Enhancer Predictions for the Encyclopedia - II

Anurag Sethi

AWG call

October 2015

Part I

Which histone marks help predict enhancers the most?

Plan is to estimate this without training.

Tested over the VISTA database

Forebrain - signal based models - H3K27ac and/or p300/other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.79	0.42
P300 signal	0.49	0.16
H3K27ac + P300 signal	0.79	0.42
H3K4me1 signal	0.70	0.26
H3K27ac + H3K4me1 signal	0.78	0.41
H3K4me2 signal	0.73	0.29
H3K27ac + H3K4me2 signal	0.79	0.42
H3K4me3 signal	0.68	0.42
H3K27ac + H3K4me3 signal	0.79	0.44
H3K9ac signal	0.64	0.23
H3K27ac + H3K9ac	0.81	0.46
GC	0.45	0.15
H3K27ac + GC	0.79	0.42
Null model	0.50	0.17

Tested over the VISTA database

Heart - signal based models - H3K27ac and/or p300/other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.90	0.48
P300 signal	0.58	0.10
H3K27ac + P300 signal	0.90	0.48
H3K4me1 signal	0.85	0.31
H3K27ac + H3K4me1 signal	0.89	0.48
H3K4me2 signal	0.80	0.16
H3K27ac + H3K4me2 signal	0.90	0.51
H3K4me3 signal	0.75	0.13
H3K27ac + H3K4me3 signal	0.90	0.50
H3K9ac signal	0.83	0.24
H3K27ac + H3K9ac	0.89	0.51
GC	0.54	0.10
H3K27ac + GC	0.90	0.48
Null model	0.50	0.07

Tested over the VISTA database

Midbrain - signal based models - H3K27ac and/or p300/other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.77	0.42
P300 signal	0.55	0.18
H3K27ac + P300 signal	0.77	0.41
H3K4me1 signal	0.75	0.30
H3K27ac + H3K4me1 signal	0.78	0.40
H3K4me2 signal	0.74	0.29
H3K27ac + H3K4me2 signal	0.77	0.42
H3K4me3 signal	0.69	0.25
H3K27ac + H3K4me3 signal	0.77	0.43
H3K9ac signal	0.69	0.26
H3K27ac + H3K9ac	0.78	0.44
GC	0.46	0.15
H3K27ac + GC	0.76	0.41
Null model	0.50	0.15

Tested over the VISTA database

Limb - signal based models - H3K27ac and/or p300/other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.78	0.28
P300 signal	0.57	0.15
H3K27ac + P300 signal	0.77	0.28
H3K4me1 signal	0.74	0.21
H3K27ac + H3K4me1 signal	0.77	0.21
H3K4me2 signal	0.69	0.16
H3K27ac + H3K4me2 signal	0.78	0.29
H3K4me3 signal	0.61	0.13
H3K27ac + H3K4me3 signal	0.79	0.30
H3K9ac signal	0.63	0.14
H3K27ac + H3K9ac	0.80	0.32
GC	0.44	0.28
H3K27ac + GC	0.78	0.28
Null model	0.50	0.10

Tested over the VISTA database

Hindbrain - signal based models - H3K27ac and/or other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.71	0.27
H3K4me1 signal	0.65	0.20
H3K27ac + H3K4me1 signal	0.71	0.26
H3K4me2 signal	0.67	0.22
H3K27ac + H3K4me2 signal	0.71	0.28
H3K4me3 signal	0.63	0.19
H3K27ac + H3K4me3 signal	0.71	0.28
H3K9ac signal	0.63	0.19
H3K27ac + H3K9ac	0.72	0.30
GC	0.48	0.13
H3K27ac + GC	0.70	0.27
Null model	0.50	0.13

Tested over the VISTA database

Neural tube - signal based models - H3K27ac and/or other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.72	0.21
H3K4me1 signal	0.67	0.14
H3K27ac + H3K4me1 signal	0.72	0.21
H3K4me2 signal	0.67	0.15
H3K27ac + H3K4me2 signal	0.71	0.22
H3K4me3 signal	0.63	0.13
H3K27ac + H3K4me3 signal	0.71	0.22
H3K9ac signal	0.63	0.14
H3K27ac + H3K9ac	0.72	0.23
GC	0.48	0.09
H3K27ac + GC	0.71	0.21
Null model	0.50	0.09

Tested over the Cross Validation dataset

Forebrain - signal based models - H3K27ac and/or p300/other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.69	0.64
P300 signal	0.61	0.66
H3K27ac + P300 signal	0.71	0.69
H3K4me1 signal	0.35	0.39
H3K27ac + H3K4me1 signal	0.78	0.79
H3K4me2 signal	0.26	0.40
H3K27ac + H3K4me2 signal	0.84	0.79
H3K4me3 signal	0.32	0.38
H3K27ac + H3K4me3 signal	0.80	0.73
H3K9ac signal	0.35	0.39
H3K27ac + H3K9ac	0.80	0.82
GC	0.49	0.49
H3K27ac + GC	0.71	0.65
Null model	0.50	0.49

Tested over the Cross Validation dataset

Heart - signal based models - H3K27ac and/or p300/other histone marks

Method	AUROC	AUPR
H3K27ac signal	0.54	0.33
P300 signal	0.40	0.21
H3K27ac + P300 signal	0.65	0.46
H3K4me1 signal	0.53	0.25
H3K27ac + H3K4me1 signal	0.53	0.42
H3K4me2 signal	0.51	0.24
H3K27ac + H3K4me2 signal	0.61	0.31
H3K4me3 signal	0.60	0.28
H3K27ac + H3K4me3 signal	0.55	0.27
H3K9ac signal	0.53	0.29
H3K27ac + H3K9ac	0.58	0.48
GC	0.26	0.26
H3K27ac + GC	0.54	0.33
Null model	0.50	0.26

Part I Conclusion

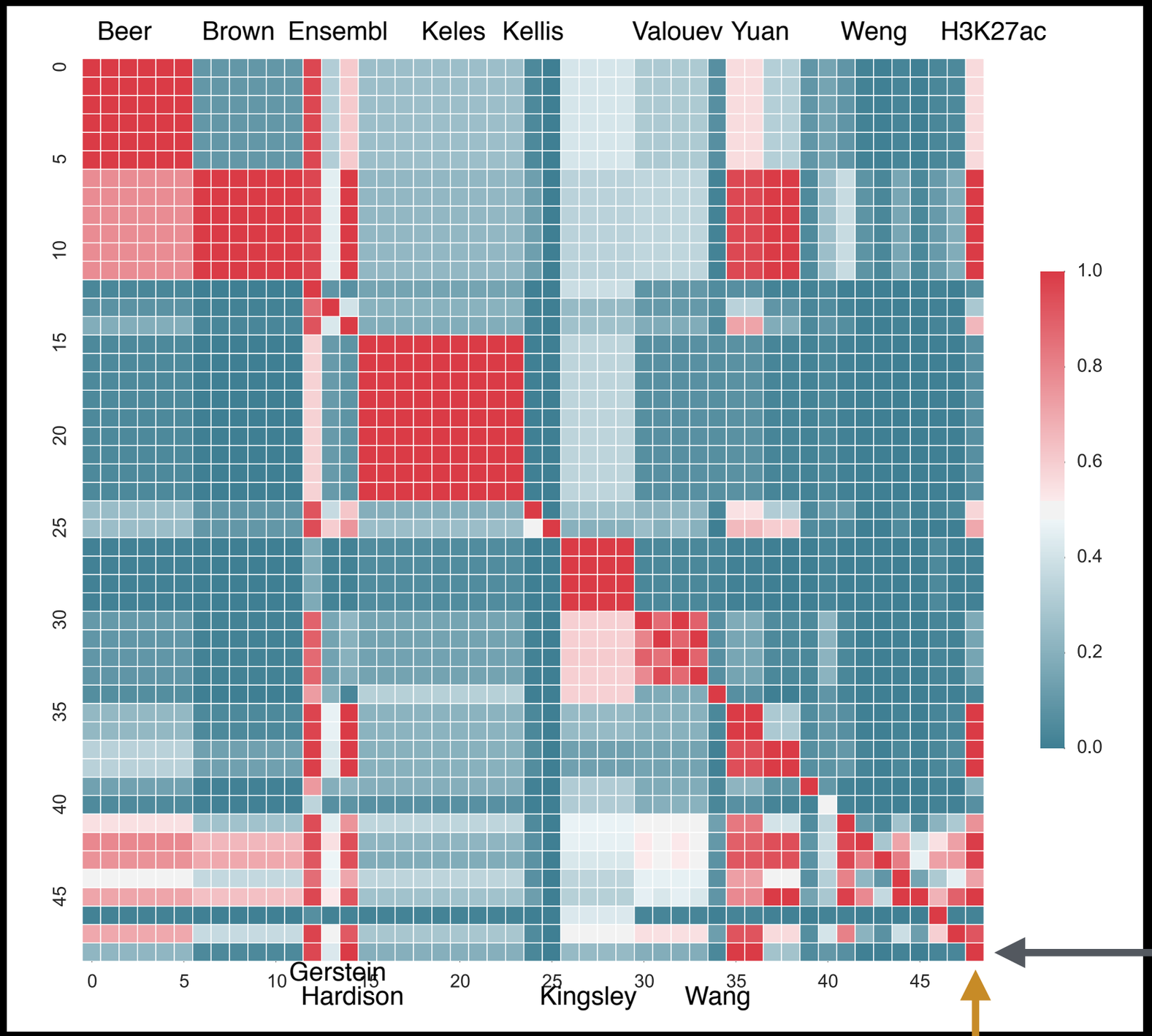
H3K27ac signal is most directly related to enhancer activity in a tissue-specific manner. It gives a AUROC of approximately 0.75 to 0.9 and AUPR of 0.45 to 0.6. There is room for improving the AUPR of these models and complex models that integrate these histone marks may perform better than simple linear regression-based models.

H3K9ac is the second most valuable histone mark and may make enhancer predictions more precise (but too little an increment) in combination with H3K27ac.

Part II

How do enhancer predictions from different people compare with H3K27ac peaks?

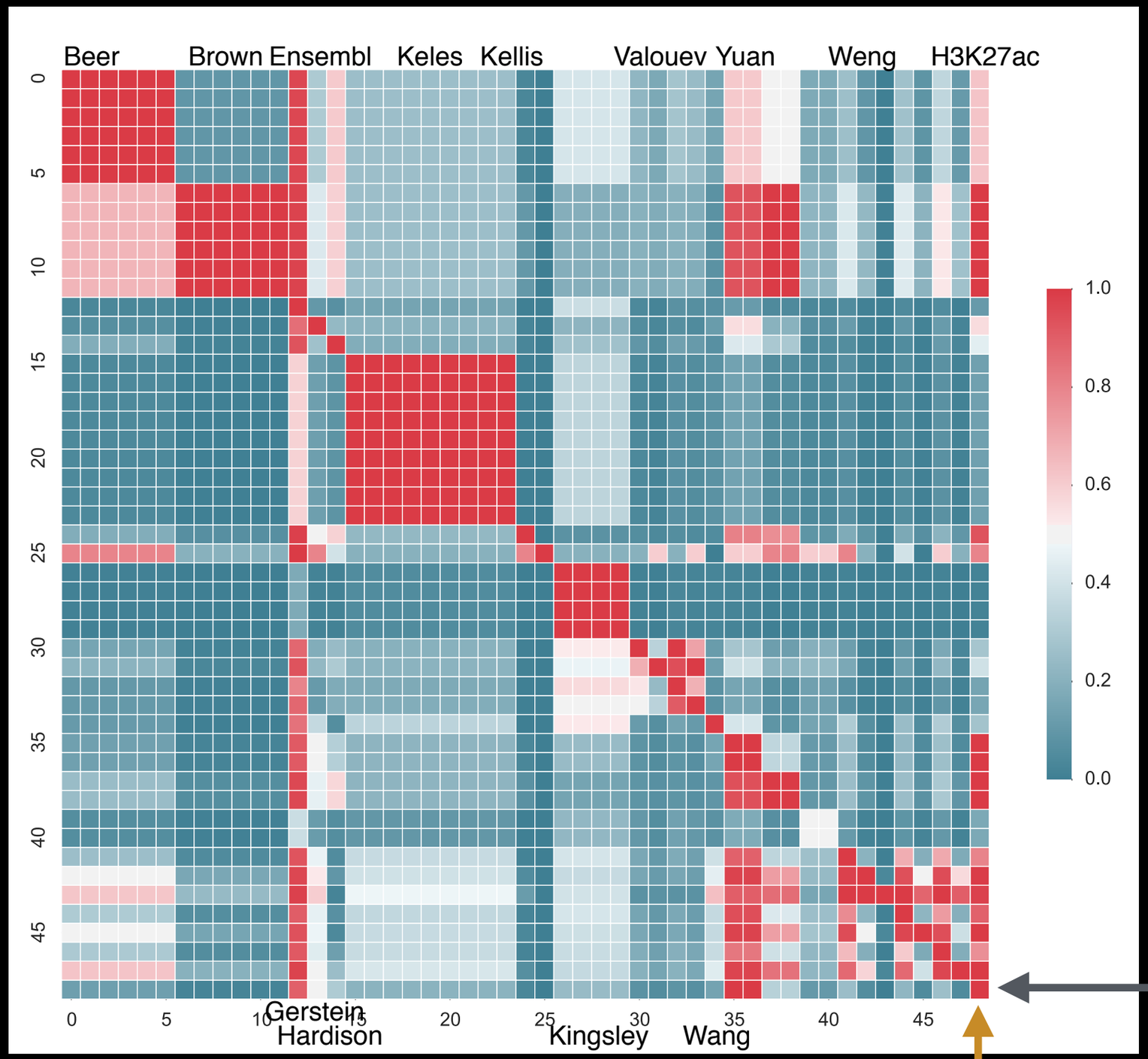
Overlap between different methods and H3K27ac peaks - forebrain



Fraction of H3K27ac peaks predicted to be an enhancer

Fraction of predictions overlapping H3K27ac peaks

Overlap between different methods and H3K27ac peaks - heart



Fraction of H3K27ac peaks predicted to be an enhancer

Fraction of predictions overlapping H3K27ac peaks

Part II Conclusion

Most of the predictions overlap to a large extent with H3K27ac peaks (last column) but not all H3K27ac peaks are predicted to be enhancers by the methods (last row). The highest rankings from the unsupervised ensemble approach will most probably be dominated by H3K27ac peaks - however the rankings of these predictions will differ from H3K27ac peak ranking and may make enhancer predictions more precise/accurate. This needs to be tested in the future.

Part III

Comparison of different unsupervised methods on the cross-validation dataset

Performance on Cross-Validation dataset

Part 1 - Forebrain H3K27ac peaks - active in any tissue

Method	AUROC	AUPR
Average	0.70	0.84
Weighted average	0.66	0.75
LRA	0.58	0.74
Markov Chain	0.73	0.86
Borda Rank	0.72	0.86
Mallows Model	0.69	0.85
BT	0.72	0.86
PL	0.72	0.86
Best	0.67	0.84
Worst	0.31	0.56

Performance on Cross-Validation dataset

Part 2 - Forebrain H3K27ac peaks - active in forebrain

Method	AUROC	AUPR
Average	0.67	0.74
Weighted average	0.59	0.60
LRA	0.61	0.55
Markov Chain	0.71	0.80
Borda Rank	0.70	0.74
Mallows Model	0.71	0.80
BT	0.70	0.79
PL	0.72	0.81
Best	0.74	0.74
Worst	0.38	0.43

Performance on Cross-Validation dataset

Part 3 - Heart H3K27ac peaks - active in any tissue

Method	AUROC	AUPR
Average	0.87	0.84
Weighted average	0.58	0.61
LRA	0.66	0.68
Markov Chain	0.89	0.85
Borda Rank	0.90	0.85
Mallows Model	0.90	0.86
BT	0.90	0.86
PL	0.88	0.85
Best	0.84	0.84
Worst	0.35	0.37

Performance on Cross-Validation dataset

Part 4 - Heart H3K27ac peaks - active in heart

Method	AUROC	AUPR
Average	0.65	0.34
Weighted average	0.66	0.34
LRA	0.52	0.37
Markov Chain	0.69	0.45
Borda Rank	0.69	0.50
Mallows Model	0.69	0.47
BT	0.69	0.46
PL	0.70	0.51
Best	0.70	0.49
Worst	0.27	0.18

Part III Conclusion

The ranking aggregation methods typically outperform score-based methods. There are quiet a few unsupervised methods with similar performance on the cross-validation dataset and one of these methods can be used for the encyclopedia