Enhancers for ENCODE Encyclopedia - Part 4

Anurag Sethi

Functional Characterization Call

Enhancers for Encyclopedia

Histone marks and/or DHS

Training data + histone marks and/or DHS

Large # of functional genomics assays

Unsupervised predictions (developing a way to combine multiple experimental datasets)

Ensemble predictions (testing)

Will use MP enhancer assays to judge which method(s) to use in this case

Part 1
Enhancers for Encyclopedia

Enhancer Predictions using epigenetic datasets (histone and/or DHS) in the absence of training data

All assessments will be performed on the VISTA database* (labeled data).

Groups are welcome to submit predictions that do not use training data and are applicable to a majority of ENCODE cell-lines/tissues.

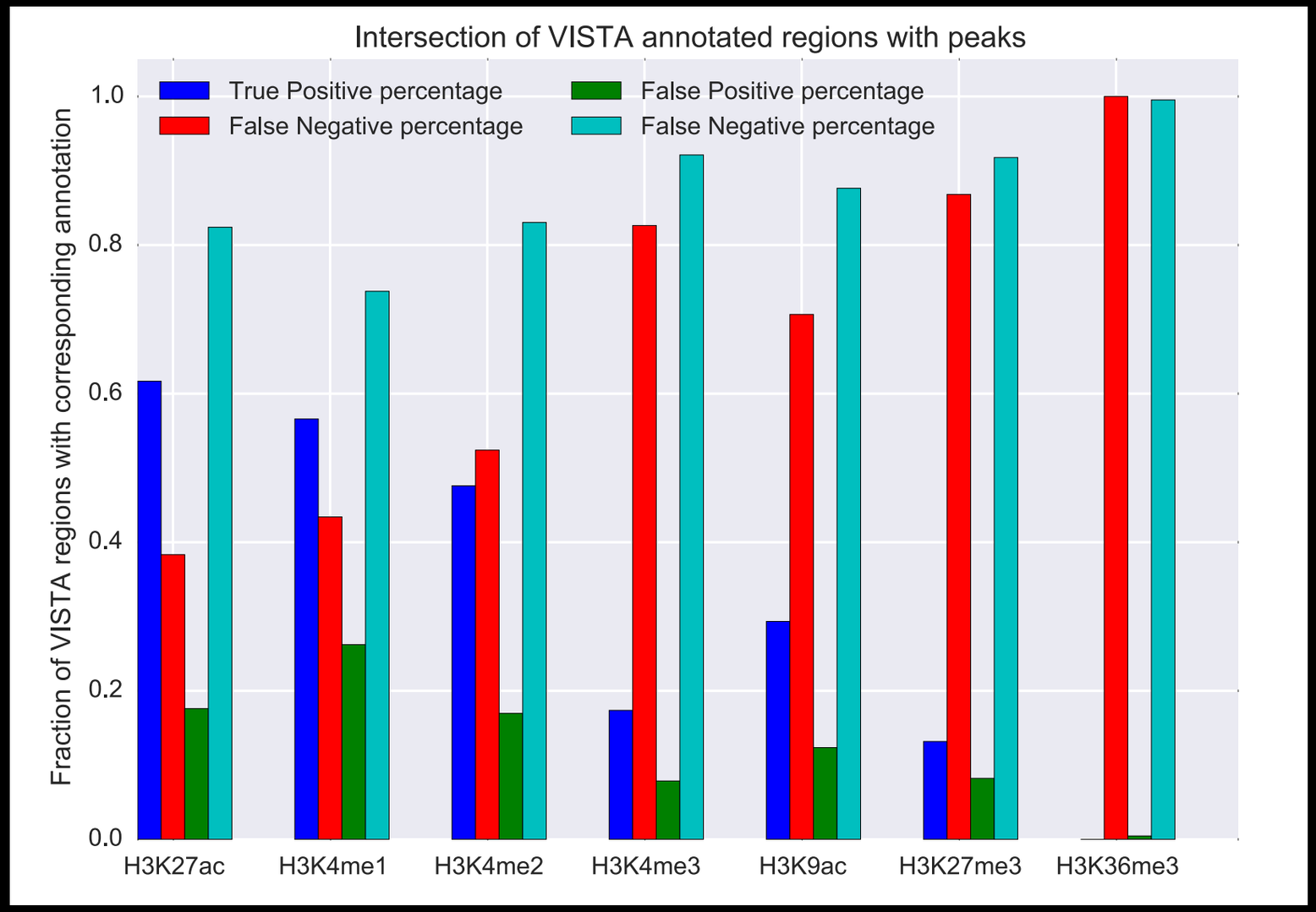
* We should always remember that the VISTA database is small and has multiple sources of bias that are not fully characterized.

Confusion matrix

		prediction outcome		
		p	***n***	**total**
actual value	***p'***	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

In imbalanced datasets, also helps to think of true positive percentage (percentage of all enhancers that are predicted to be enhancers using this method) as well as false positive percentage (percentage of non-enhancers that are predicted to be enhancers by a particular method)

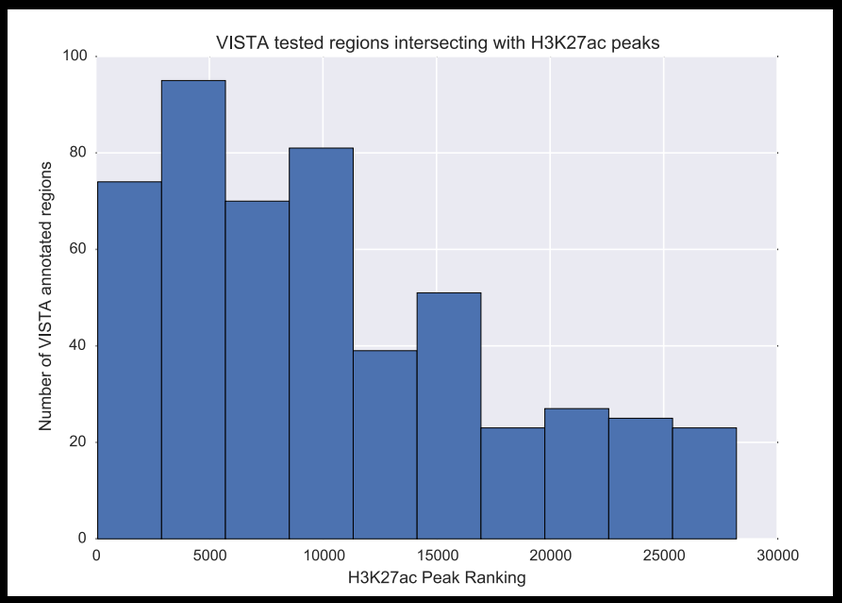
What is the coverage of VISTA database based on peaks within different epigenetic datasets (forebrain)

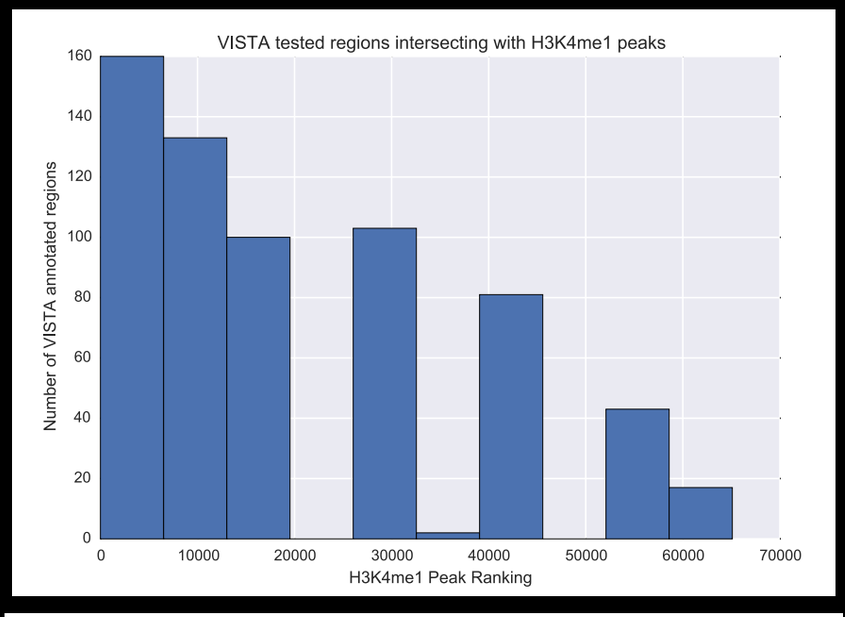


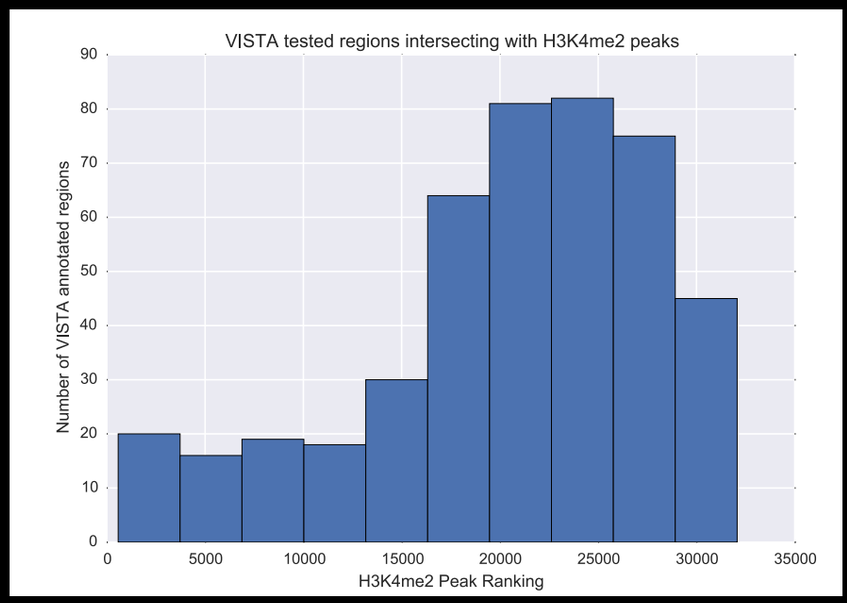
Most positives occur on H3K27ac peaks but it also has a high fraction of false positives.

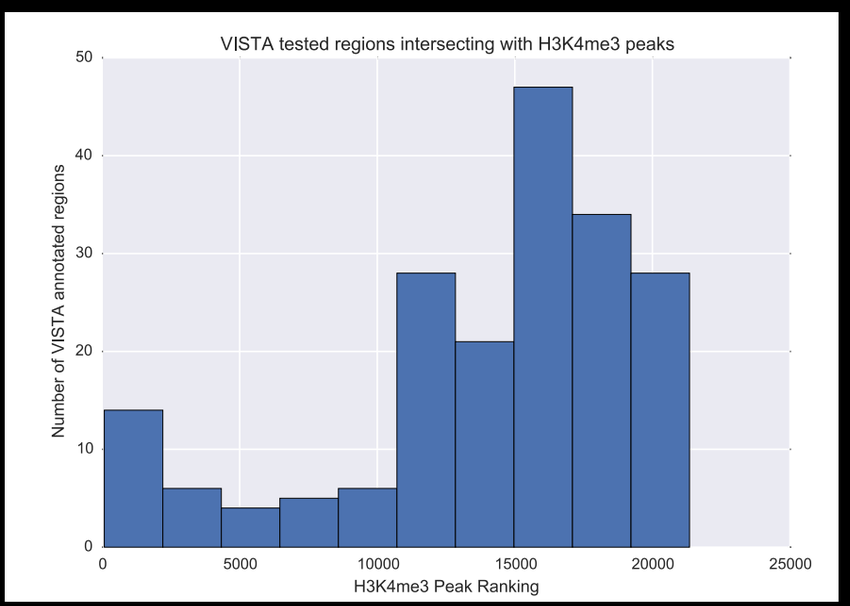
Surprised to see about 15% of enhancers also map to H3K27me3 peaks.

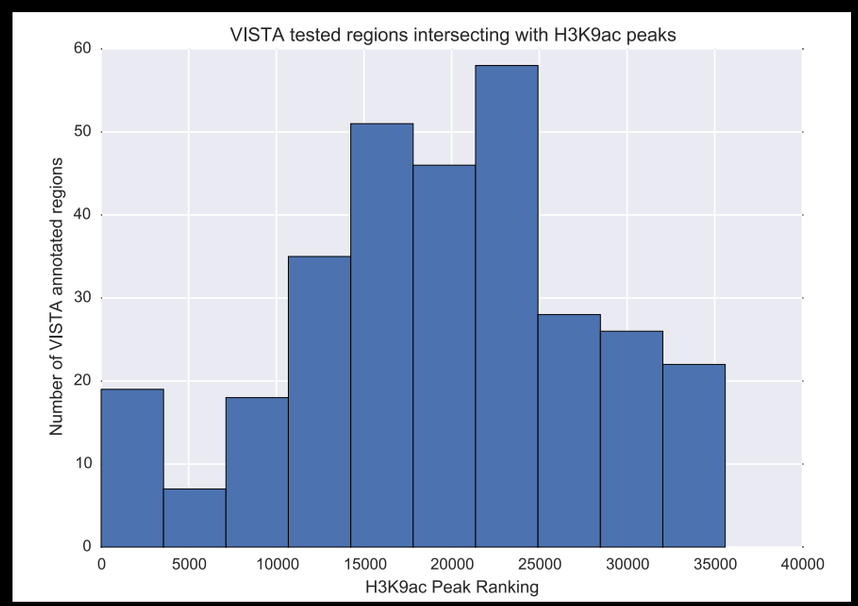
Characterizing biases in terms of peak ranking (forebrain)

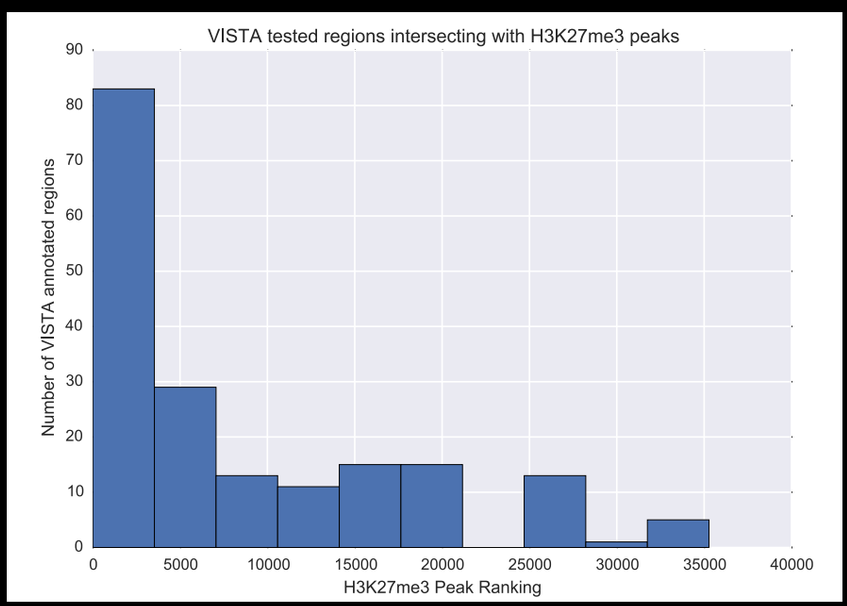










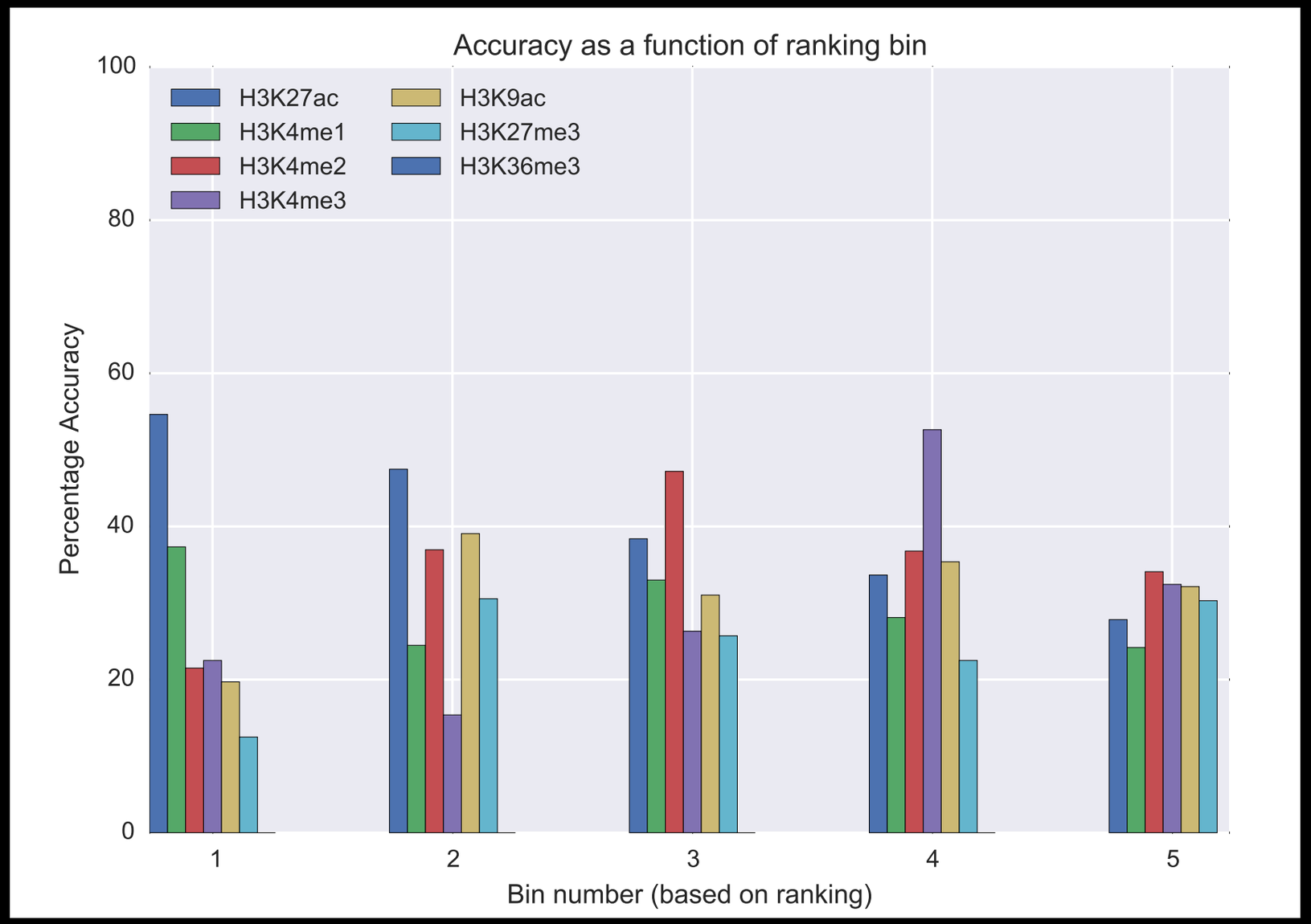


The regions tested are not even over the whole peak list.

We can also characterize bias in terms of multiple co-variates but very few data points.

Accuracy as a function of binned ranking (forebrain)

Alternative approach: Bin based on ranking but such that all intervals have equal number of tested regions



Cautious about estimating accuracy from such plots.

H3K27me3

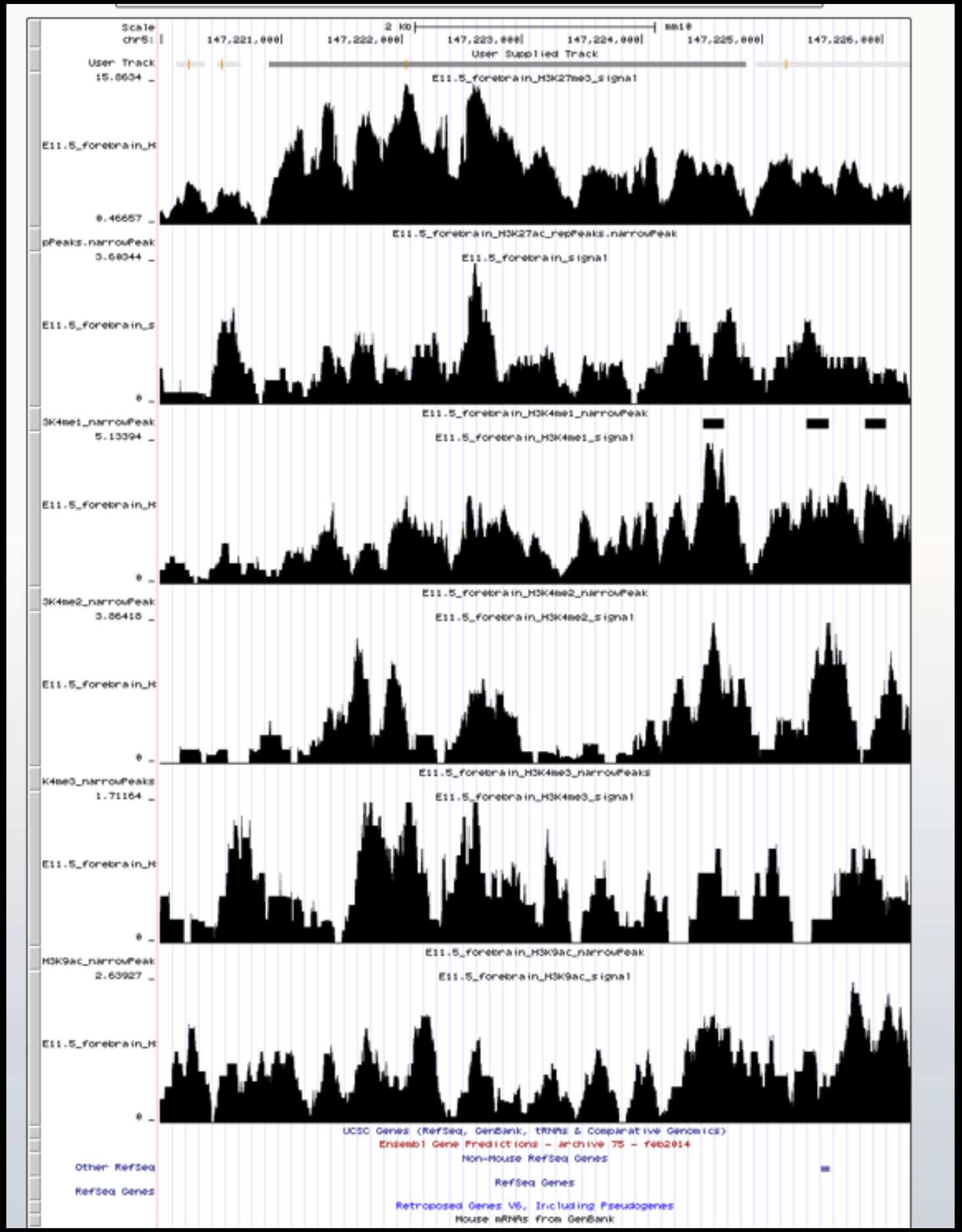
H3K27ac

H3K4me1

H3K4me2

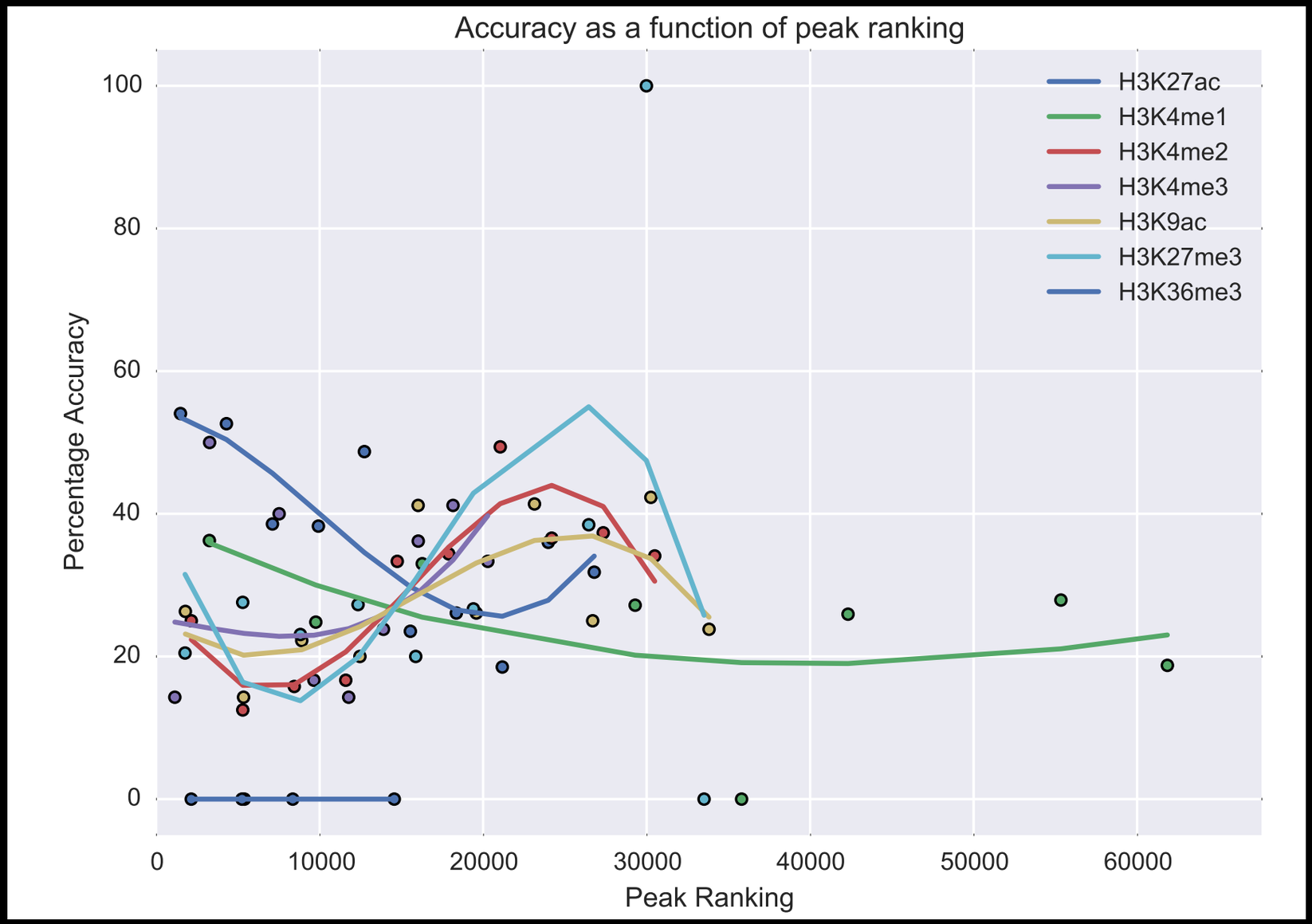
H3K4me3

H3K9ac



Some H3K27me3 peaks can also be enhancers (forebrain) ?

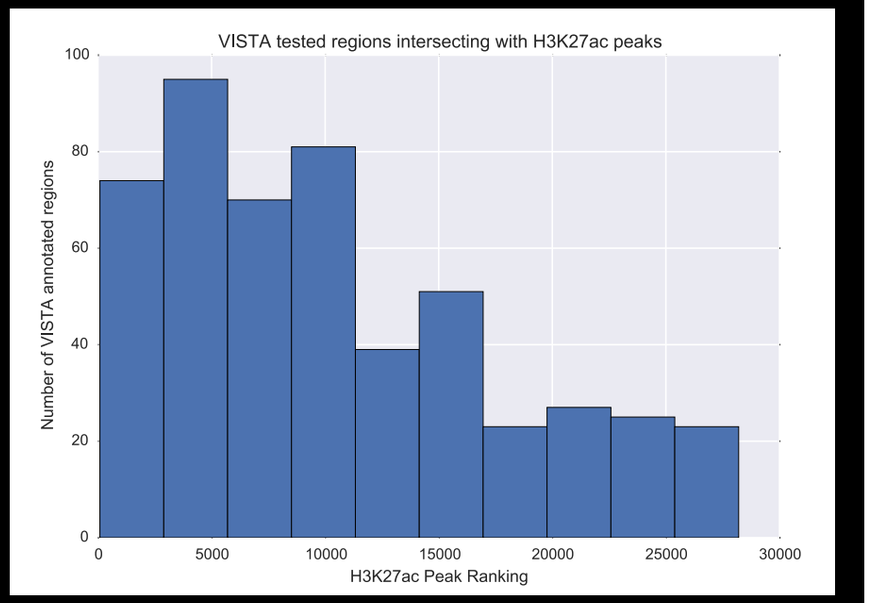
Accuracy as a function of peak ranking (forebrain)

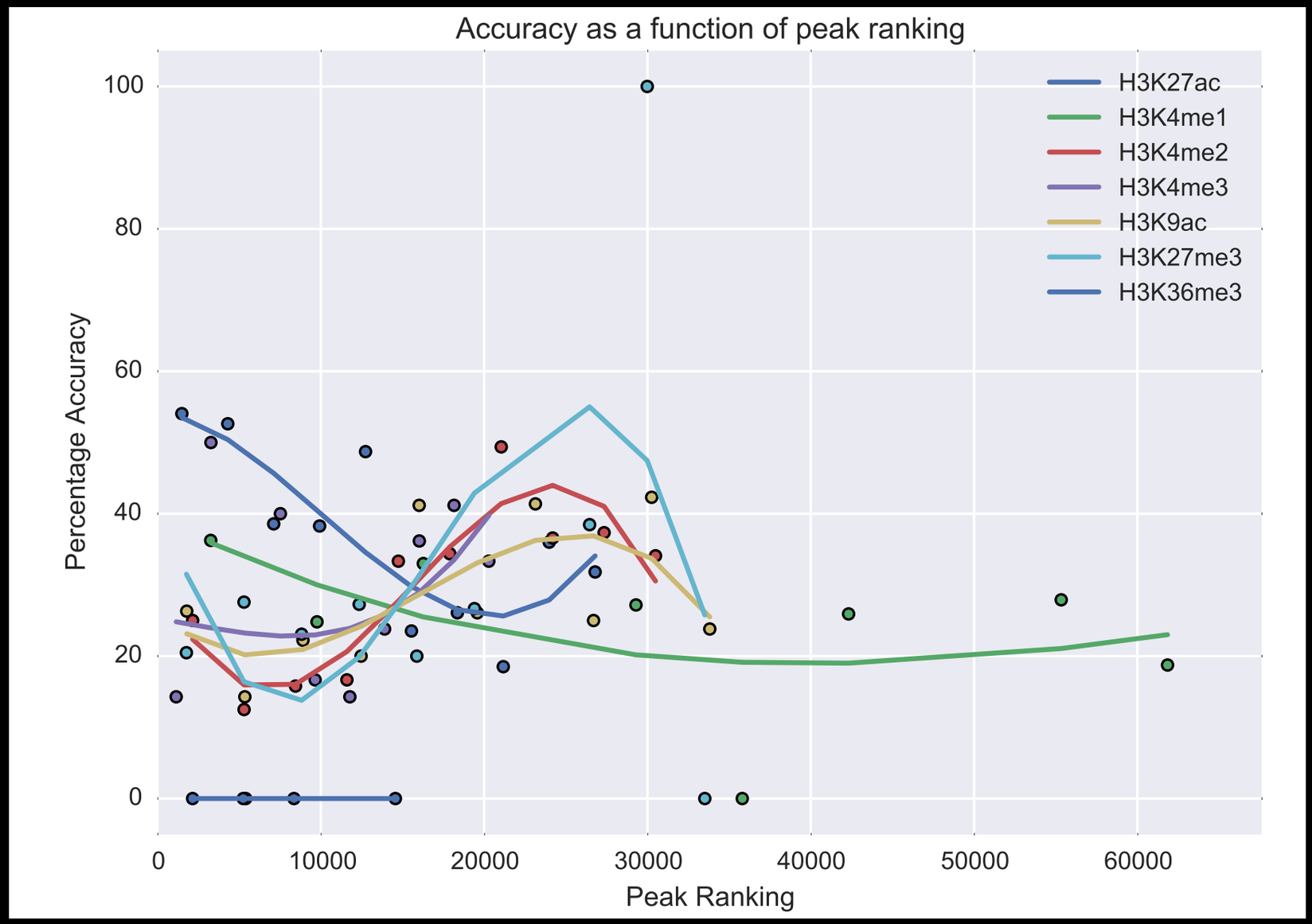


Even within the same bin, there might be additional bias over the regions tested as 20 regions are often tested from close to 5000 peaks.

We should be careful about generalizing accuracy, etc.

Accuracy as a function of peak ranking (forebrain)



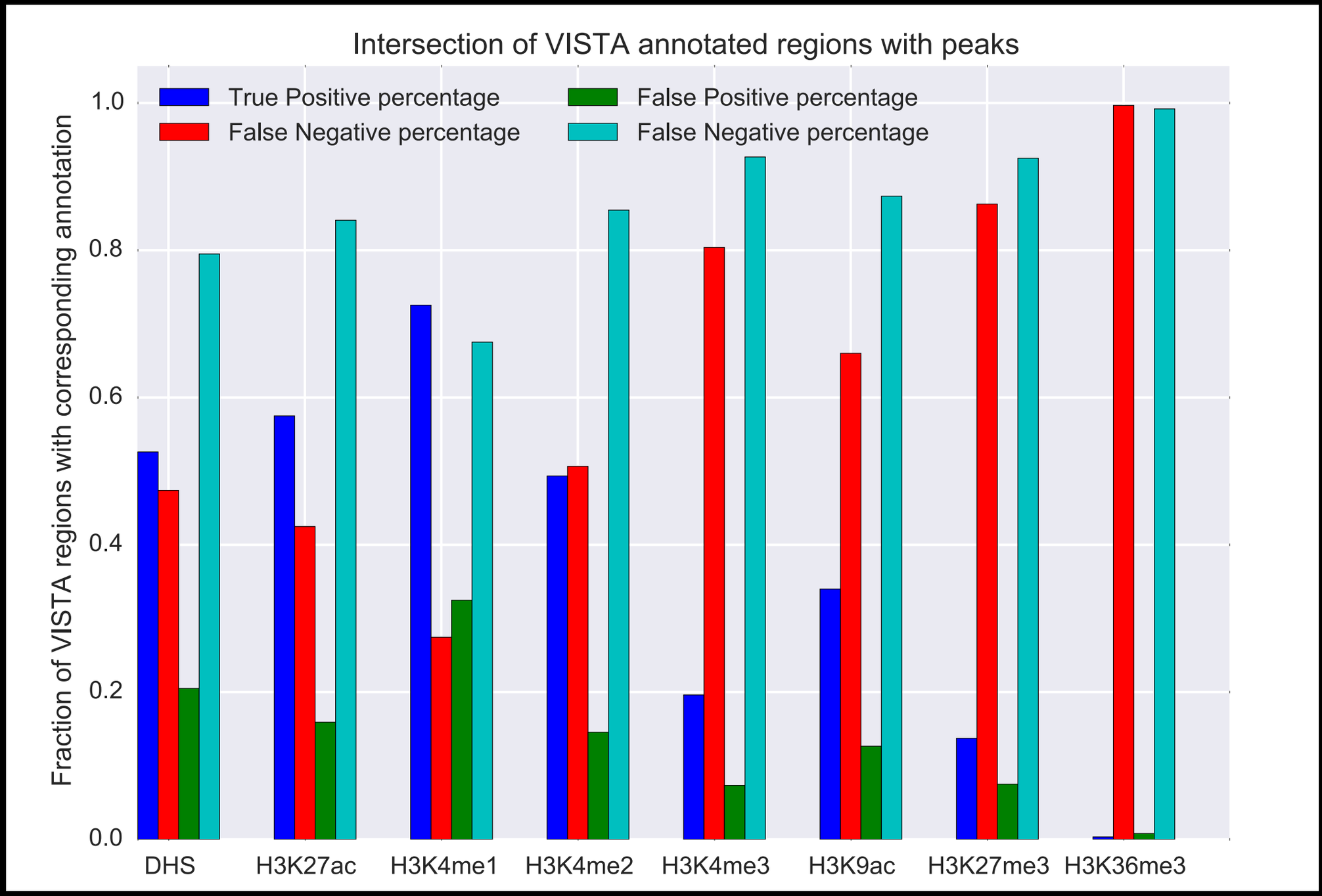


Even within the same bin, there might be additional bias over the regions tested as 20 regions are often tested from close to 5000 peaks.

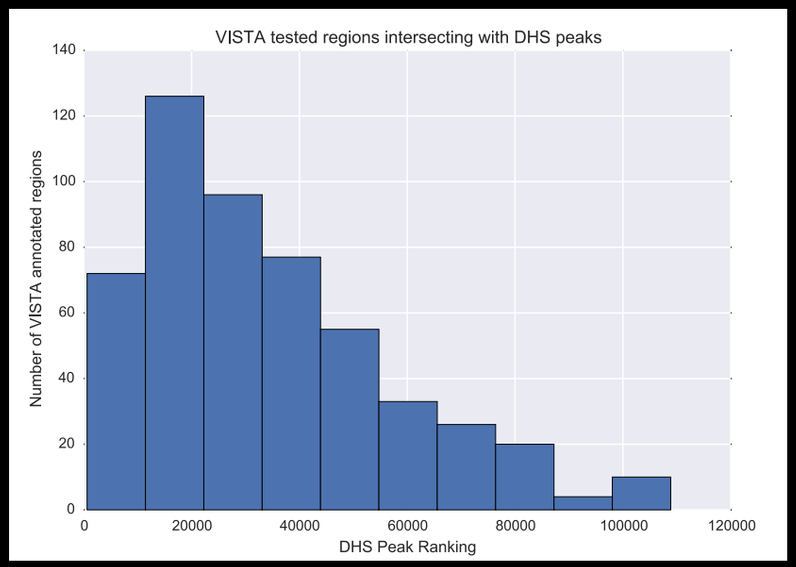
We should be careful about generalizing accuracy, etc.

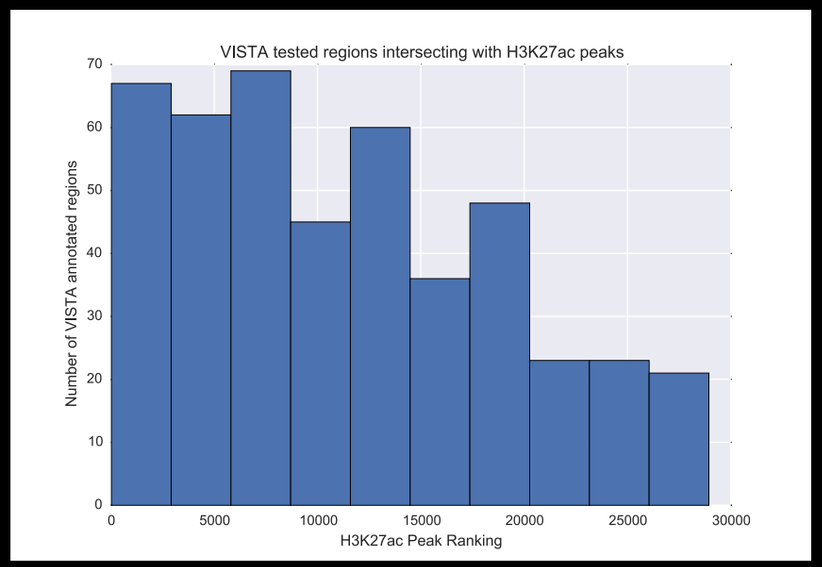
Adding DHS to the picture

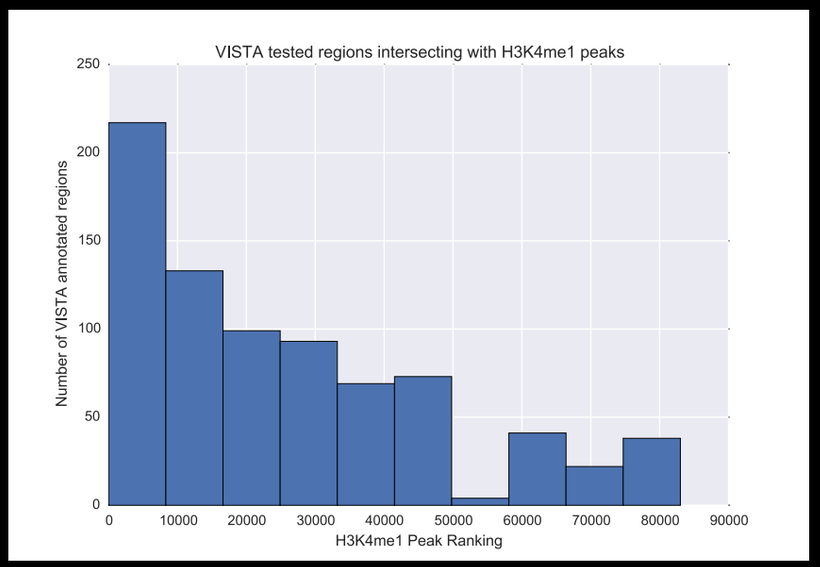
What is the coverage of VISTA database based on peaks within different epigenetic datasets (midbrain)

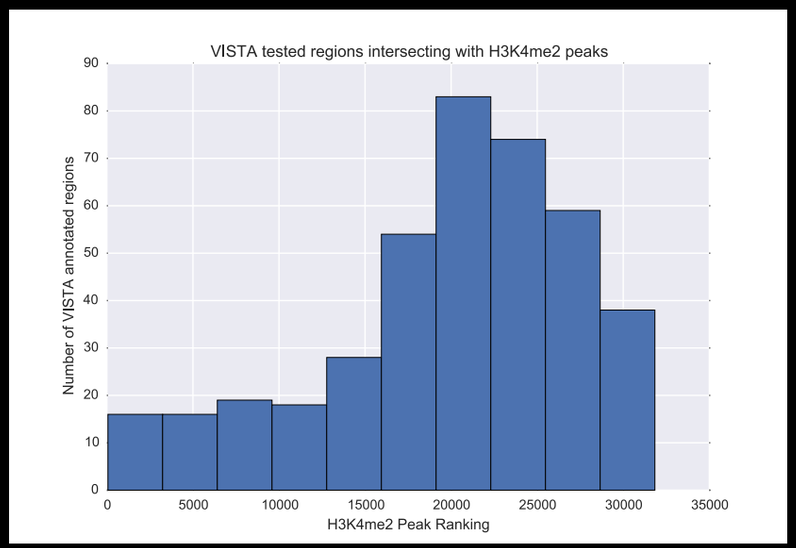


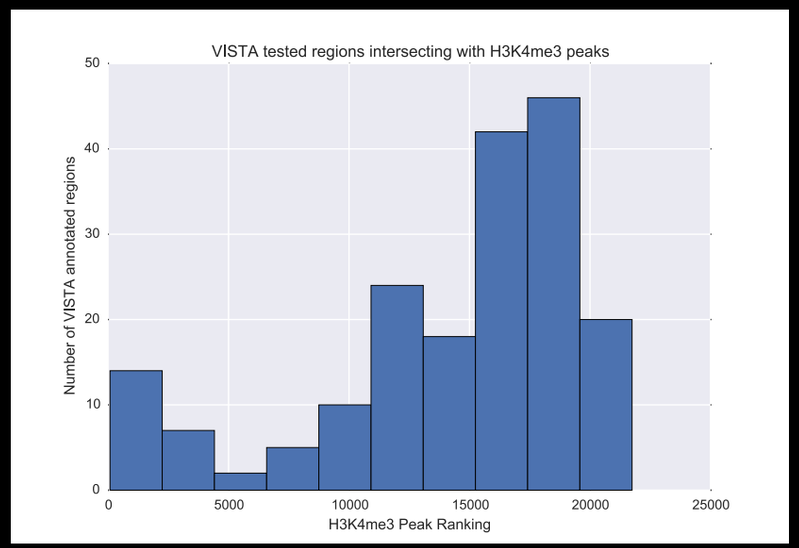
Characterizing biases in terms of peak ranking (midbrain)

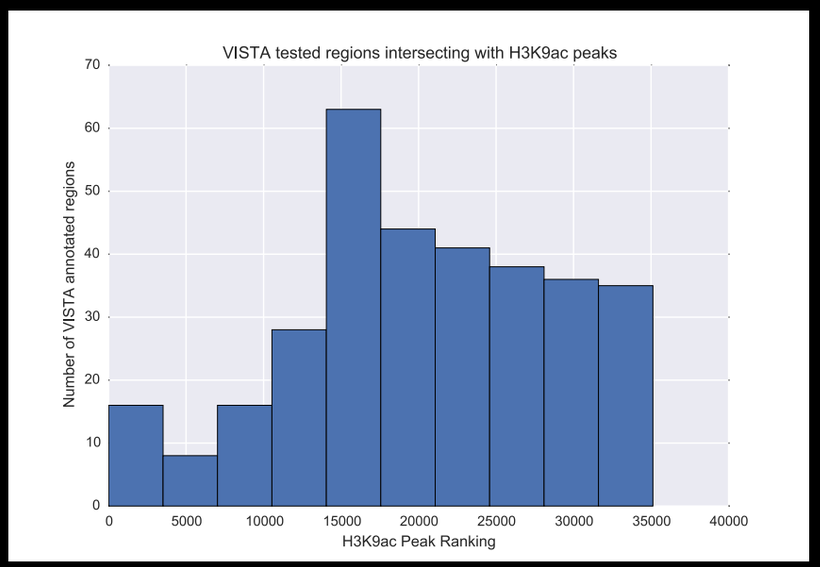


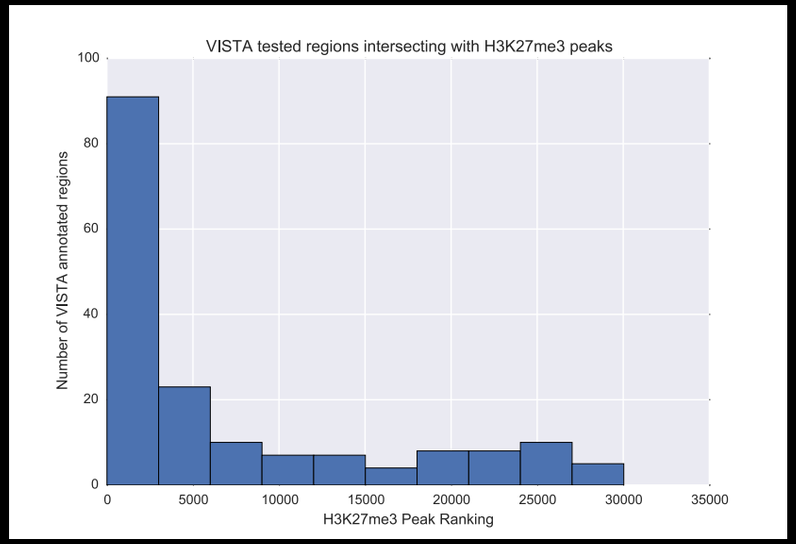








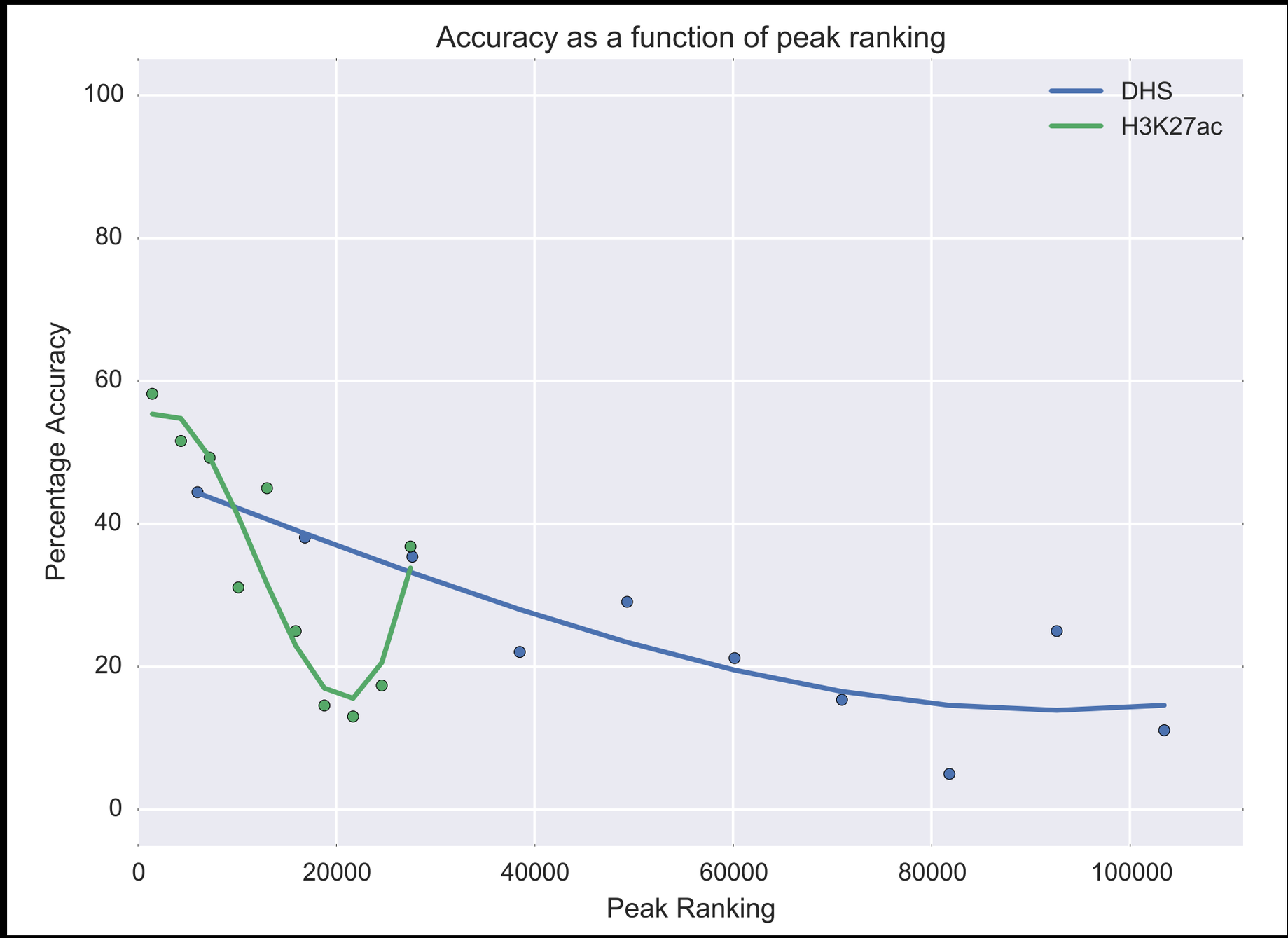




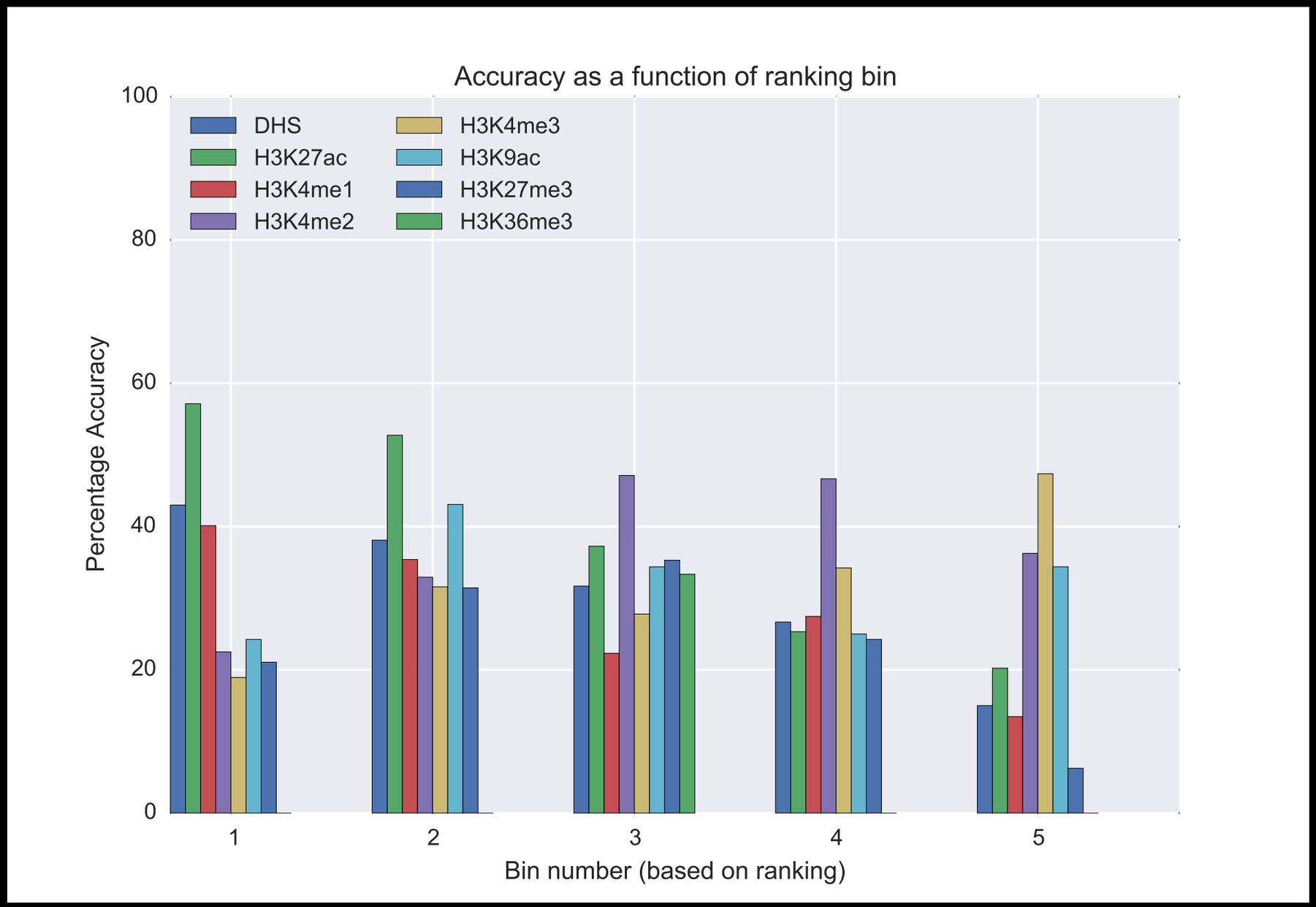
The regions tested are not even over the whole peak list.

We can also characterize bias in terms of multiple co-variates but very few data points.

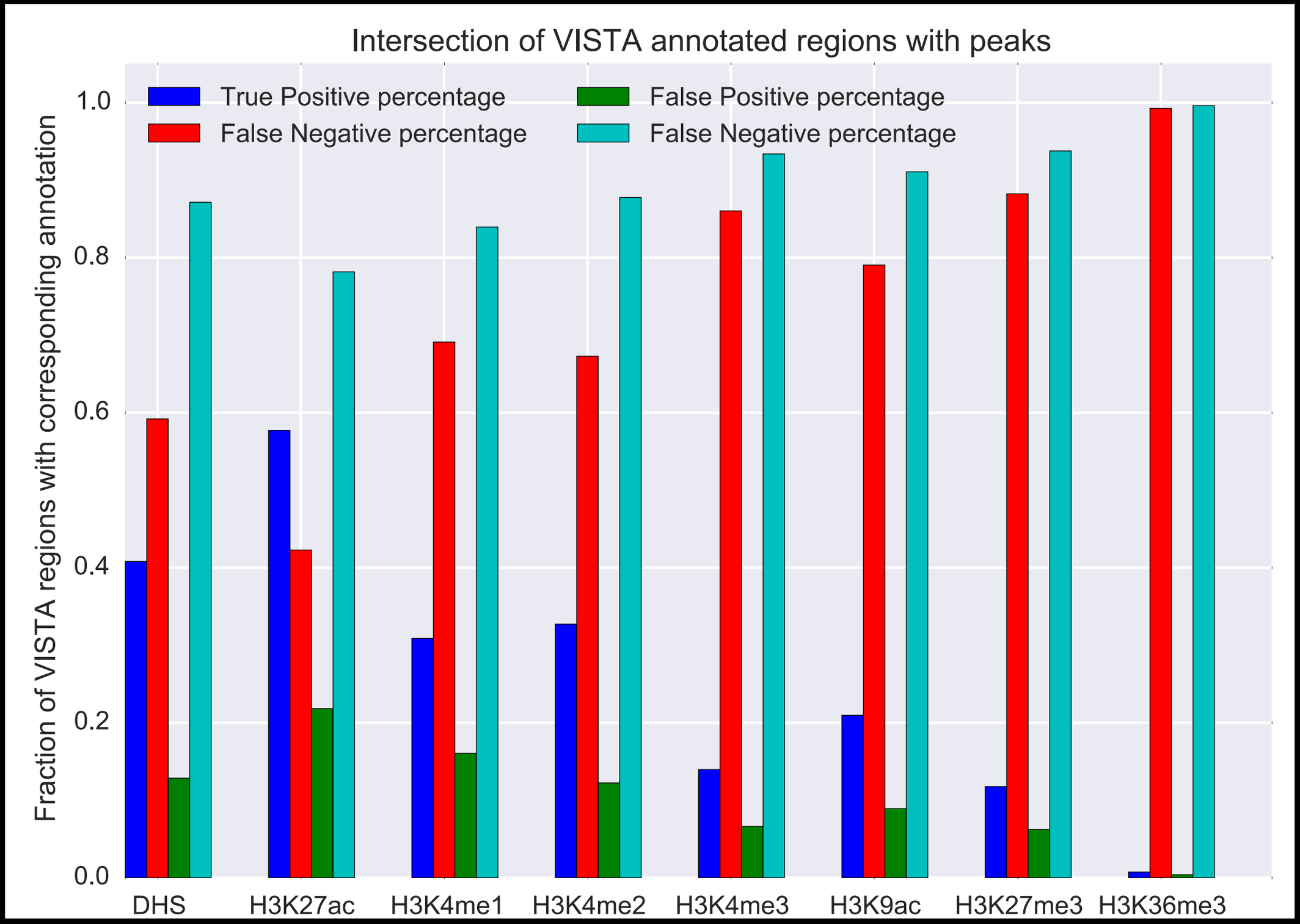
Accuracy with ranking (midbrain) - ranking based binning

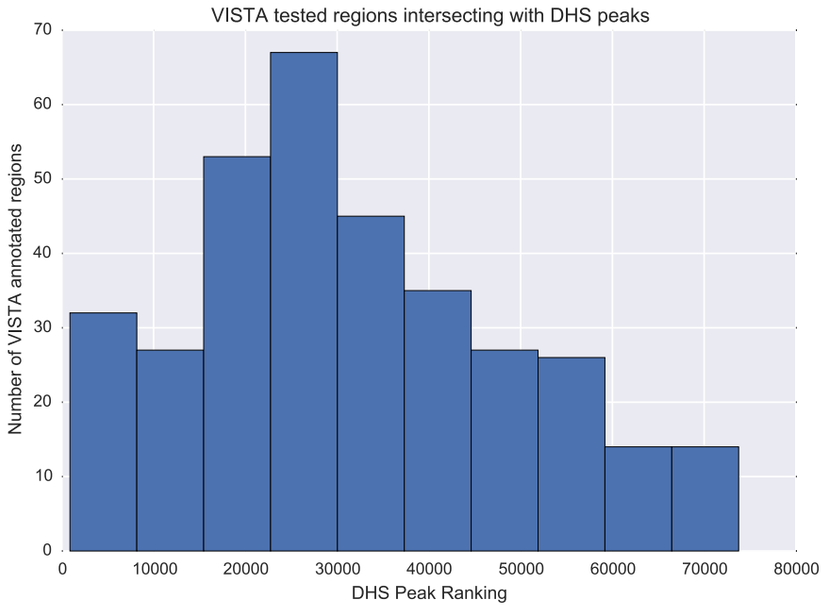


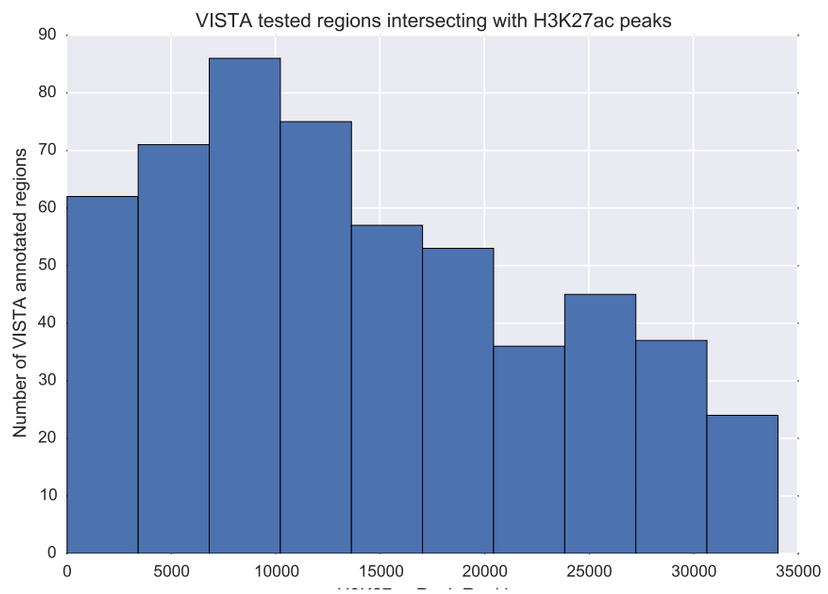
Accuracy with ranking (midbrain) - annotation-based binning

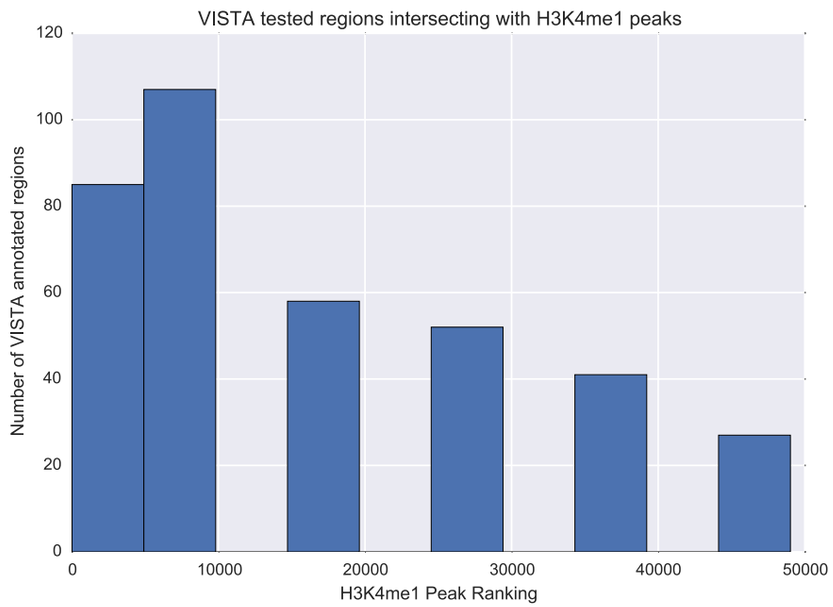


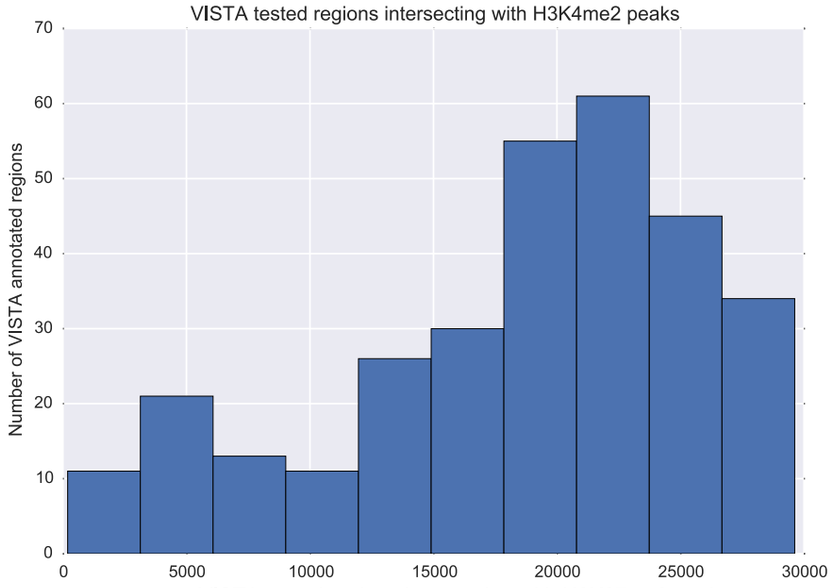
What is the coverage of VISTA database based on peaks within different epigenetic datasets (hindbrain)

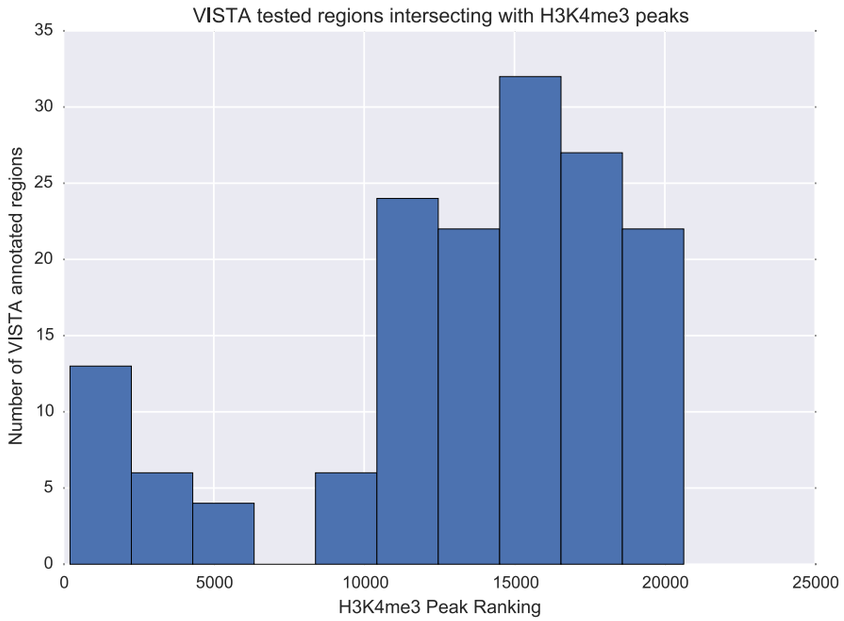


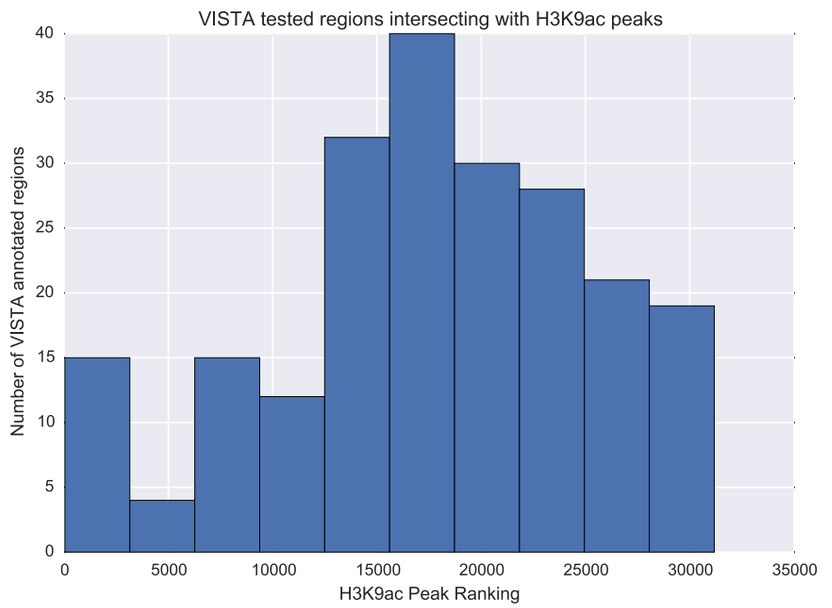


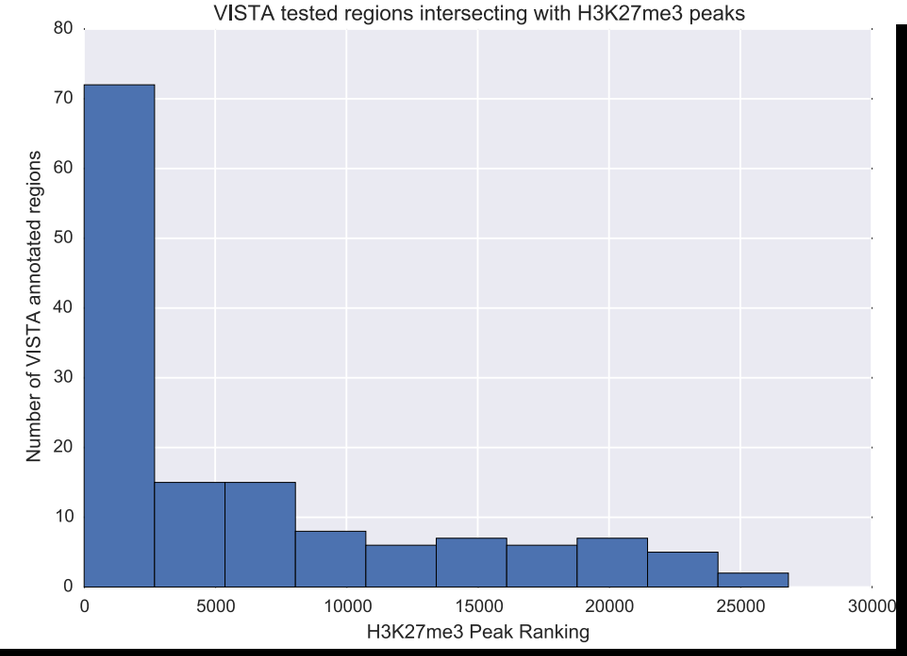










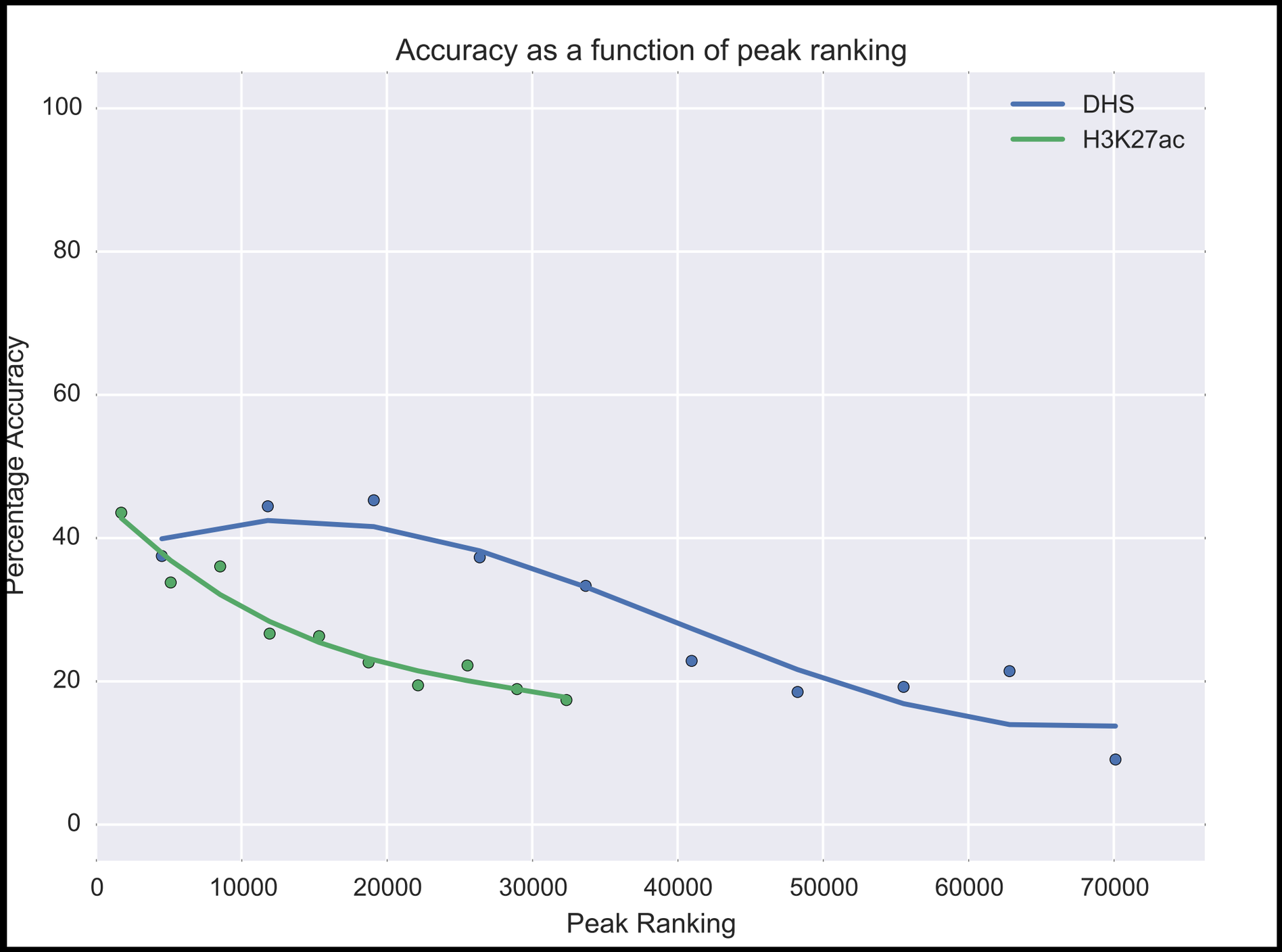


Characterizing biases in terms of peak ranking (hindbrain)

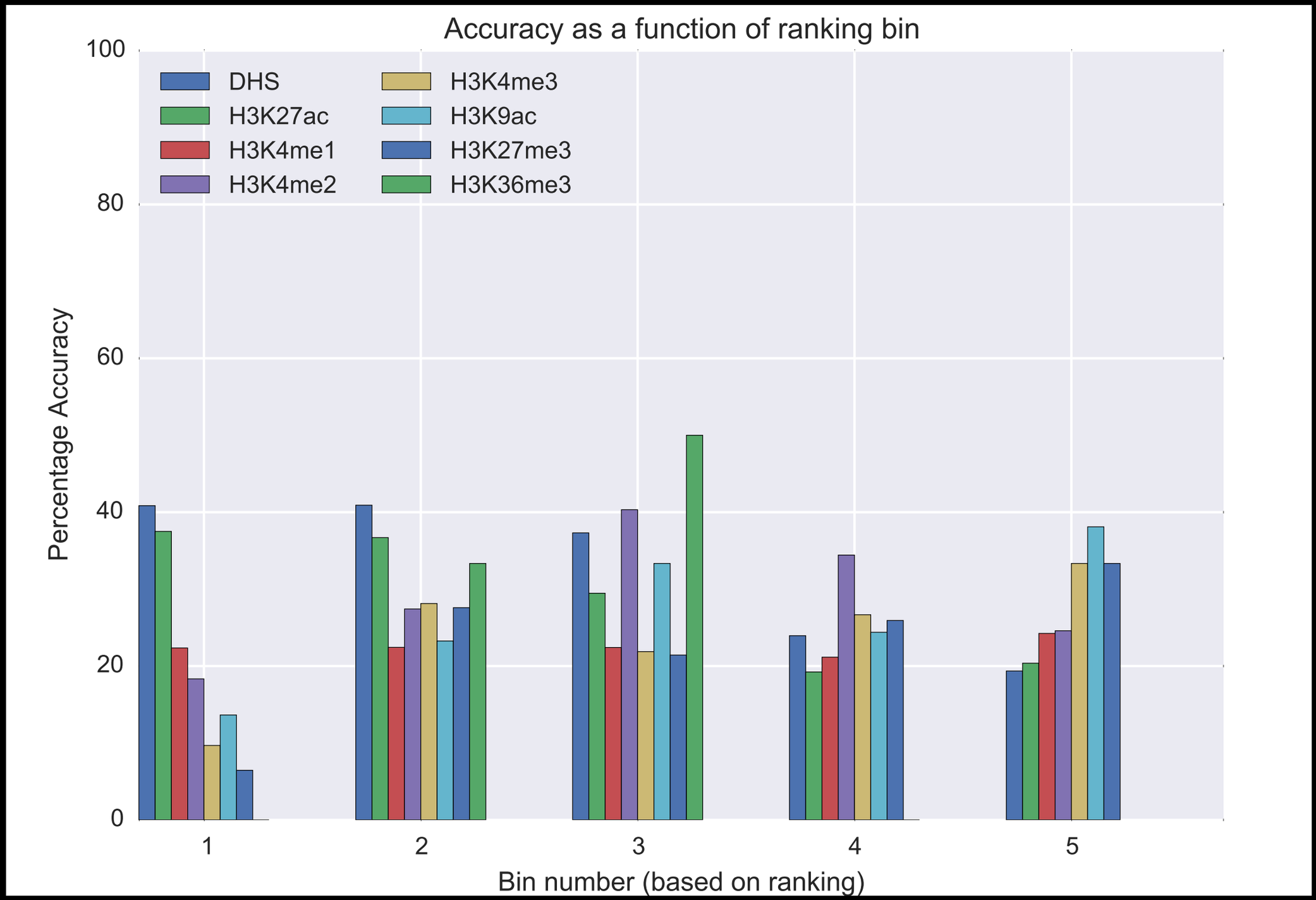
The regions tested are not even over the whole peak list.

We can also characterize bias in terms of multiple co-variates but very few data points.

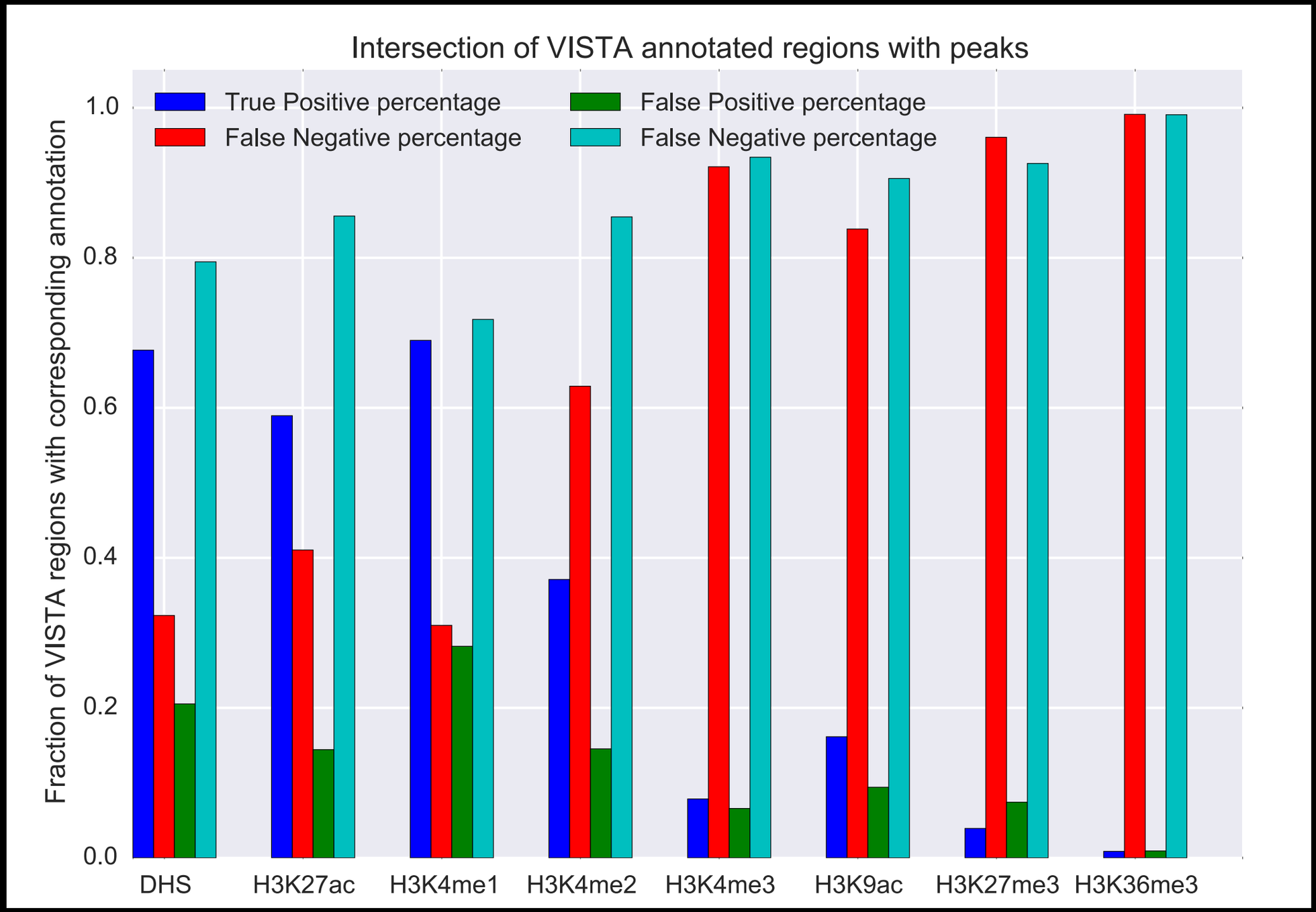
Accuracy with ranking (hindbrain) - ranking based binning



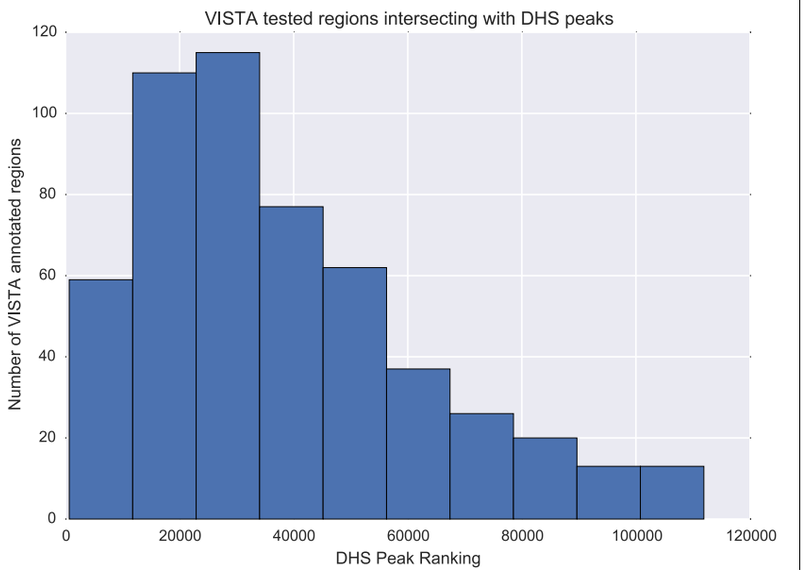
Accuracy with ranking (hindbrain) - annotation based binning

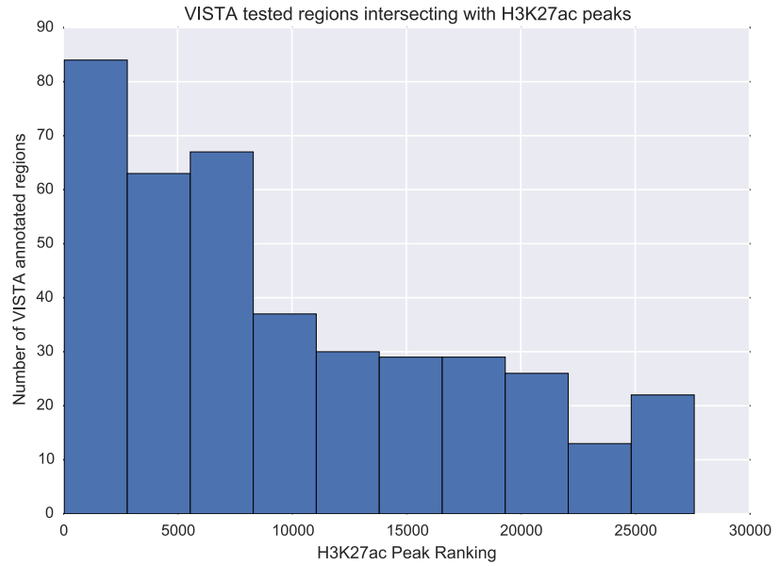


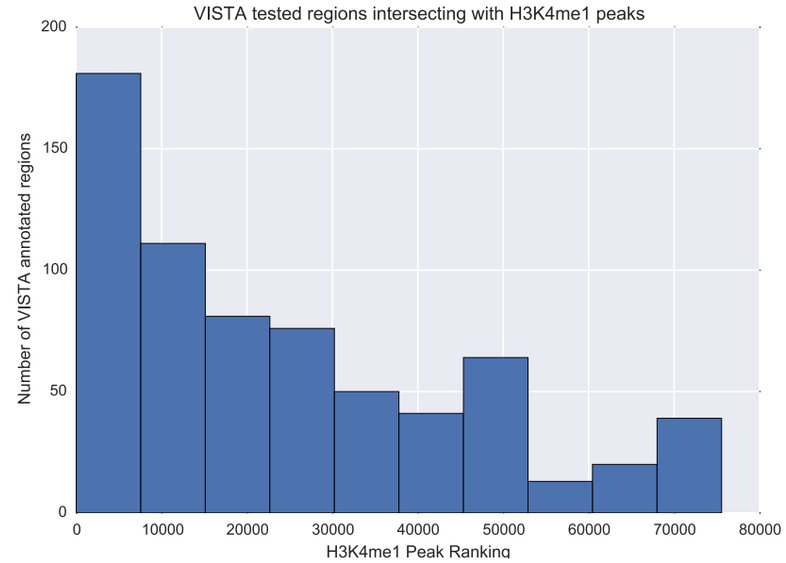
What is the coverage of VISTA database based on peaks within different epigenetic datasets (limb)

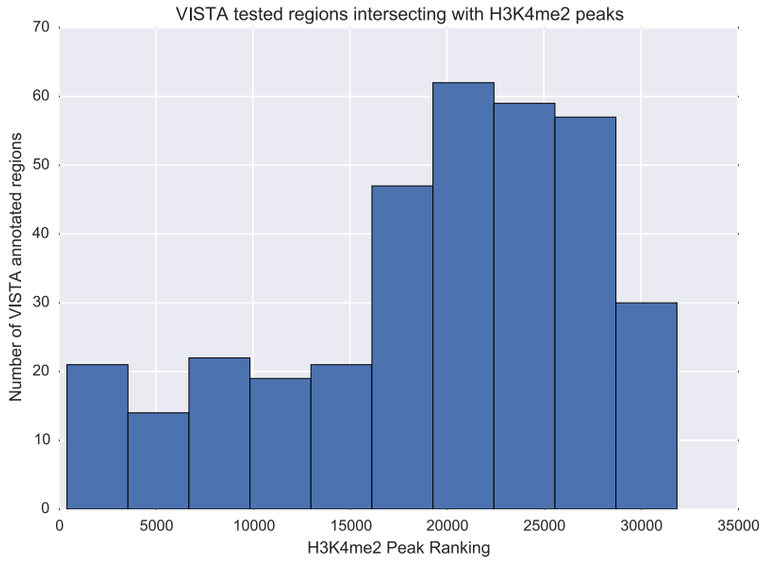


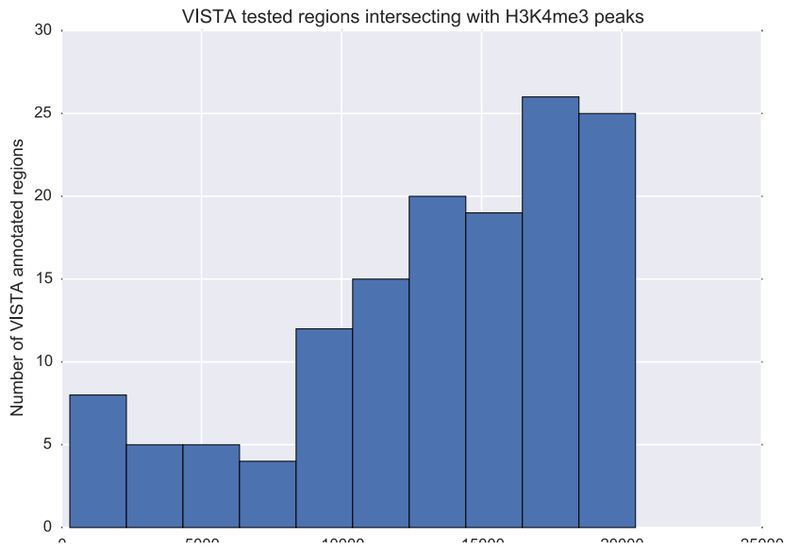
Characterizing biases in terms of peak ranking (limb)

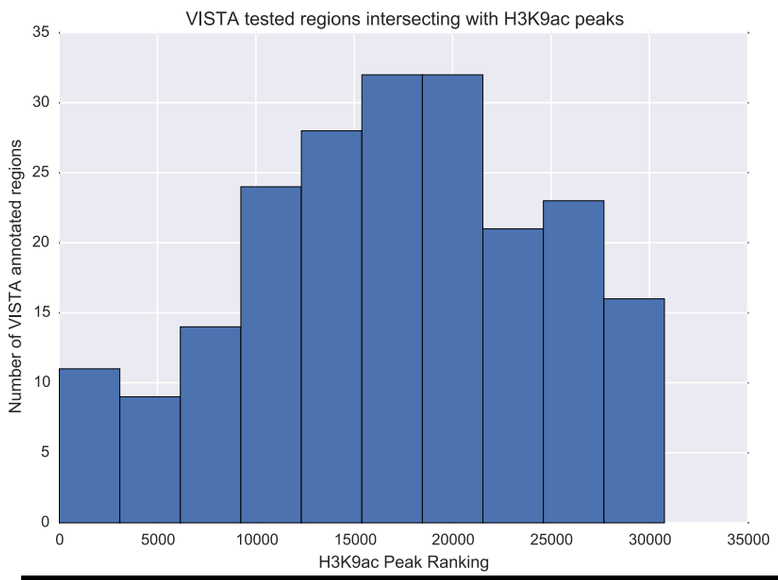


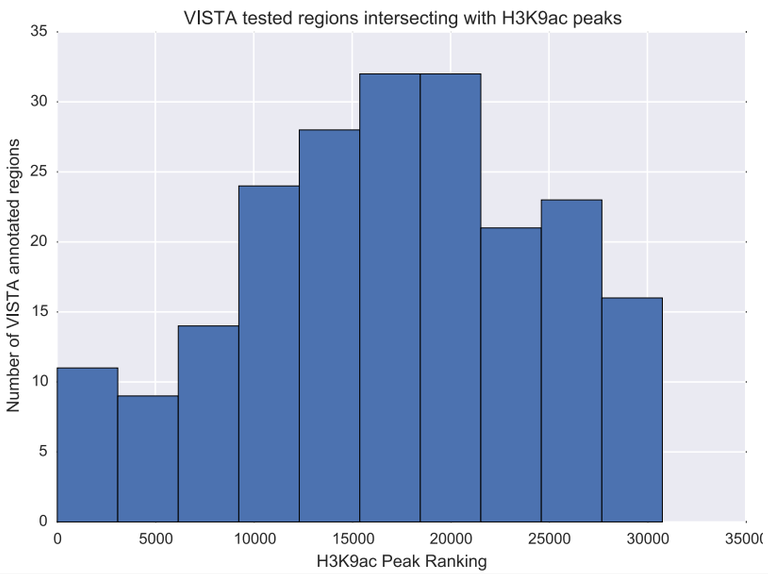


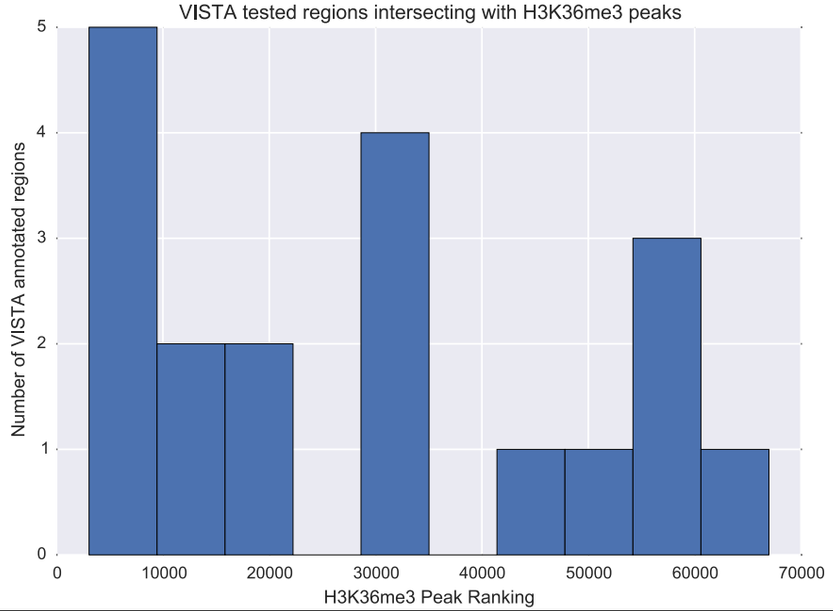




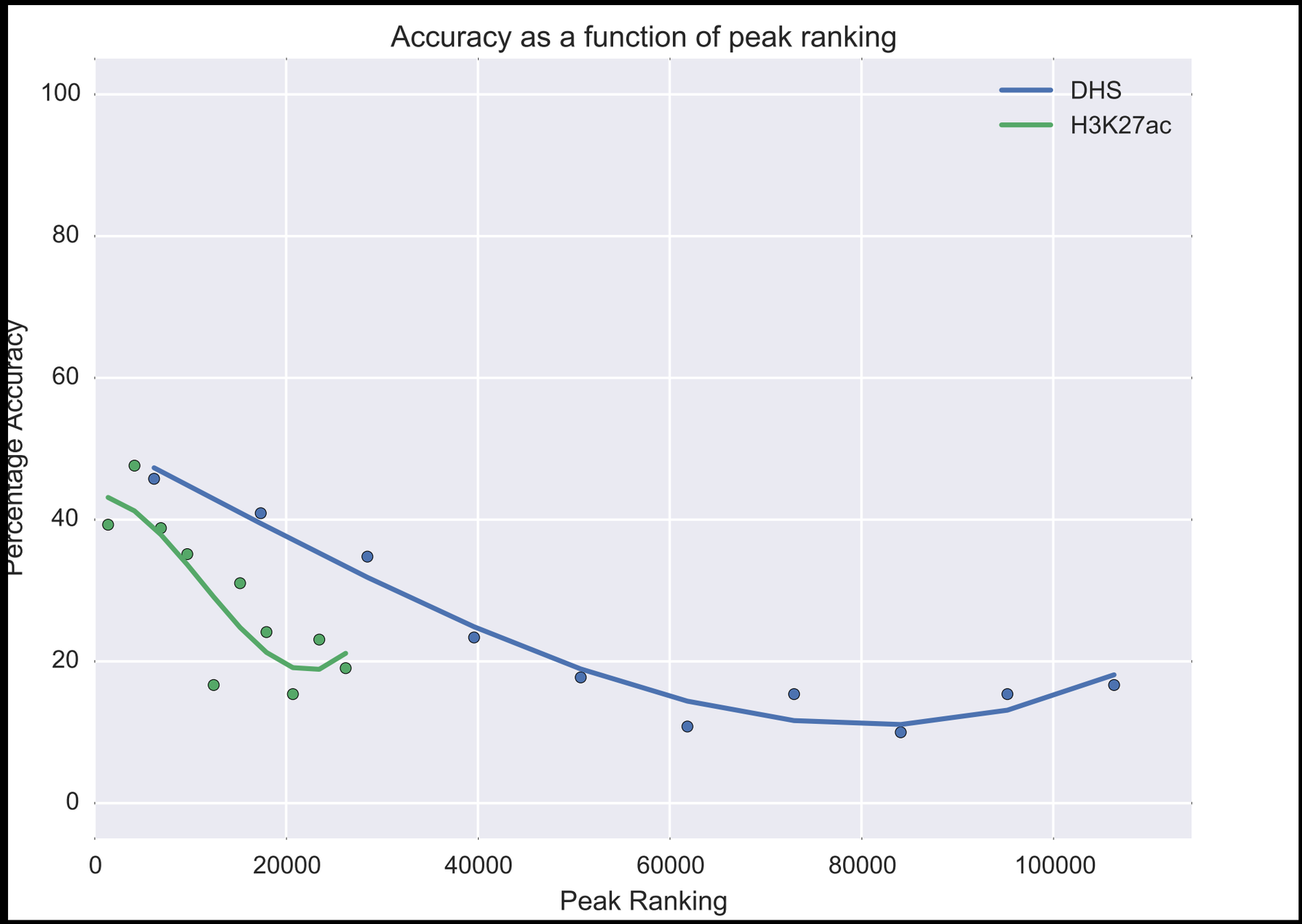




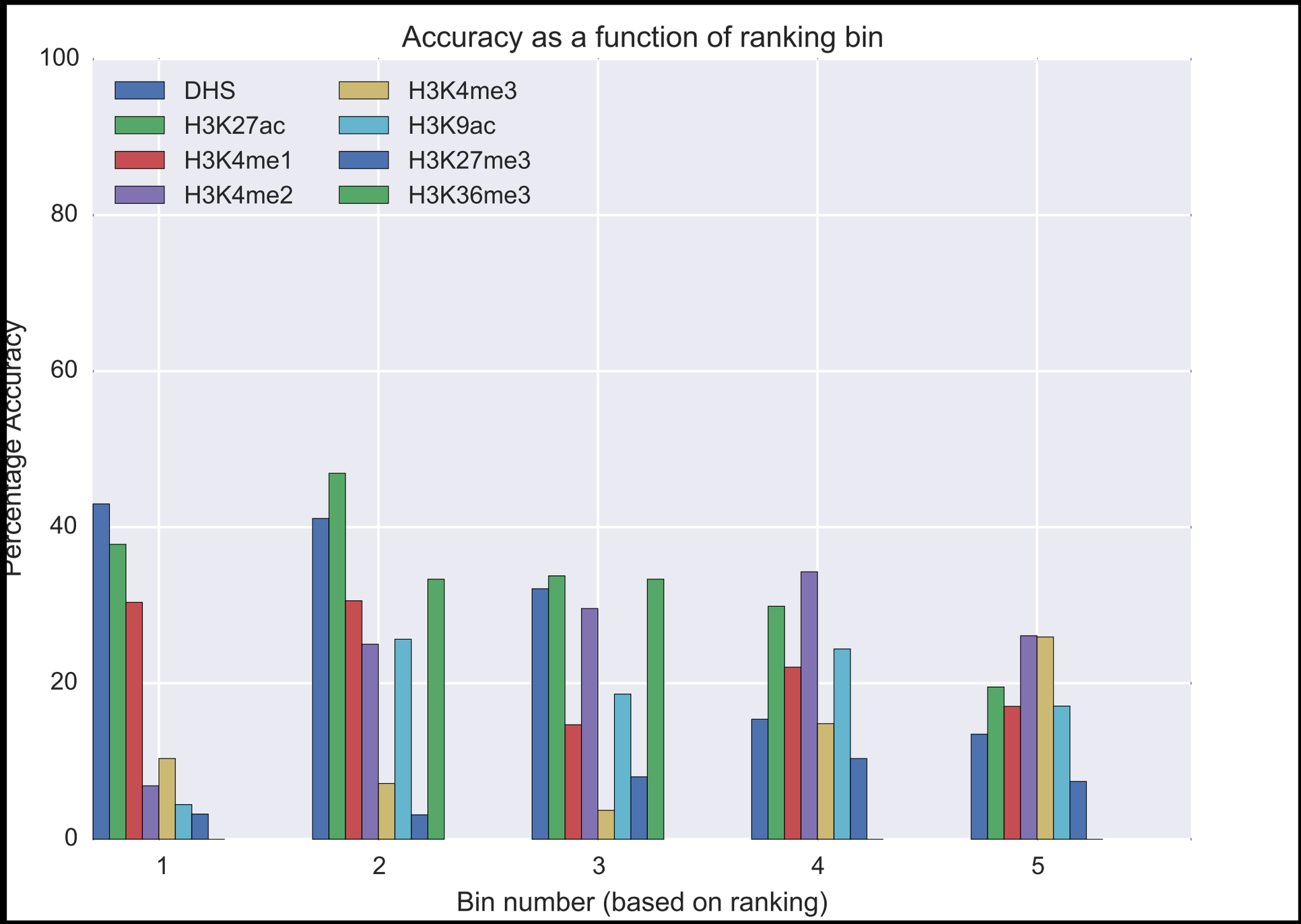




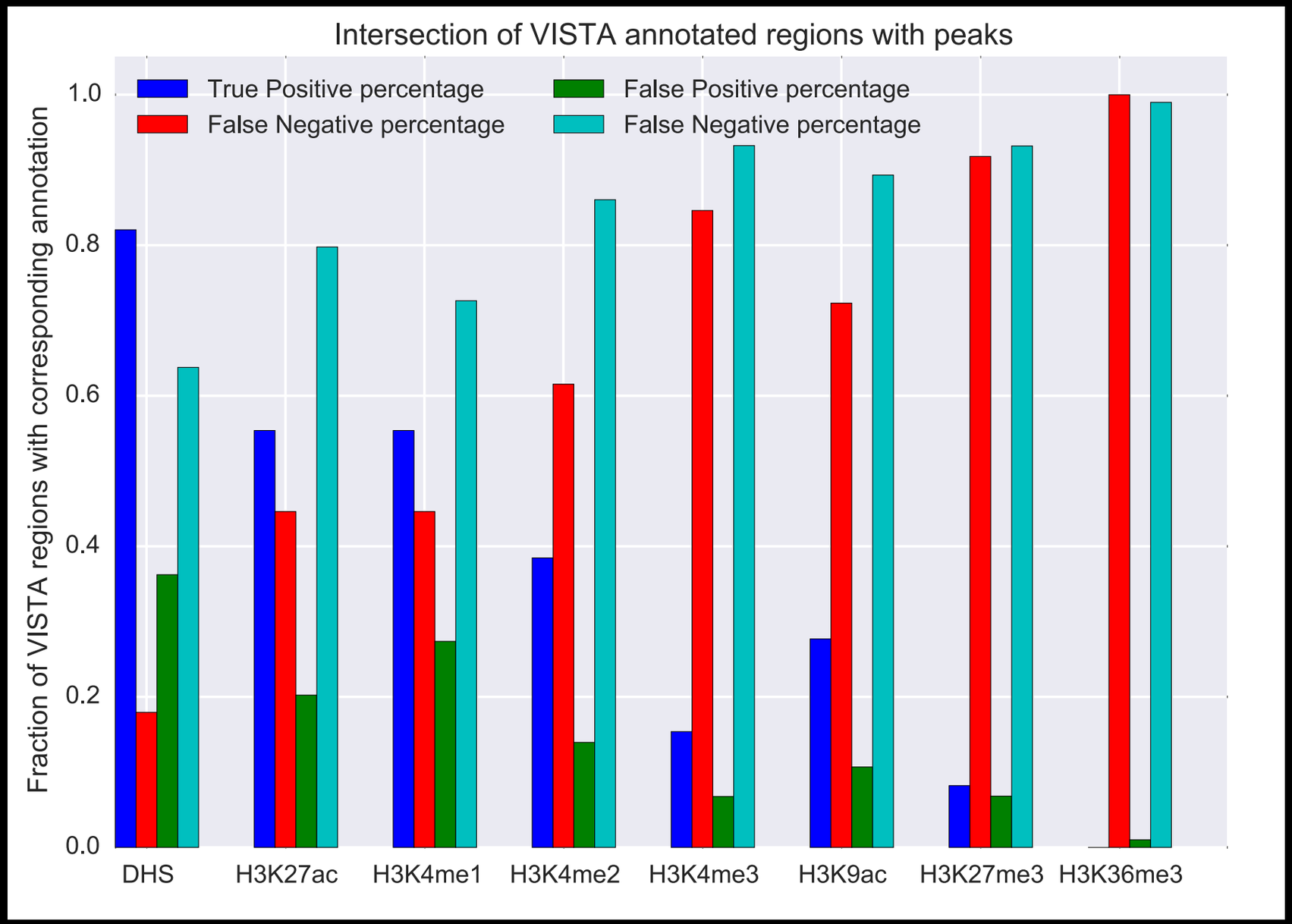
Accuracy with ranking (limb) - ranking based binning



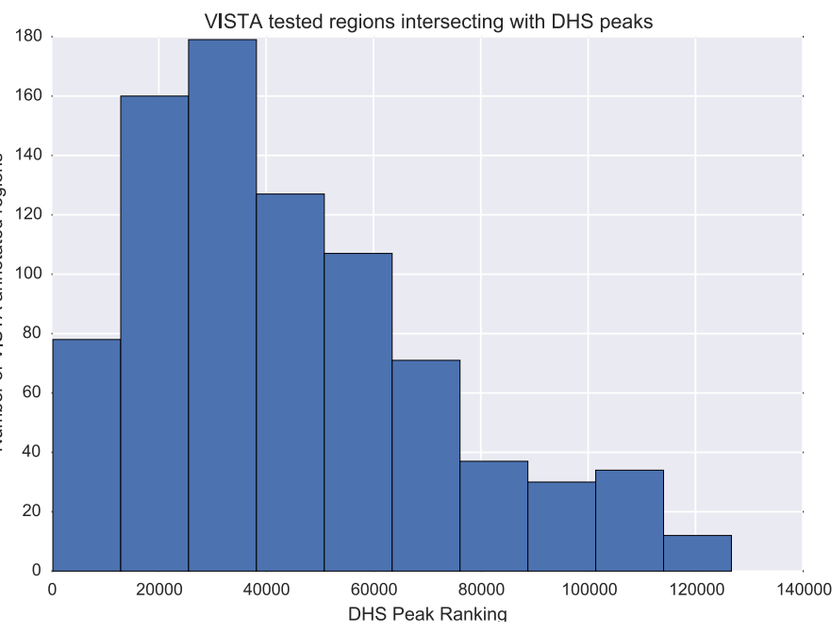
Accuracy with ranking (limb) - annotation-based binning

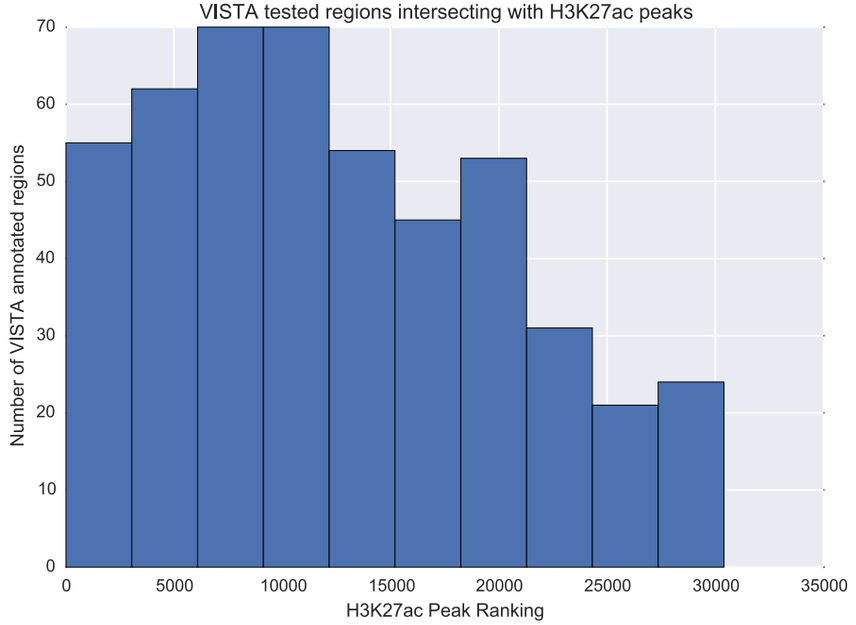


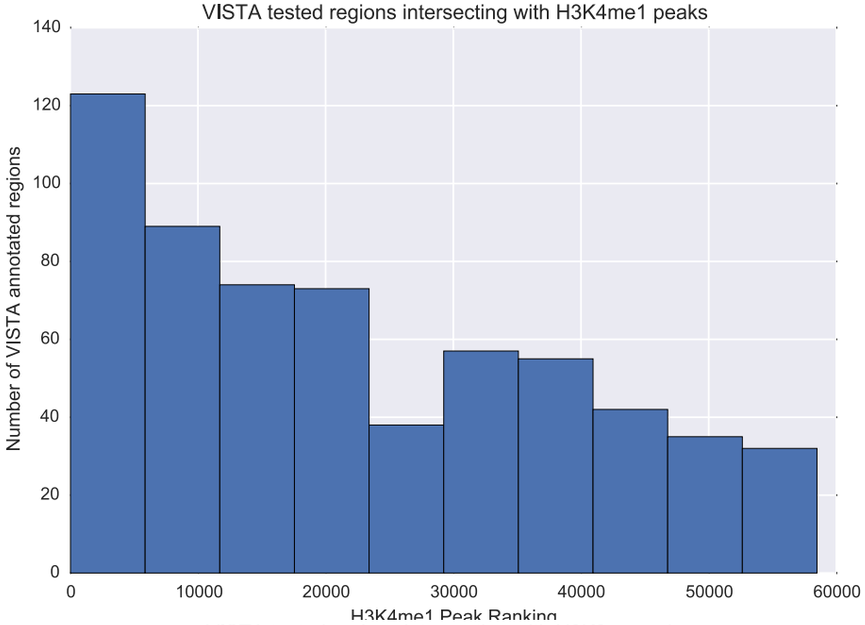
What is the coverage of VISTA database based on peaks within different epigenetic datasets (neuralTube)

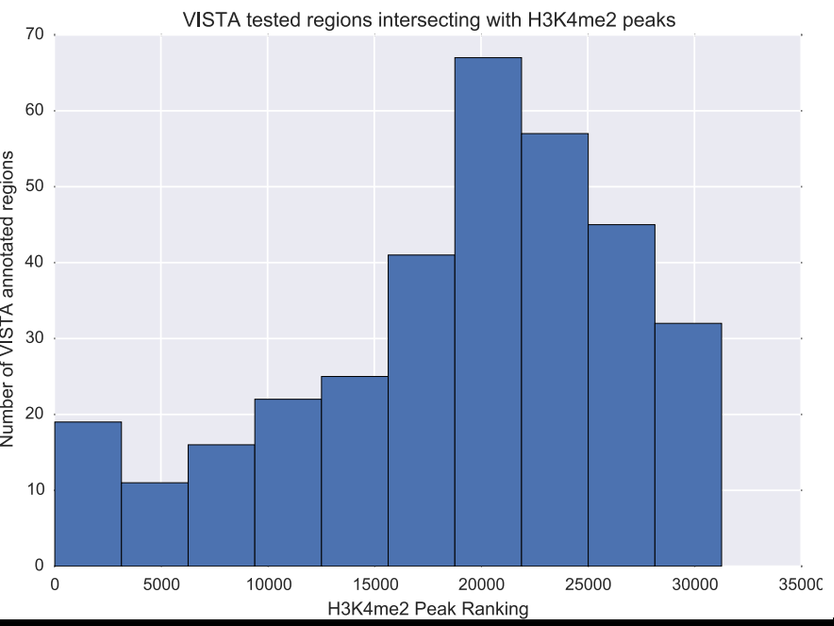


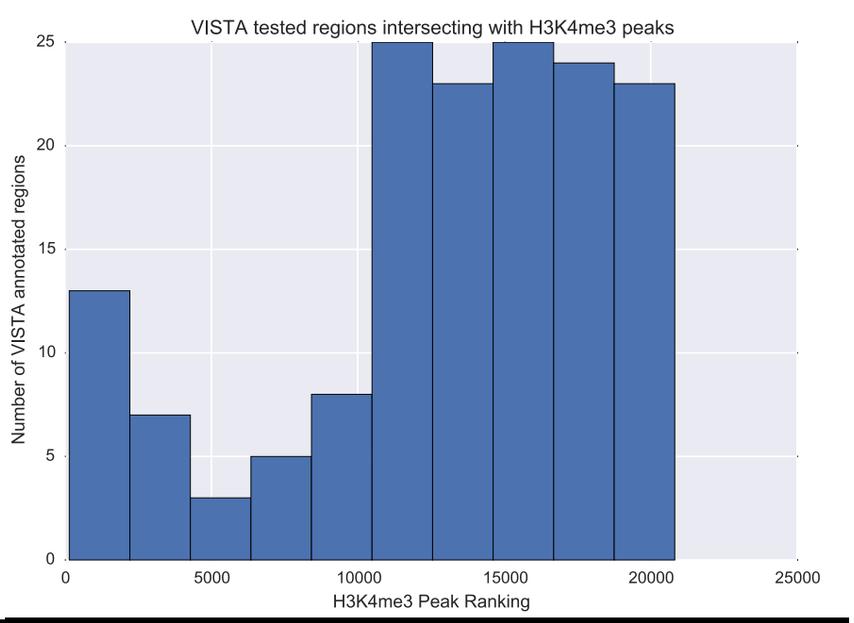
Characterizing biases in terms of peak ranking (neuralTube)

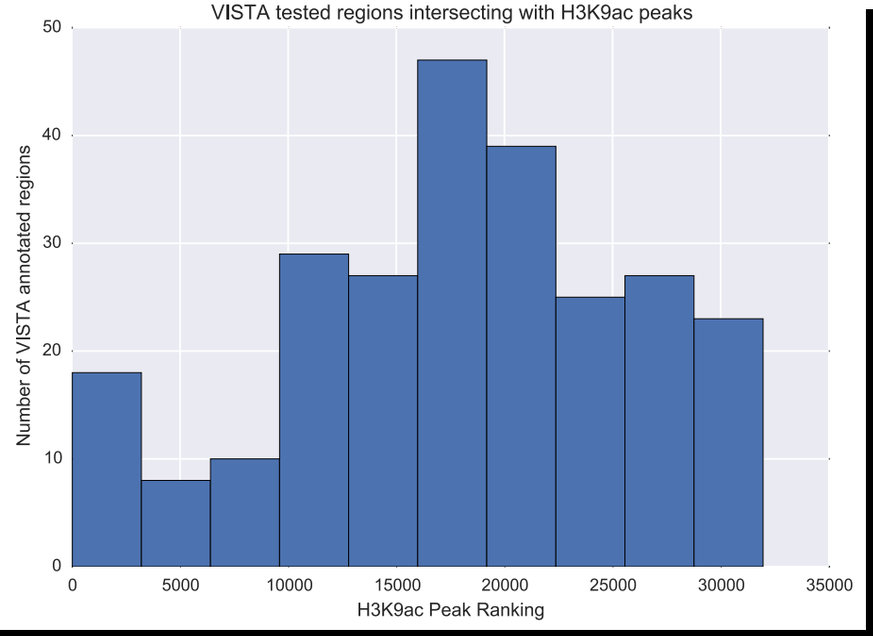






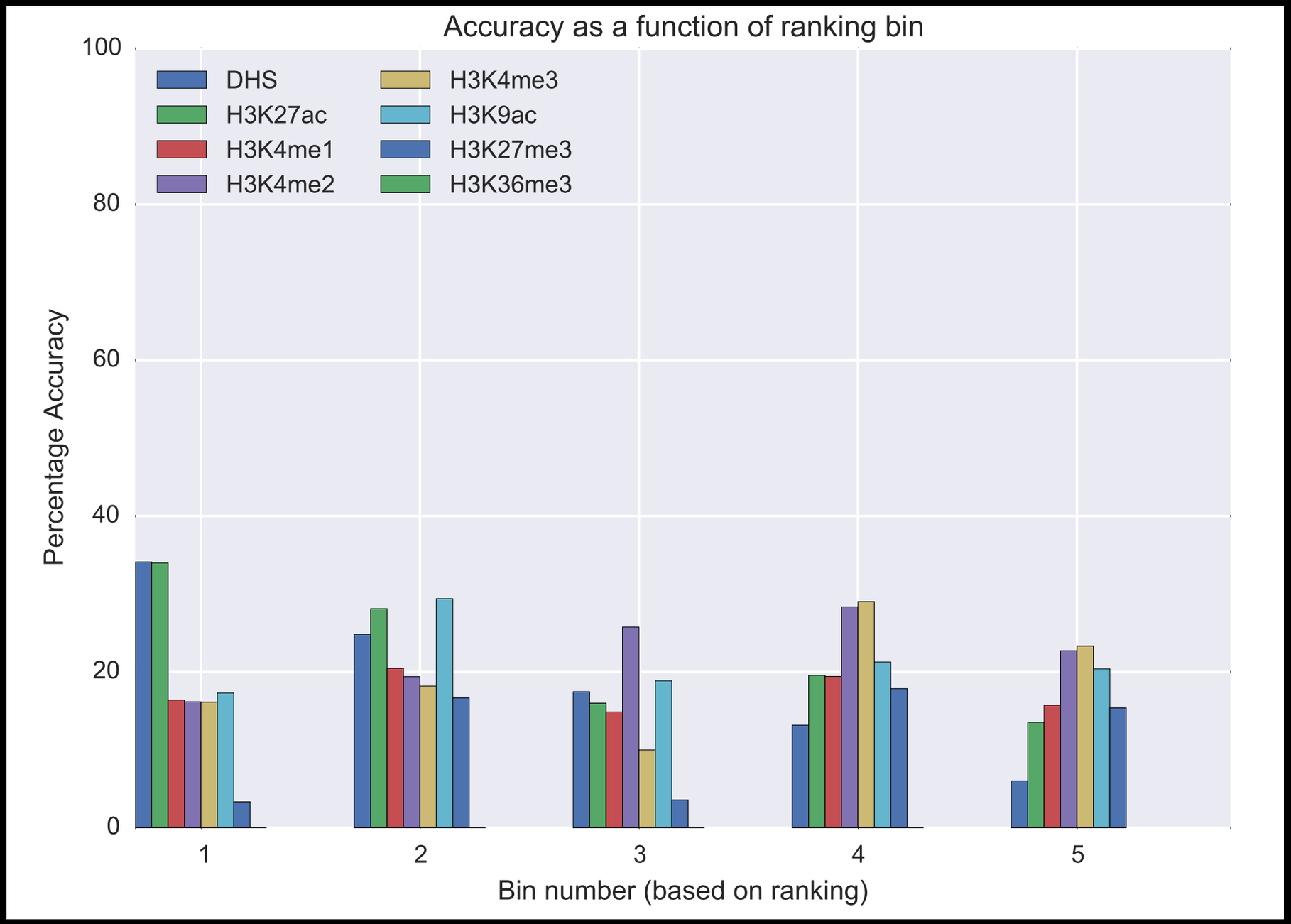






Accuracy with ranking (neuralTube) - ranking based binning

Accuracy with ranking (neuralTube) - annotation-based binning



What have we learned?

There is a large bias in the VISTA database.
We should be careful while evaluating accuracy from VISTA database.

How do we combine peaks in an unsupervised manner?

Linear/Logistic regression do not add much value - DHS and H3K27ac are most valuable marks. H3K9ac assists in the presence of H3K27ac.

Future Directions:
- We will focus on intersection of peaks and look at signal around peaks.
- We will focus on matched filter ranking and compare it to peak-based methods (with intersection).

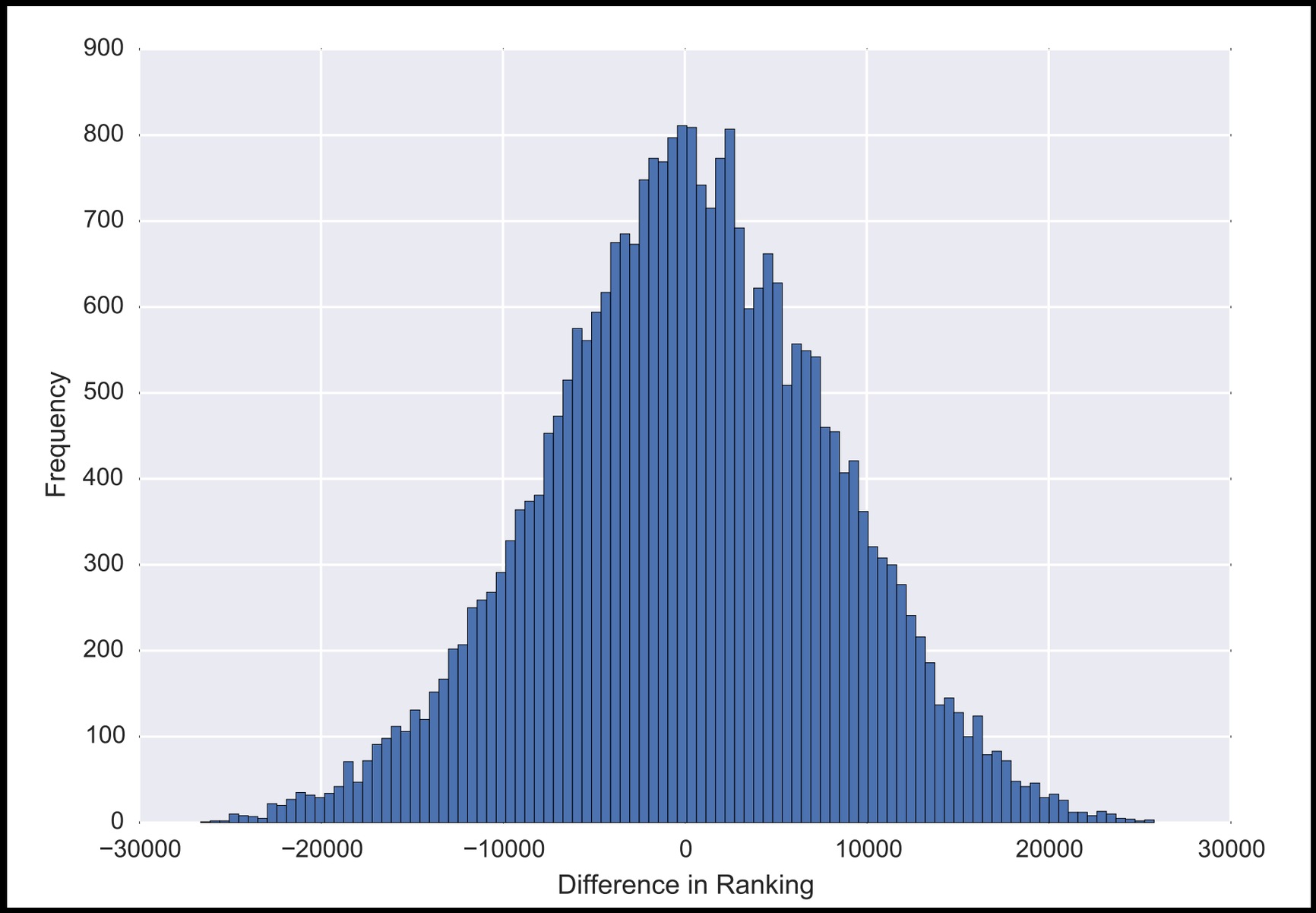
Part 2
Enhancers for Encyclopedia

Enhancer Predictions using epigenetic datasets (histone and/or DHS) in the presence of training data

All assessments will be performed on the VISTA database and new ENCODE phase 2 datasets (labeled data).

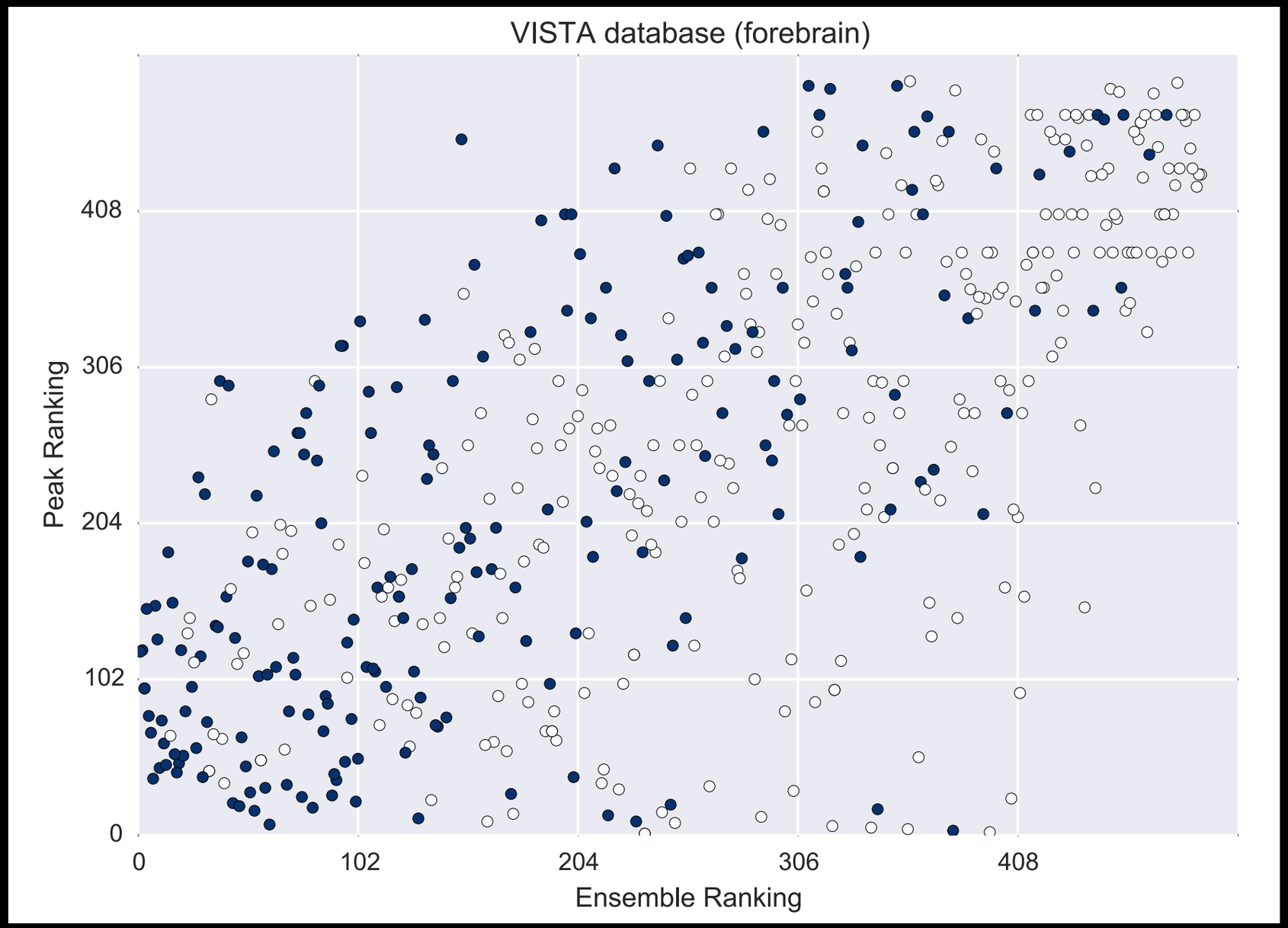
Assessing the accuracy of ensemble method by comparing with best unsupervised predictor (H3K27ac peaks - no DHS datasets in this tissue).

Differences in ranking between H3K27ac peaks and ensemble method



Concentrating on VISTA regions

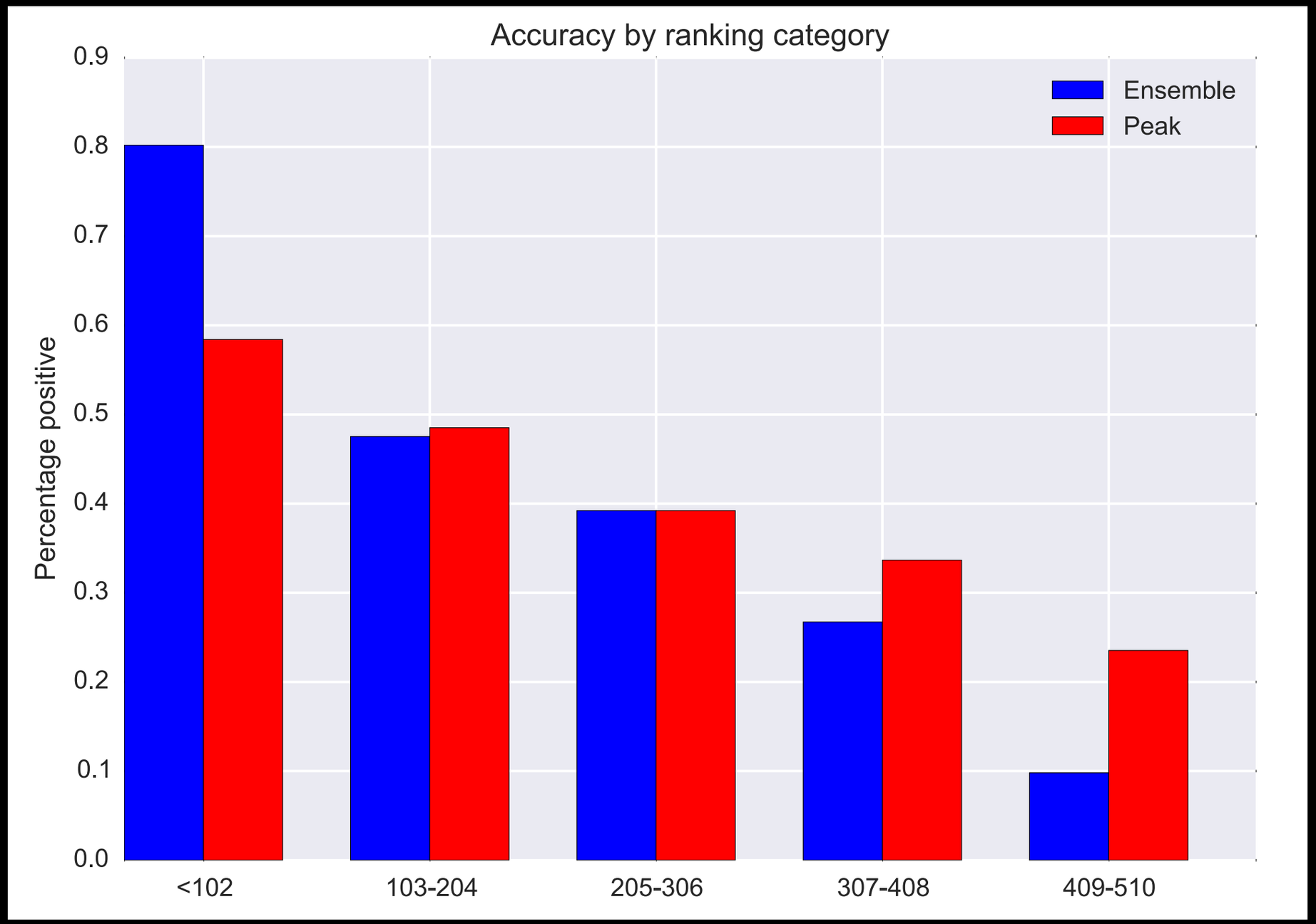
Comparing ranking of VISTA regions



positives - filled circles
negatives - empty circles

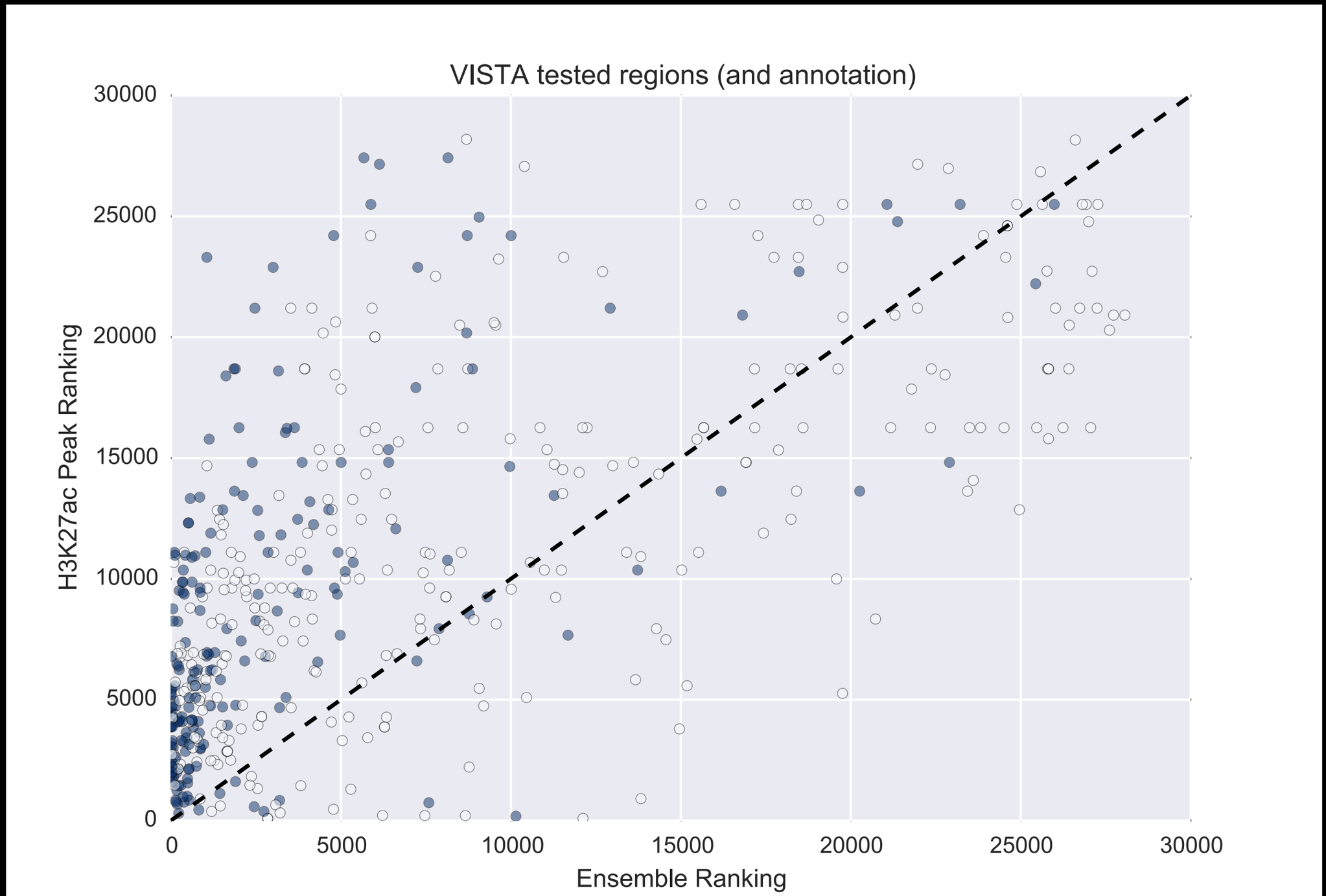
Split in to 5 bins based on ranking (grids)

Does accuracy reduce with ranking - VISTA

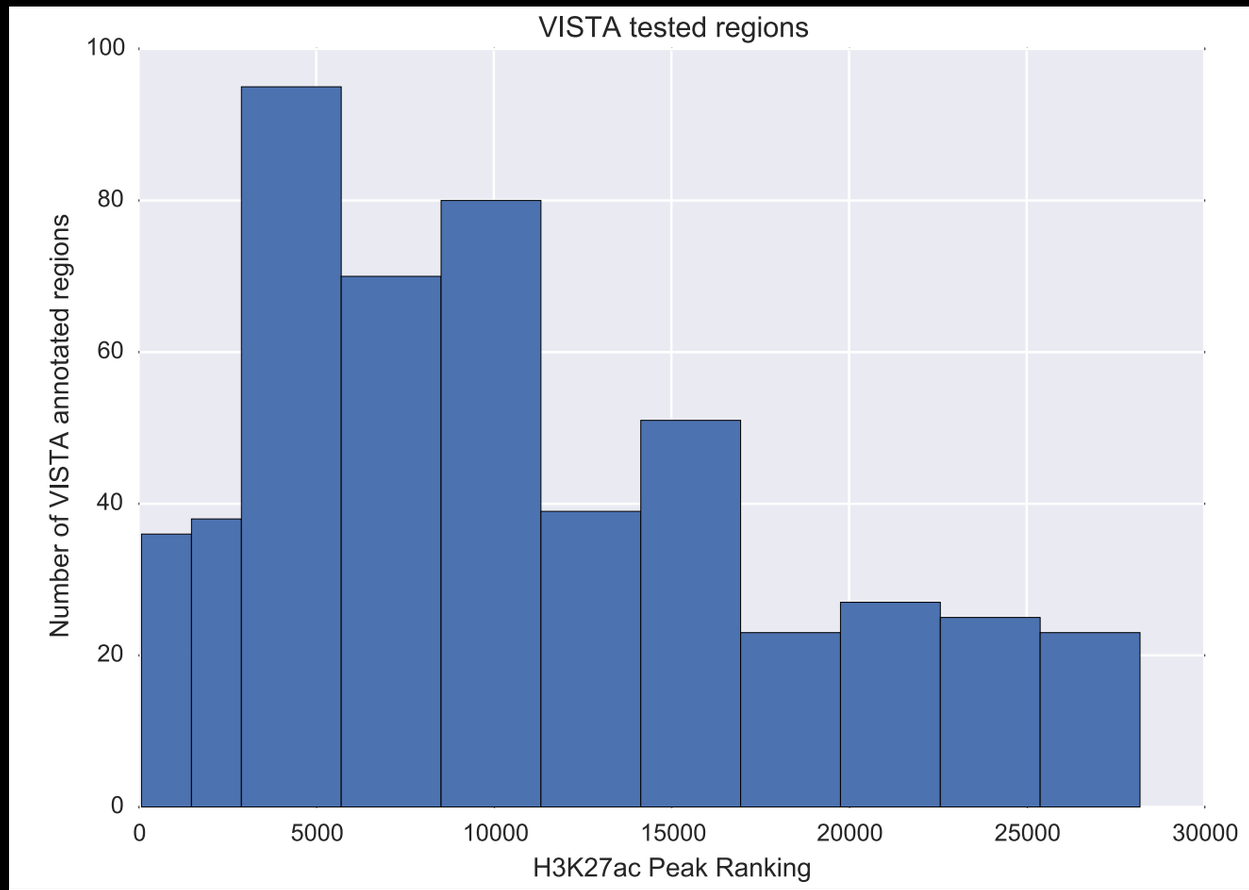


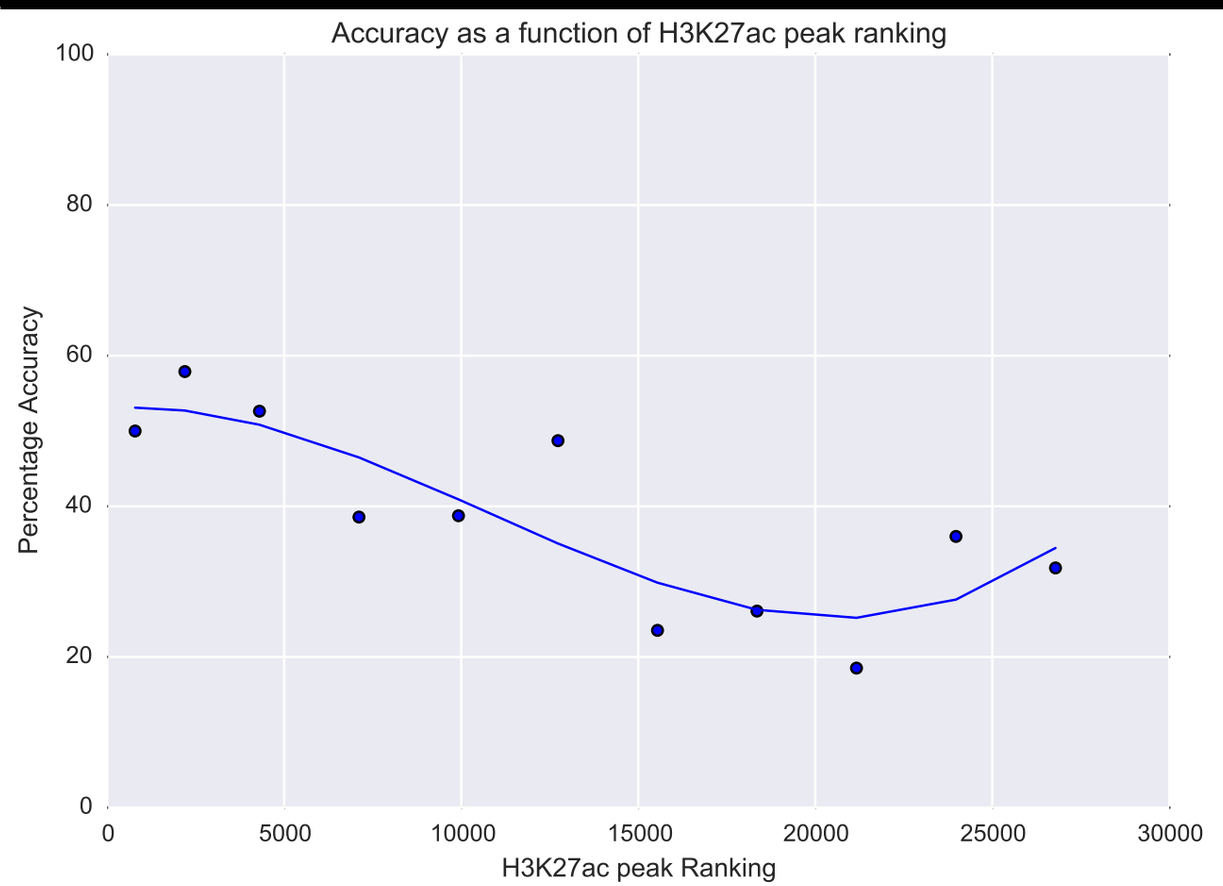
Accuracy of highest ranked VISTA regions by Ensemble method are higher

Comparing ranking of VISTA regions (full ranking)

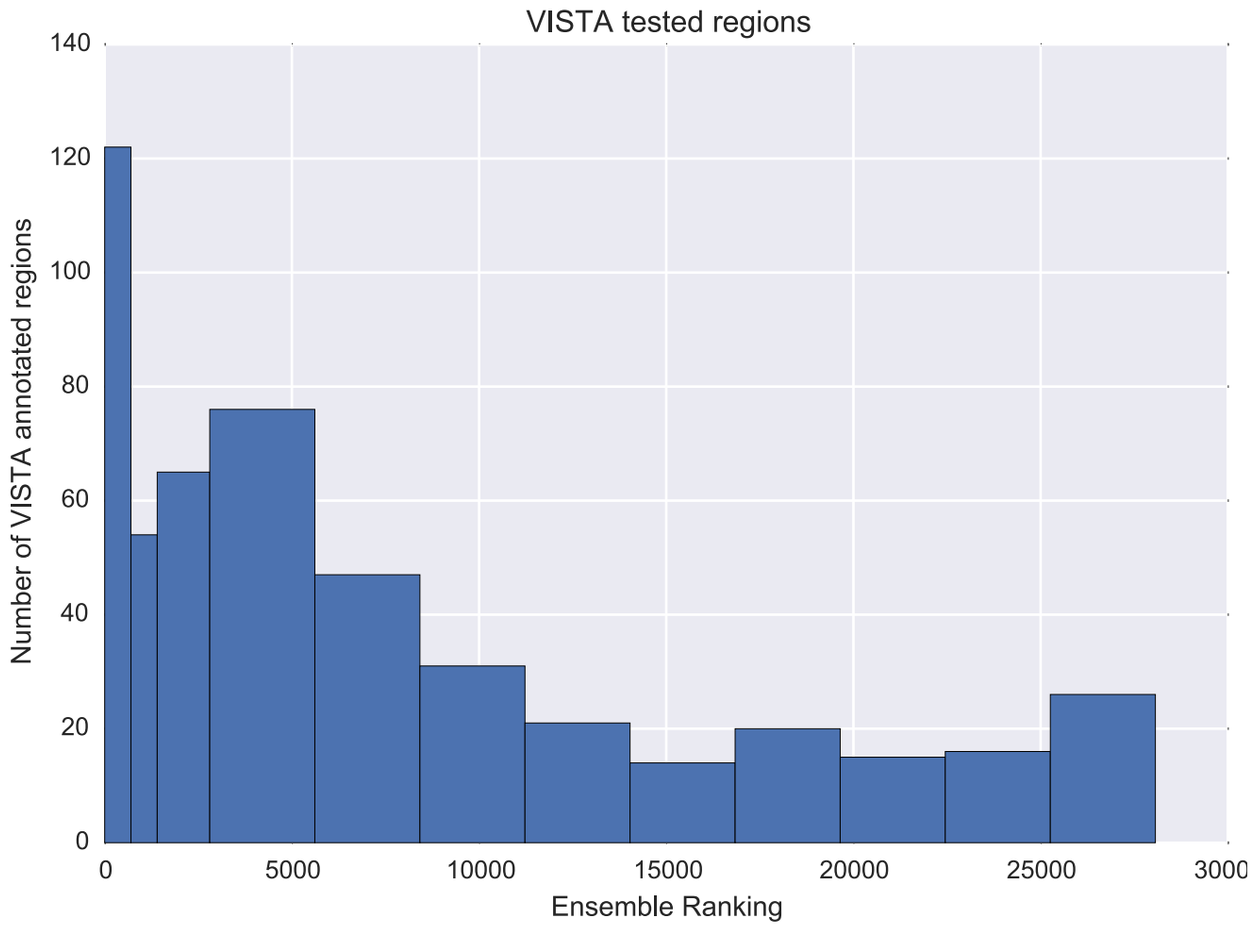


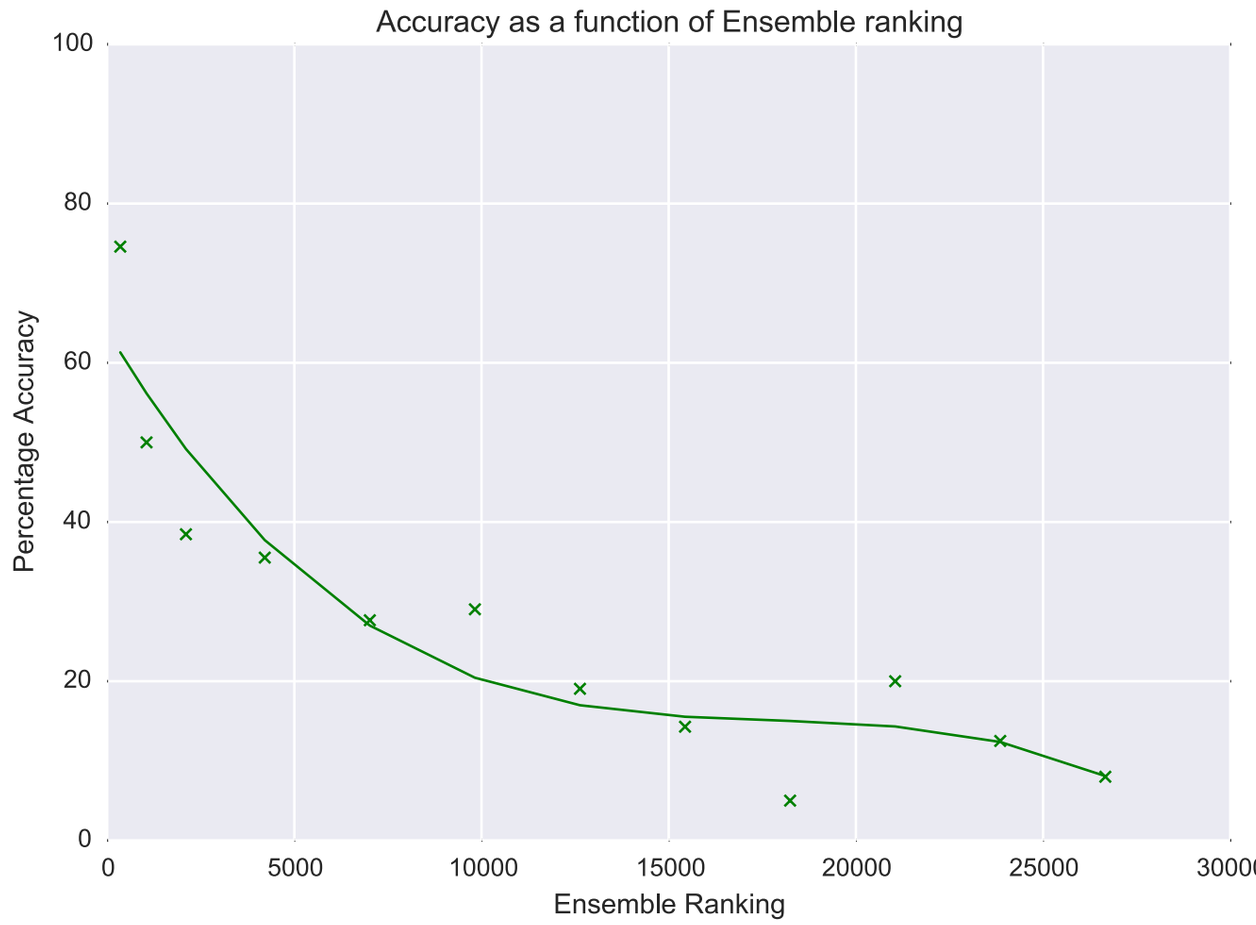
Calculating accuracy of predictions as a function of peak ranking





Calculating accuracy of predictions as a function of ensemble ranking



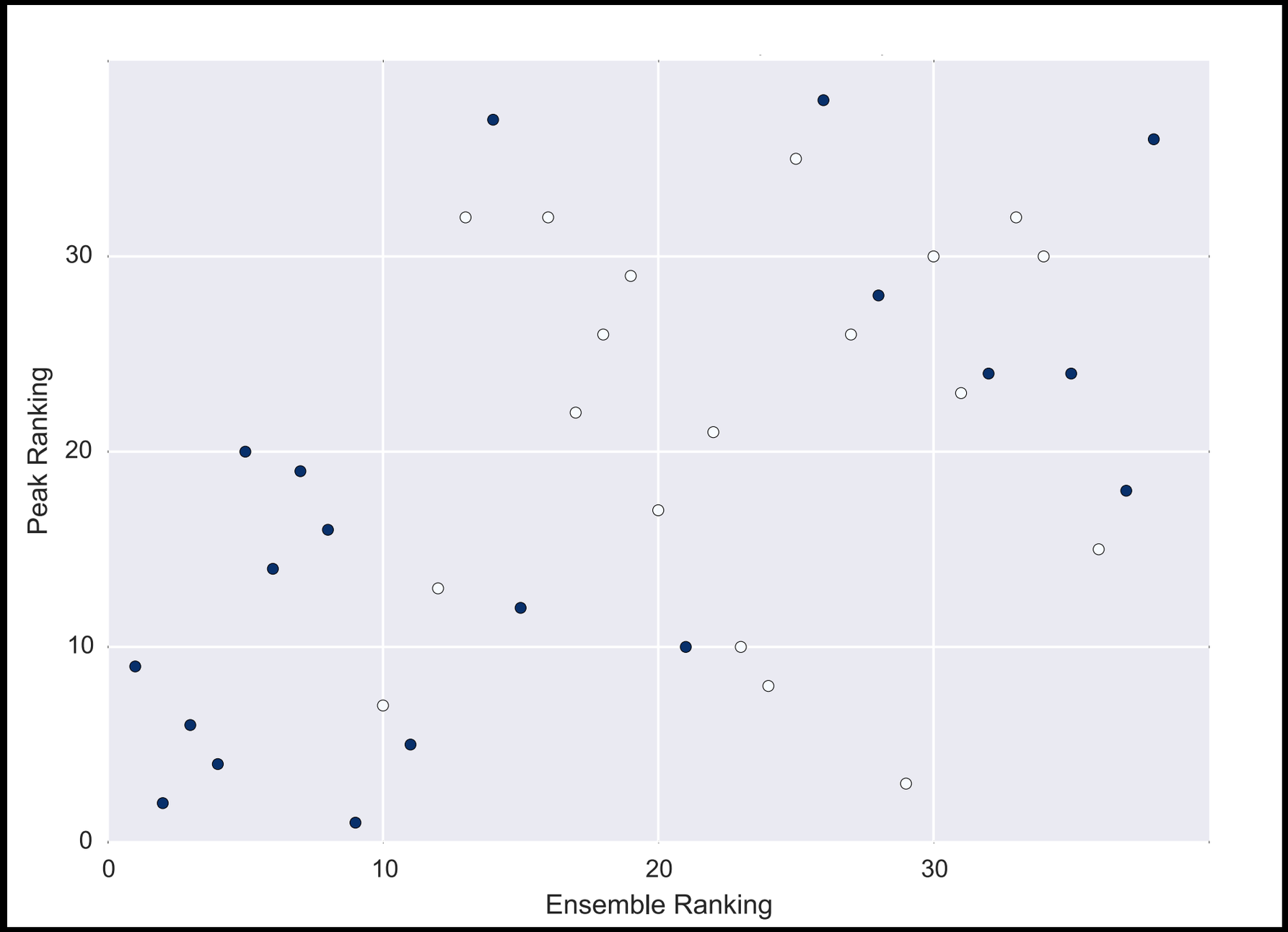


Comparing accuracy as a function of ranking (Head to head)

Concentrating on new ENCODE phase 2 (2015) results

Pros - Prospective rather than retrospective - not trained for this data.
Cons - Very few data points and results are bound to be noisy.

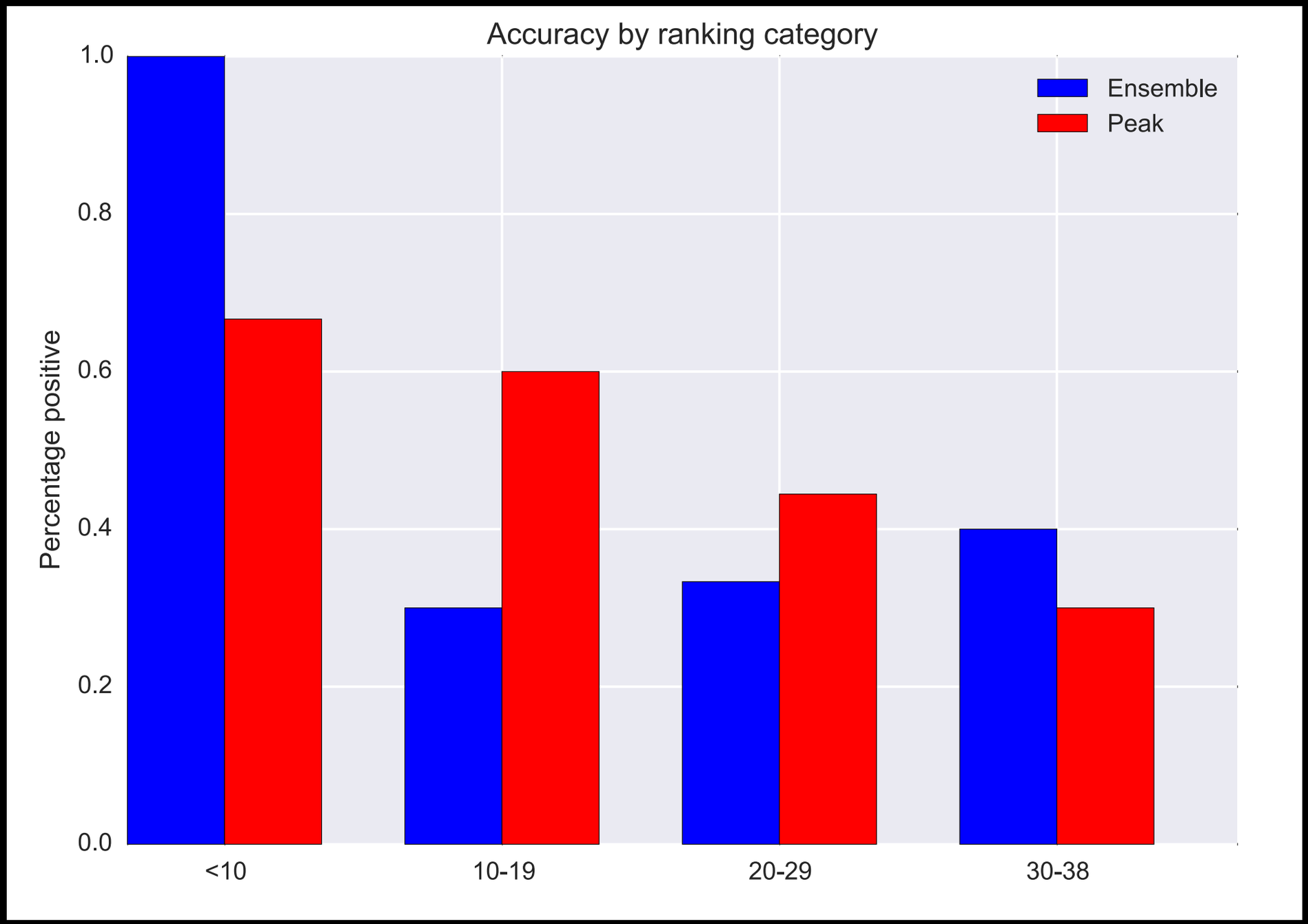
Comparing ranking of ENCODE phase 2 (2015) dataset



positives - filled circles
negatives - empty circles

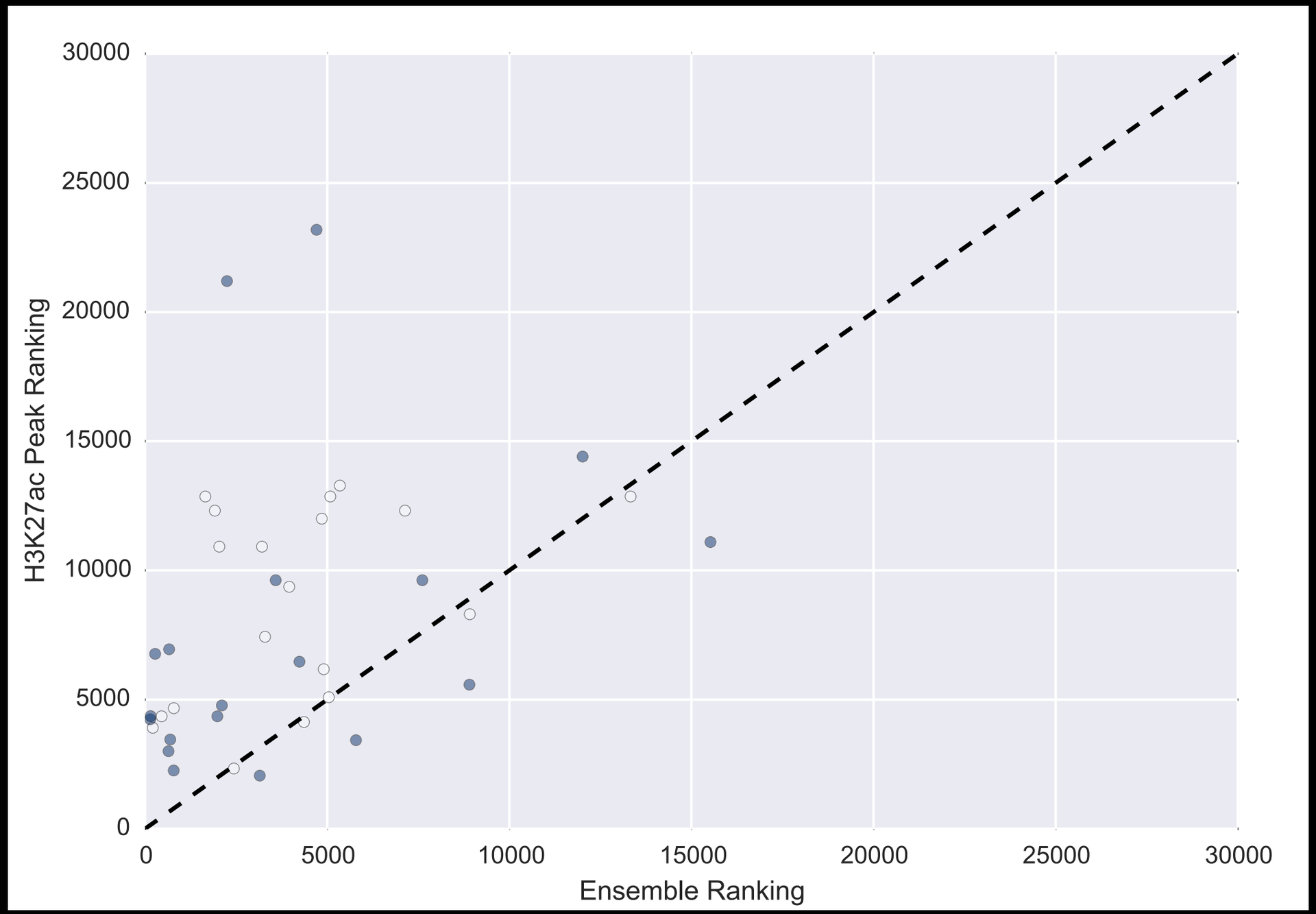
Split in to 4 bins based on ranking (grids)

Does accuracy reduce with ranking - ENCODE phase 2 (2015) dataset

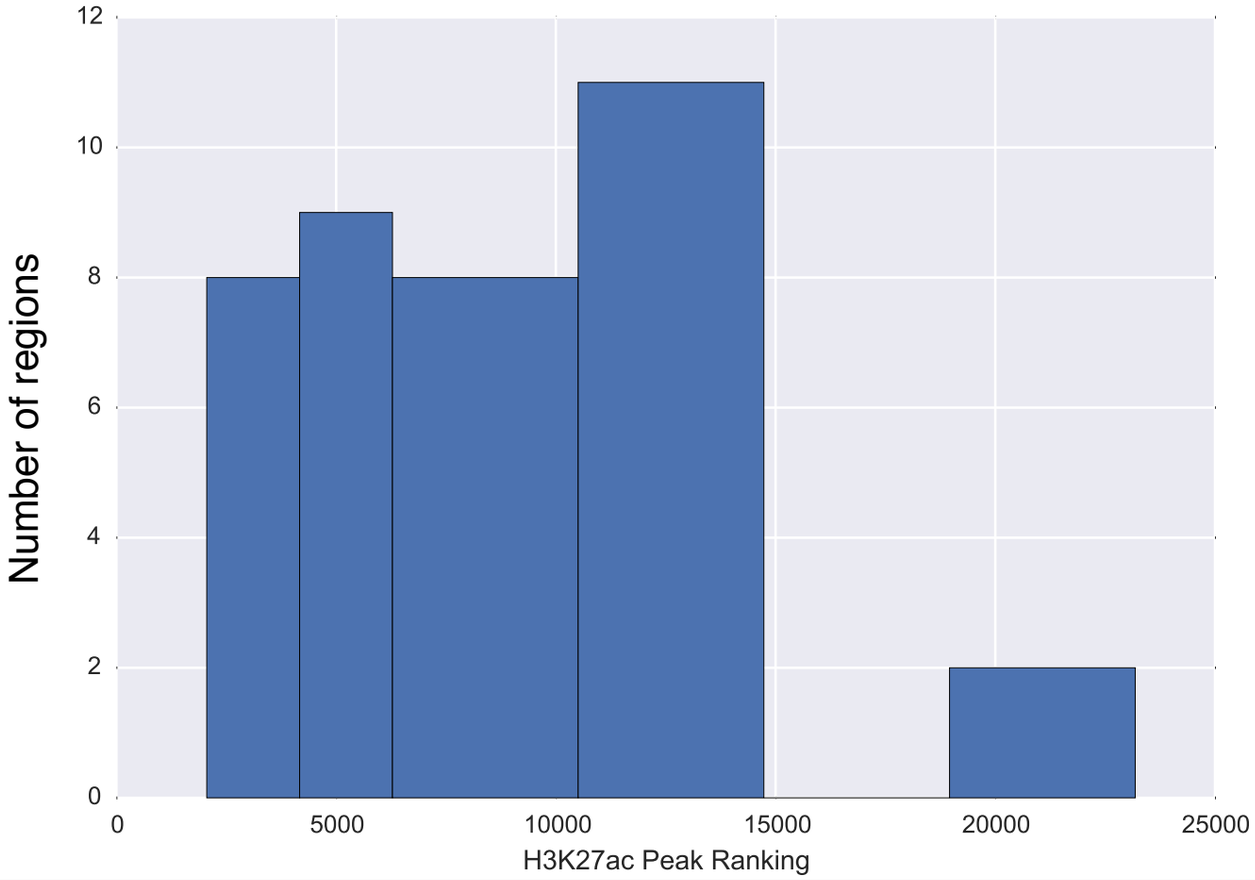


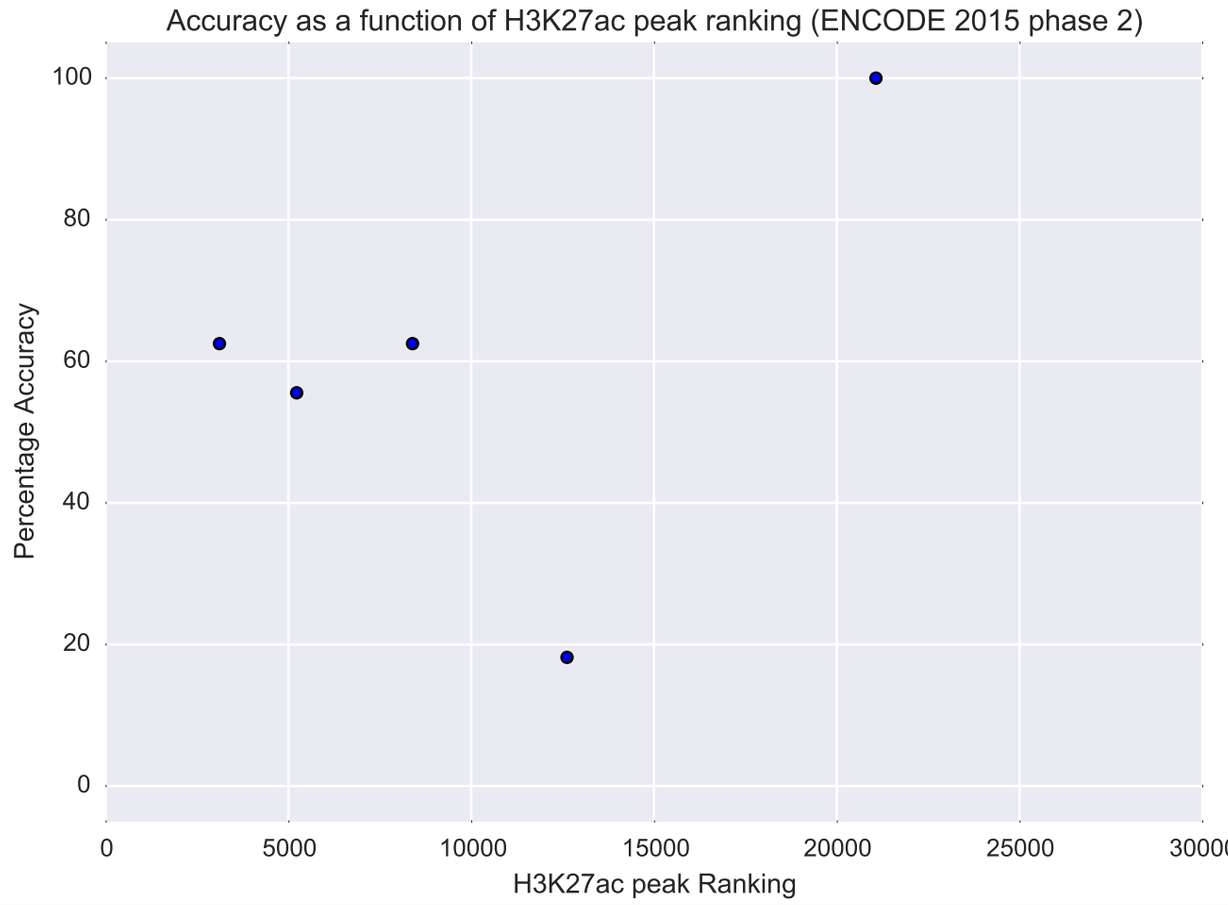
Higher accuracy for highest ranked regions by Ensemble method

Comparing ranking of ENCODE phase 2 regions (full ranking



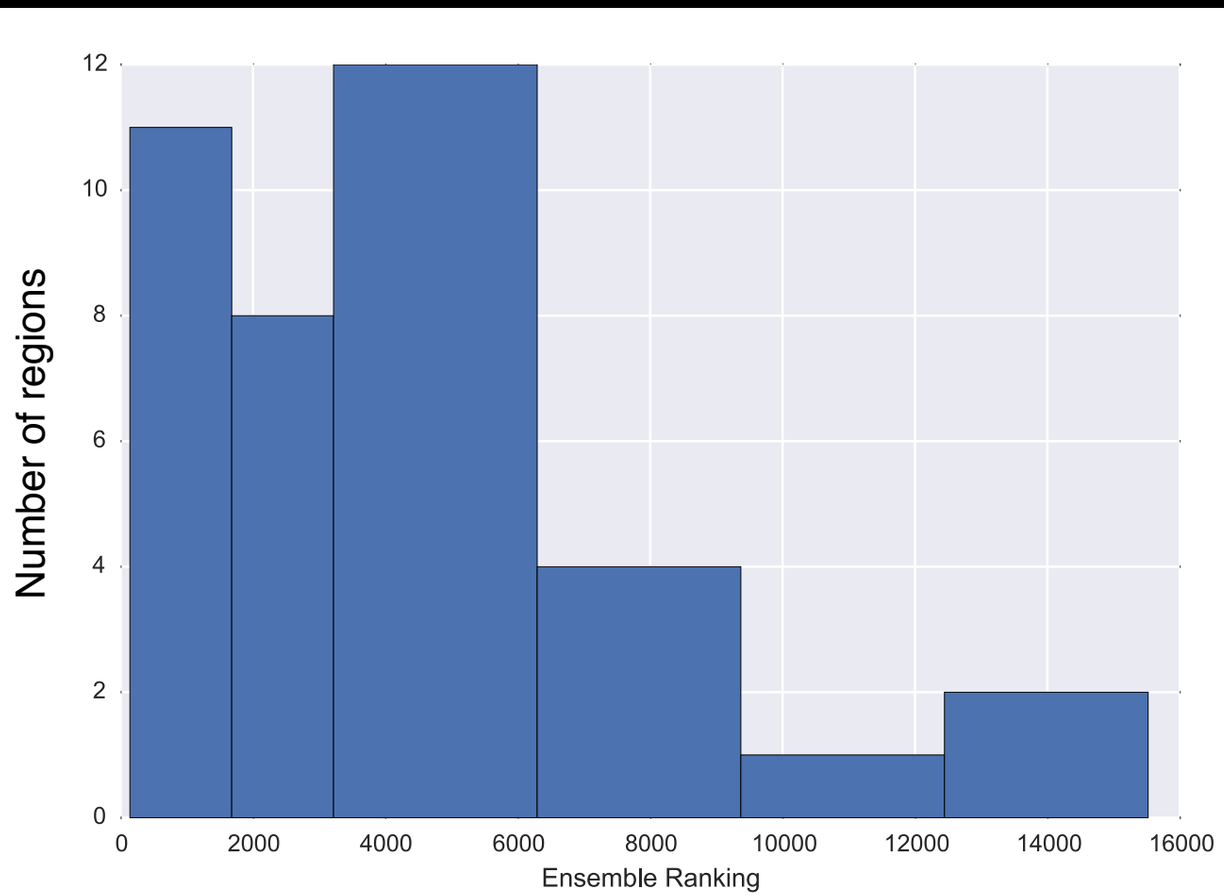
Calculating accuracy of predictions as a function of H3K27ac peak ranking (ENCODE 2015)

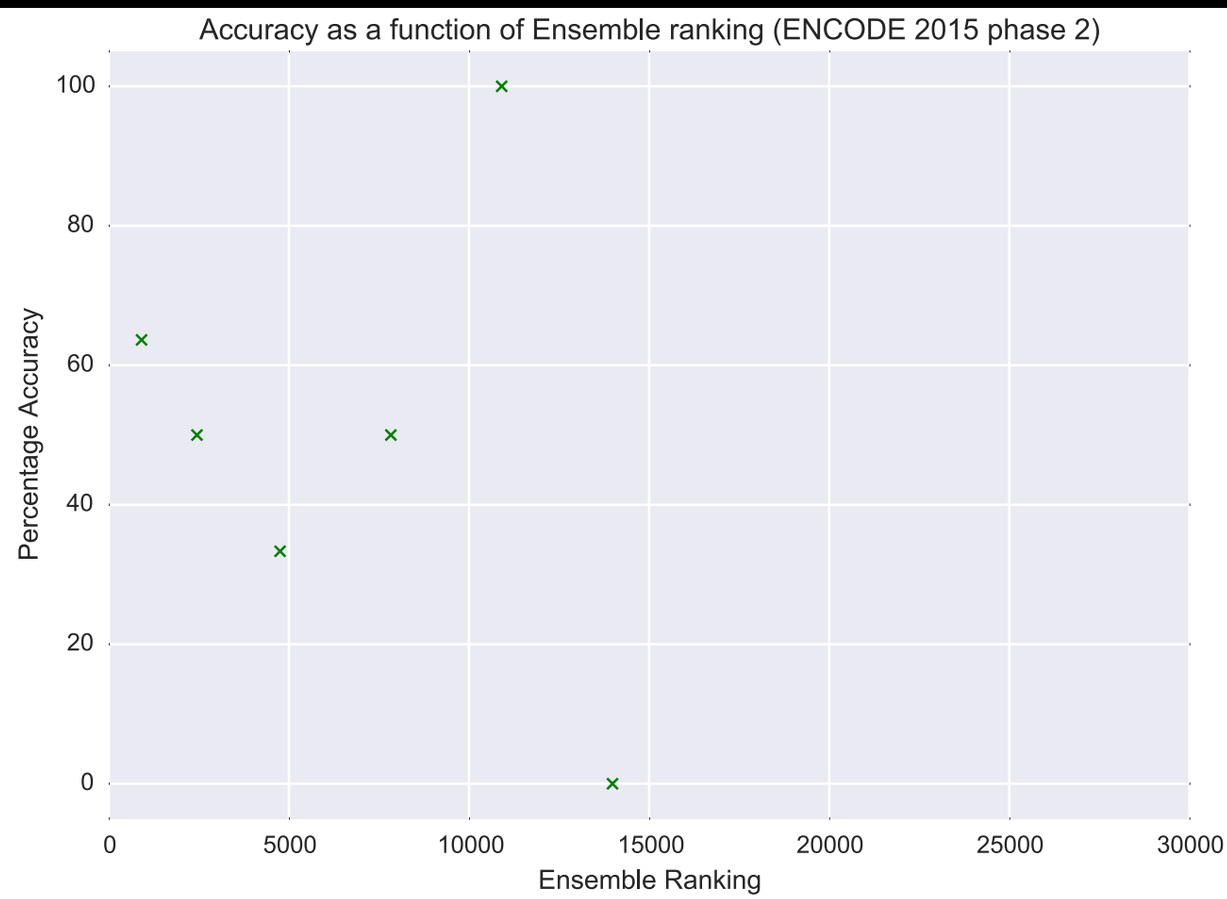




Too noisy to make conclusions except that some of the best ranking predictions are good

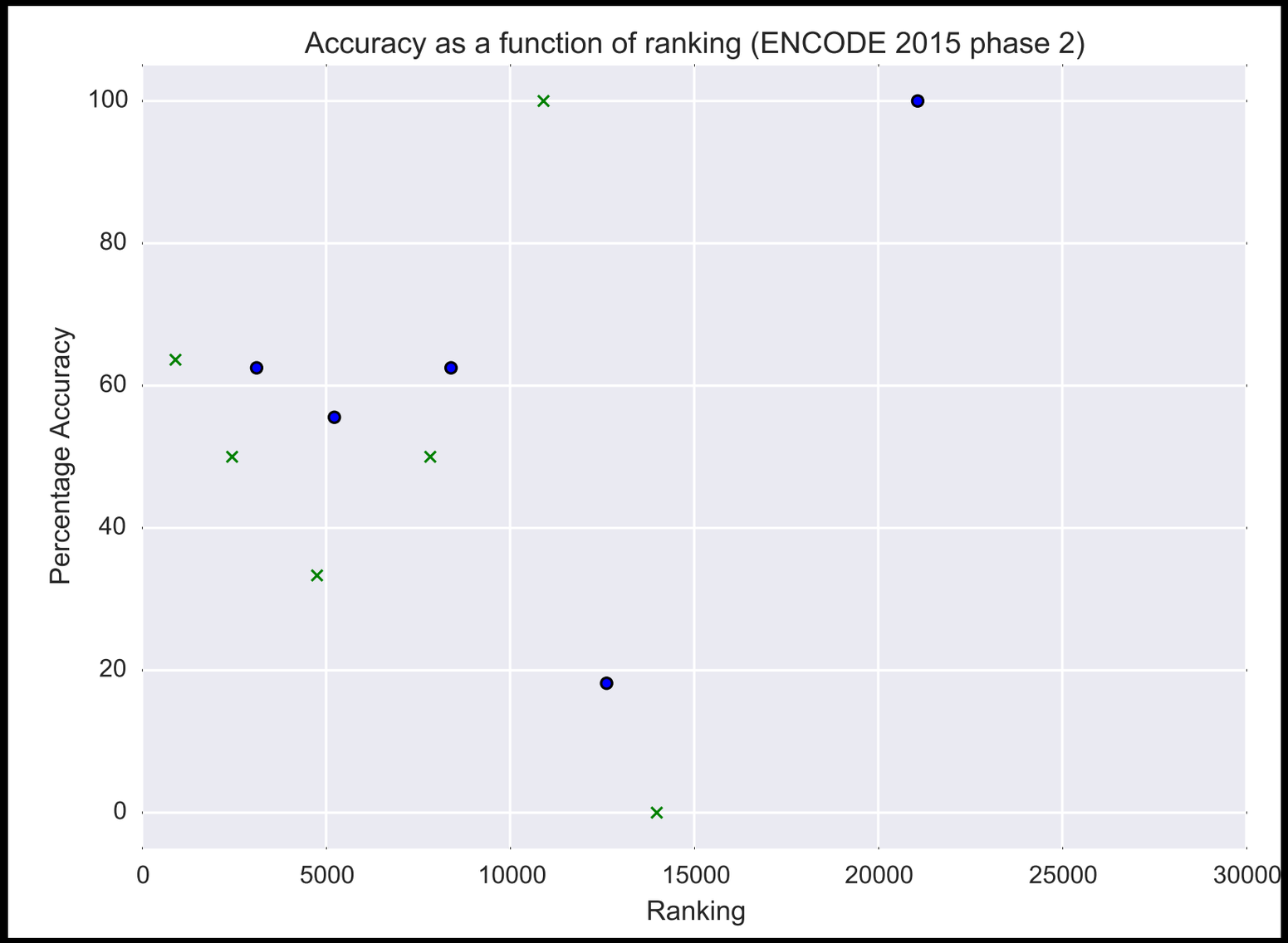
Calculating accuracy of predictions as a function of ensemble ranking (ENCODE 2015)





Too noisy to make conclusions except that some of the best ranking predictions are good

Comparing accuracy as a function of ranking (Head to head - ENCODE 2015)



Too noisy to make conclusions except that some of the best ranking predictions (both methods) are good

Conclusions:

The ensemble method seems to have higher accuracy in the higher ranked elements.

The ensemble method has higher accuracy than H3K27ac peaks in the higher ranked elements.

We can find a few elements to test experimentally (highly ranked by ensemble but mid ranking by H3K27ac and vice versa) to have a head-to-head comparison between the two methods.