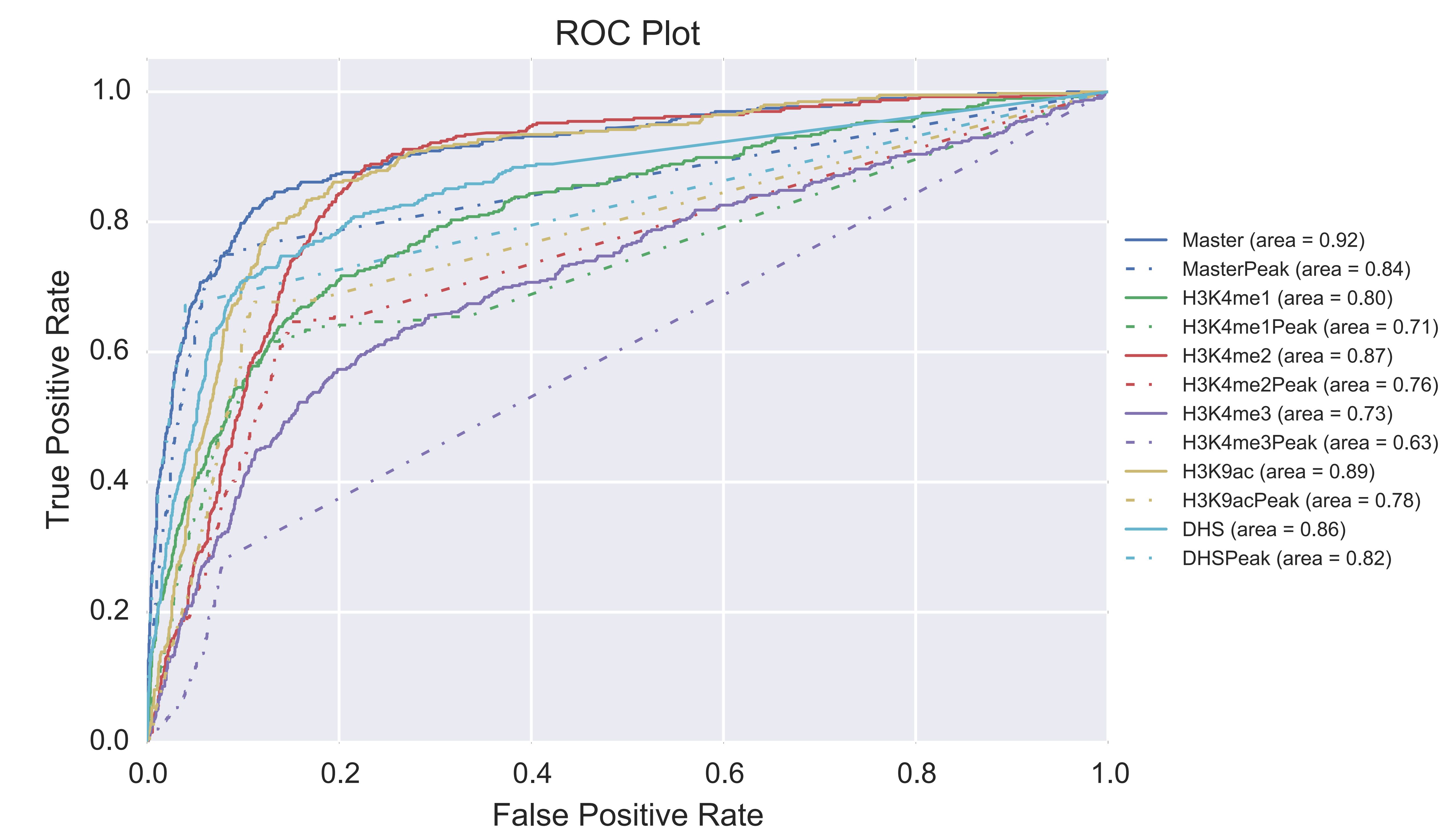
Comparison of performance of averaged and asymmetric marks (10-fold cross validation)



H3K27ac > DHS > H3K9ac > H3K4me1 > H3K4me2 > H3K4me3

Asymmetric profiles have similar AUROC/AUPR as symmetric profiles

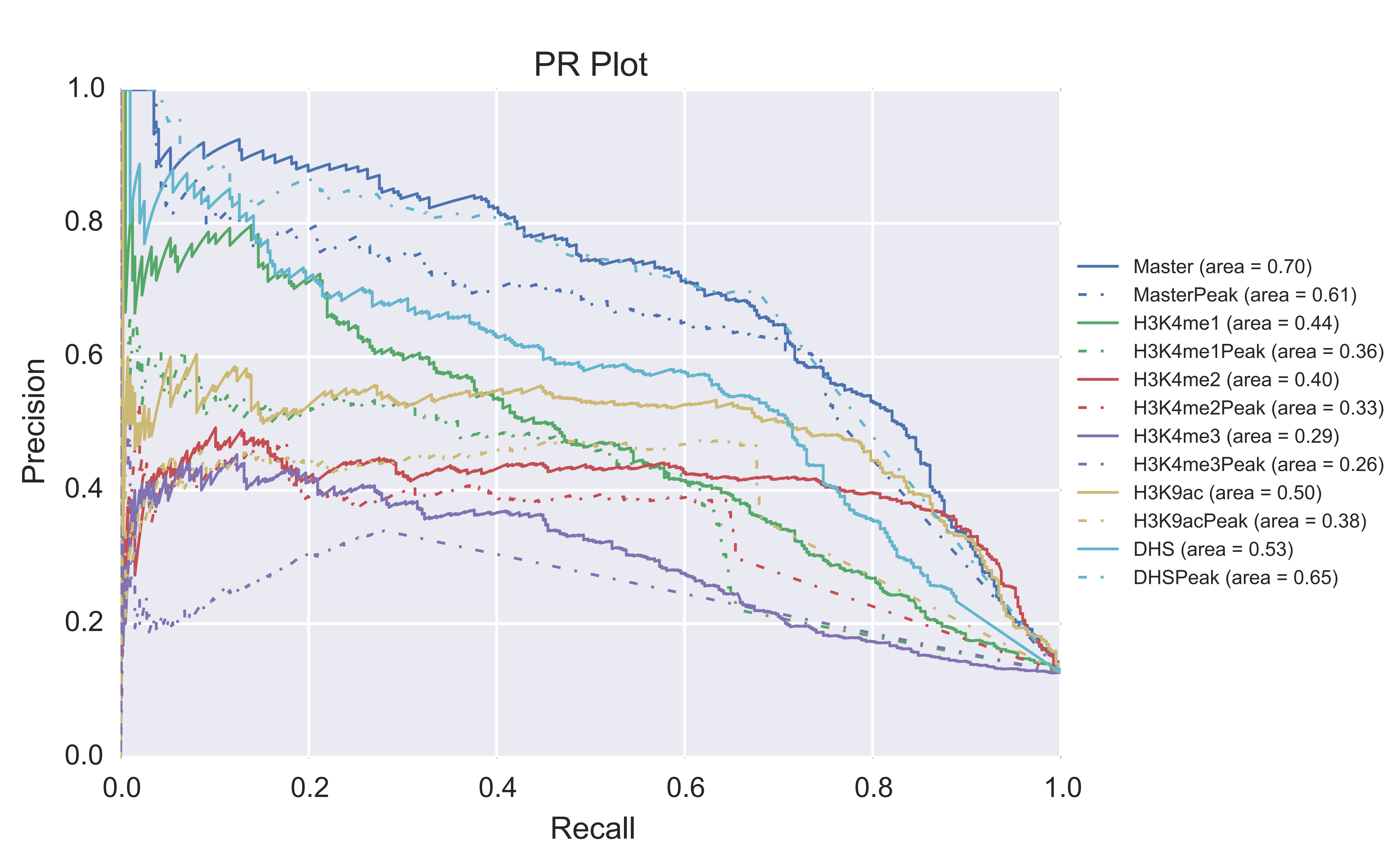
Comparison of performance of matched filter and peaks



peak order: H3K27ac > DHS > H3K9ac > H3K4me2 > H3K4me3 > H3K4me1

Metaprofiles work better for histone marks than for identifying regulatory elements from DHS signal.

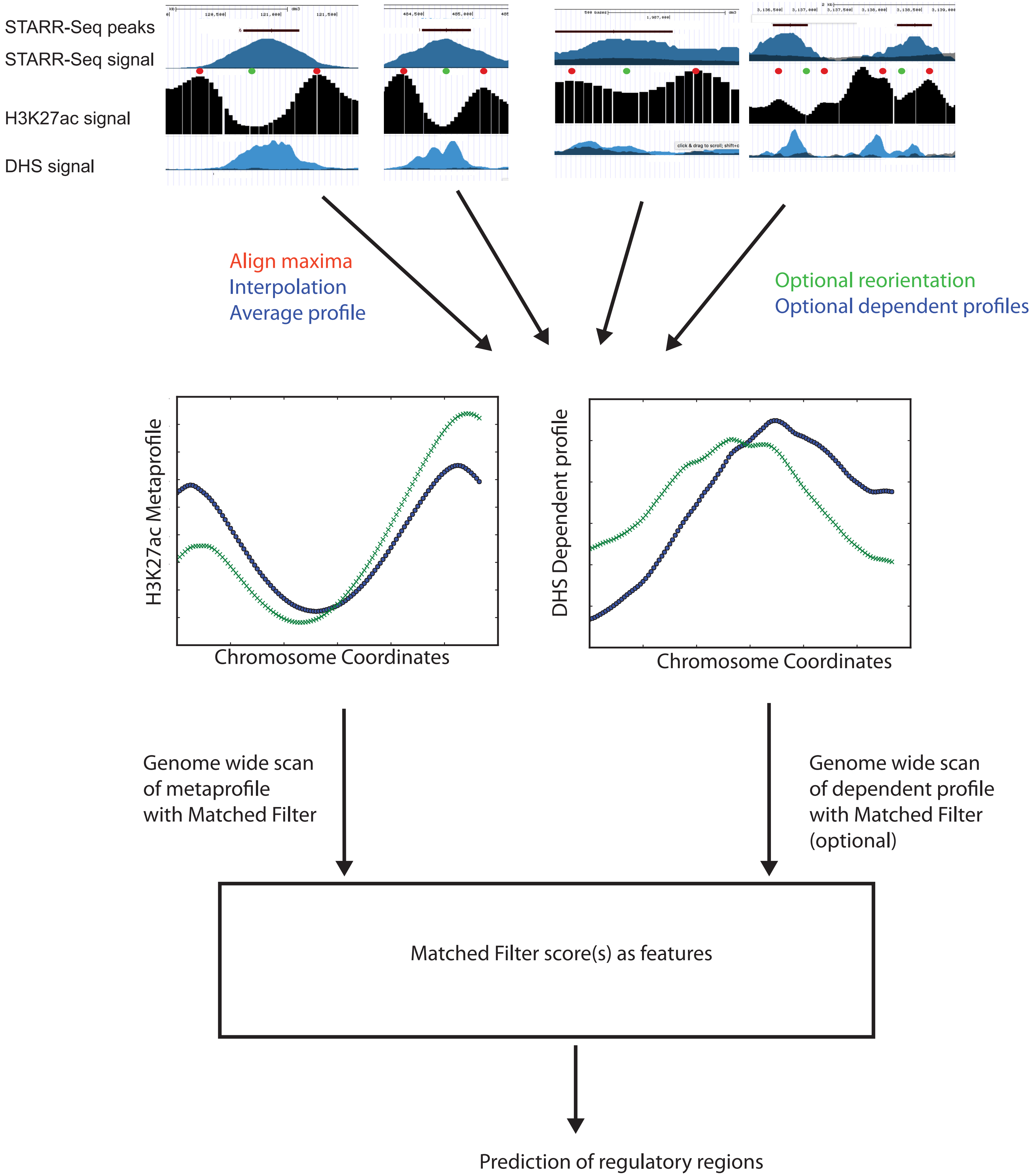
Comparison of performance of matched filter and peaks



peak order: DHS > H3K27ac > H3K9ac > H3K4me1 > H3K4me2 > H3K4me3

Metaprofiles work better for histone marks than for identifying regulatory elements from DHS signal.

Massively parallel assay for regulatory activity



Align maxima
Interpolation
Average profile

Optional reorientation
Optional dependent profiles

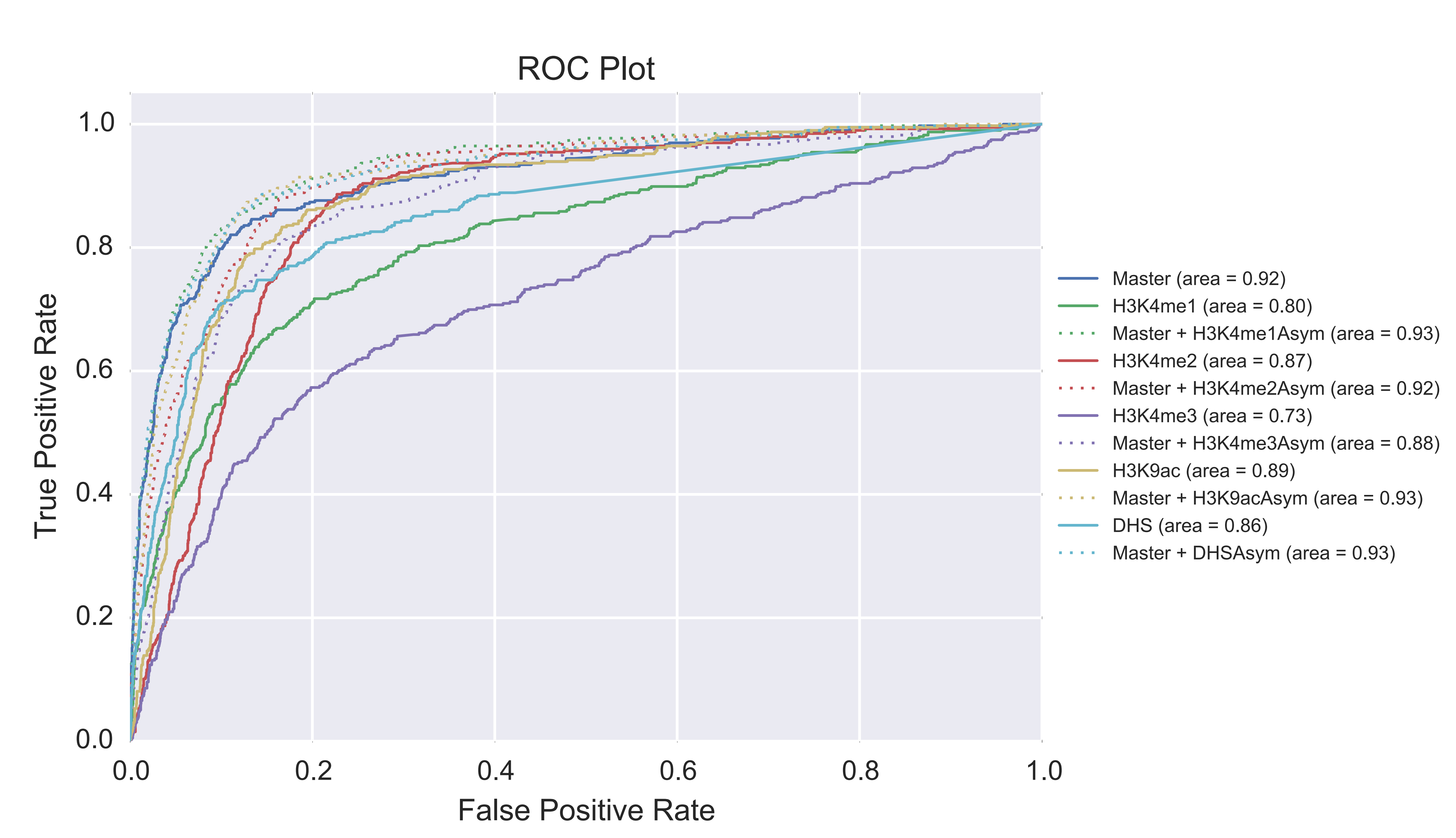
Genome wide scan
of metaprofile
with Matched Filter

Genome wide scan
of dependent profile
with Matched Filter
(optional)

Matched Filter score(s) as features

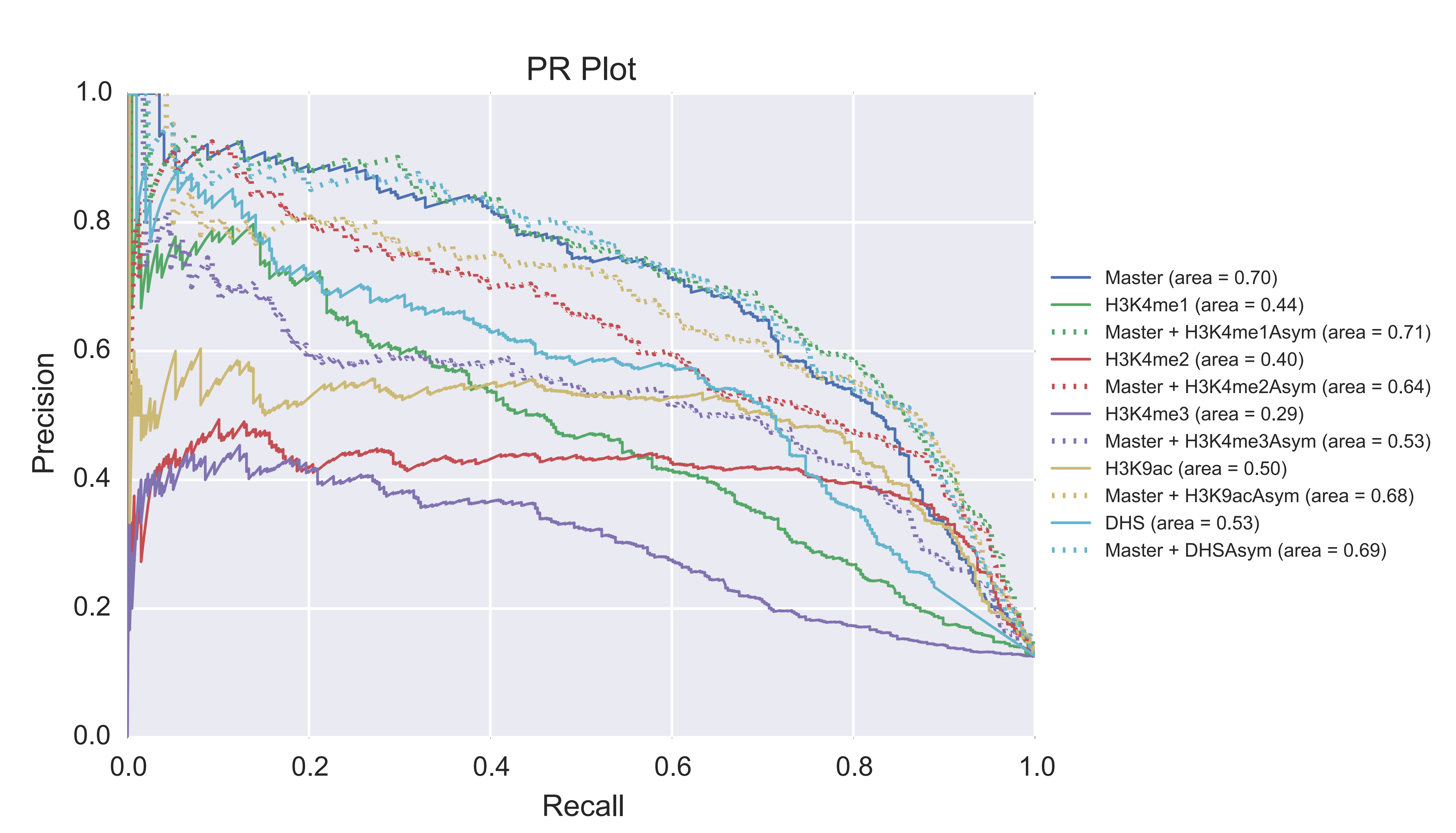
Prediction of regulatory regions

Comparison of performance of linear regression models with Matched filter scores



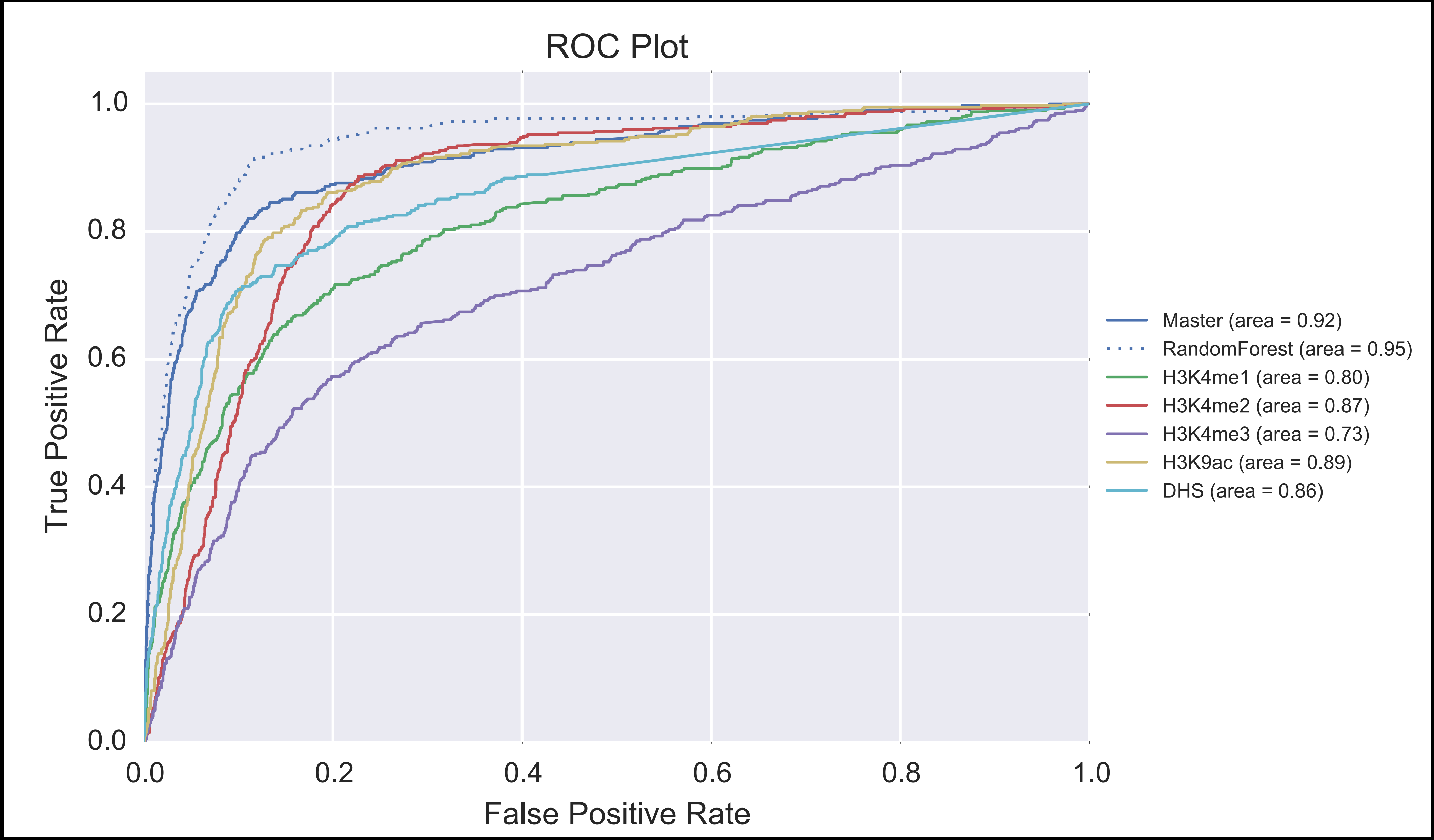
Too close to call in ROC curves (0.92 vs 0.93 but H3K27ac combines with a number of marks to give similar accuracy on ROC).
ROC curve doesn't vary much between asymmetric and symmetric version of matched filter.

Comparison of performance of linear regression models with Matched filter scores



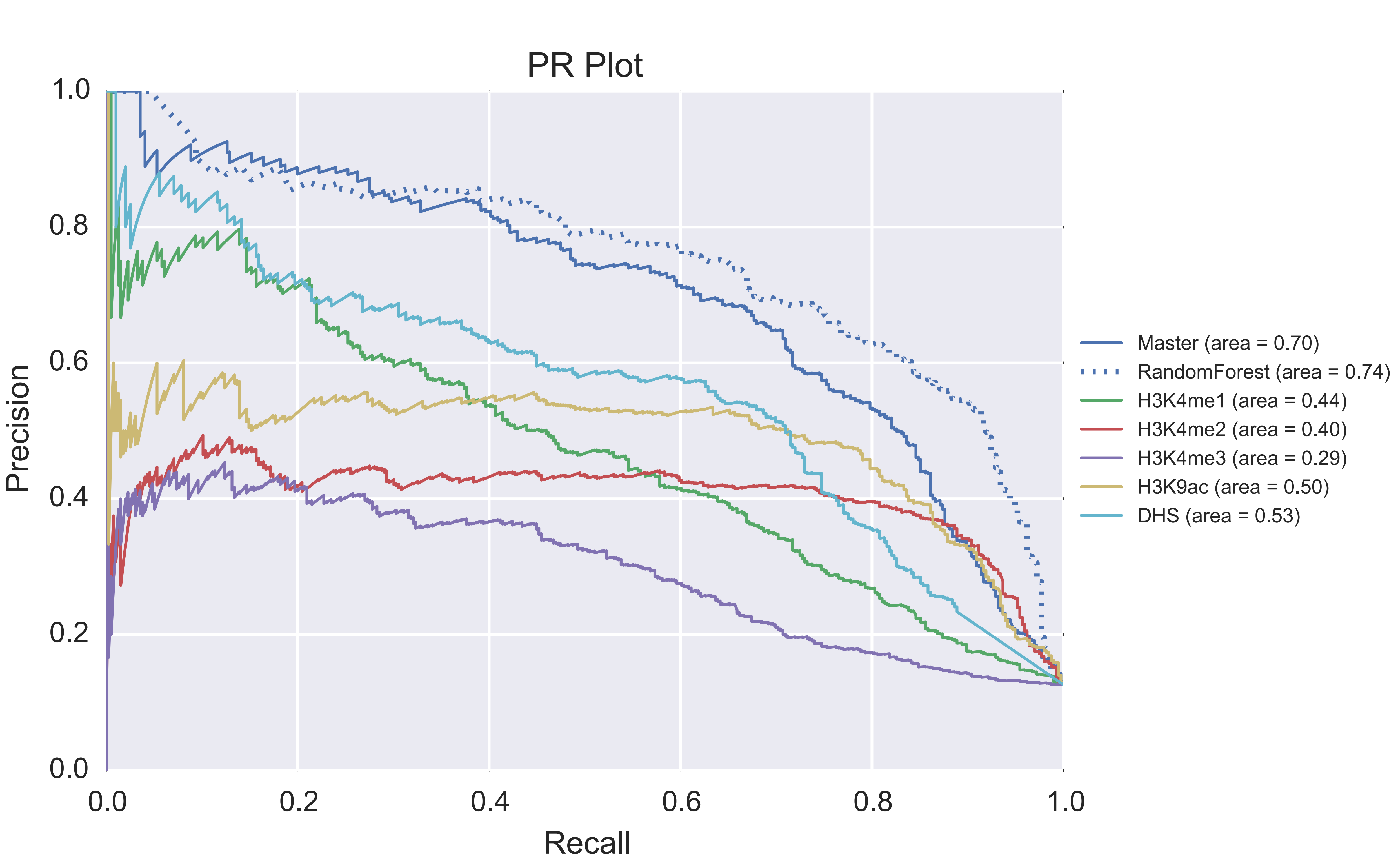
Too close to call on ROC curves (0.69-0.71 range but H3K27ac combines with H3K4me1 or DHS to give similar accuracy on PR).

Combining all marks using random forest does improve the accuracy



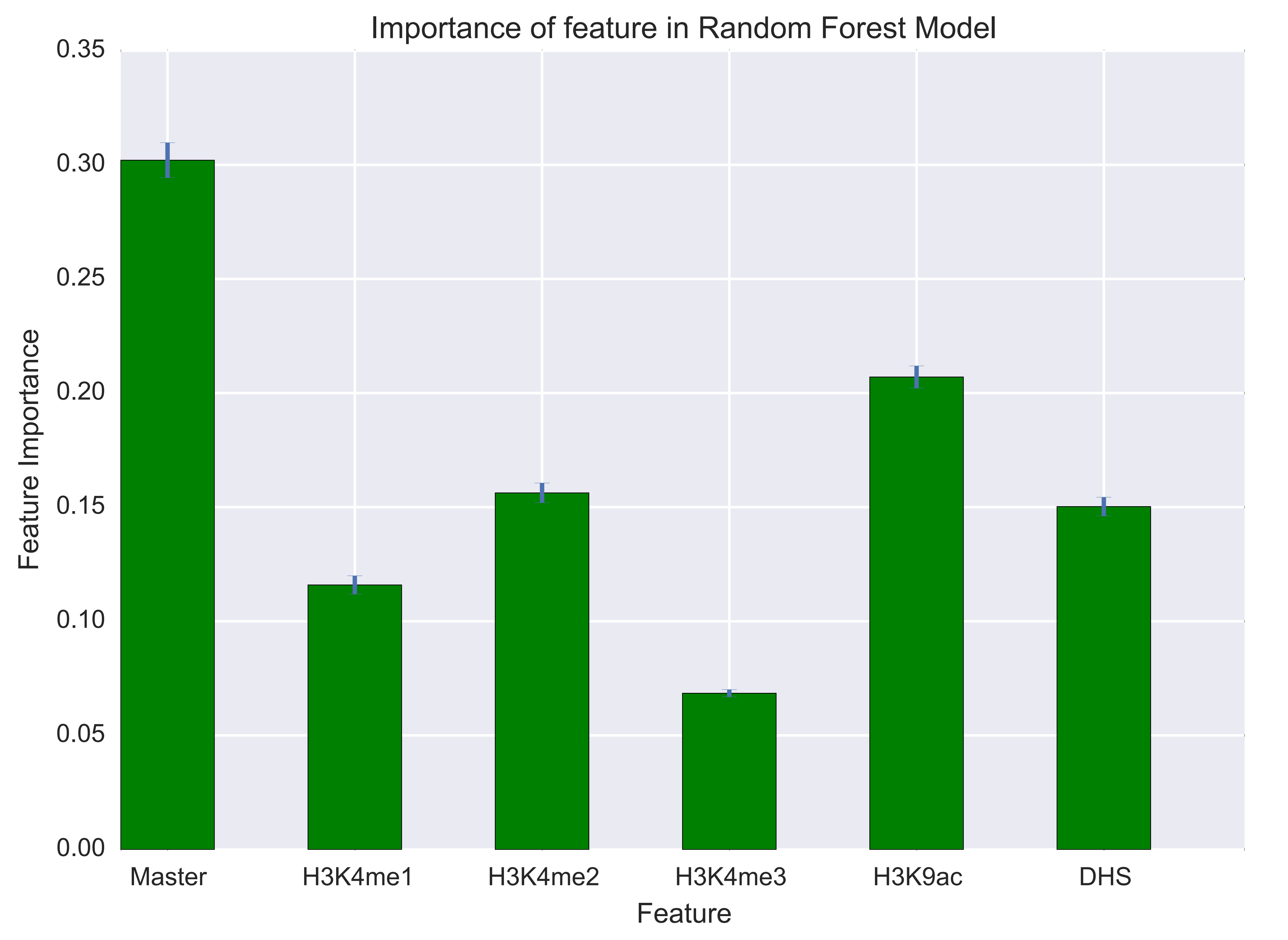
Random forest performs the best.

Combining all marks using random forest does improve the accuracy

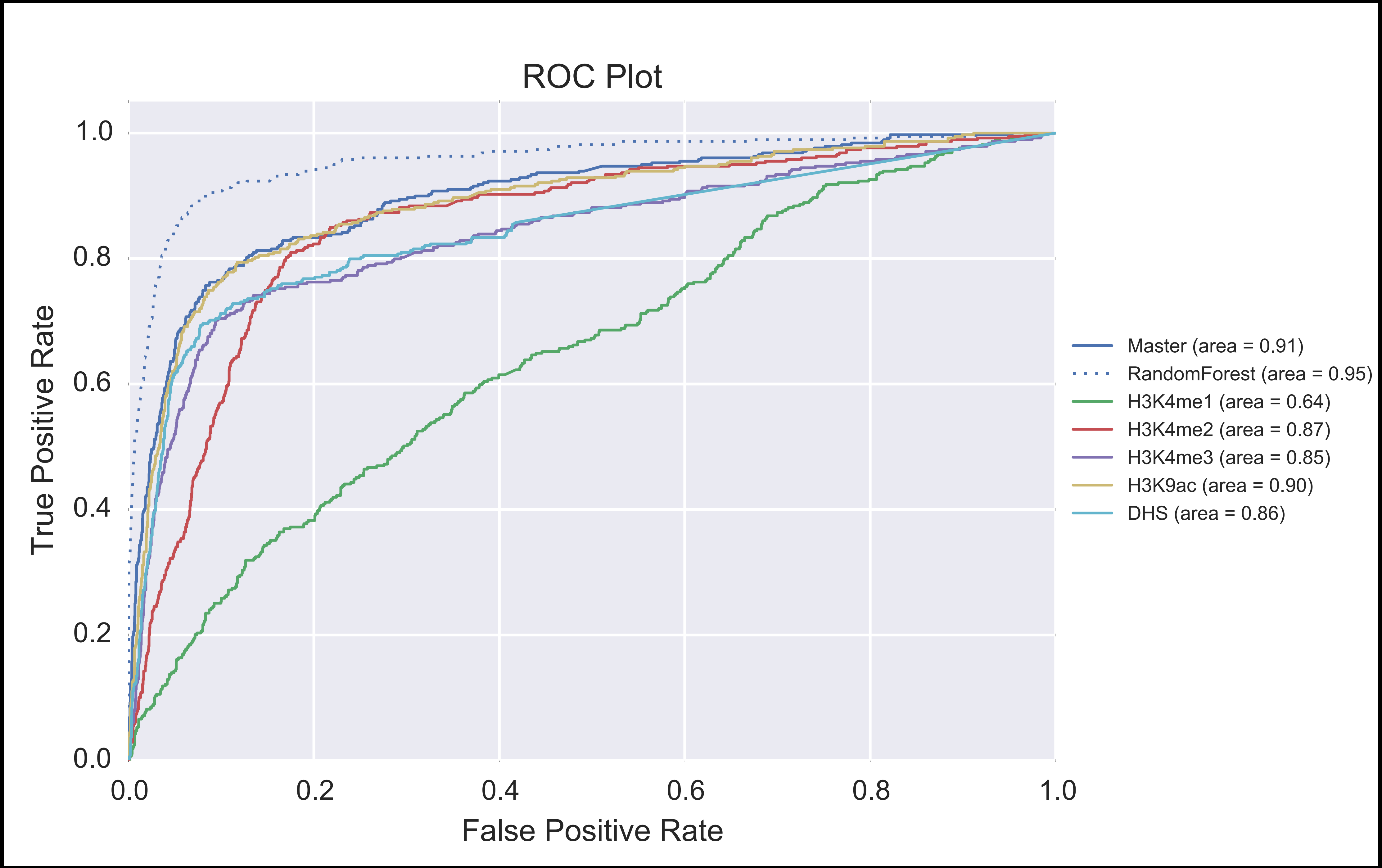


Most of the improvement comes around recall > 0.4 which indicates that additional information in other marks are more useful at moderate to lower strength H3K27ac matched filter regions.

Importance of features indicates acetylations are best indicators of regulatory regions followed by DHS and H3K4me2

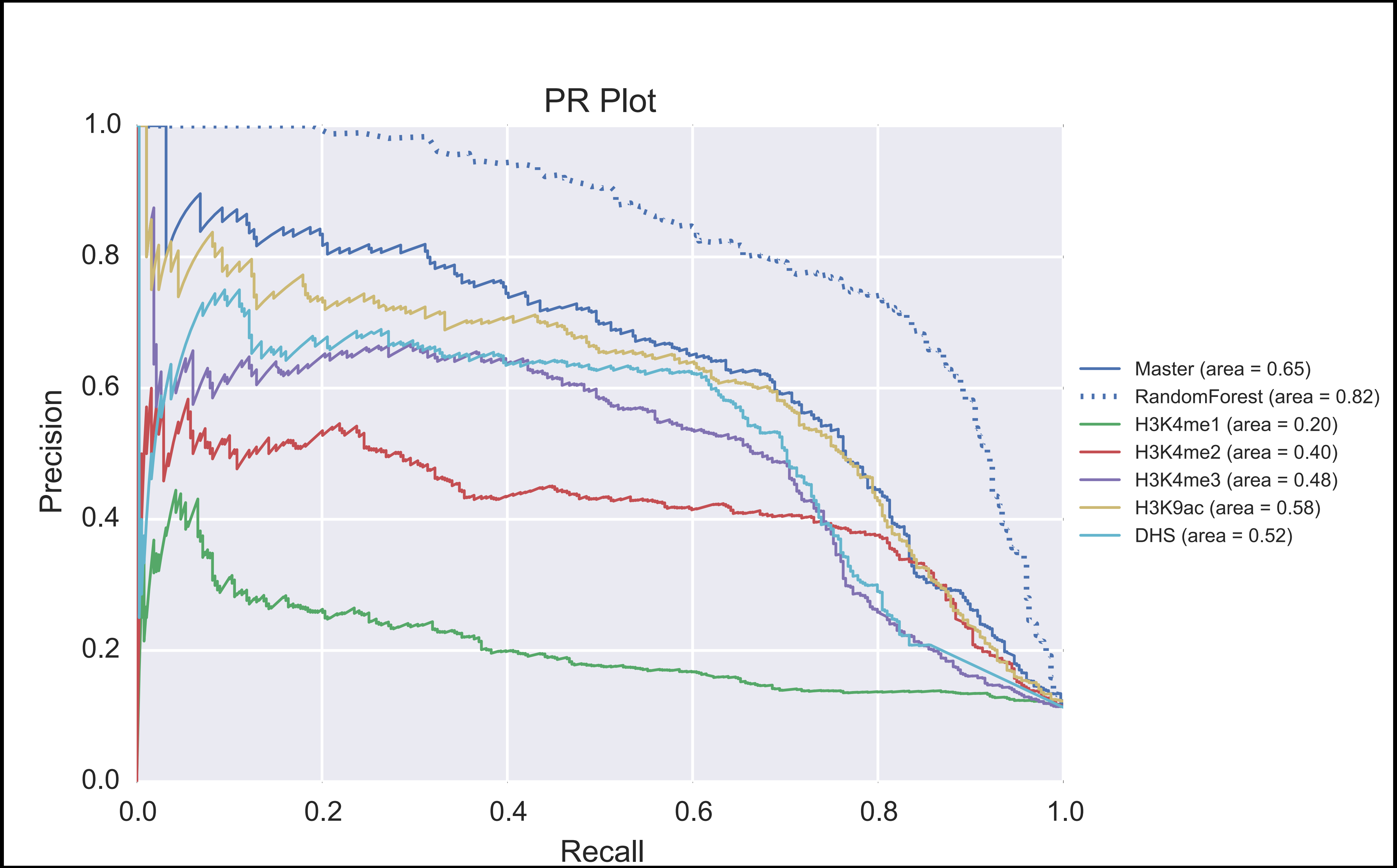


Distal versus proximal STARR-Seq peaks (proximal)



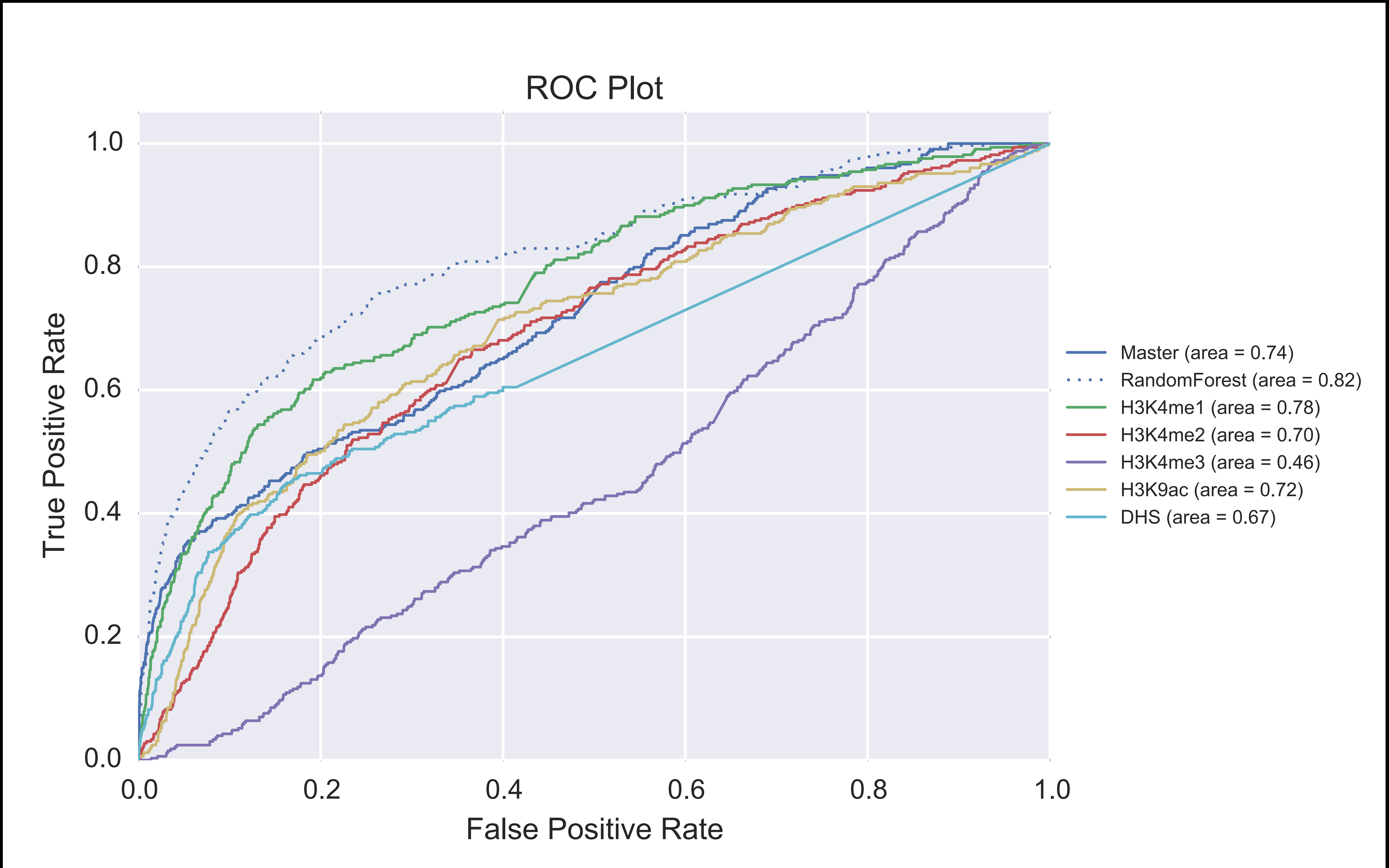
Performance of different marks is similar to previous results

Distal versus proximal STARR-Seq peaks (proximal)



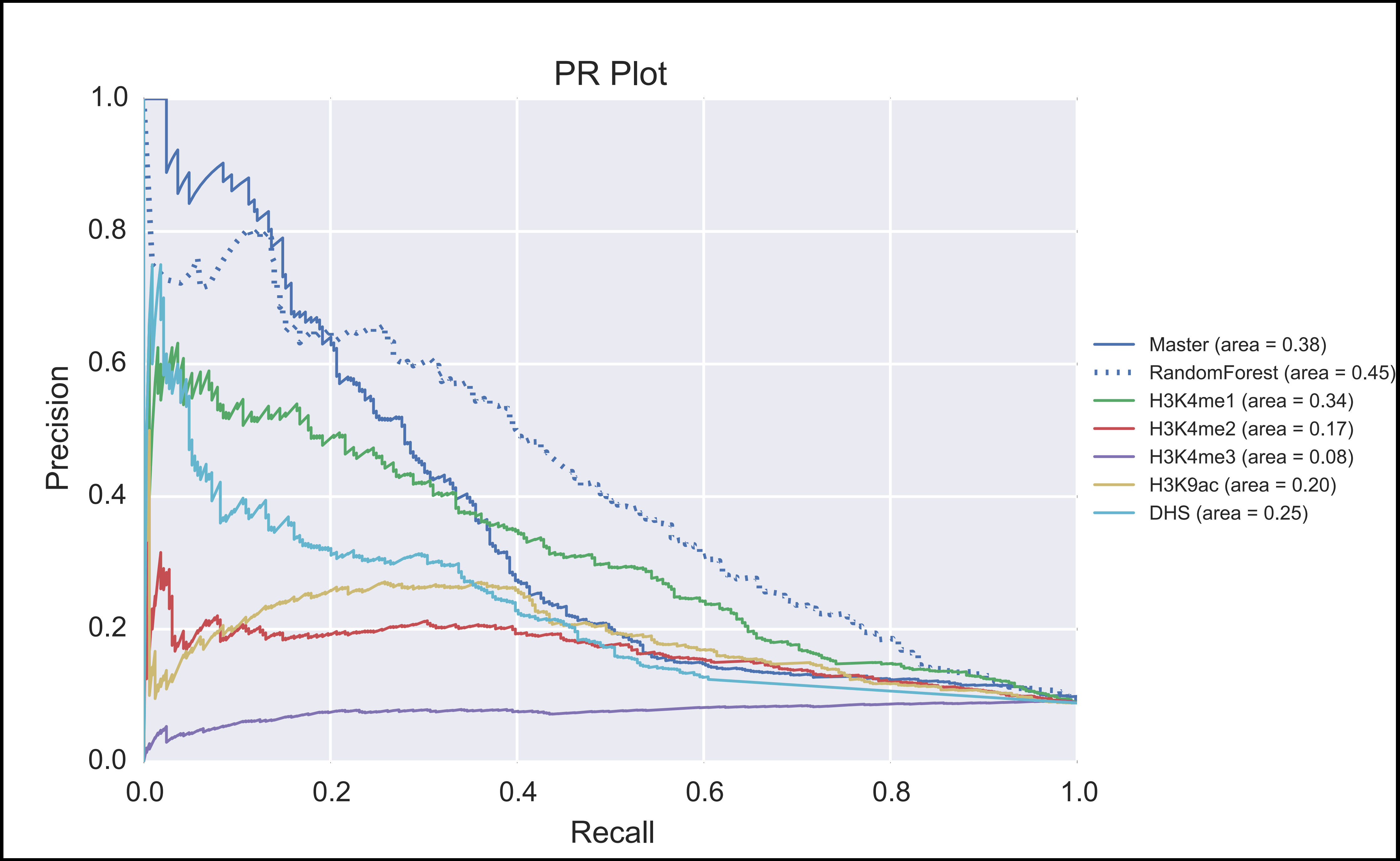
Reduction in precision of different marks but Random forest performs well.

Distal versus proximal STARR-Seq peaks (distal)



Reduction in accuracy of different marks for distal predictions (results closer in AUROC/AUPR to the results from VISTA)

Distal versus proximal STARR-Seq peaks (distal)

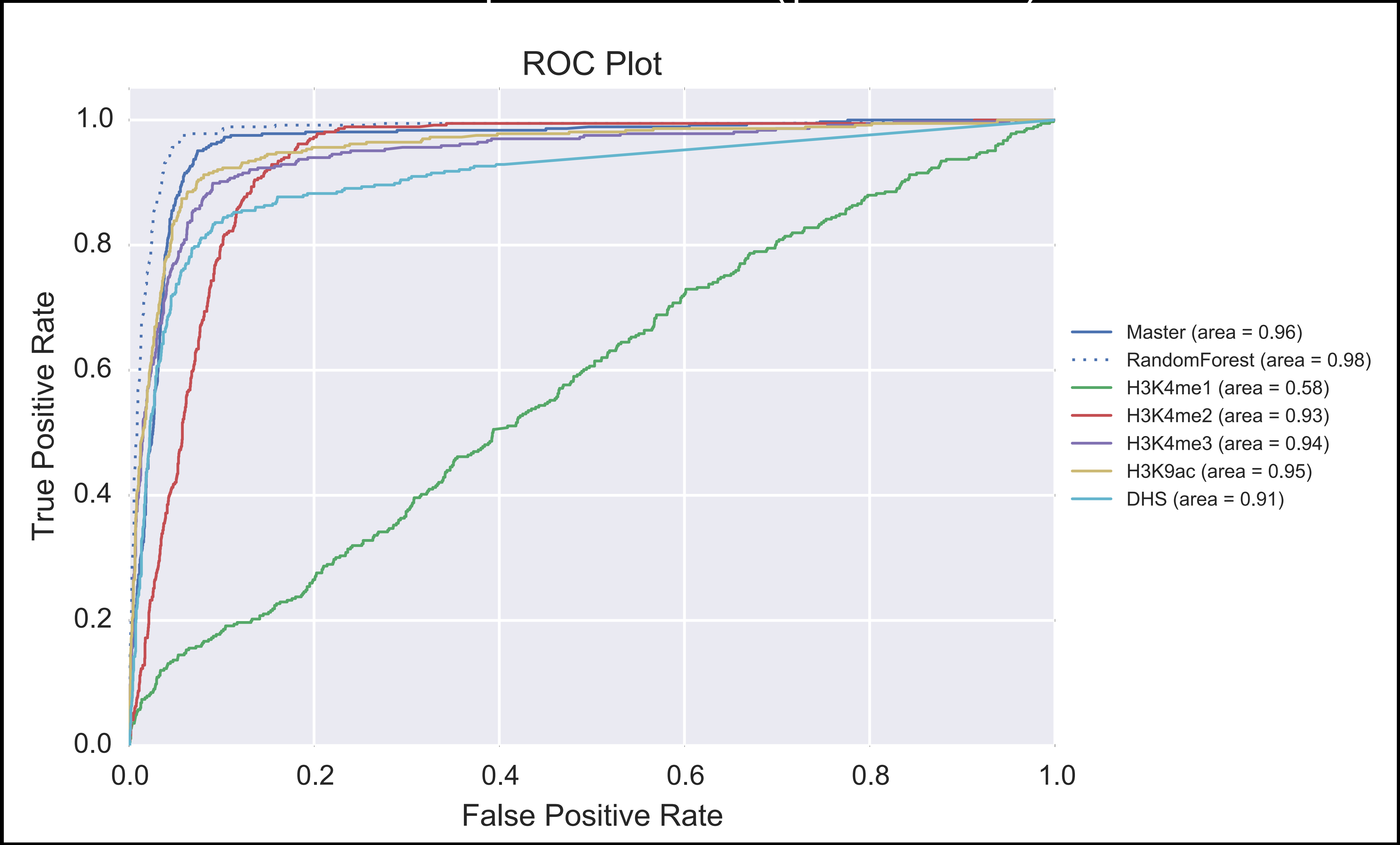


Reduction in accuracy of different marks for distal predictions (results closer in AUROC/AUPR to the results from VISTA)

But the enhancers in STARR-Seq are promoter-specific. Our predictions are not promoter-specific. How do we know this is an issue?

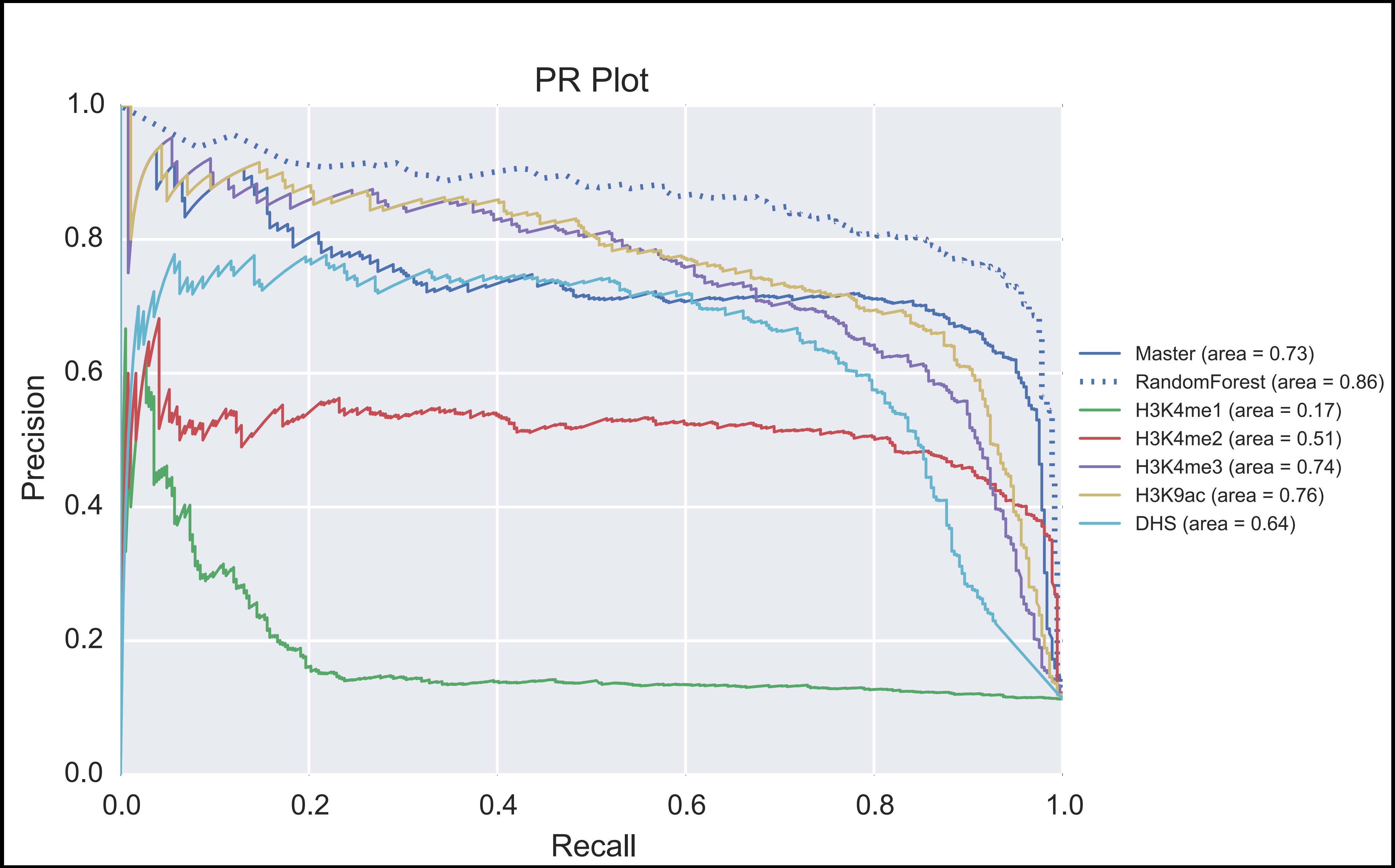
Combine STARR-seq peaks from different experiments and compare results

Distal versus proximal STARR-Seq peaks from multiple experiments (proximal)



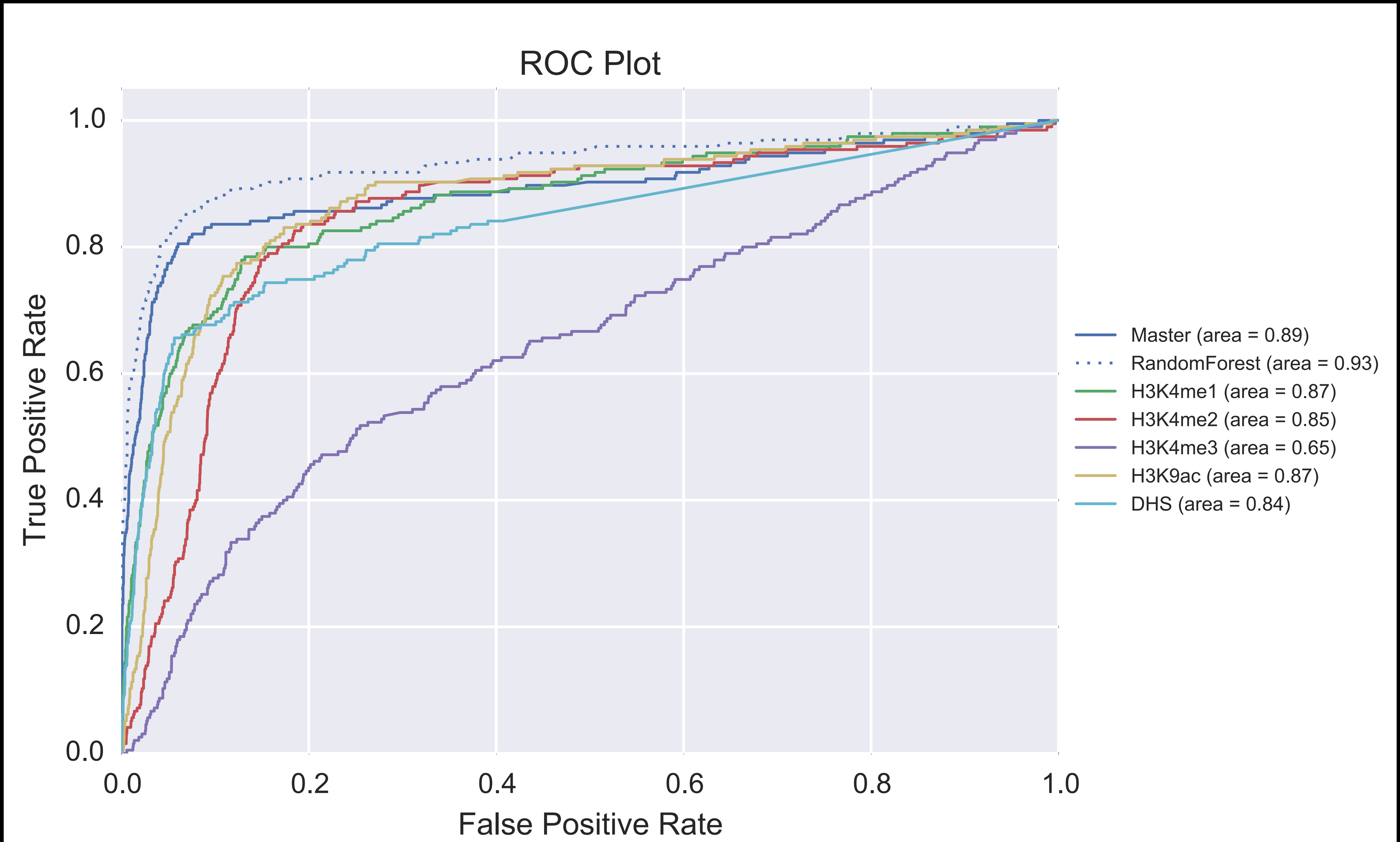
There is slight improvement in accuracy with matched filter predictions when considering the union of STARR-seq experiments.

Distal versus proximal STARR-Seq peaks from multiple experiments (proximal)



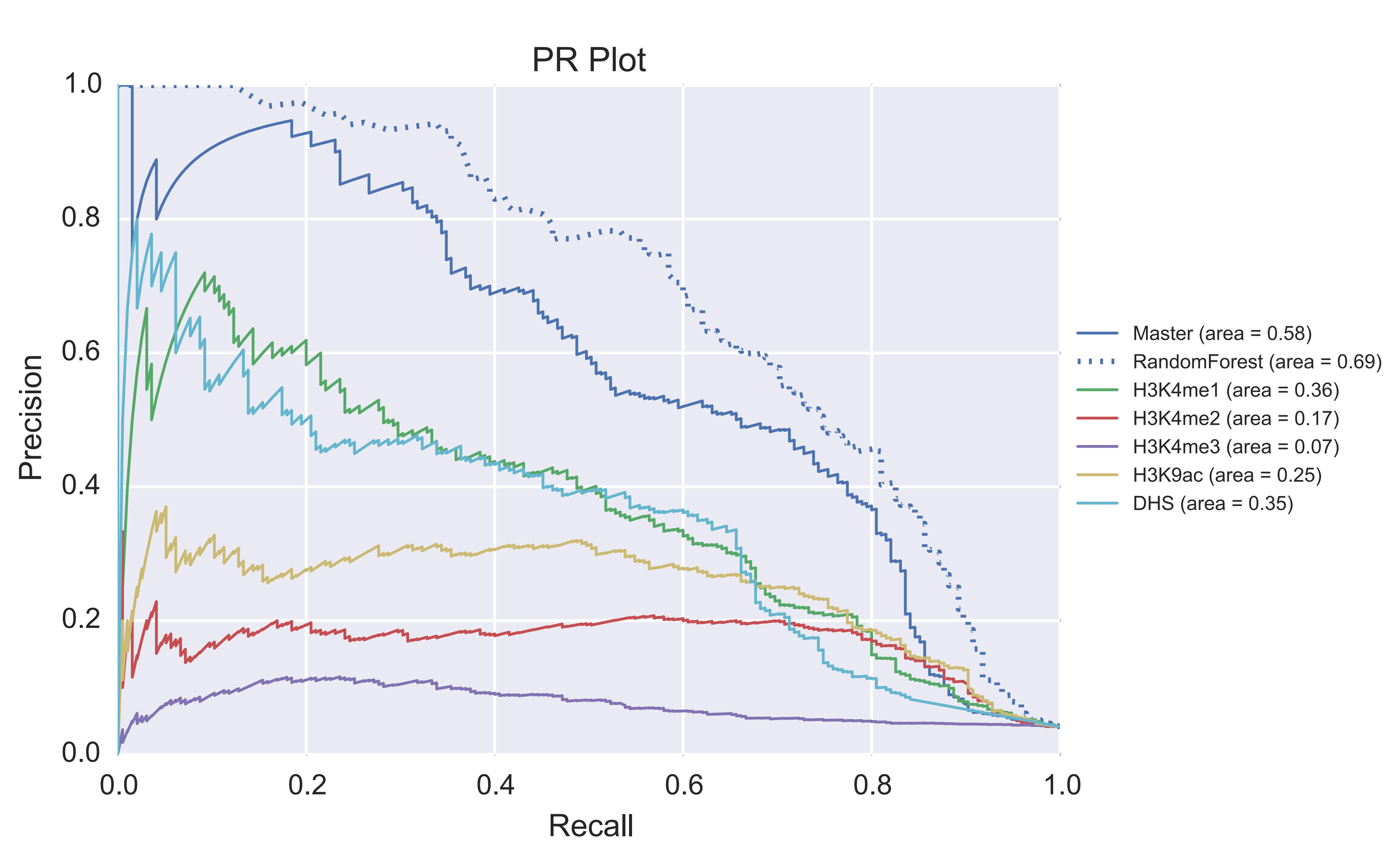
There is slight improvement in accuracy with matched filter predictions when considering the union of STARR-seq experiments.

Distal versus proximal STARR-Seq peaks from multiple experiments (distal)



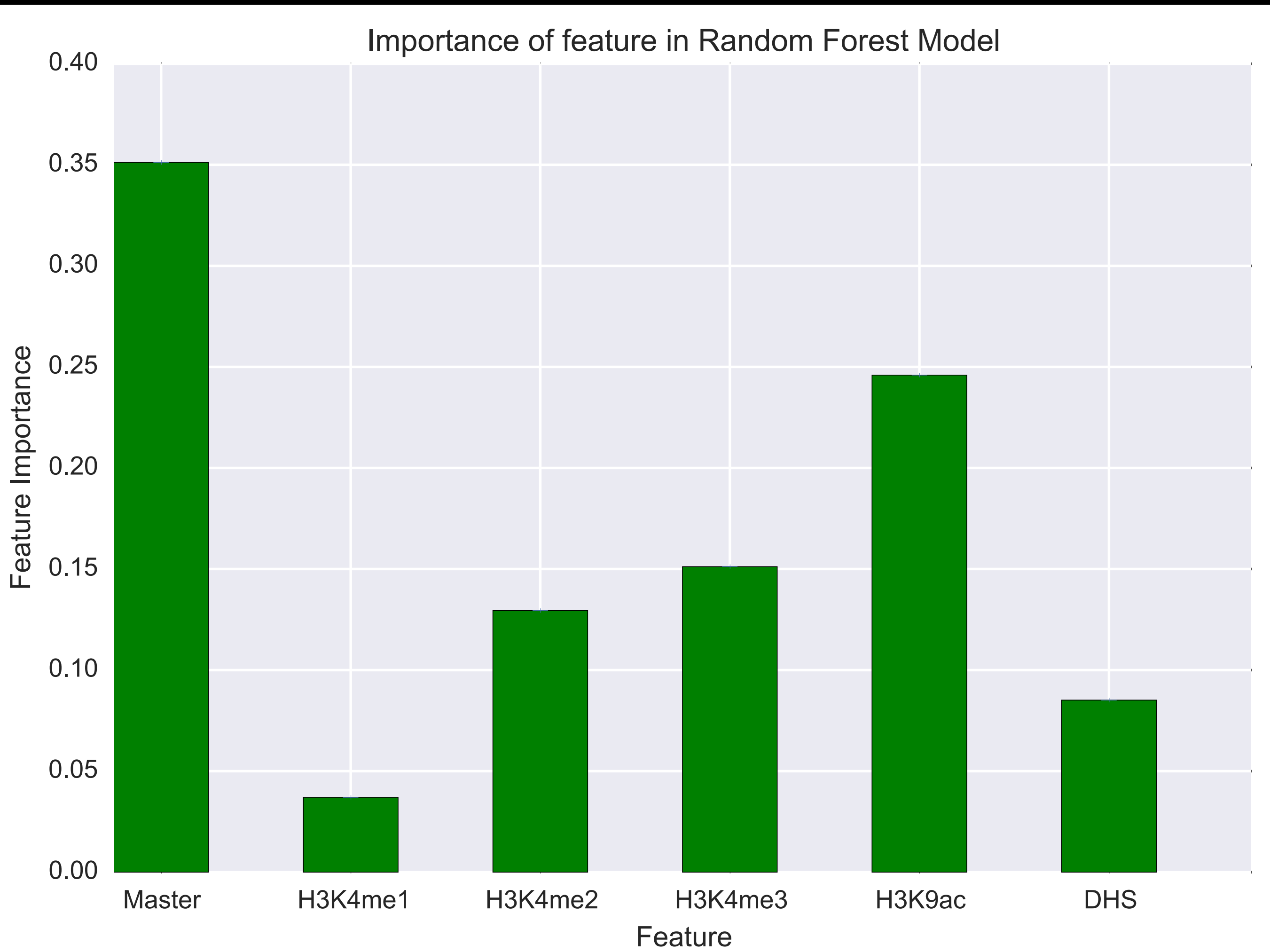
Improvement in predictions for distal enhancers when considering the union of multiple experiments

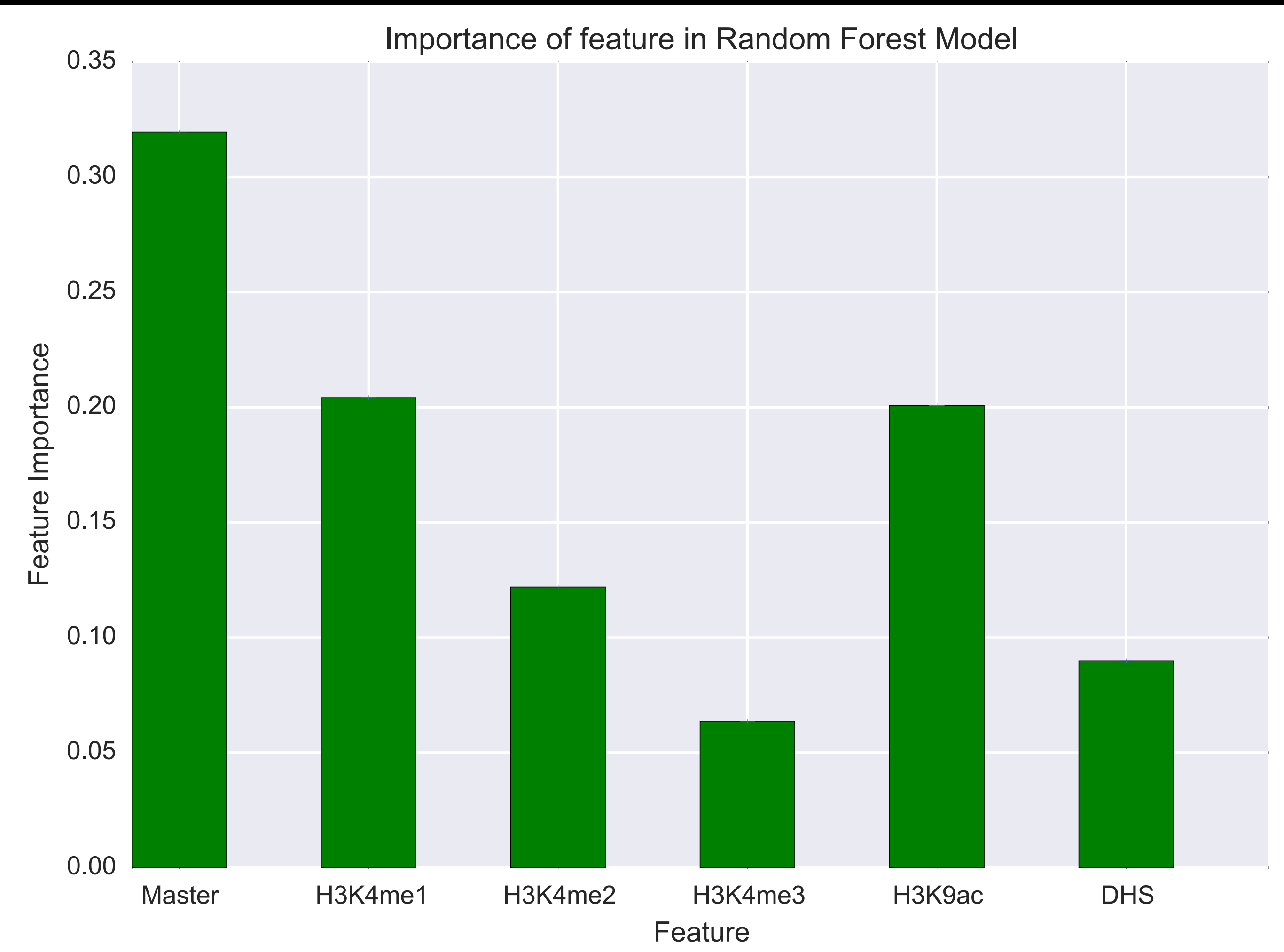
Distal versus proximal STARR-Seq peaks from multiple experiments (distal)



Improvement in predictions for distal enhancers when considering the union of multiple experiments

Comparison of important features





H3K27ac/H3K9ac matched filters contain most independent information. H3K4me1 is an important mark for enhancers while H3K4me3 is the next mark for calling promoters.

Questions to consider

Will the Random forest model work across cell-lines/tissues/species?

H3K27ac matched filter predictions could also indicate CTCF binding sites (insulators) and nuclear pore complex binding sites (localization of super enhancers near nuclear membrane) - maybe look for these motifs and term them as different categories during enhancer prediction could help.

Including information about known motifs will improve the accuracy of these models.