Evaluation of ENCODE Enhancer Challenge Predictions

Mengting Gu

Gerstein Lab
Yale University

April 23,2018

Challenge set up

- Goal:
  - To test computation algorithms for predicting enhancers that can be validated through transgenic mouse enhancer assay
- Setup:
  - For each round of challenge, a total of 150 elements in the mouse genome will be tested using transgenic mouse enhancer essay. Experimental results will include
    1. general call if an element is an *in vivo* enhancer at e12.5,
    2. list of tissues in which the enhancer showed reproducible activity,
    3. number of transgenic animals that showed the pattern for each tissue (reproducibility).

Challenge set up

	Forebrain	Midbrain	Hindbrain	Heart	Limb	
E11.5	*(last round)*	20 top tier 15 mid tier 15 low tier	20 top tier 15 mid tier 15 low tier	*(last round)*	20 top tier 15 mid tier 15 low tier	*150 elements*
E12.5	20 top tier 15 mid tier 15 low tier			20 top tier 15 mid tier 15 low tier	20 top tier 15 mid tier 15 low tier	*150 elements*
	50 elements	*50 elements*	*50 elements*	*50 elements*	*100 elements*	

By Tier:

120 top tier	*Rank **1-20** from respective list**
90 mid tier	*Rank **1501-1515** from respective list**
90 low tier	*Rank **3001-3015** from respective list**
300 total	

** if an overlapping element has been previously tested, the next available untested rank will be tested instead*

- Data
  - Uniformly processed mappings to mm10 and peak calls are being generated by the ENCODE Analysis Pipelines deployed by the ENCODE DCC. The computational groups are free to use all available ENCODE data and other publicly available data.
  - Candidate elements to be tested in each region is available Nucleotide sequence to be tested can be downloaded

Challenge set up

- Computational predictions:
  - We invite computational predictions on these regulatory elements to the DCC, prior to the release of experimental testing results.
- Submission includes:
  - Probability between 0 and 1 the element being active in the corresponding tissue
  - Probability between 0 and 1 the element being active in any tissue
  - Description of the method
- Evaluation:
  - Each set of submission will be evaluated by the DAC against the experimental results by computing the area under ROC curve (AUROC) and the area under the PR curve (AUPR).

Submitted predictions

Group	Method	Datasets
A0	RF, KNN, LR (SGD)	ChIP, Methyl, RNA
A1	RF, KNN, LR (SGD)	ChIP, Methyl, RNA
A2	RF, KNN, LR (SGD)	ChIP, Methyl, RNA, KSM
A3	RF, KNN, LR (SGD)	ChIP, Methyl, RNA, KSM
A4	LR	ChIP, Methyl, RNA, KSM
A5	RF, KNN, LR (SGD)	ChIP, Methyl, RNA
A6	RF, KNN, LR (SGD)	ChIP, Methyl, RNA, KSM
A7	RF, KNN, LR (SGD)	ChIP, Methyl, RNA, KSM
A8	RF, KNN, LR (SGD)	ChIP, Methyl, RNA
A9	RF, KNN, LR (SGD)	ChIP, Methyl, RNA
B0	gkm-SVM	DHS/H3K27ac/p300
B1	gkm-SVM	DHS/H3K27ac/p300
B2	gkm-SVM	DHS/H3K27ac/p300
B3	gkm-SVM	DHS/H3K27ac/p300

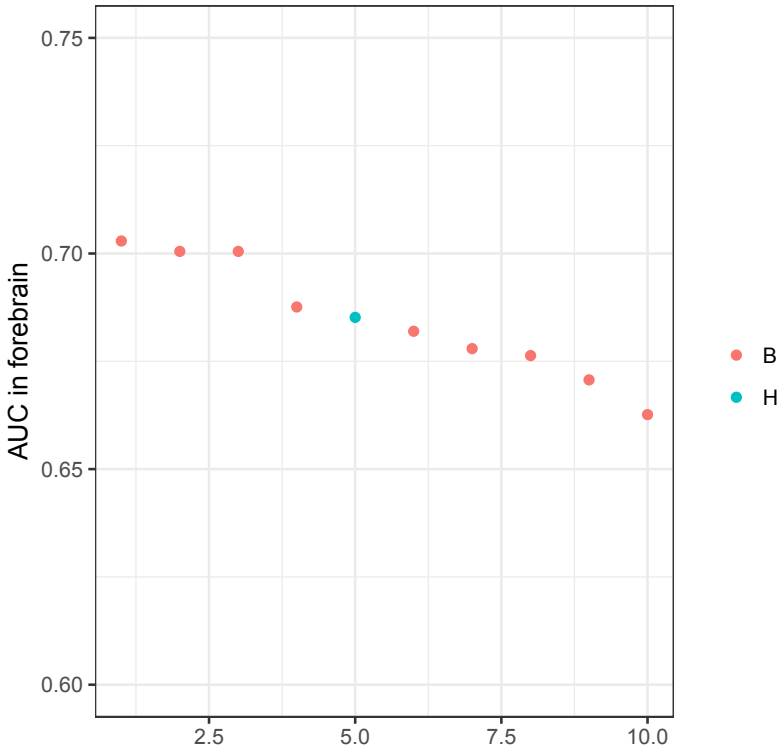
Group	Method	Datasets
B4	gkm-SVM	DHS/H3K27ac/p300
B5	gkm-SVM	DHS/H3K27ac/p300
B6	gkm-SVM	DHS/H3K27ac/p300
B7	gkm-SVM	DHS/H3K27ac/p300
B8	gkm-SVM	DHS/H3K27ac/p300
C0	diHMM	Histone modification
C1	diHMM	Histone modification
D0	Random Forest clustering	Motif, conservation
D1	Random Forest clustering	Motif, conservation
E0	REPTILE	DHS/H3K27ac/p300
F0	Matched Filter	Histone, DHS
G0	Hierarchical clustering	Histone
H0	Random Forest	DHS, Histone, CTCF

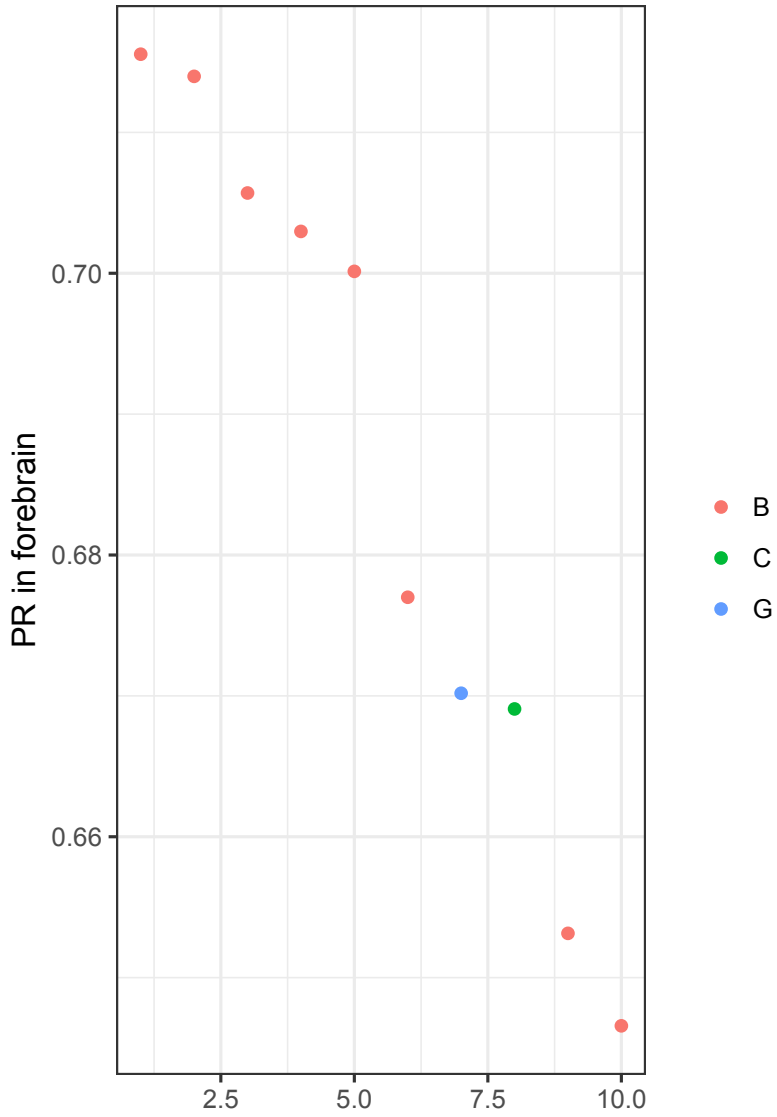
Forebrain

Group	AUC tissue	PR tissue
A0	0.51851852	0.58392145
A1	0.54750403	0.50153067
A2	0.47181965	0.43506739
A3	0.40740741	0.40109957
A4	0.43961353	0.44435883
A5	0.52012882	0.60985578
A6	0.5136876	0.54979894
A7	0.42673108	0.4596689
A8	0.57809984	0.52624192
A9	0.55394525	0.46958334
B0	0.67793881	0.65314231
B1	0.68760064	0.71397775
B2	0.70048309	0.7155503
B3	0.6763285	0.646578

Group	AUC tissue	PR tissue
B4	0.68196457	0.70013468
B5	0.70048309	0.70569791
B6	0.67069243	0.63741411
B7	0.6626409	0.67700063
B8	0.70289855	0.70297054
C0	0.54186795	0.44305432
C1	0.64251208	0.66907232
D0	0.54589372	0.56938755
D1	0.57165862	0.55276673
E0	0.63768116	0.59182598
F0	0.59259259	0.61109551
G0	0.63123994	0.67018667
H0	0.68518519	0.60033106

Forebrain


AUC in forebrain
0.75
0.70
0.65
0.60
2.5
5.0
7.5
10.0
B
H

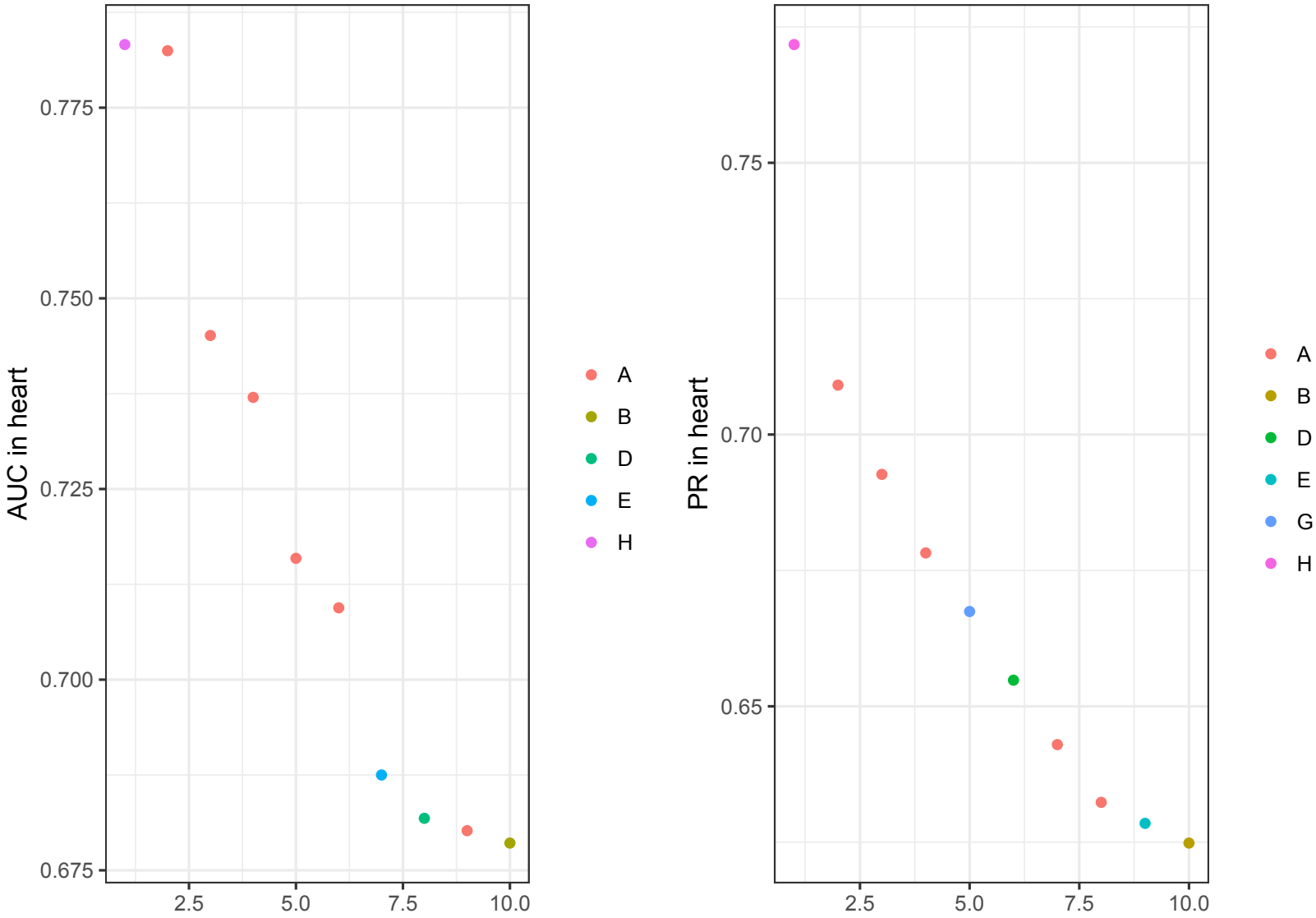

PR in forebrain
0.70
0.68
0.66
2.5
5.0
7.5
10.0
B
C
G

Heart

Group	AUC tissue	PR tissue
A0	0.78246753	0.70909644
A1	0.70941558	0.64295321
A2	0.6461039	0.63232731
A3	0.66883117	0.59398912
A4	0.67532468	0.58338774
A5	0.74512987	0.67821483
A6	0.68019481	0.60945453
A7	0.73701299	0.69267067
A8	0.71590909	0.57192009
A9	0.64448052	0.55091175
B0	0.67857143	0.62483654
B1	0.61201299	0.45824296
B2	0.6737013	0.52194708
B3	0.66964286	0.59838628

Group	AUC tissue	PR tissue
B4	0.62418831	0.46963215
B5	0.66396104	0.51369919
B6	0.6599026	0.58907044
B7	0.61769481	0.4739063
B8	0.66964286	0.51979061
C0	0.32548701	0.37381175
C1	0.43831169	0.45745843
D0	0.67694805	0.65480604
D1	0.68181818	0.54867124
E0	0.6875	0.62846148
F0	0.66314935	0.59688765
G0	0.63636364	0.66743799
H0	0.78327922	0.77175593

Heart

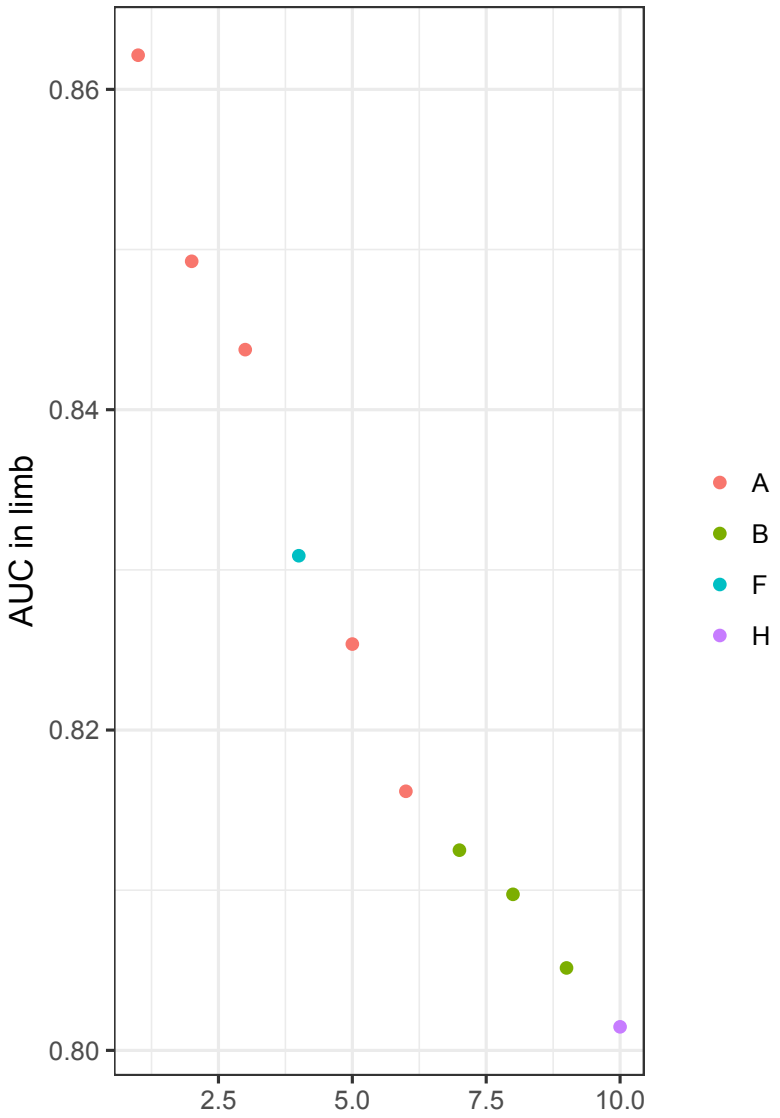

AUC in heart
0.775
0.750
0.725
0.700
0.675
2.5
5.0
7.5
10.0
A
B
D
E
H
PR in heart
0.75
0.70
0.65
2.5
5.0
7.5
10.0
A
B
D
E
G
H

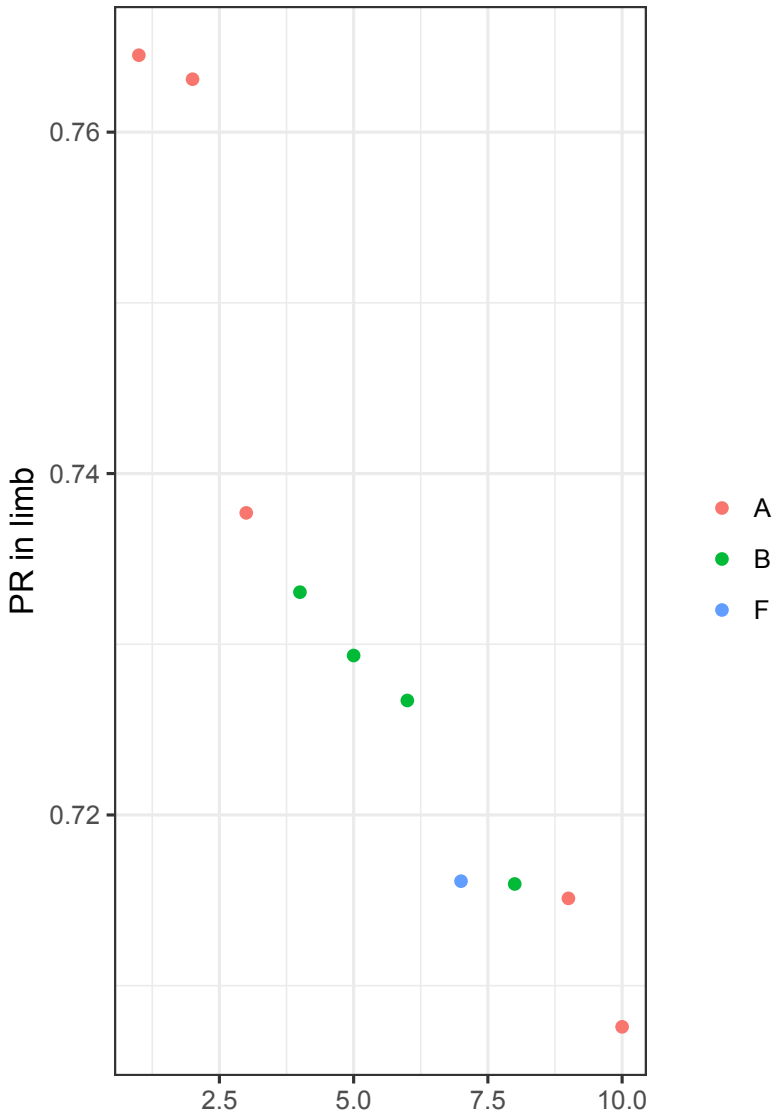
Limb

Group	AUC tissue	PR tissue
A0	0.68933824	0.57213286
A1	0.79044118	0.66267391
A2	0.81617647	0.73769658
A3	0.75	0.61858271
A4	0.72058824	0.57495036
A5	0.86213235	0.71511402
A6	0.77757353	0.67924366
A7	0.84375	0.76310113
A8	0.84926471	0.7645072
A9	0.82536765	0.70758842
B0	0.75735294	0.69740968
B1	0.77389706	0.56386396
B2	0.8125	0.72670578
B3	0.75919118	0.70715088

Group	AUC tissue	PR tissue
B4	0.76838235	0.63549266
B5	0.80974265	0.72934046
B6	0.76102941	0.71595681
B7	0.74724265	0.63052409
B8	0.80514706	0.7330513
C0	0.55606618	0.44844498
C1	0.59926471	0.46709922
D0	0.74264706	0.51763545
D1	0.68933824	0.58377138
E0	0.72058824	0.48017205
F0	0.83088235	0.71612112
G0	0.70588235	0.68438606
H0	0.80147059	0.63716763

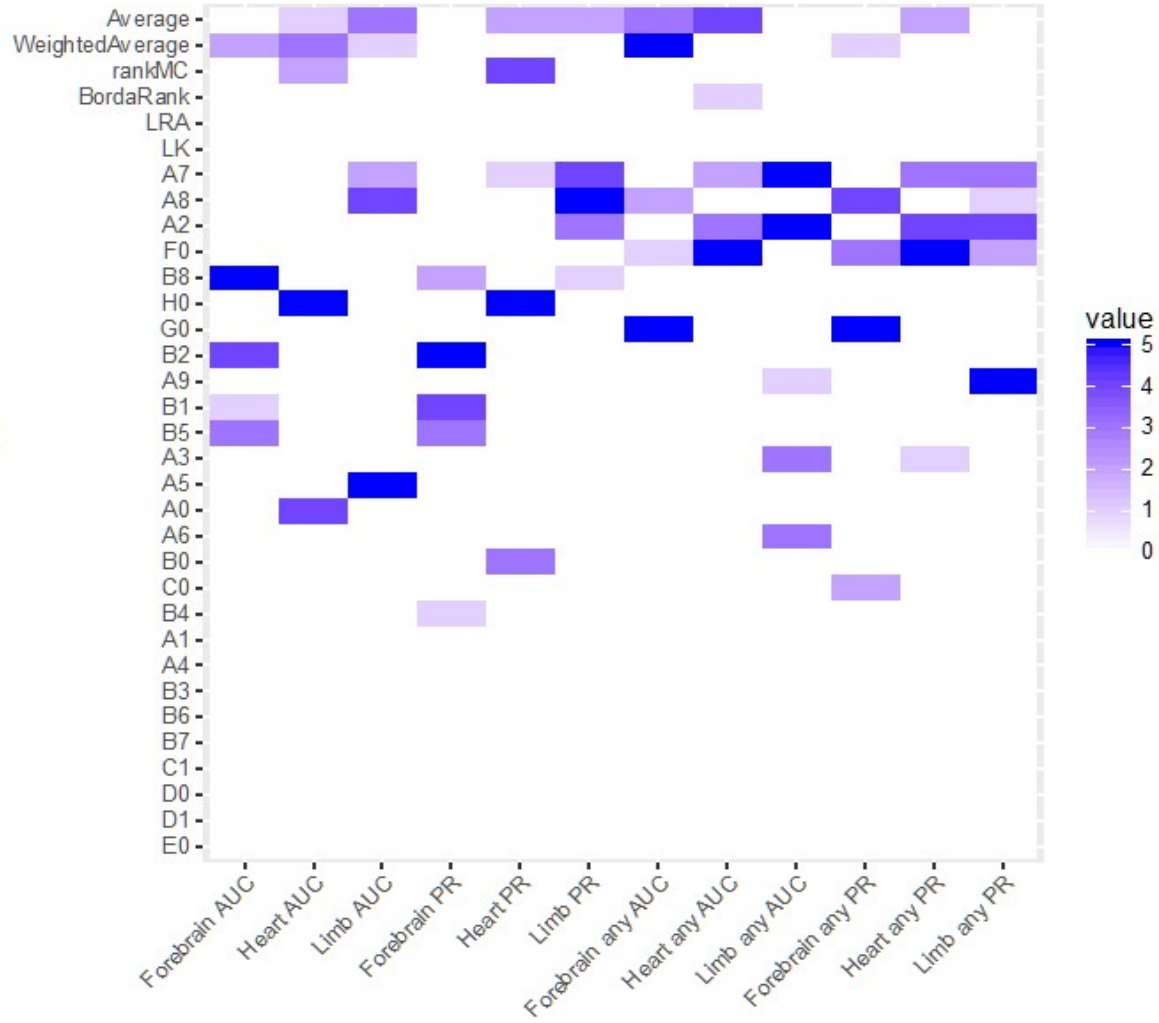
Limb


AUC in limb
0.86
0.84
0.82
0.80
2.5
5.0
7.5
10.0
A
B
F
H


PR in limb
0.76
0.74
0.72
2.5
5.0
7.5
10.0
A
B
F

Any

Ensemble methods generally have higher overall performance



Sensitivity