Element tracks (version 1.1)

ENCODE Elements group

(prepared by Kevin Yip)

9th June 2011

Tracks

- Predicted active promoters
- Predicted active enhancers
- High occupancy of TF (HOT) regions
- Low occupancy of TF (LOT) regions

* See comparison of features at these regions at the end of this set of slides

Common properties (1)

- Based on datasets from the analysis (January 2011) freeze
- Human genome build: hg19
- Cell-line specific
- Blacklist regions (http://www.ebi.ac.uk/~anshul/public/encodeRawData/blacklists/wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz) filtered in all analyses
- Gene annotations based on Gencode v.3c levels 1 and 2

Common properties (2)

- Considered cell lines with binding data for at least 20 TFs
  - GM12878, H1-hESC, HeLa-S3, HepG2, K562
- TF peaks called by PeakSeq, using “optimal” setting

Track #1

PREDICTED ACTIVE PROMOTERS

Goal

- Setup a standard procedure for calling extended promoters

Strategies

- Instead of using fixed annotations or defining rules, we learned features of active promoters based on a conservative set of positive examples and multiple sets of negative examples with different properties
- Using these examples, we built statistical models that tell the likelihood of a region being an active promoter based on its open chromatin, histone modifications and TF binding features
  - Expected the examples to contain errors $\rightarrow$ the learning methods were allowed to not trust examples that appear different from the others

Methods (1)

- Divided the whole genome into 100bp bins
- Defined the example sets:
  - Positives: Bins within 100bp upstream from an expressed Gencode TSS (RPM > 1)
  - Negatives: equal number of
    1. Bins not within TF binding peaks (for learning the differences between binding and non-binding regions)
    2. Non-Pol2 TF binding bins at least 10,000bp away from any transcript (for learning the differences between promoter and non-promoter TF binding regions)
    3. Non-TF binding bins within +/- 5,000bp from the TSS of a transcript (to avoid the learning algorithm to use open chromatin as the only feature for identifying promoters)

Methods (2)

- Model training:
  - Examples: 5,000 random positive and 5,000 random negative
  - Features: all open chromatin, histone modifications and TF binding datasets of a cell line
  - Learners: Bayes net, linear regression, random forest, support vector machines (SVM)

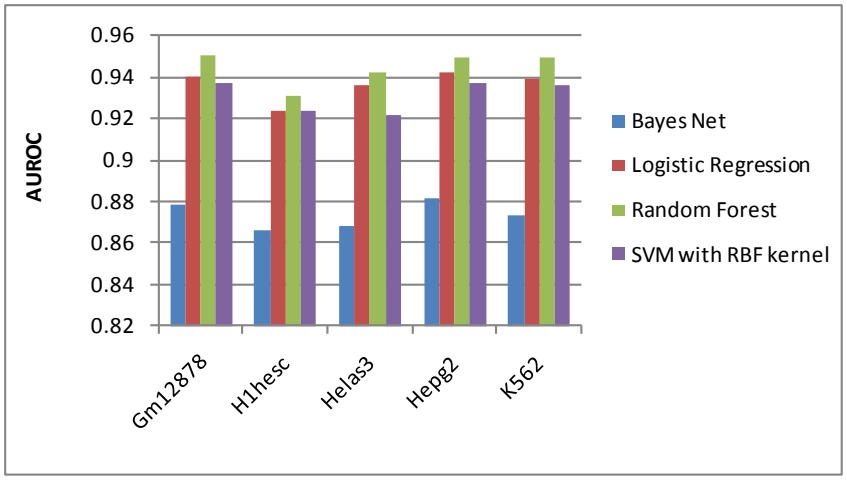
Methods (3)

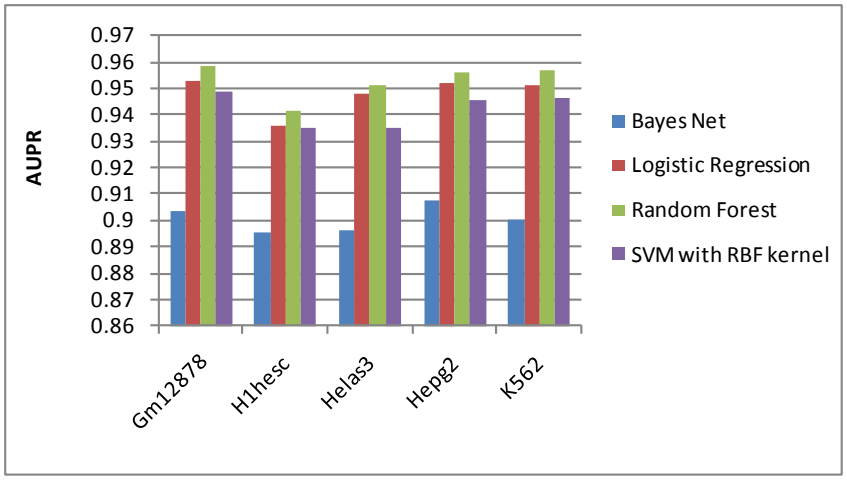
- Model evaluation:
  - Area under receiver operator characteristics (AUROC) and precision-recall (AUPR) curves for **hold-out non-training** examples in the example sets
  - Repeated 10 times and take average
- Model selection:
  - Selected the learner with the highest AUROC
- Model application:
  - Predicted an “active promoter score” for each bin in the whole genome using the selected learner

Methods (4)

- Calling active promoters:
  - 1% “FDR” threshold: 1% of the negative hold-out examples had a predicted active promoter score larger than the threshold
  - Used the threshold to find all potential active promoters in the whole genome
  - Filter out bins with a phastCons primate score < 0.1
  - Merge adjacent bins

Results – holdout AUROC and AUPR

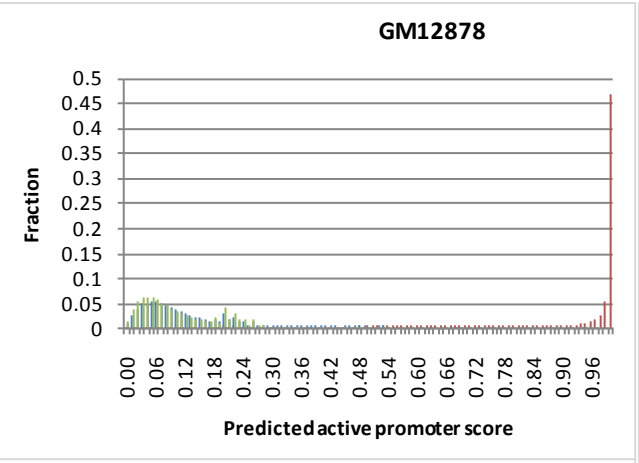


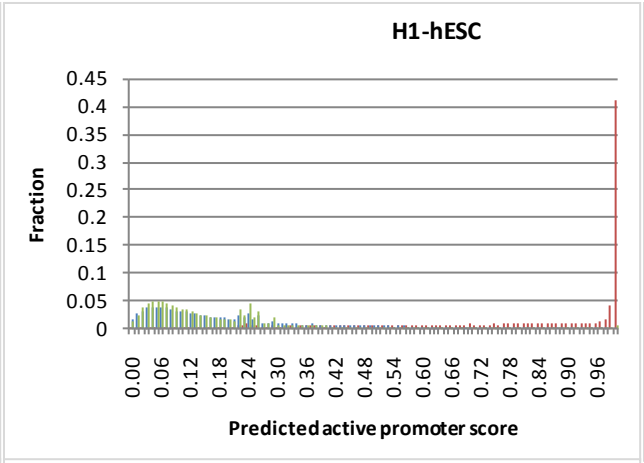


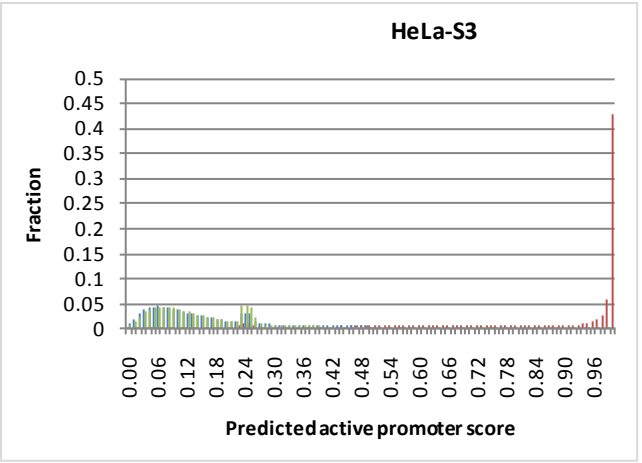
- Selected learner: random forest
- AUROC and AUPR are not direct measures of prediction accuracy, since
  - The positive and negative sets likely contain errors
  - The positive examples are only the most conservative ones

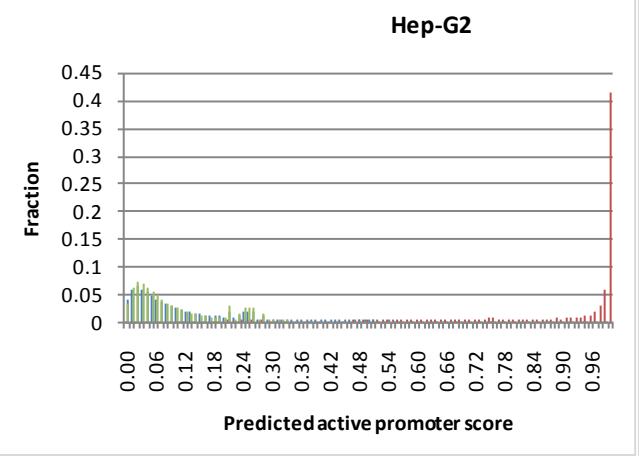
  But a good AUROC/AUPR does indicate that a good model can be constructed, using the open chromatin, histone modification and TF binding features, to distinguish between the positive and negative examples

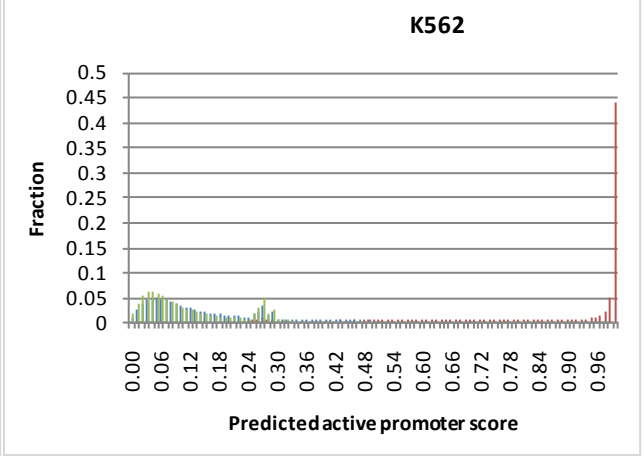
Results – distribution of active promoter scores

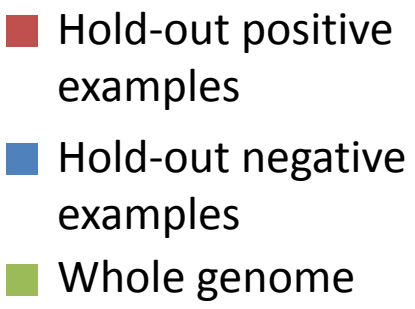










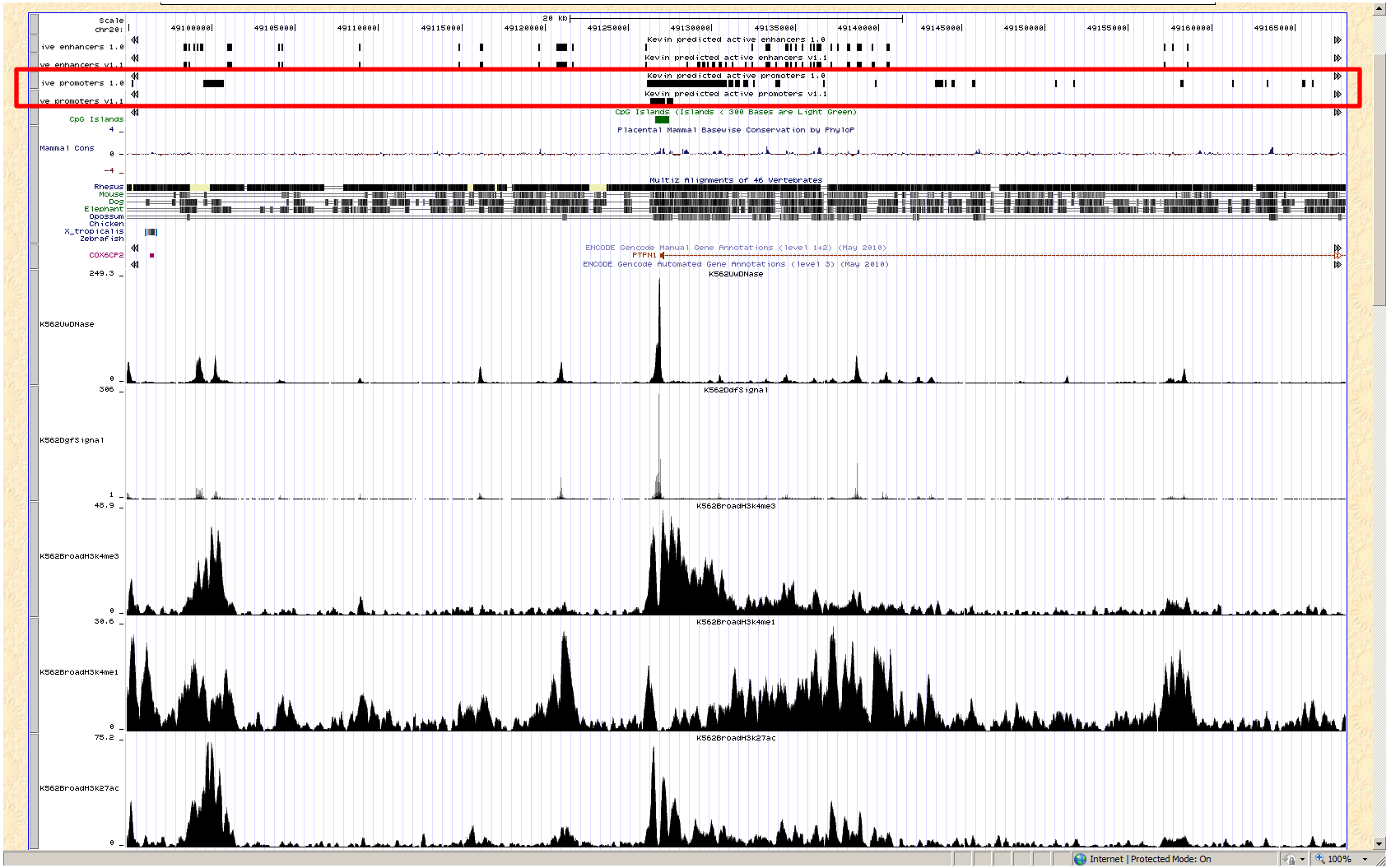


- Clear separation of positive and negative examples

Results – thresholds, recalls and counts

	GM12878	H1-hESC	HeLa-S3	HepG2	K562
Score threshold at 1% “FDR”	0.76	0.85	0.78	0.82	0.76
“Recall rate” (fraction of pos. hold-out examples with predicted score > threshold)	0.70	0.59	0.65	0.65	0.67
#Predicted active promoter 100-bp bins in whole genome before conservation filtering	613,929	290,720	430,846	324,504	556,713
#Predicted active promoter 100-bp bins in whole genome after conservation filtering	265,139	137,901	202,616	155,517	251,383
#Predicted active promoter regions by merging adjacent bins	111,763	51,233	84,694	62,652	105,541

Results – version 1.0 vs. 1.1



Version 1.0 → 1.1

- FDR: 5% → 1%
- Pol2: used to define examples → used as features
- Conservation: added phastCons mammals filter

Results – version 1.0 vs. 1.1

Version 1.0 → 1.1

- FDR: 5% → 1%
- Pol2: used to define examples → used as features
- Conservation: added phastCons mammals filter

Files provided

- PredictedActivePromoters_<cell line>.bed
  - <cell line>: {Gm12878, H1hesc, Helas3, Hepg2, K562}

Things to be done in future versions

- Increase prediction resolution (100bp for current version)
- Anshul's shape analysis
- Max-gap-min-run type of bin merging
- Predict high-GC and low-GC promoters separately
- High confidence vs. high coverage sets

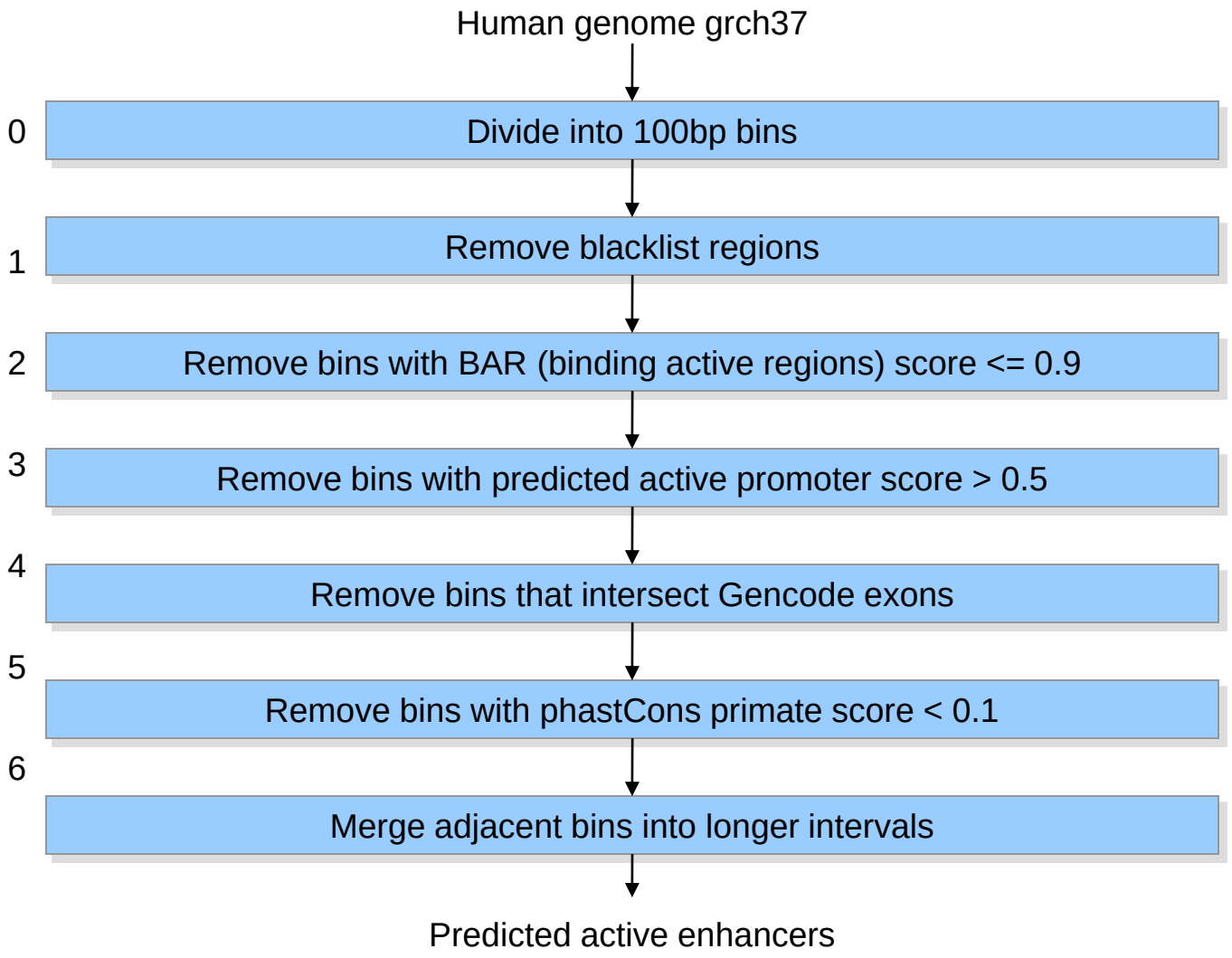
Track #2

PREDICTED ACTIVE ENHANCERS

Strategies

- Since there are no large-scale gold-standard examples, we used a step-by-step filtering of regions unlikely to be active enhancers (or possibly also other elements):
  - Not within binding active regions
  - Within CTCF binding peaks
  - Within potential promoter regions
  - Within exons
  - Have low conservation
- H3K4me1, H3K27ac and p300 were only used to check resulting predictions

Methods (1)


Human genome grch37
0
Divide into 100bp bins
1
Remove blacklist regions
2
Remove bins with BAR (binding active regions) score <= 0.9
3
Remove bins with predicted active promoter score > 0.5
4
Remove bins that intersect Gencode exons
5
Remove bins with phastCons primate score < 0.1
6
Merge adjacent bins into longer intervals
Predicted active enhancers

Methods (2)

- Binding active regions (BAR):
  - Used a procedure similar to the one for promoter prediction
  - Positive examples: TF binding bins
  - Negative examples: non-TF binding bins
  - Features: open chromatin and histone modification signals

Results - counts

	GM12878	H1-hESC	HeLa-S3	Hep-G2	K562
#Bins after step 0 (binning)			30,956,951		
#Bins after step 1 (blacklist)			30,840,721		
#Bins after step 2 (BAR)	1,041,102	712,156	819,967	827,509	923,811
#Bins after step 3 (promoters)	528,559	363,937	472,452	442,136	506,118
#Bins after step 4 (exons)	506,608	331,725	457,156	424,643	481,343
#Bins after step 5 (conservation)	204,867	146,776	191,657	170,151	191,655
#Regions after step 6 (merging)	118,398	80,352	101,196	94,323	107,172

Files provided

- PredictedActiveEnhancers_<cell line>.bed
  - <cell line>: {Gm12878, H1hesc, Helas3, Hepg2, K562}

Things to be done in future versions

- Merge with Anshul's list
- Increase prediction resolution (100bp for current version)
- Perform motif and shape analyses
- Define high-confidence and high-coverage sets
- Filter CTCF binding sites (?)

Tracks #3,#4

HIGH AND LOW TF OCCUPANCY REGIONS

Methods (1)

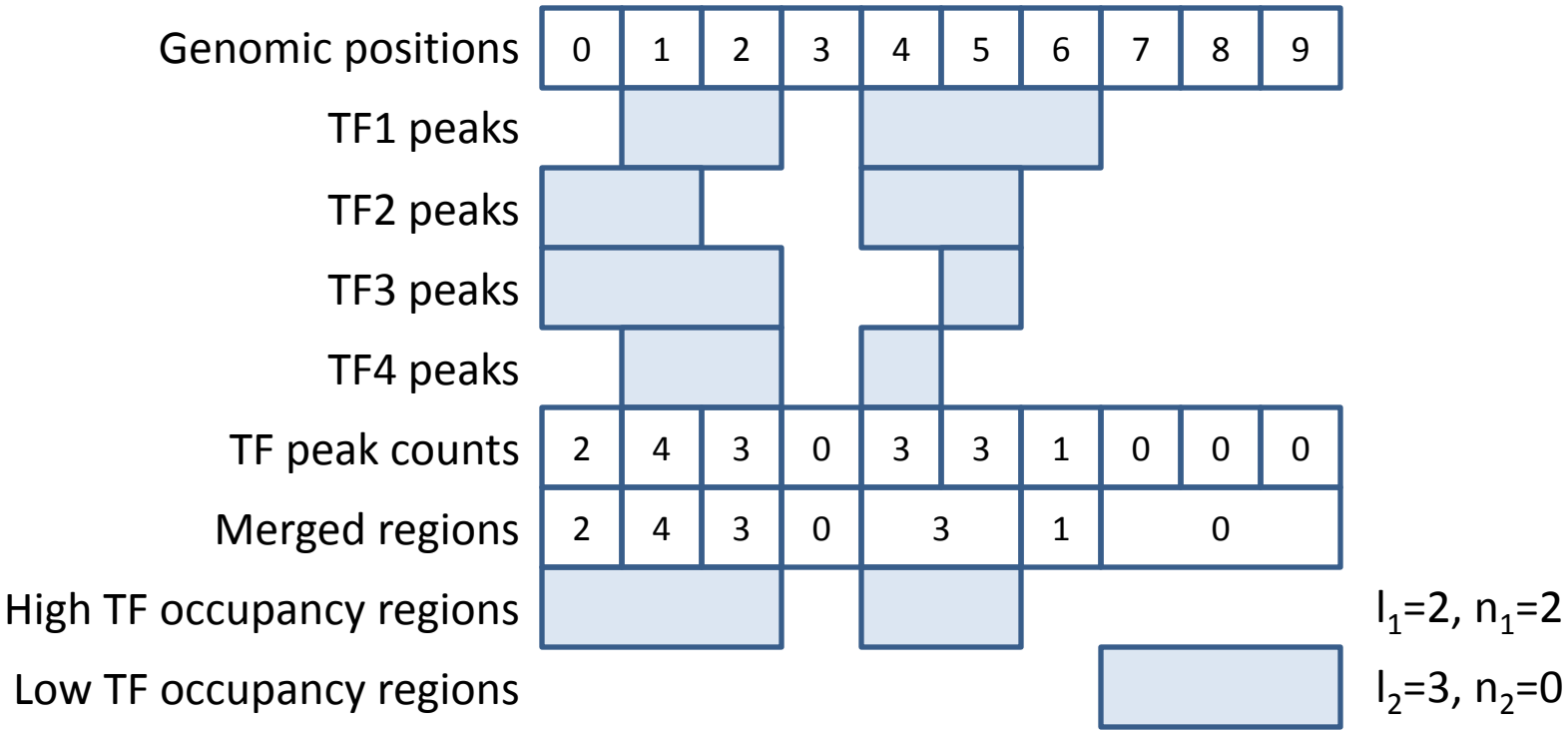
- For each of the five cell lines
  - Obtained the binding peaks of each TF
    - If there were multiple datasets for one TF (due to different labs/antibodies, etc.), took their union
  - For each base pair, counted the number of TFs with a binding peak covering it
  - Merged adjacent base pairs with the same number of TF peak count

Methods (2)

- A high TF occupancy (HOT) region is defined as a region with length $\geq l_1$ bp with each base pair covered by binding peaks of $\geq n_1$ TFs
- A low TF occupancy (LOT) region is defined as a region with length $\geq l_2$ bp with each base pair covered by binding peaks of $\leq n_2$ TFs

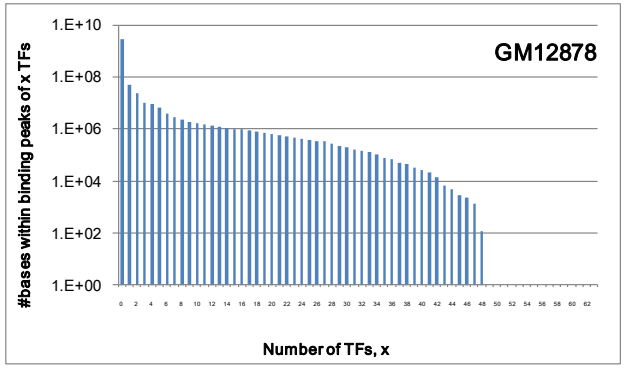
Illustration

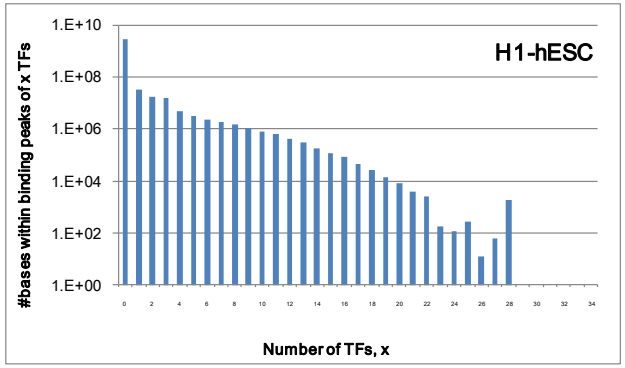
- l_1=2, n_1=2, l_2=3, n_2=0

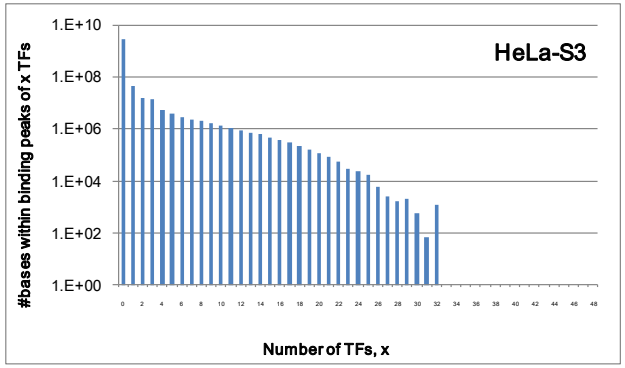


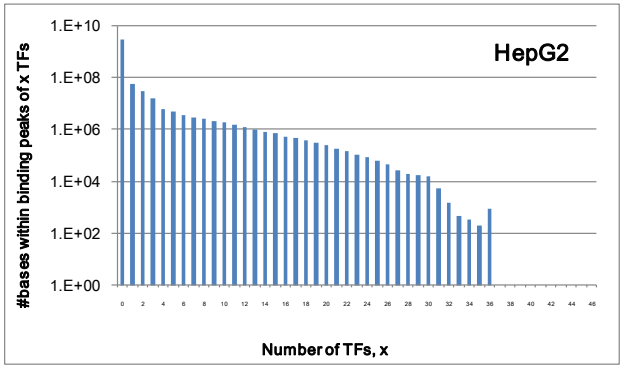
Results - histograms

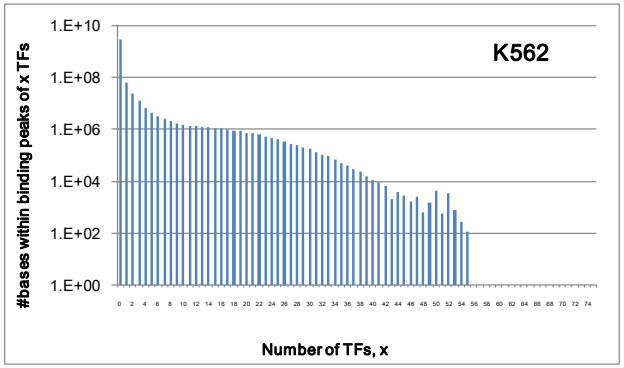
- Whole genome





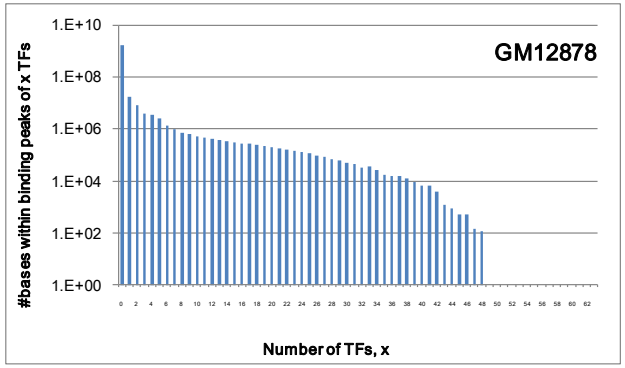


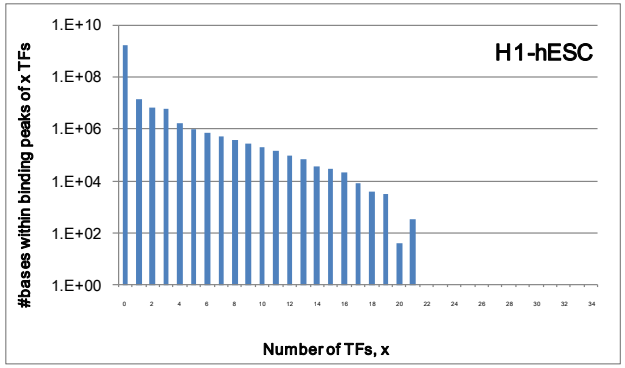


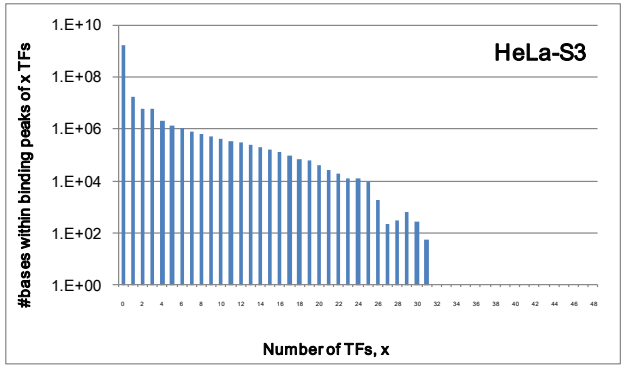


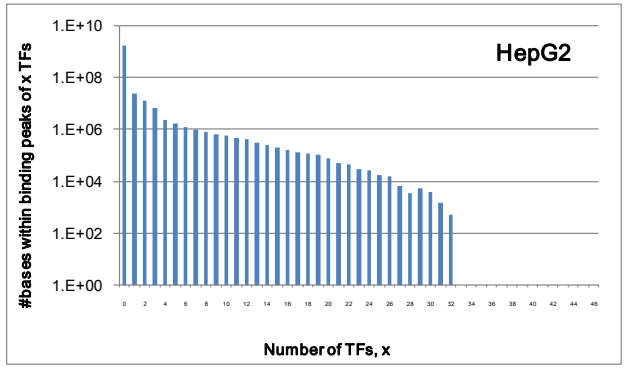
Results - histograms

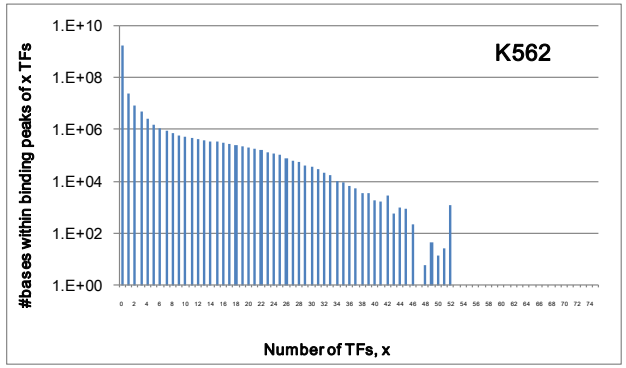
- Distal regions (>10,000bp from known transcripts)











Files provided

1. Merged region files
   - TFCounts_<cell line>_<region>.bed
     - <cell line>: {Gm12878, H1hesc, Helas3, Hepg2, K562}
     - <region>: {all, distal}
     - The score field contains the TF count
2. HOT and LOT files
   - l_1=500, n_1=30% total number of TFs with binding data, l_2=100,000, n_2=0
   - <track>_<cell line>_<region>.bed
     - <track>: {HOT, LOT}

Things to be done in future versions

- Evaluate statistical significance of HOT and LOT regions (working with Ben Brown, Nathan Boley and Peter Bickel on it)
- Remove potential active promoter regions from the distal regions

Track comparisons

FEATURE VALUE DISTRIBUTIONS

GM12878

Positive promoter examples

Negative promoter examples

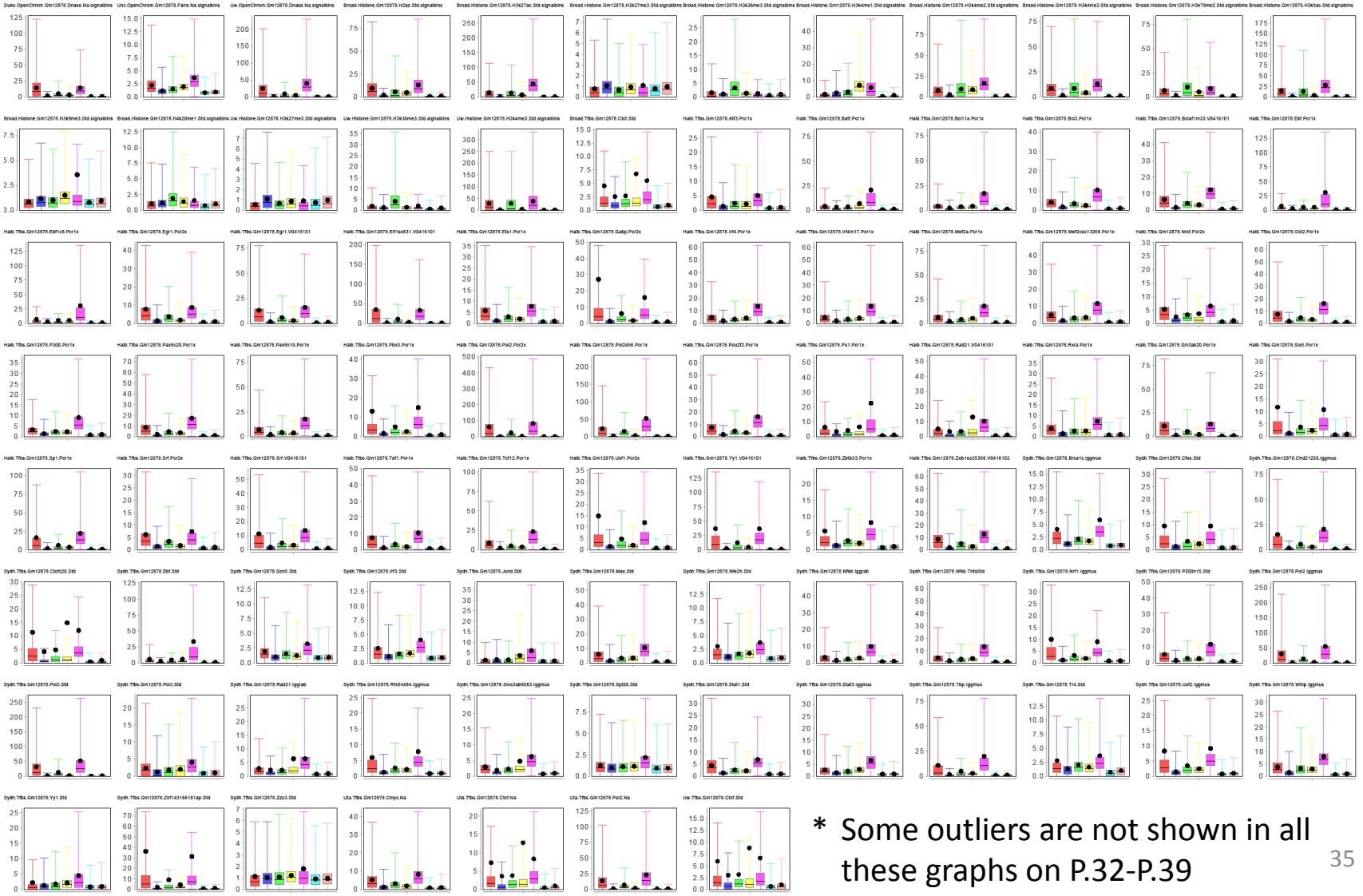
Predicted active promoters

Predicted active enhancers

HOT regions

LOT regions

Whole genome



* Some outliers are not shown in all these graphs on P.32-P.39

H1-hESC

Positive promoter examples

Negative promoter examples

Predicted active promoters

Predicted active enhancers

HOT regions

LOT regions

Whole genome

HeLa-S3

Positive promoter examples

Negative promoter examples

Predicted active promoters

Predicted active enhancers

HOT regions

LOT regions

Whole genome

HepG2

Positive promoter examples

Negative promoter examples

Predicted active promoters

Predicted active enhancers

HOT regions

LOT regions

Whole genome

K562

Positive promoter examples

Negative promoter examples

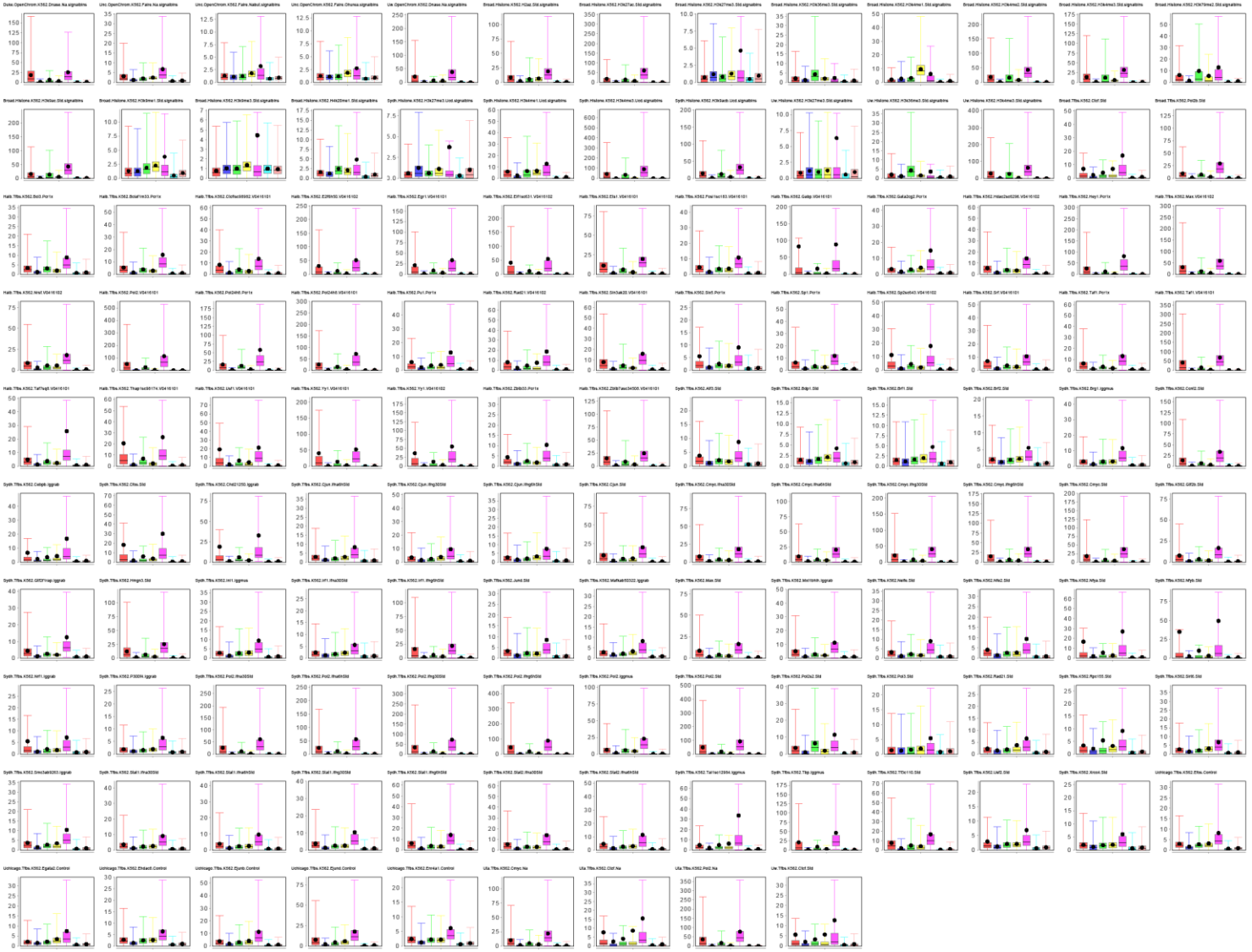
Predicted active promoters

Predicted active enhancers

HOT regions

LOT regions

Whole genome



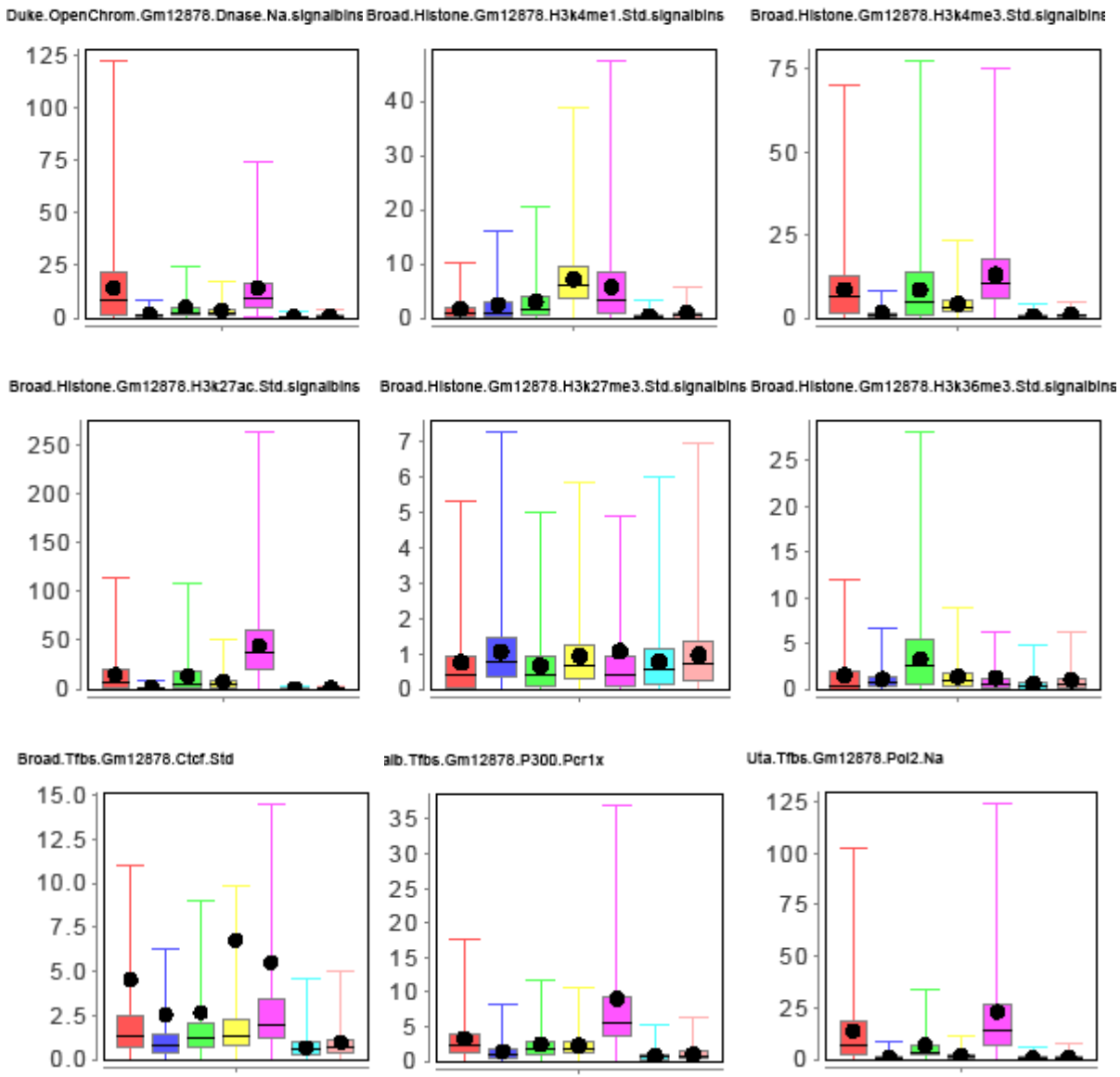
Summary (1)

- Predicted active promoters:
  - Strong signals from DNase, FAIRE, H2A.Z, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, P300, Pol2 and some TFs (e.g., Gabp)
- Predicted active enhancers:
  - Strong signals from DNase, FAIRE, H2A.Z, H3K4me1, H3K27ac (but not as strong as promoters), P300 and some TFs (e.g., Gata2 and Jund)

Summary (2)

- HOT regions:
  - Stronger signals than other regions from most experiments, except for H3K9me1, H3K9me3, H3K27me3, H3K36me3 and H4K20me1
- LOT regions:
  - Lower signals than all other regions from most experiments, except for H3K9me3 and H3K27me3

Selected examples (from GM12878)



Positive promoter examples

Negative promoter examples

Predicted active promoters

Predicted active enhancers

HOT regions

LOT regions

Whole genome

Comments and suggestions are welcome

yuklap.yip@yale.edu