

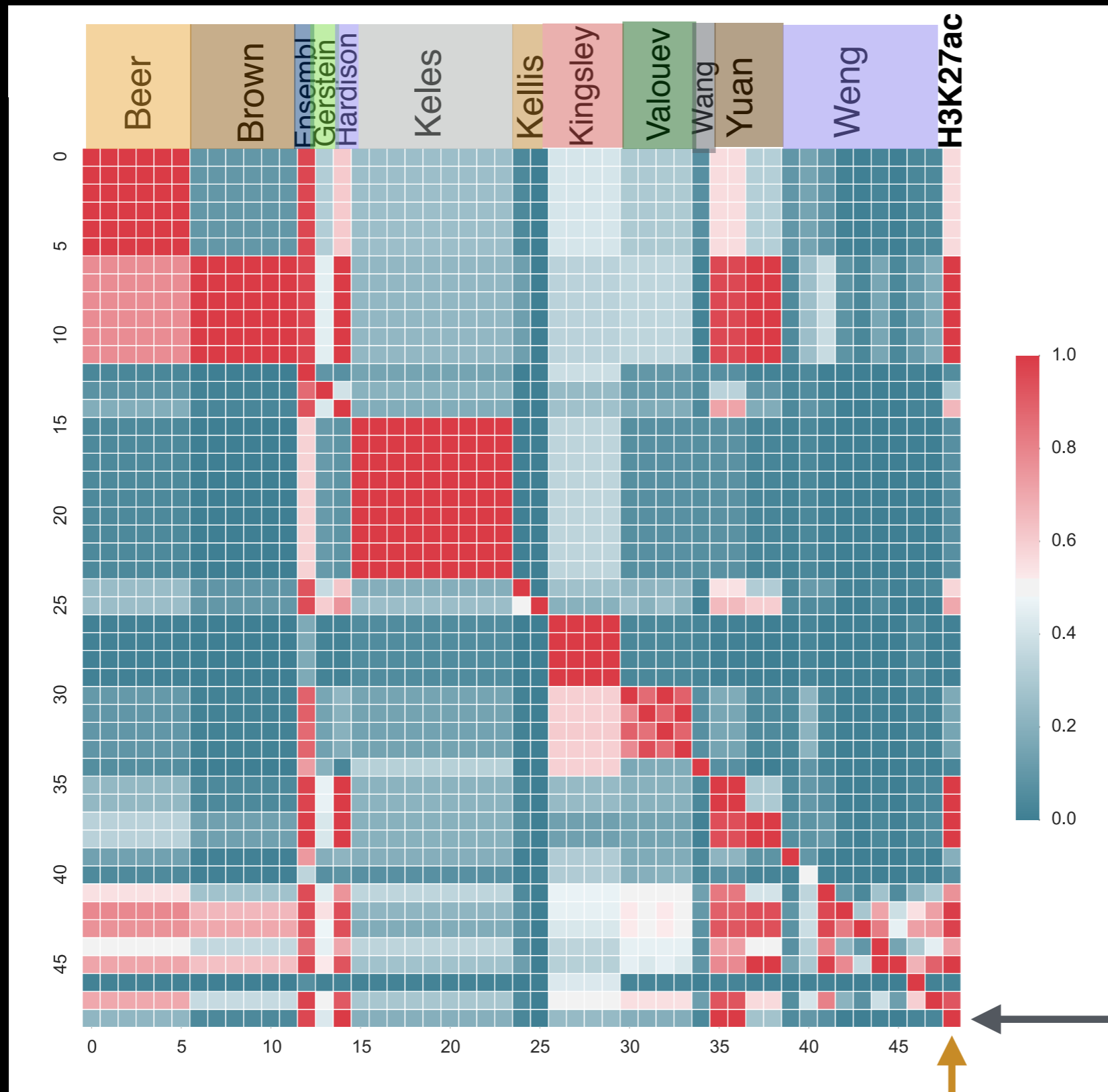
Enhancer Predictions for the Encyclopedia - part 3

Anurag Sethi (joint effort between Gerstein and Weng labs)
Functional Characterization Call

Encyclopedia Predictions in the Presence of Training Data

Similarity between Ensemble ranking and all prediction methods + H3K27ac peak ranking

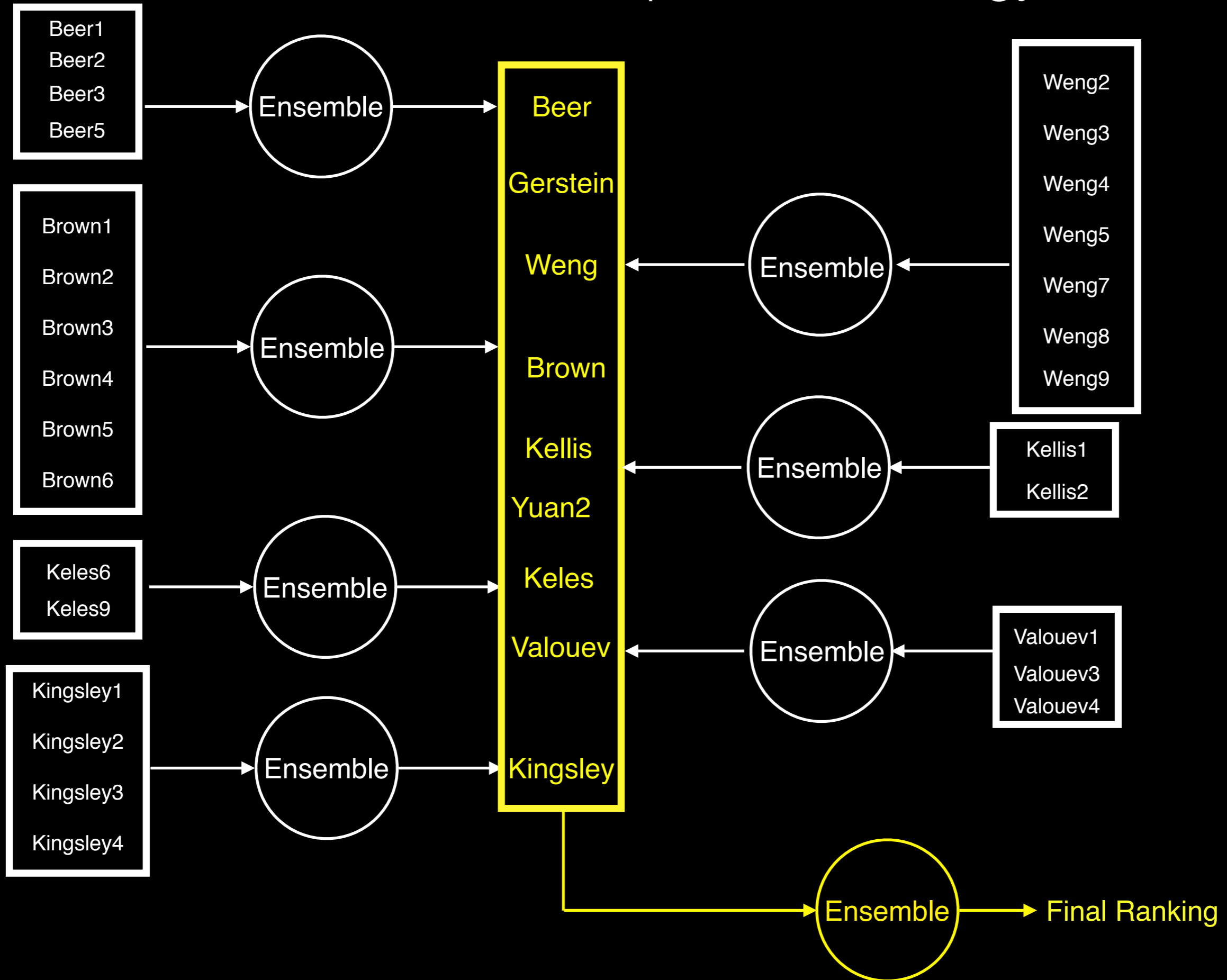
Overlap between different methods and H3K27ac peaks - forebrain



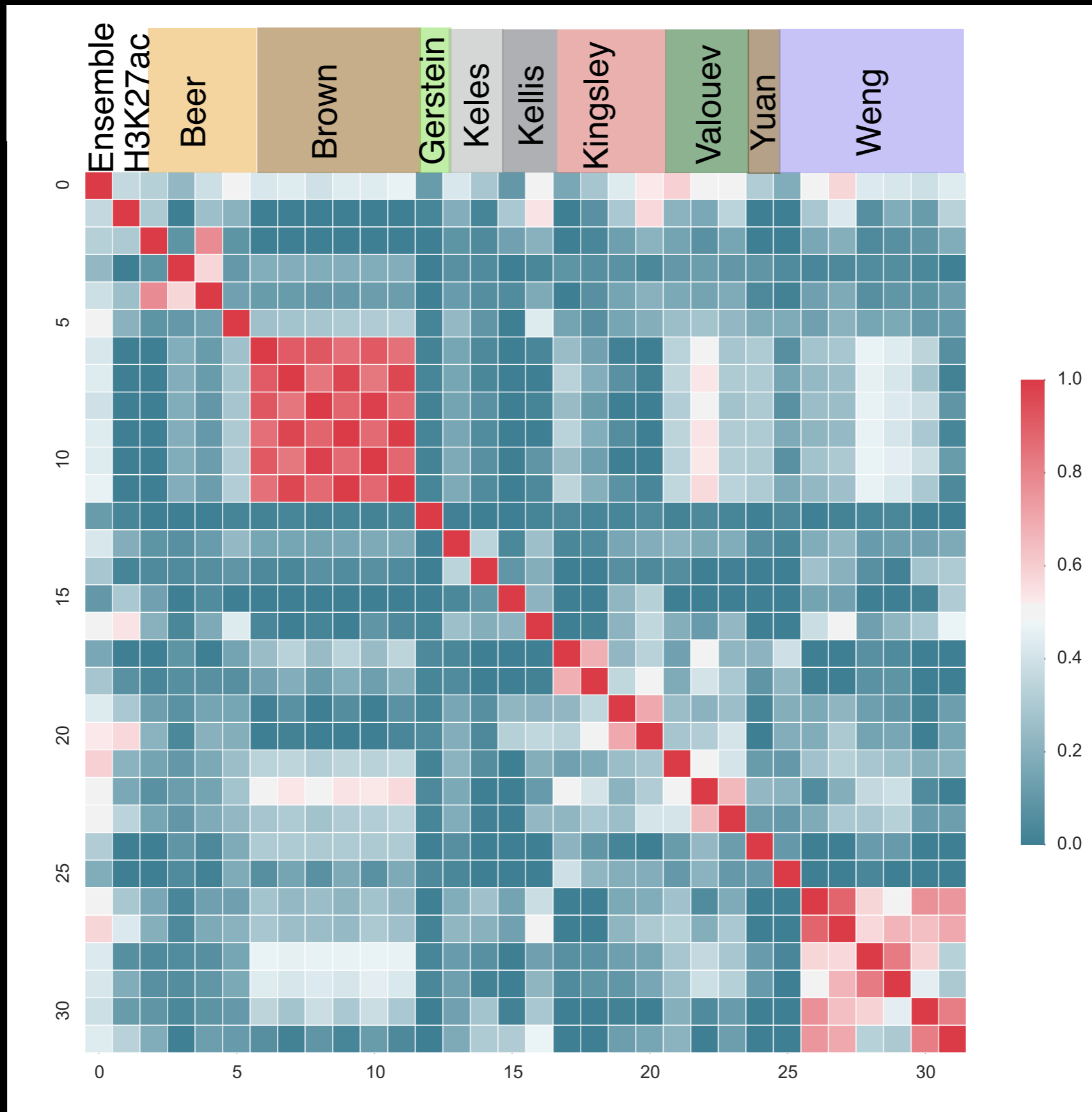
Fraction of H3K27ac peaks predicted to be an enhancer

Fraction of predictions overlapping H3K27ac peaks

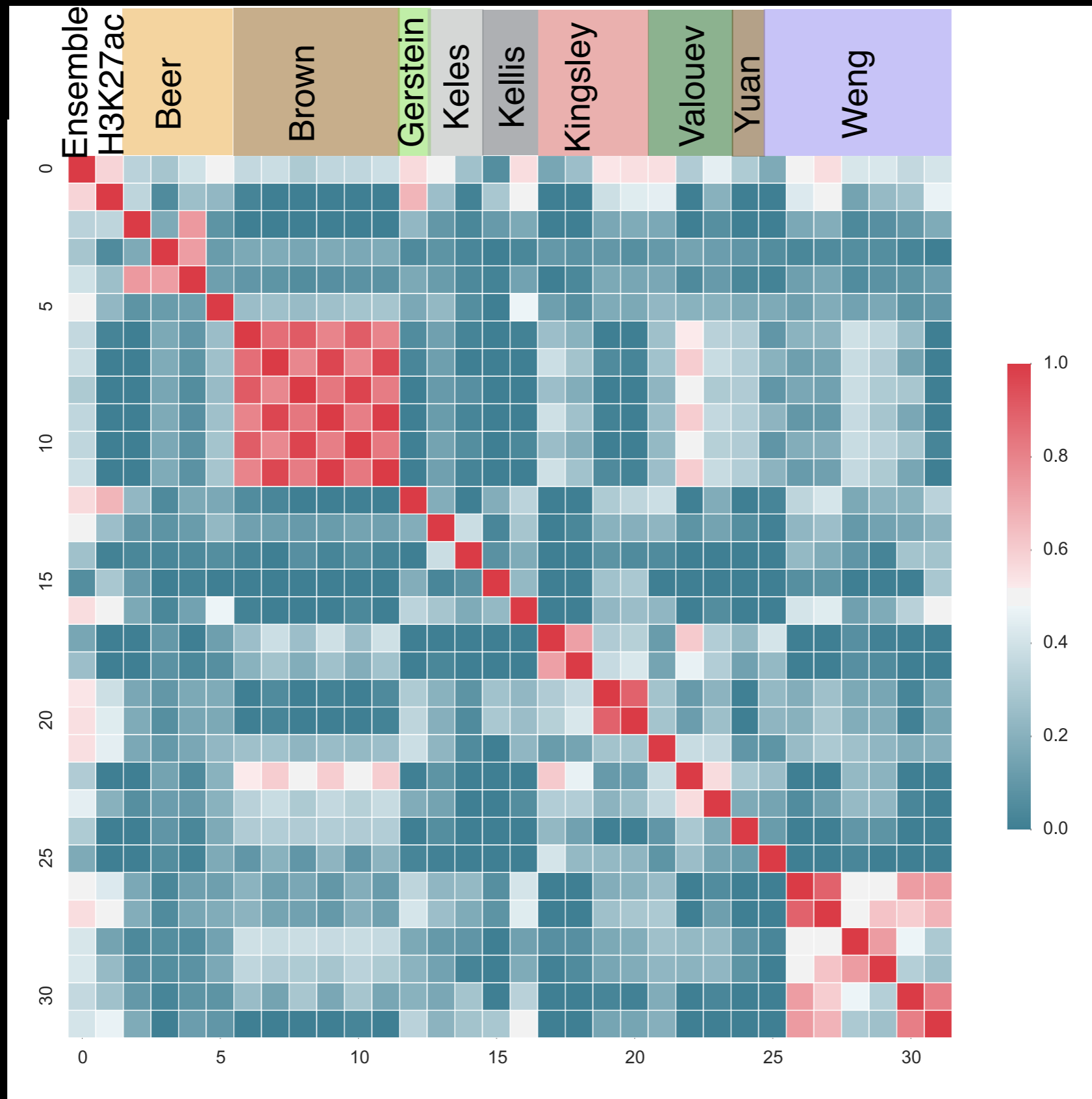
The final unsupervised strategy



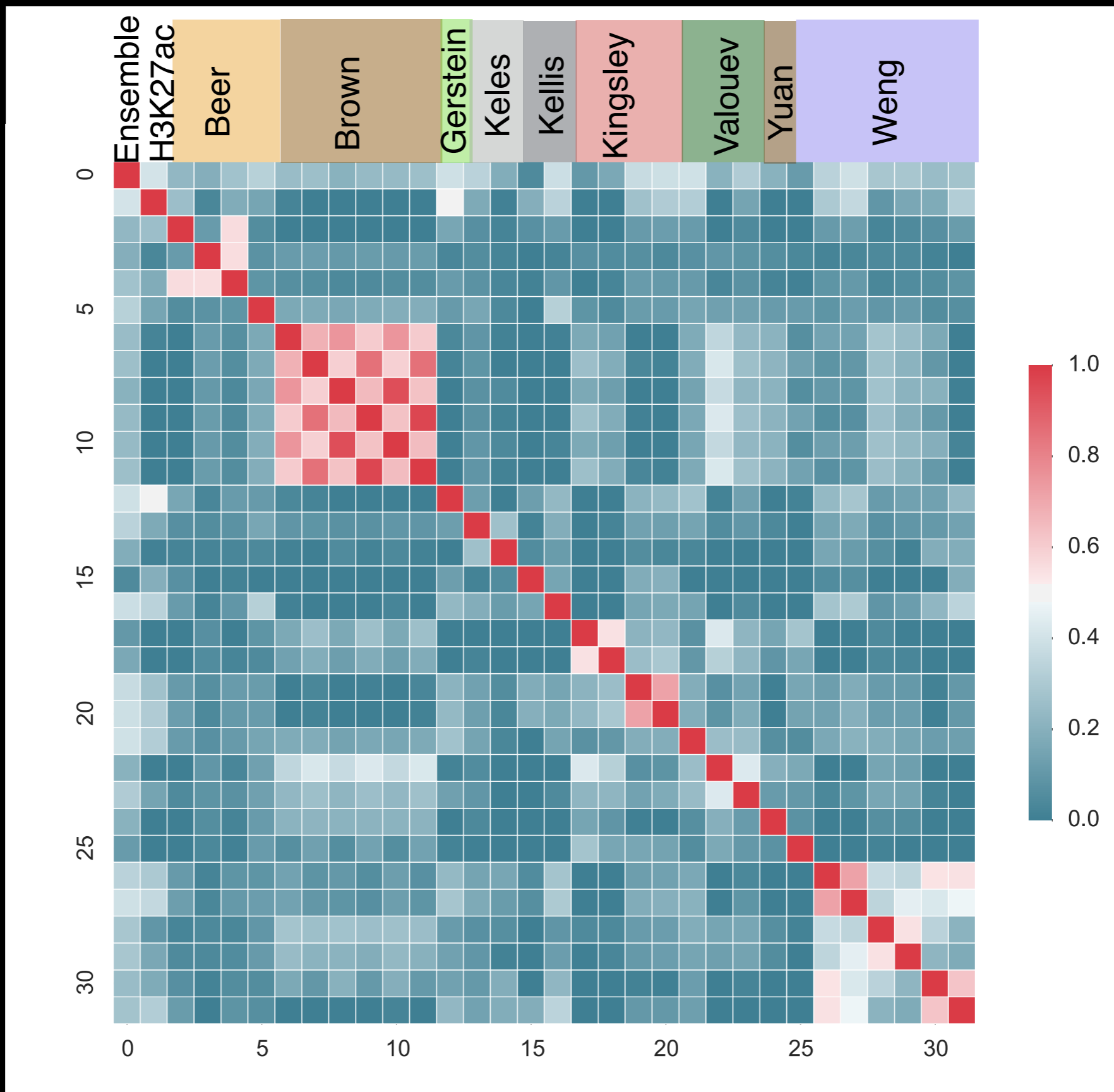
Pearson Correlation between different groups for H3K27ac peaks



Spearman rank correlation between all methods for H3K27ac peaks

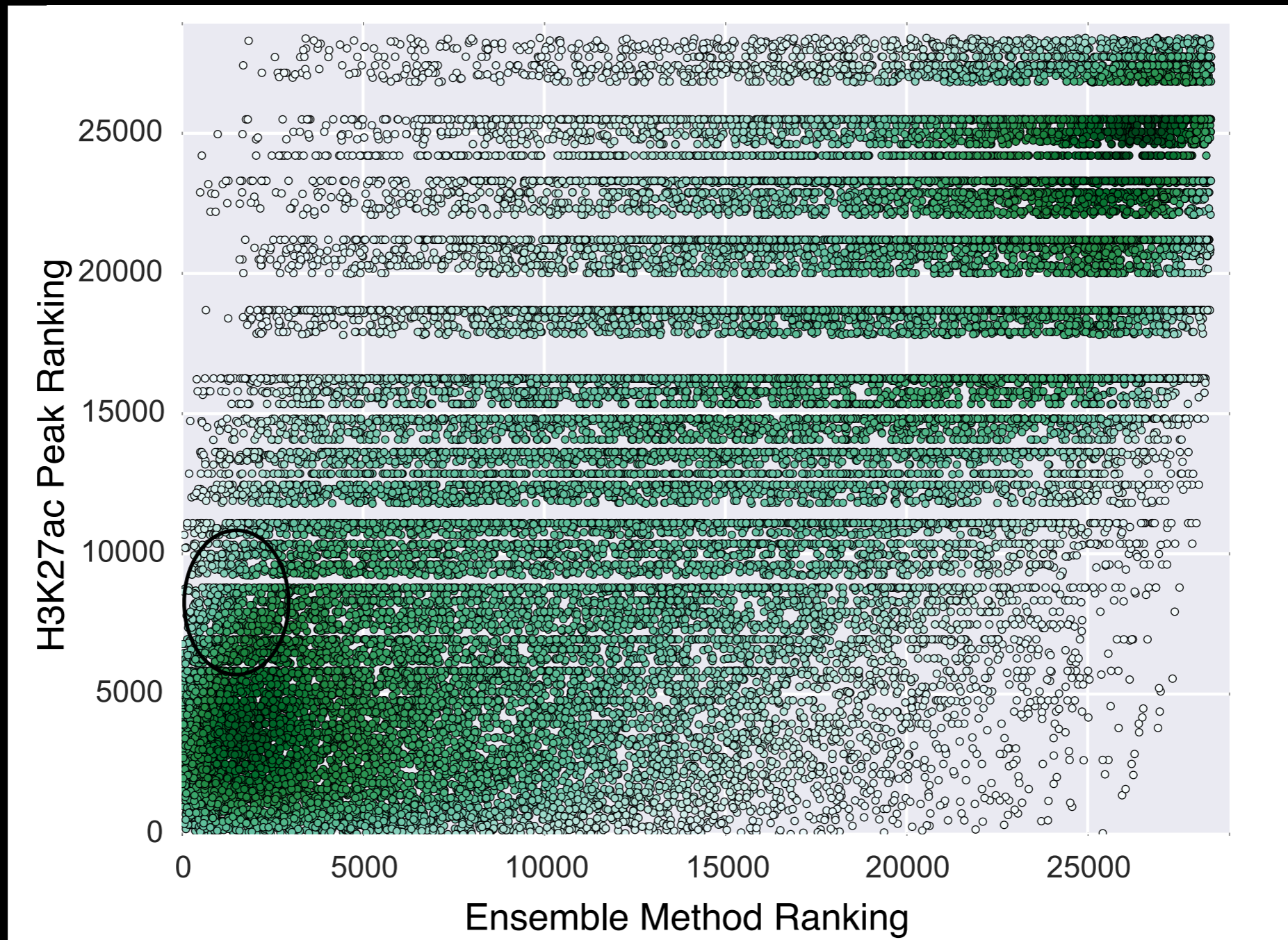


Kendall's tau correlation



Fraction of pairs with concordant ordering

Comparison of ranking between H3K27ac peaks and Ensemble method



Spearman Ranking Correlation = 0.58

Encyclopedia Predictions in the Presence of Training Data

There is significant concordance between ranking orders of H3K27ac peaks and the ensemble rankings.

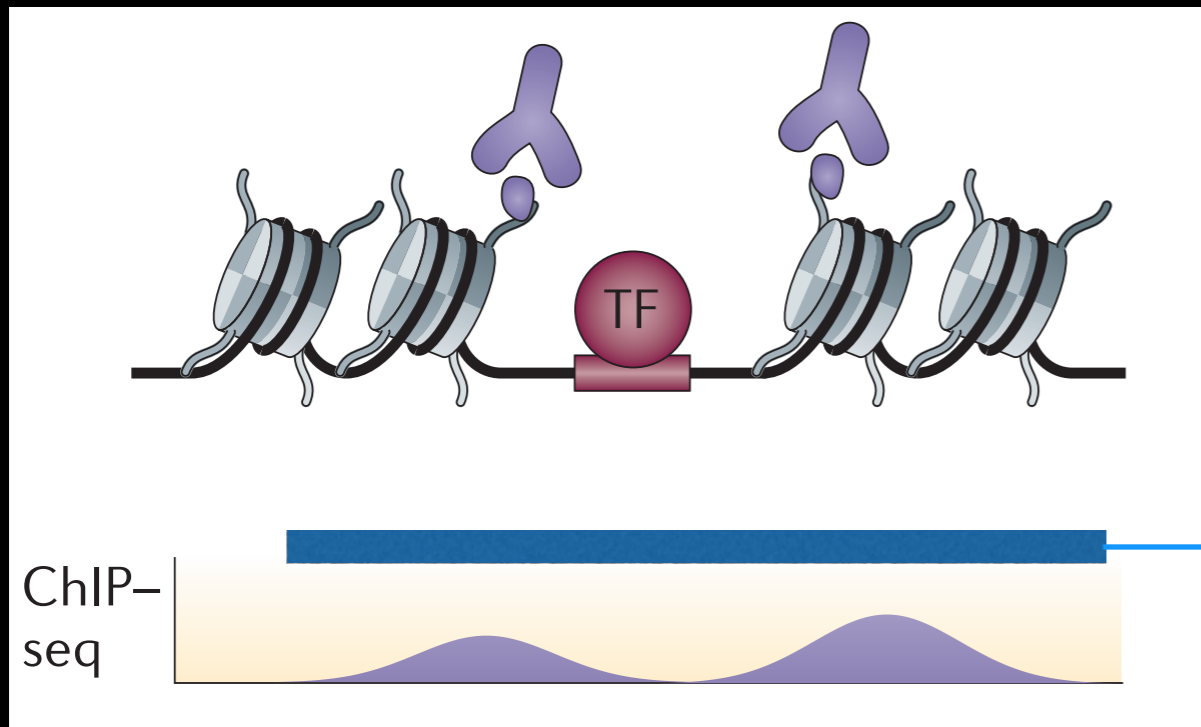
We can use the ensemble rankings to identify a few regions that can be tested experimentally in the next phase.

Encyclopedia Predictions in the Absence of Training Data

Simple Models

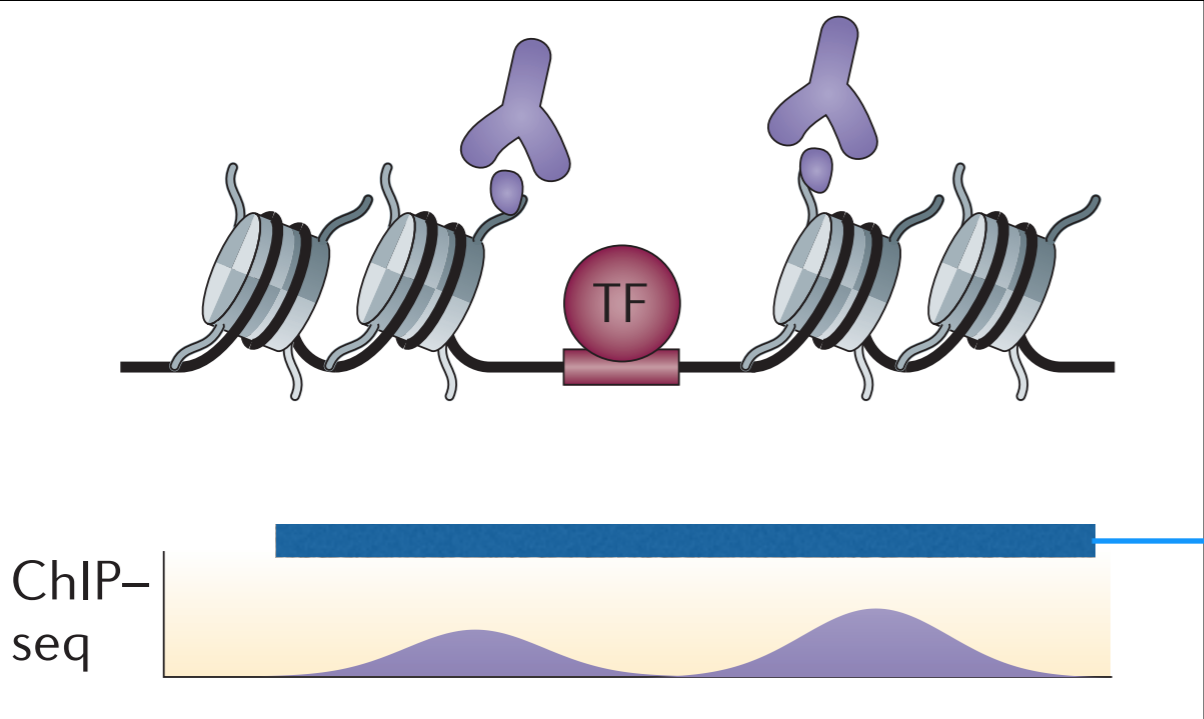
- Signal based model (average model)
- Peak based model (q-value of peak)
- Linear Regression models for combining two marks
- DNase-Seq signal is made by averaging two biological replicates (each biological replicate is actually made by combining 4-6 technical replicates).
- DNA-me is only CpG signal in one of the replicates for that tissue.

Peak-based predictions -distinguish between multiple peaks and single peak overlap



Experiment

Peak-based predictions -distinguish between multiple peaks and single peak overlap

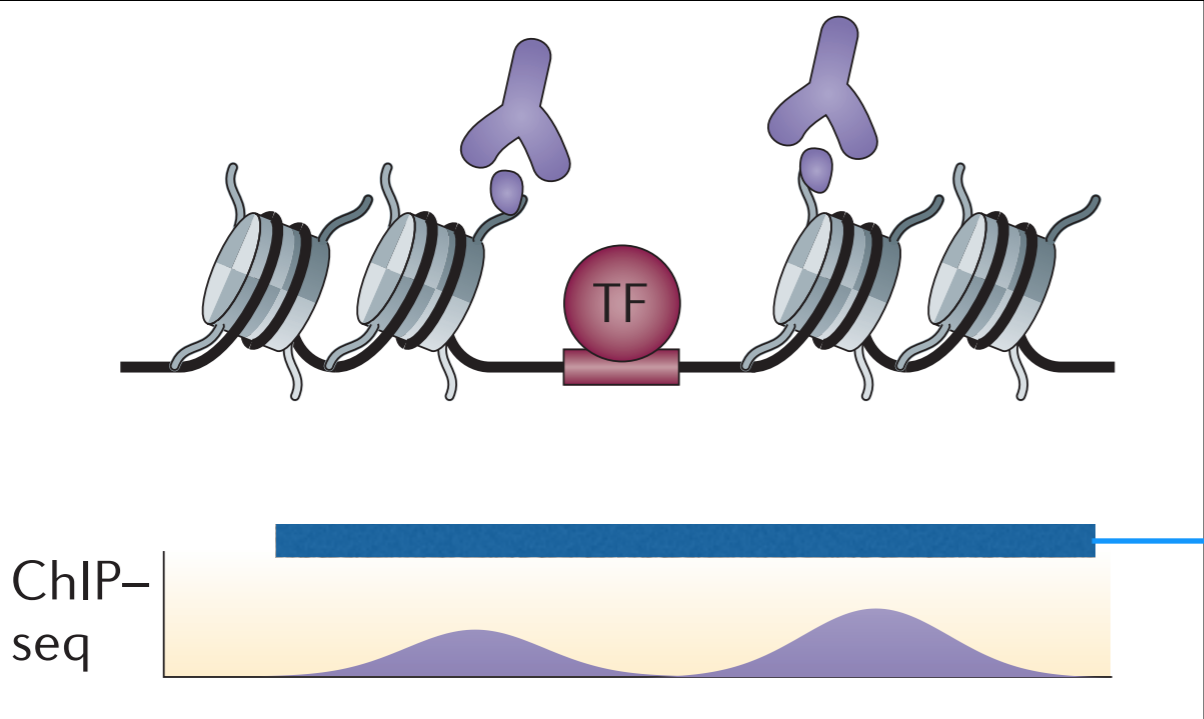


Experiment



Single Histone Peak

Peak-based predictions -distinguish between multiple peaks and single peak overlap



Experiment

Single Histone Peak

Multi Histone Peak

Accuracy of different experimental datasets to predict **forebrain** enhancers

Method	Signal		Peak - Multi(Single)	
	AUROC	AUPR	AUROC	AUPR
H3K27ac	0.81	0.45	0.73 (0.62)	0.41 (0.39)
P300 signal	0.48	0.18		
H3K27ac + P300 linear	0.81	0.45		
H3K4me1	0.72	0.28	0.66 (0.51)	0.30 (0.20)
H3K27ac + H3K4me1 lin	0.80	0.44	0.74 (0.62)	0.37 (0.38)
H3K4me2	0.74	0.31	0.65 (0.55)	0.28 (0.26)
H3K27ac + H3K4me2 lin	0.81	0.45	0.71 (0.63)	0.41 (0.37)
H3K4me3	0.70	0.27	0.55 (0.51)	0.23 (0.20)
H3K27ac + H3K4me3 lin	0.81	0.47	0.73 (0.62)	0.42 (0.39)
H3K9ac	0.65	0.24	0.58 (0.5)	0.25 (0.18)
H3K27ac + H3K9ac lin	0.83	0.50	0.74 (0.62)	0.43 (0.41)
DNA-me	0.72	0.29		
H3K27ac + DNA-me lin	0.82	0.46		
GC	0.45	0.15		
H3K27ac + GC lin	0.81	0.45		
Null model	0.50	0.16	0.50	0.16
Matched Filter			0.80	0.42

Accuracy of different experimental datasets to predict **heart** enhancers

Method	Signal		Peak - Multi(Single)	
	AUROC	AUPR	AUROC	AUPR
H3K27ac	0.89	0.50	0.86 (0.80)	0.44 (0.43)
P300 signal	0.61	0.12		
H3K27ac + P300 lin	0.89	0.50		
H3K4me1	0.81	0.20	0.80 (0.63)	0.33 (0.29)
H3K27ac + H3K4me1 lin	0.89	0.52	0.86 (0.80)	0.43 (0.42)
H3K4me2	0.76	0.17	0.72 (0.59)	0.17 (0.15)
H3K27ac + H3K4me2 lin	0.90	0.51	0.86 (0.80)	0.45 (0.44)
H3K4me3	0.76	0.13	0.57 (0.52)	0.12 (0.11)
H3K27ac + H3K4me3 lin	0.90	0.51	0.85 (0.80)	0.47 (0.45)
H3K9ac	0.83	0.27	0.74 (0.56)	0.24 (0.19)
H3K27ac + H3K9ac lin	0.89	0.52	0.85 (0.80)	0.47 (0.45)
DNA-me	0.37	0.06		
H3K27ac + DNA-me lin	0.89	0.50		
GC	0.53	0.11		
H3K27ac + GC lin	0.89	0.50		
Null model	0.50	0.07	0.50	0.07
Matched Filter			0.88	0.48

Accuracy of different experimental datasets to predict **midbrain** enhancers

Method	Signal		Peak - Multi(Single)	
	AUROC	AUPR	AUROC	AUPR
H3K27ac	0.78	0.43	0.72 (0.61)	0.41 (0.37)
P300	0.52	0.18		
H3K27ac + P300 lin	0.78	0.43		
H3K4me1	0.76	0.30	0.73 (0.56)	0.33 (0.24)
H3K27ac + H3K4me1 lin	0.79	0.41	0.75 (0.64)	0.37 (0.32)
H3K4me2	0.75	0.30	0.67 (0.57)	0.29 (0.27)
H3K27ac + H3K4me2 lin	0.78	0.43	0.70 (0.63)	0.40 (0.35)
H3K4me3	0.70	0.26	0.56 (0.52)	0.24 (0.23)
H3K27ac + H3K4me3 lin	0.78	0.434	0.71 (0.60)	0.41 (0.37)
H3K9ac	0.69	0.26	0.61 (0.53)	0.26 (0.25)
H3K27ac + H3K9ac lin	0.79	0.46	0.72 (0.60)	0.42 (0.37)
DNase	0.78	0.37	0.68 (0.52)	0.32 (0.36)
H3K27ac + DNase lin	0.79	0.43	0.74 (0.61)	0.38 (0.38)
GC	0.46	0.14		
H3K27ac + GC lin	0.77	0.41		
Null model	0.50	0.15	0.50	0.15
Matched Filter			0.78	0.42

Accuracy of different experimental datasets to predict **limb** enhancers

Method	Signal		Peak - Multi(Single)	
	AUROC	AUPR	AUROC	AUPR
H3K27ac	0.78	0.31	0.73 (0.60)	0.31 (0.27)
P300 signal	0.55	0.15		
H3K27ac + P300 lin	0.77	0.31		
H3K4me1	0.74	0.23	0.72 (0.53)	0.25 (0.16)
H3K27ac + H3K4me1 lin	0.78	0.30	0.75 (0.61)	0.27 (0.24)
H3K4me2	0.79	0.19	0.61 (0.53)	0.18 (0.15)
H3K27ac + H3K4me2 lin	0.78	0.32	0.72 (0.59)	0.32 (0.27)
H3K4me3	0.62	0.15	0.51 (0.5)	0.12 (0.11)
H3K27ac + H3K4me3 lin	0.79	0.33	0.73 (0.60)	0.33 (0.27)
H3K9ac	0.63	0.16	0.53 (0.50)	0.14 (0.12)
H3K27ac + H3K9ac lin	0.80	0.35	0.74 (0.60)	0.35 (0.29)
DNase	0.75	0.34	0.76 (0.52)	0.33 (0.26)
H3K27ac + DNase lin	0.82	0.37	0.79 (0.61)	0.34 (0.28)
DNA-me	0.64	0.16		
H3K27ac + DNA-me lin	0.78	0.31		
GC	0.44	0.10		
H3K27ac + GC lin	0.78	0.32		
Null model	0.50	0.10		
Matched Filter			0.79	0.32

Accuracy of different experimental datasets to predict **hindbrain** enhancers

Method	Signal		Peak - Multi(Single)	
	AUROC	AUPR	AUROC	AUPR
H3K27ac	0.73	0.30	0.69 (0.60)	0.29 (0.28)
H3K4me1	0.66	0.20	0.58 (0.50)	0.20 (0.18)
H3K27ac + H3K4me1 lin	0.73	0.29	0.69 (0.60)	0.26 (0.28)
H3K4me2	0.68	0.23	0.60 (0.53)	0.23 (0.20)
H3K27ac + H3K4me2 lin	0.73	0.30	0.70 (0.60)	0.29 (0.27)
H3K4me3	0.64	0.19	0.54 (0.51)	0.19 (0.17)
H3K27ac + H3K4me3 lin	0.73	0.31	0.68 (0.59)	0.30 (0.28)
H3K9ac	0.64	0.19	0.56 (0.51)	0.20 (0.17)
H3K27ac + H3K9ac lin	0.73	0.33	0.69 (0.60)	0.32 (0.29)
DNase	0.69	0.29	0.65 (0.50)	0.29 (0.26)
H3K27ac + DNase lin	0.76	0.34	0.72 (0.60)	0.31 (0.29)
GC	0.47	0.13		
H3K27ac + GC lin	0.72	0.30		
Null model	0.50	0.13	0.50	0.13
Matched Filter			0.73	0.31

Accuracy of different experimental datasets to predict **neuralTube** enhancers

Method	Signal		Peak - Multi(Single)	
	AUROC	AUPR	AUROC	AUPR
H3K27ac	0.74	0.25	0.69 (0.58)	0.24 (0.22)
H3K4me1	0.69	0.15	0.64 (0.51)	0.16 (0.11)
H3K27ac + H3K4me1 lin	0.71	0.24	0.71 (0.58)	0.21 (0.23)
H3K4me2	0.69	0.17	0.62 (0.52)	0.18 (0.13)
H3K27ac + H3K4me2 lin	0.74	0.25	0.71 (0.57)	0.24 (0.22)
H3K4me3	0.65	0.14	0.54 (0.52)	0.15 (0.13)
H3K27ac + H3K4me3 lin	0.74	0.26	0.68 (0.58)	0.25 (0.23)
H3K9ac	0.64	0.15	0.59 (0.51)	0.17 (0.14)
H3K27ac + H3K9ac lin	0.75	0.27	0.69 (0.58)	0.27 (0.24)
DNase	0.78	0.27	0.78 (0.57)	0.27 (0.22)
H3K27ac + DNase lin	0.77	0.26	0.79 (0.62)	0.28 (0.23)
GC	0.48	0.09		
H3K27ac + GC lin	0.73	0.25		
Null model	0.50	0.09	0.50	0.09
Matched Filter			0.74	0.26

Encyclopedia Predictions in the Absence of Training Data

Histone matched filters (especially double peaks) are better than peaks at identifying putative enhancers.

We are currently working on coming up with better ways to combine the matched filter scores without training data.