Merging Enhancer Predictions

Anurag Sethi
Gerstein Laboratory
Yale University

AWG Call
May 2014

Transgenic mouse assays in ENCODE 3

Overall Goal: To develop and assess different methodologies for predicting enhancers active in different tissues for human/ mouse.

Phase I: Test 100-150 predicted enhancers in transgenic mouse Enhancer assays.
Tissues chosen : Heart (maybe forebrain later in phase I).
Developmental stage: E14.5

Enhancer Validation Strategy

Genomic Datasets (Flexible)

Histone Marks
Open Chromatin
TF binding
conservation

↓

Models (Flexible)

Supervised (trained using known enhancers - VISTA)
Unsupervised (trained on features typical of enhancers)

↓

11 sets of Predictions

↓

Merge

↓

Perform validations in transgenic mice

Rules for Submissions

Two tissue specific lists of up to 10000 enhancers active in human heart and forebrain (with validation in corresponding mouse tissues).

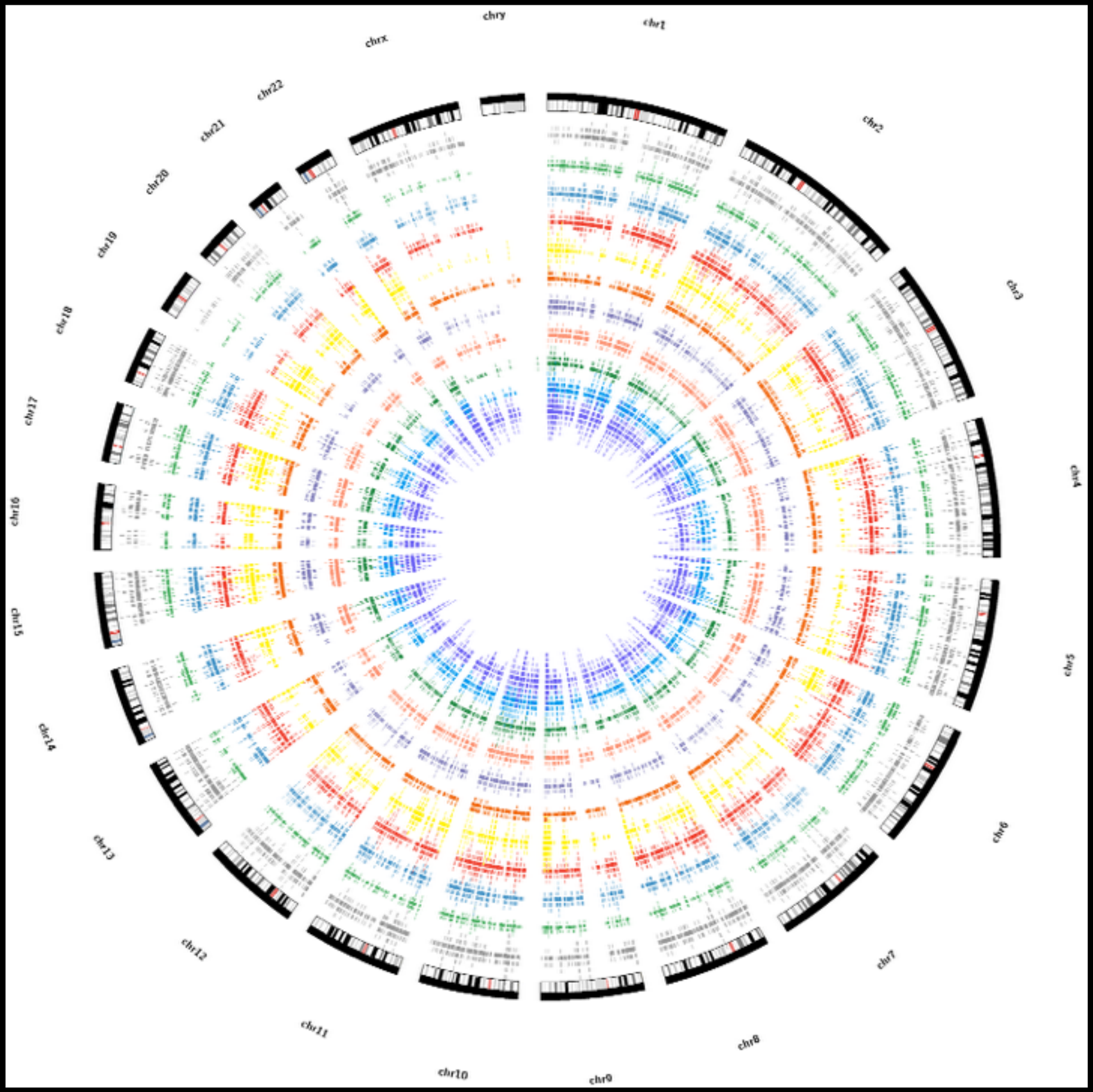
Predicted enhancers were all 1 kb in width.

Validated regions (VISTA) and promoters (+/- 2kb of GENCODE 19 TSS) were removed from the predicted enhancers.

Different prediction strategies and datasets utilized

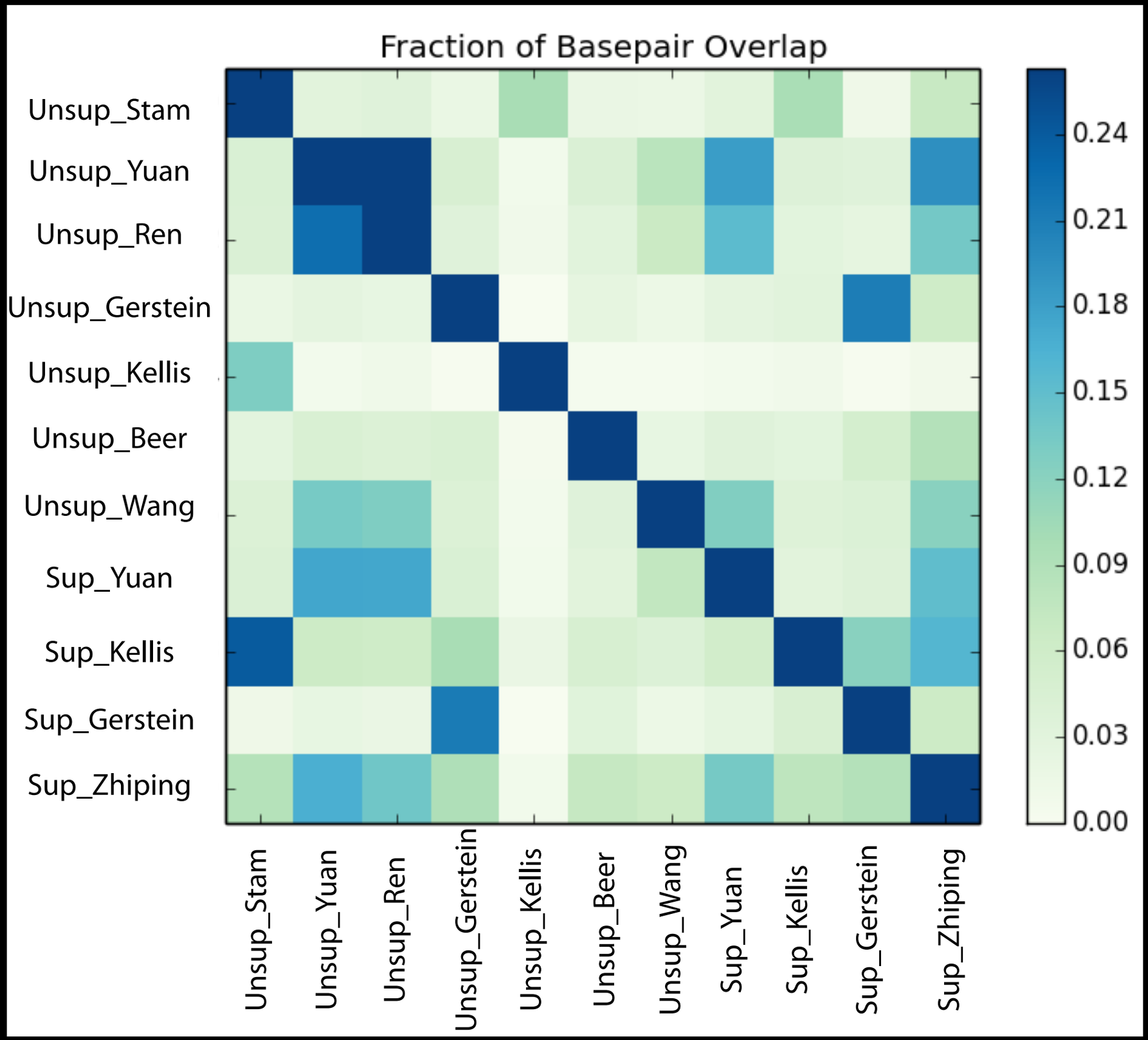
Group	Method	Datasets	#prediction heart → after filtering
Stam	k-means clustering	DNase HS	9875
Yuan	HMM	histone (mouse)	7075
Ren	Random Forest	P300, histone (mouse)	8345
Gerstein	Random Forest	TF, histone, DNase HS conservation, motifs	9990
Kellis	Empirical evidence	H3K27ac, DNase HS, chromatin state	7614
Beer	k-mer SVM	Sequence, DNase HS, P300, H3K27ac	6616
Wei Wang	HMM	histone, P300, known enhancer	4325
Yuan	SVM	VISTA, motif,H3K27ac	7363
Kellis	SimpleLogistic	VISTA, H3K27ac, DNase HS	4037
Gerstein	Random Forest	VISTA, TF, histone, DNase conservation, motifs	9989
Zhiping Weng	Combination	VISTA, P300, histone, DNase HS	8105

4 supervised
7 unsupervised

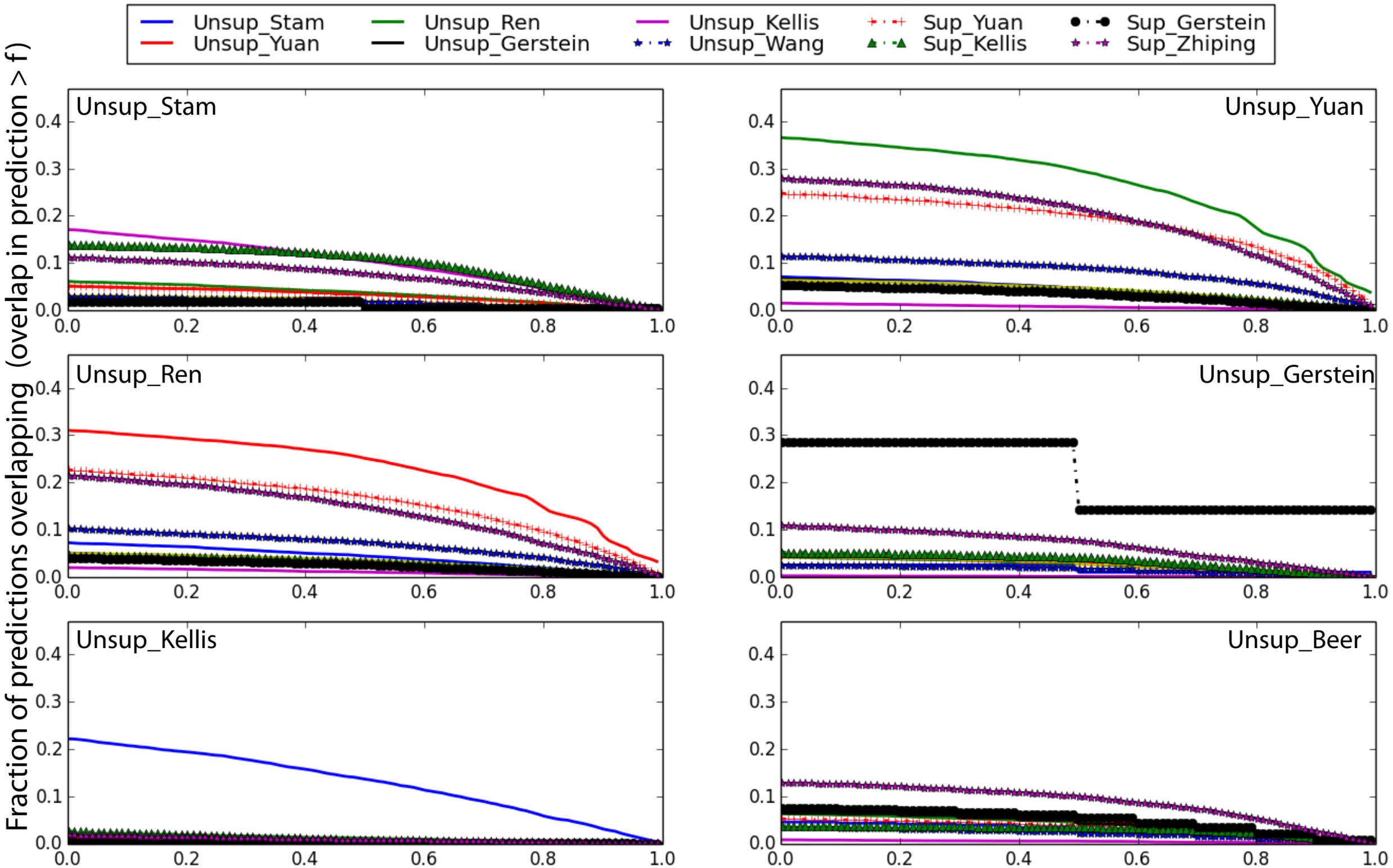


The predictions are distributed over the whole genome.

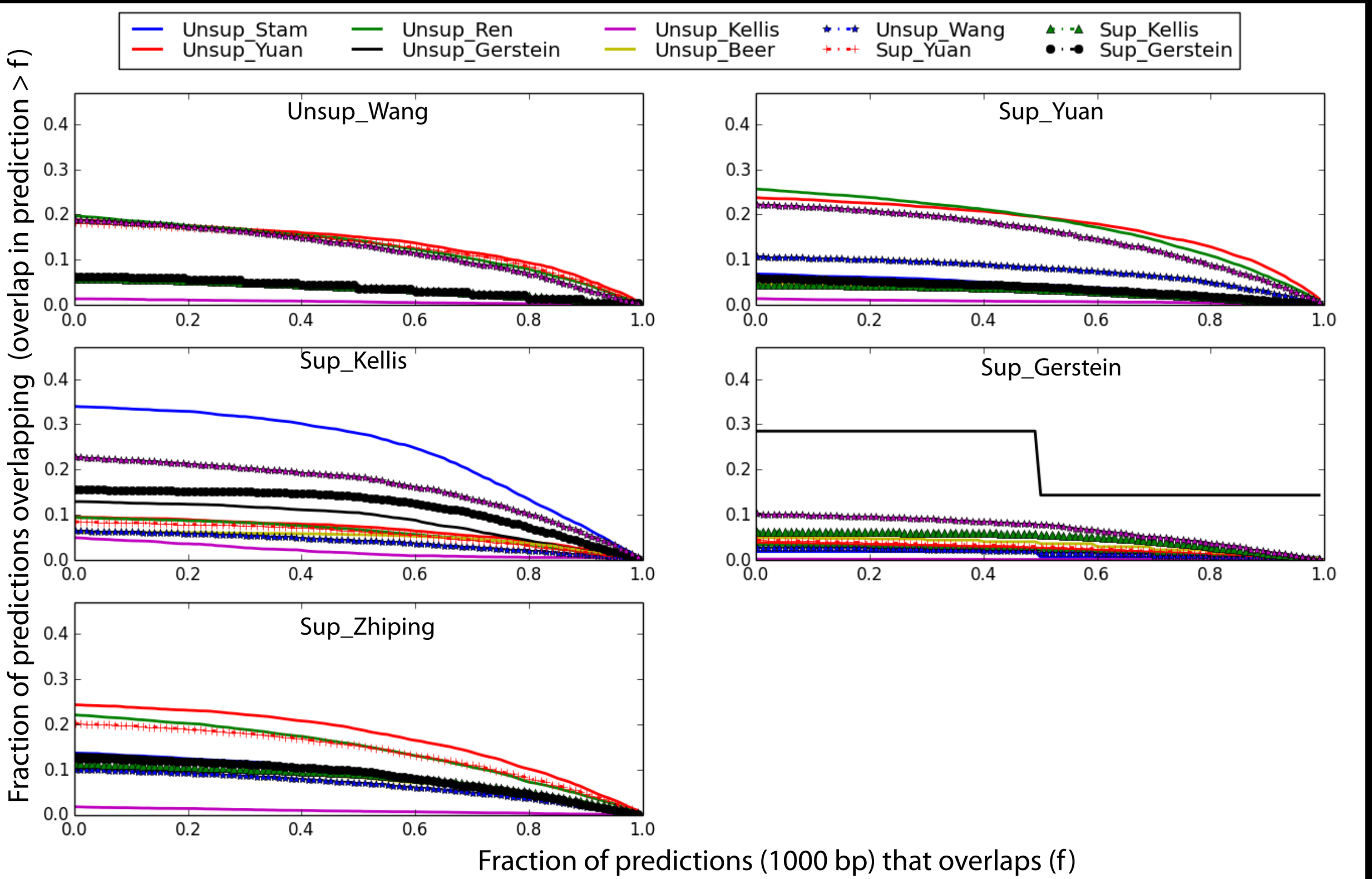
Methods that utilize similar datasets make more similar predictions. Very different predictions when different datasets used.



Methods that use similar datasets make more similar predictions.



Methods that use similar datasets make more similar predictions.



Strategy for merge

Predictions divided into regions based on overlap of different methods.

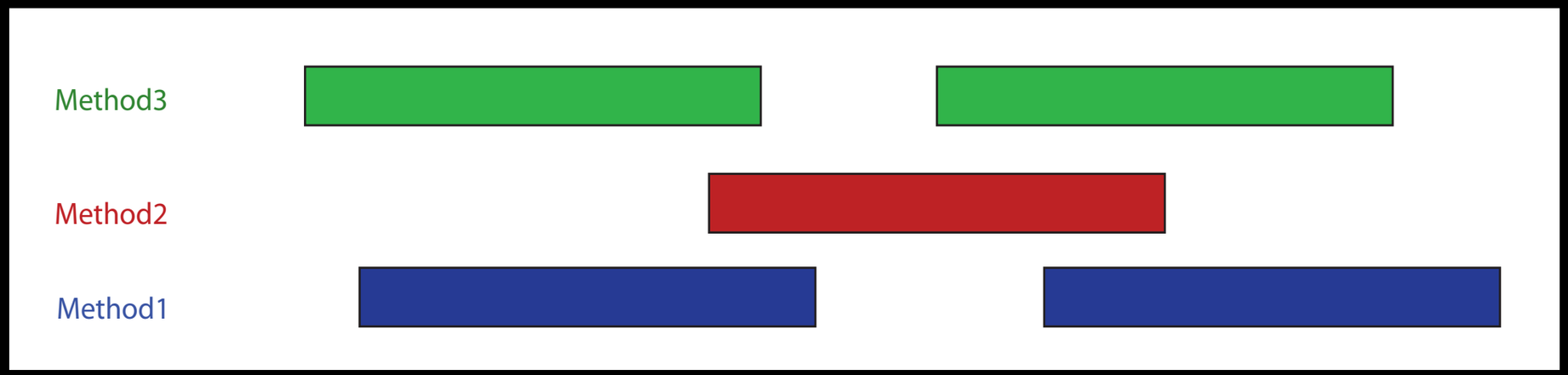
Regions are scored based on quantile normalization (score = 0 if region not in a particular method's predictions before quantile normalization). Overall score of a region is the sum of these quantile normalized scores.

Regions are then ranked based on number of methods that predict this region to be an enhancer and the overall score of these predictions.

Highest ranked regions used as seed for merge regions and the predictions are then expanded to 1 kb width based on these rankings.

None of the predictions can be within 5 kb of each other.

Analysis of common regions in predictions from multiple methods

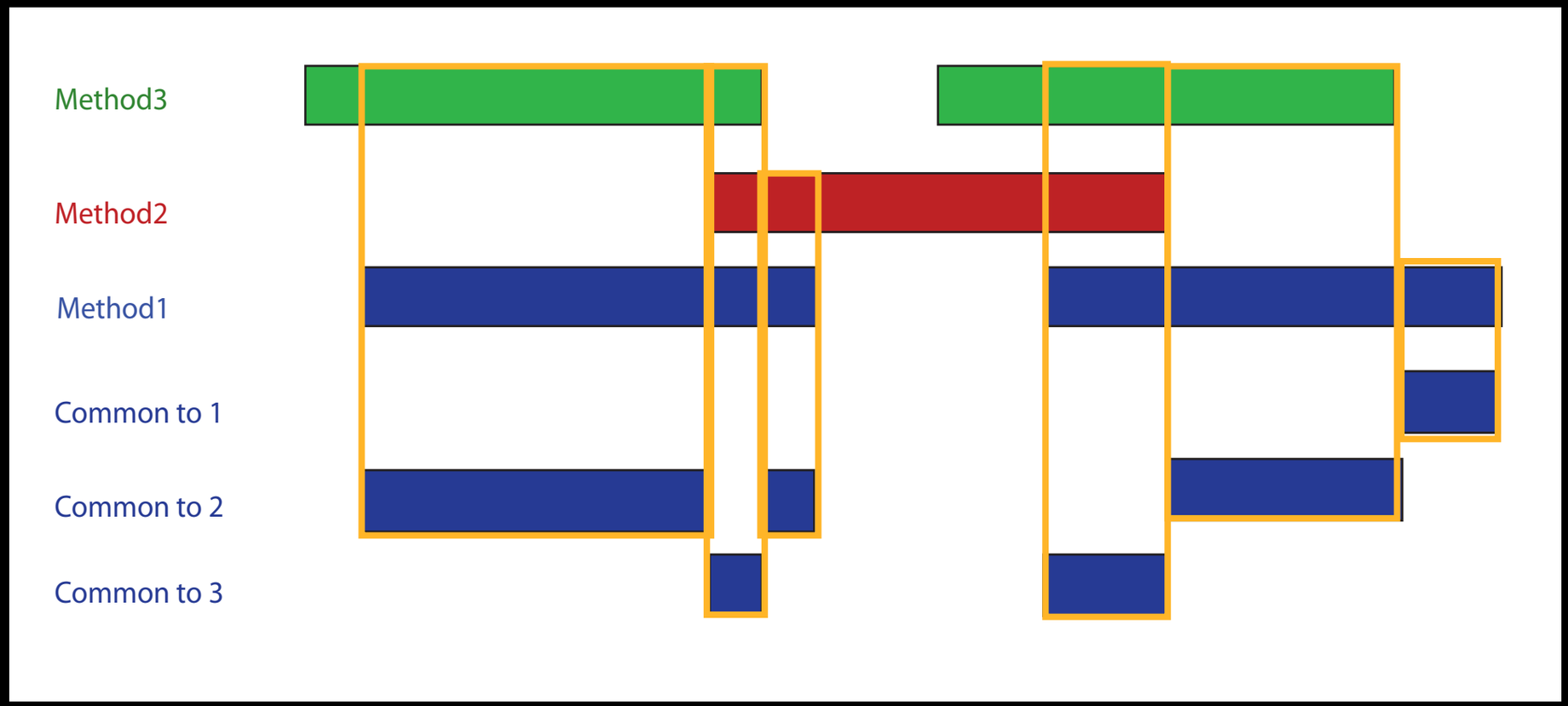


A single prediction gets split into multiple regions based on overlap with predictions made by other groups.

Priority for ordering of regions:

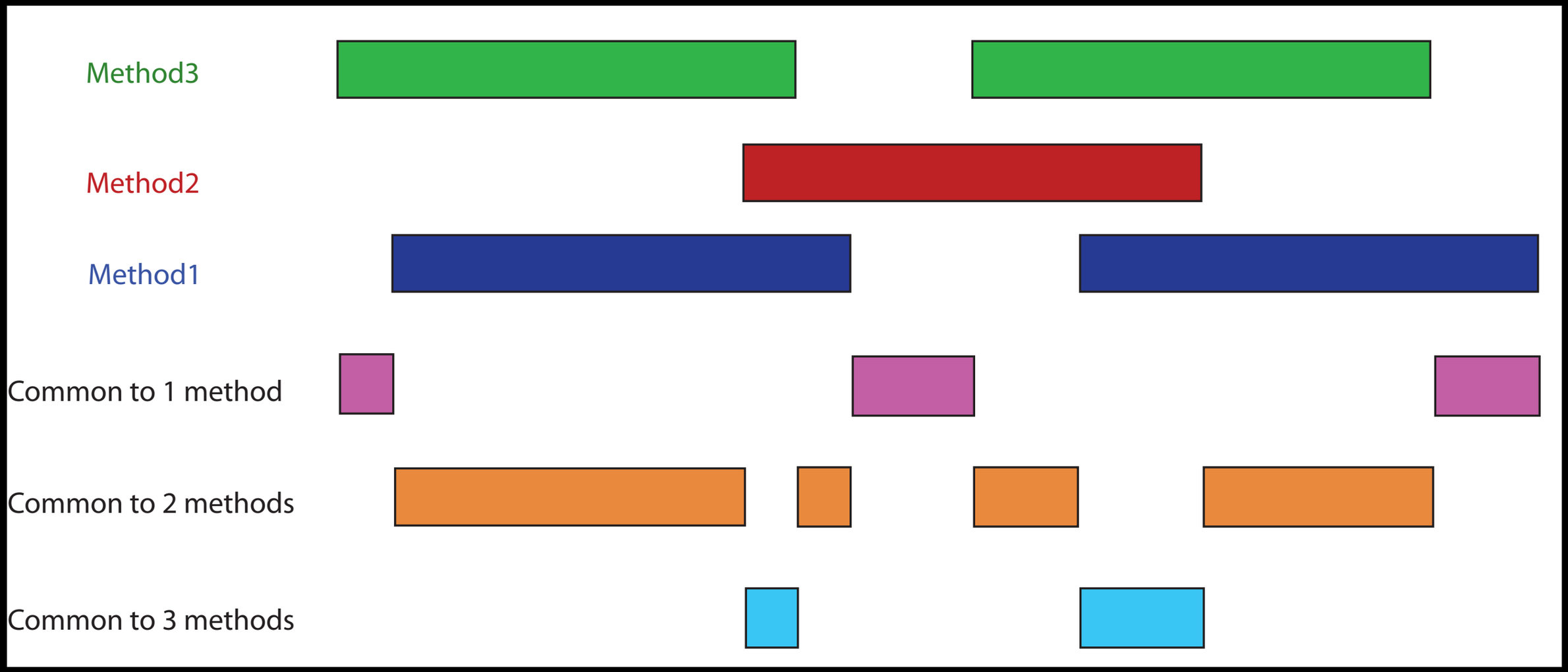
1) Number of predictions common to that region
2) Quantile normalized score for regions that are tied according to condition 1.

Analysis of common regions in predictions from multiple methods



A single prediction gets split into multiple regions based on overlap with predictions made by other groups.

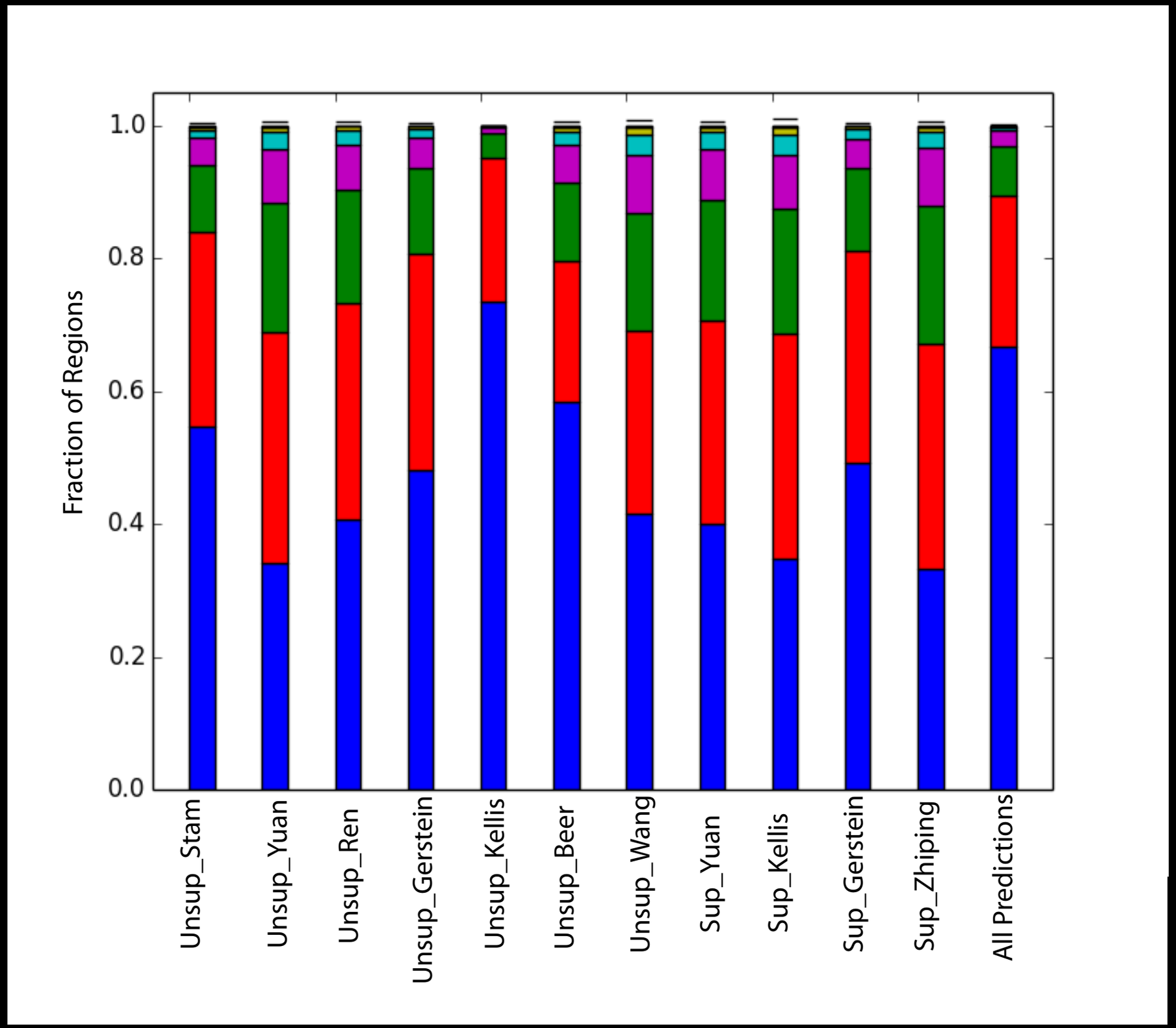
Analysis of common regions in predictions from multiple methods



Priority for ordering of regions:

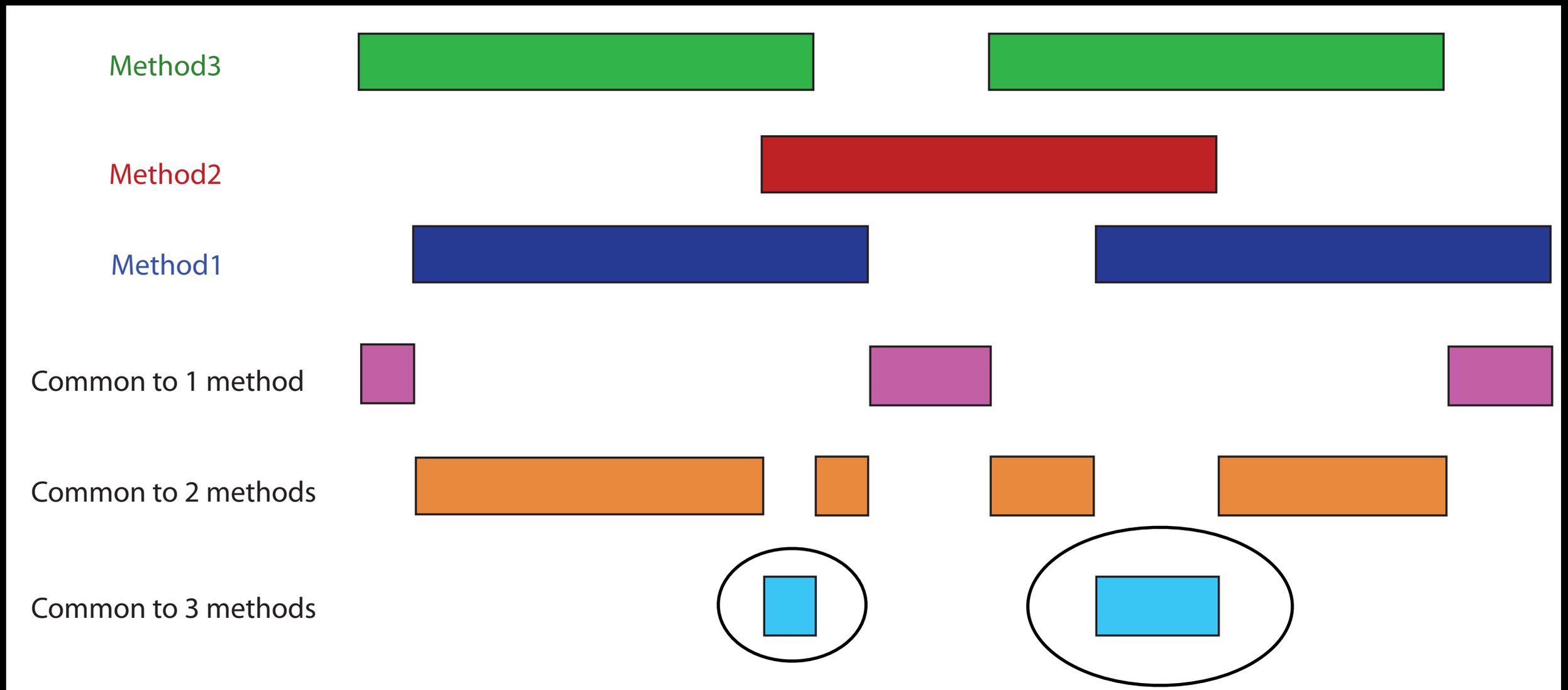
1) Number of predictions common to that region
2) Quantile normalized score for regions that are tied according to condition 1.

There are a lot of regions that overlap in 5 or larger number of prediction lists



Number of Methods	Number of regions	Average Width
9	1	343
8	4	151
7	29	282.7
6	149	261.7
5	626	284.2
4	2433	297
3	7828	341.3
2	23891	450
1	70320	707

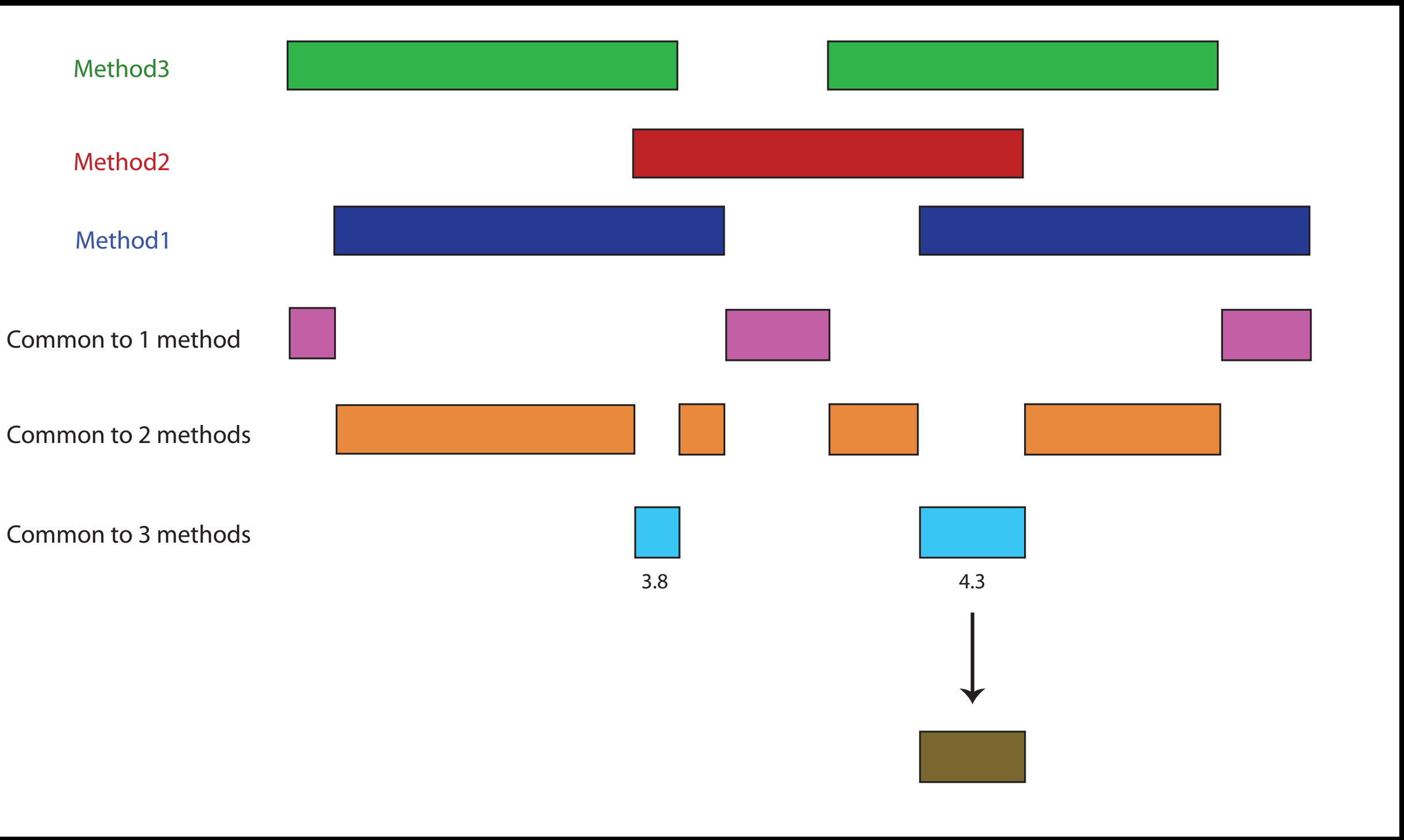
Choose highest ranked regions as seeds for regions chosen for enhancer validation



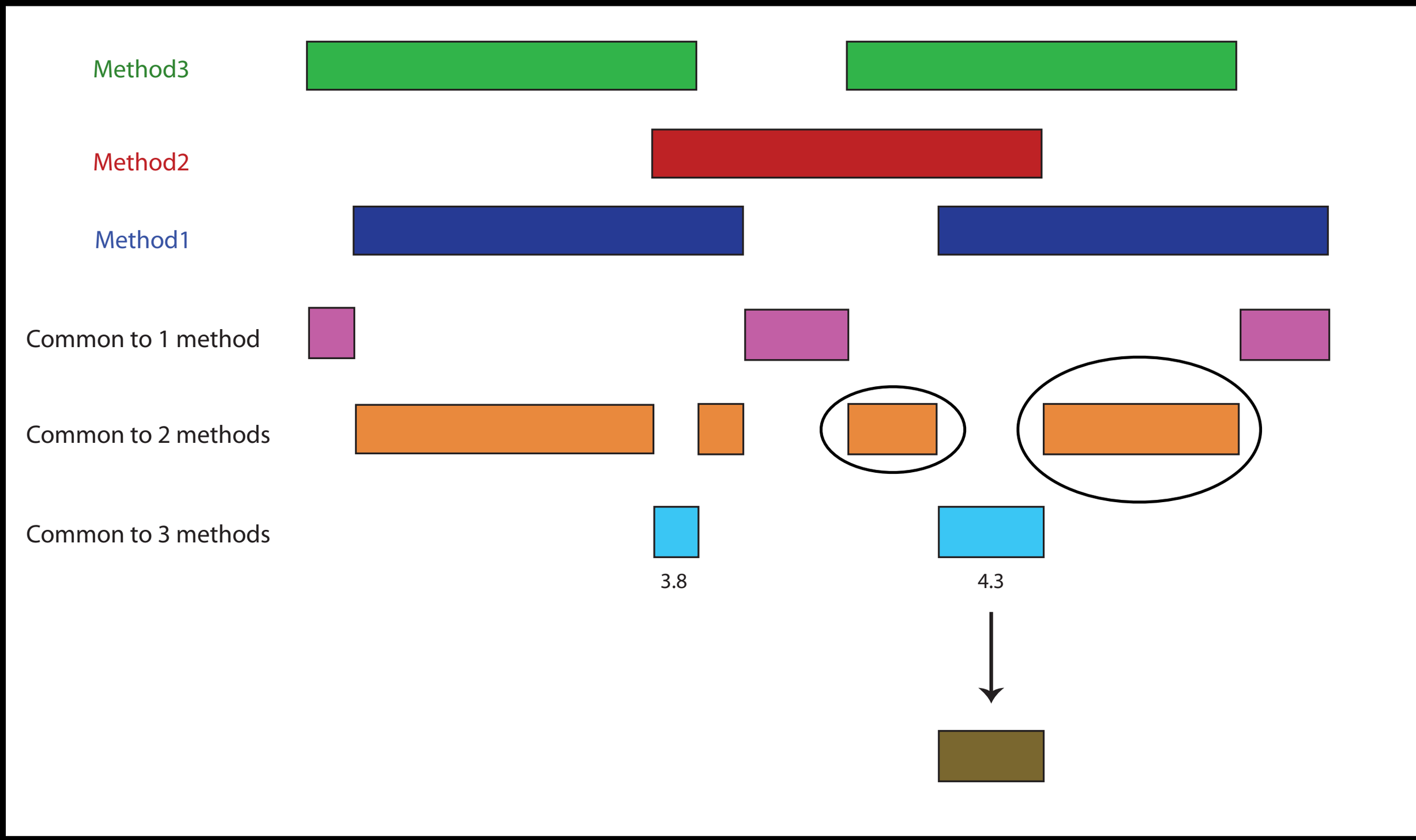
Ranking of regions based on two criteria:

1) Number of methods that predict a region (cyan > orange > magenta).
2) Sum of quantile normalized scores of different regions when tied by criteria 1.

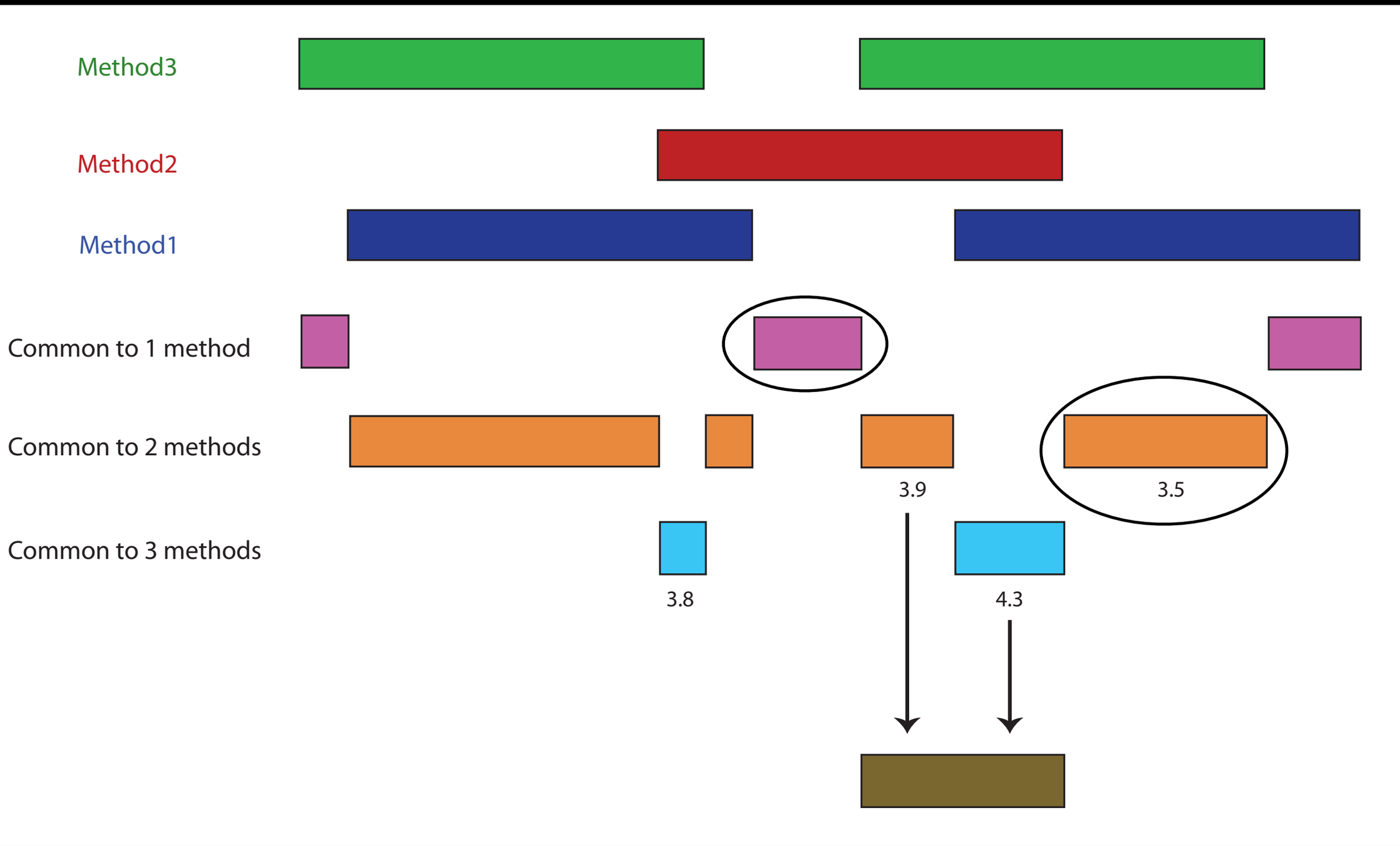
Choose highest ranked regions as seeds for regions chosen for enhancer validation



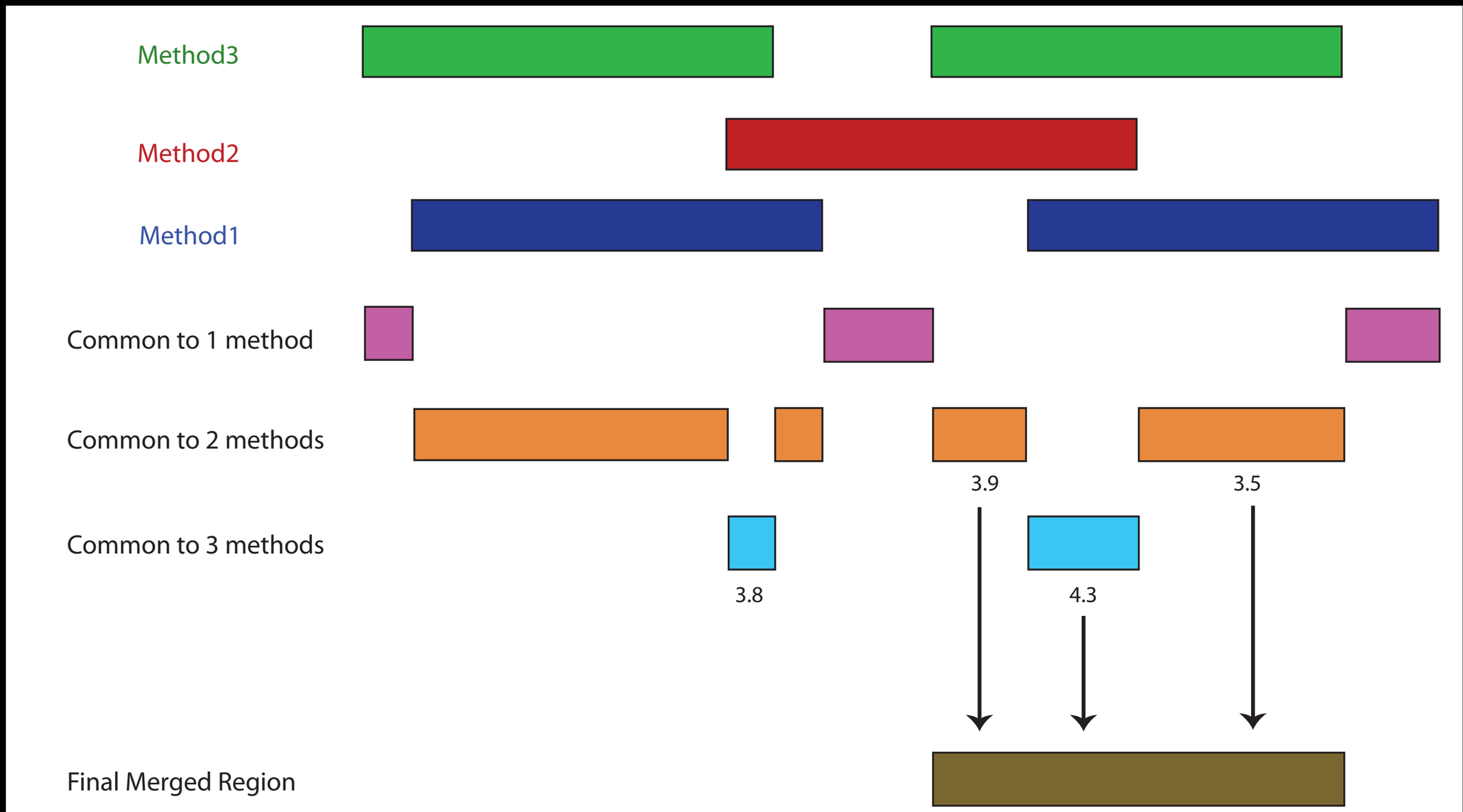
Expansion of seed regions based on ranking of neighboring regions



Expansion of seed regions based on ranking of neighboring regions



Expansion of merged regions continues until the merged bin is 1000 bp in width



All methods represented - only one method highly under-represented

Method	Number of predictions
Unsup_Stam	73
Unsup_Yuan	99
Unsup_Ran	116
Unsup_Gerstein	99
Unsup_Kellis	9
Unsup_Beer	69
Unsup_Wang	94
Sup_Yuan	112
Sup_Kellis	85
Sup_Gerstein	72
Sup_Zhiping	126

Future Work:

Statistical analysis of overlap between different methods.

Similar strategy may be applied for enhancers predicted to be active in mouse forebrain.

Acknowledgements

Mark Gerstein
Kevin Yip
Joel Rozowsky
Sushant Kumar
Jing Zhang

Len Pennacchio
Manolis Kellis
Zhiping Weng
Members of DAC for submissions