Using gkm-SVM to Detect Conserved Enhancers Where Alignment Fails

Mike Beer
Dept of Biomedical Engineering
Institute of Genetic Medicine
Sept 2, 2017

Jinwoo Oh
Amy Xiao
Wang Xi
Nico Eng
Kendrick Hougen

Felix Yu
Gianluca Silva Croso
Justin Shigaki
Dongwon Lee
M. Ghandi

validation experiments:
Dave Gorkin
Sarah McClymont
Andy McCallion

- ~25,000 non-coding genetic variants (SNPs) associated with common diseases, most in cell specific regulatory regions as detected by chromatin accessibility

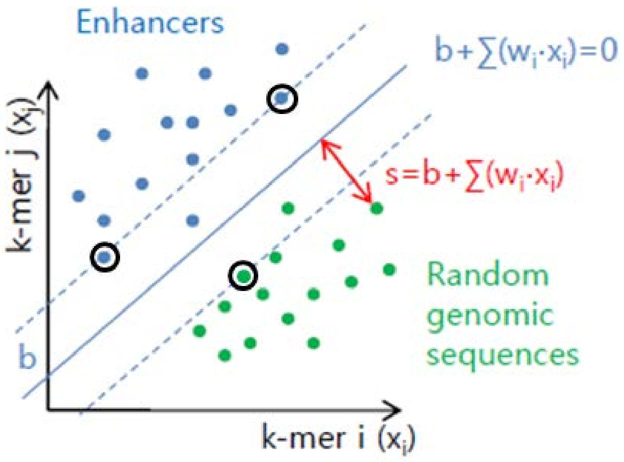
- gapped kmer SVM (gkm-SVM) classifier trained on open chromatin can predict impact of GWAS SNPs

- relevant tissue/cell-type and orthologous mouse regulatory regions for functional evaluation often unknown

- gkm-SVM kernel can be used to detect conserved orthologous regions where alignment fails

gkm-SVM and deltaSVM Method Overview

- Define a set of cell type specific enhancers using functional genomics data: Dnase-seq ATAC-seq
- Train gkm-SVM to learn regulatory TF binding site vocabulary for given cell-type
- Calculate how each SNP changes gkm-SVM score to predict impact (deltaSVM)

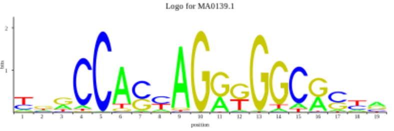


2017 Beer, Human Mutation
2016 gkm-SVM R package, Bioinformatics
2015 Lee - Beer Nature Genetics (deltaSVM)
2014 Lee, Ghandi, Beer, PLOS Comp Bio (gapped kmers)
2013 Fletez-Brant, Lee, Beer, NAR
2013 Ghandi, Beer, J Math Biol (gapped kmers)
2011 Lee, Karchin, Beer, Genome Research (kmers)

2004 Noble, Leslie

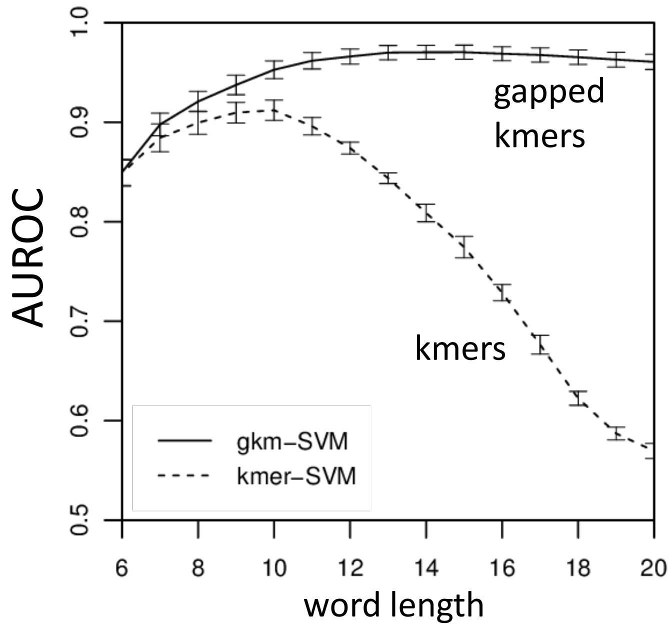
k-mer counts:
CACCAGGGGG
CCACCAGGGG
CCACCTGGTG
CCACCAGGTG

gapped kmer counts:
CA--AG--GG
CC-C--G-GG
CA-C-G--GG
CCA---GG-G



CTCF len=15

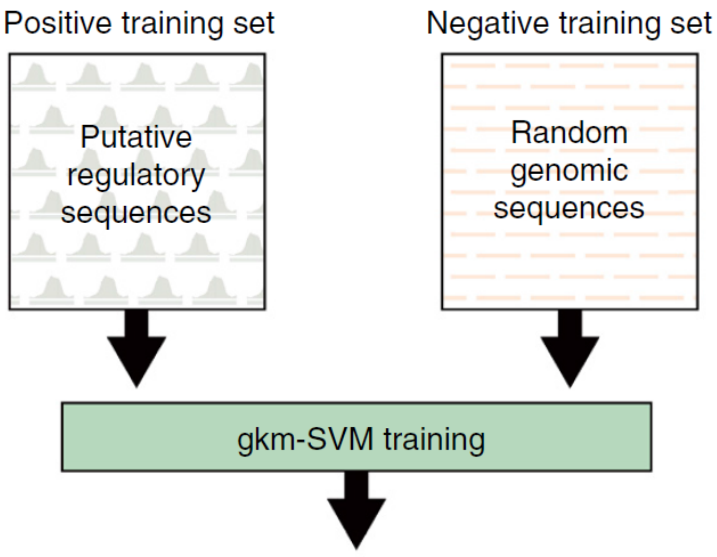
$$K(S_1,S_2)=\frac{\langle f^{S_1},f^{S_2}\rangle}{\|f^{S_1}\|\|f^{S_2}\|}$$



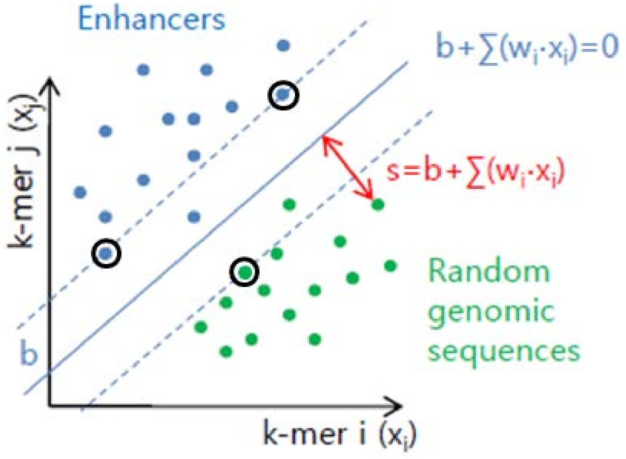
Generate GC matched negative set, Train SVM to Identify Regulatory Features

Lee, Karchin, Beer Genome Research Dec 2011

$y_j = +1$ $y_j = -1$



$$\hat{y}_j = \sum_i w_i x_{i,j} + b$$



Regulatory sequence vocabulary

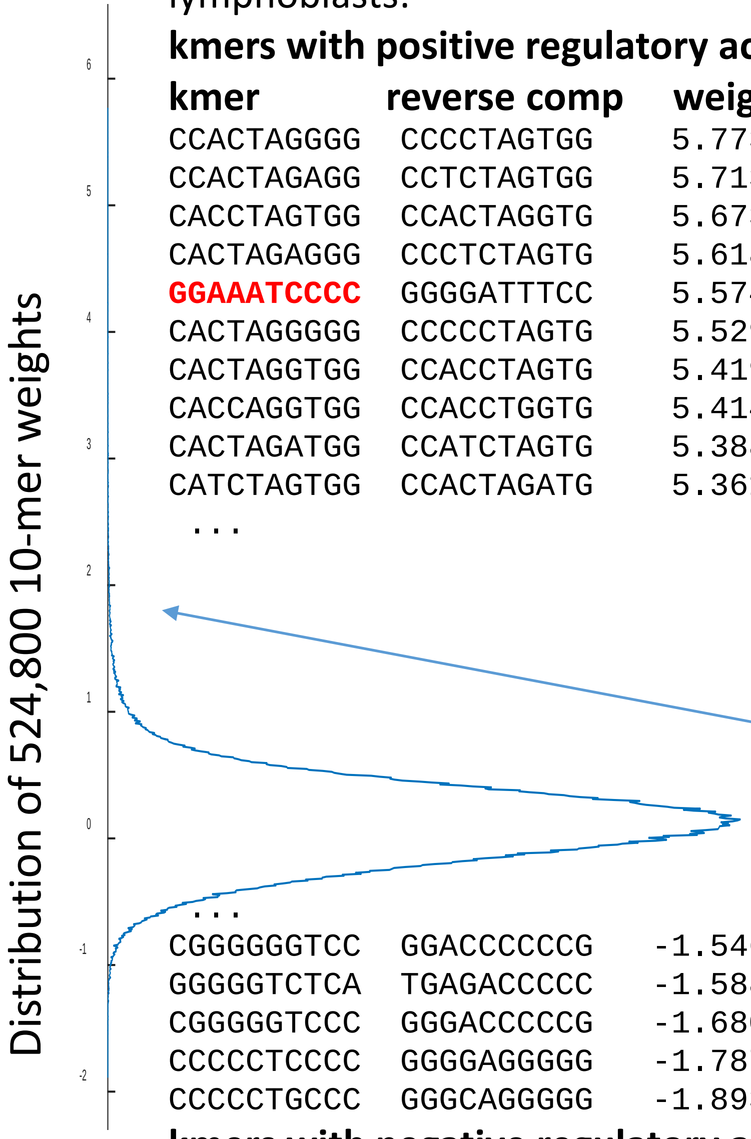
All unique 10-mers	SVM weights	
ATGACTCATC	3.275	Positive regulatory activity
ATGAGTCATC	3.147	
ATCATGTGAC	3.091	
GTCACATGAC	2.992	
⋮	⋮	
ACGAGAAACA	0.0002	Neutral
ACTATAACCA	0.0001	
ATTGCTAAGC	–0.0001	
GGATAAAATA	–0.0001	
⋮	⋮	
CAGGTGTGAG	–1.171	Negative regulatory activity
ATCACACCTG	–1.183	
ACACACCTGT	–1.253	
AATCCAGGTG	–1.282	

deltaSVM calculation

WT 10-mer	Weight	Variant 10-mer	Weight	Difference
TCTCTGCAA**C**	0.012	TCTCTGCAA**A**	0.298	0.286
CTCTGCAA**C**A	0.082	CTCTGCAA**A**A	0.104	0.021
TCTGCAA**C**AA	0.280	TCTGCAA**A**AA	0.114	–0.166
CTGCAA**C**AAA	0.330	CTGCAA**A**AAA	–0.029	–0.359
TGCAA**C**AAAG	0.441	TGCAA**A**AAAG	–0.025	–0.466
GCAA**C**AAAGA	0.784	GCAA**A**AAAGA	0.008	–0.776
CAA**C**AAAGAC	1.031	CAA**A**AAAGAC	0.109	–0.922
AA**C**AAAGACA	0.545	AA**A**AAAGACA	–0.453	–0.998
A**C**AAAGACAG	0.671	A**A**AAAGACAG	–0.516	–1.187
CAAAGACAGA	–0.036	**A**AAAGACAGA	–0.478	–0.442
▲		▲		

deltaSVM = –5.007

gkm-SVM Weights Encode Co-factor TFs Required to Predict Cell-specific Binding

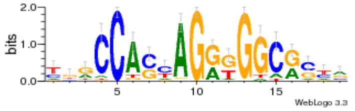


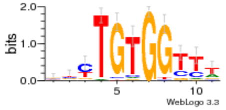
lymphoblasts:

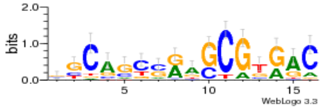
kmers with positive regulatory activity

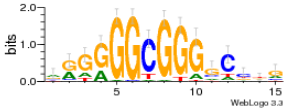
kmer	**reverse comp**	**weight**
CCACTAGGGG	CCCCTAGTGG	5.773
CCACTAGAGG	CCTCTAGTGG	5.713
CACCTAGTGG	CCACTAGGTG	5.673
CACTAGAGGG	CCCTCTAGTG	5.618
GGAAATCCCC	GGGGATTTCC	5.574
CACTAGGGGG	CCCCCTAGTG	5.529
CACTAGGTGG	CCACCTAGTG	5.419
CACCAGGTGG	CCACCTGGTG	5.414
CACTAGATGG	CCATCTAGTG	5.388
CATCTAGTGG	CCACTAGATG	5.361
...		
...		
CGGGGGGTCC	GGACCCCCCG	-1.546
GGGGGTCTCA	TGAGACCCCC	-1.588
CGGGGGTCCC	GGGACCCCCG	-1.680
CCCCCTCCCC	GGGGAGGGGG	-1.787
CCCCCTGCCC	GGGCAGGGGG	-1.893

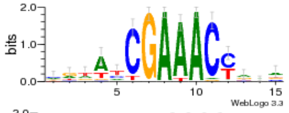
kmers with negative regulatory activity

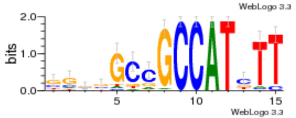
CTCF


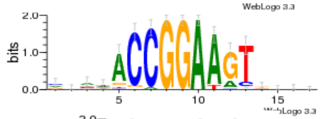
Bcl11a/Runx1


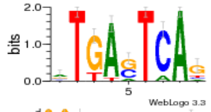
Pax5


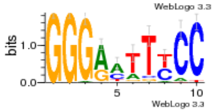
SP1


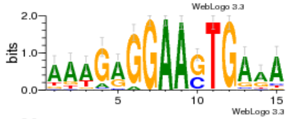
IRF4


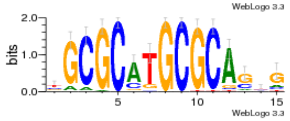
YY1


Gabpa


Batf/AP-1


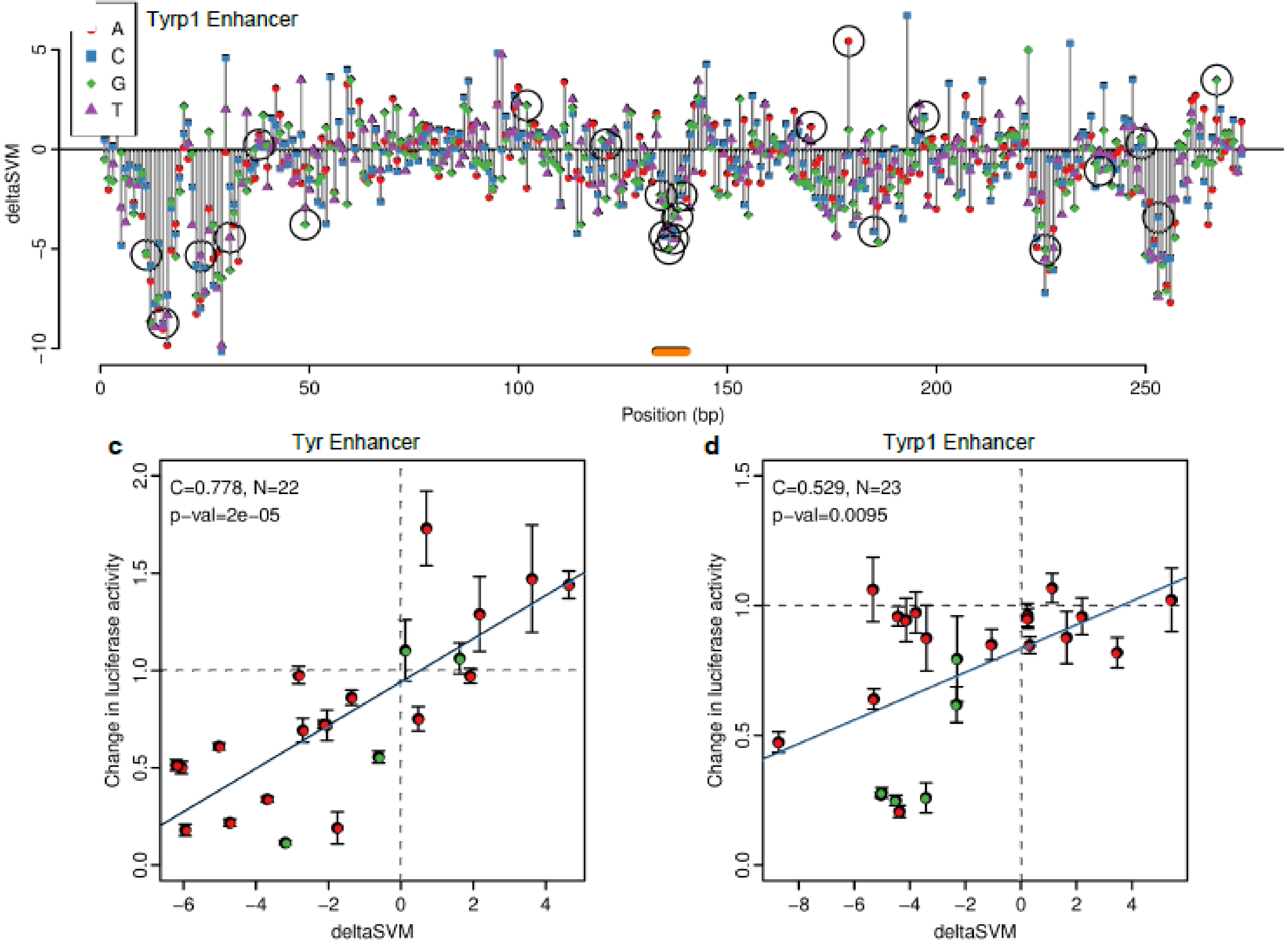
NFkB


PU1


NRF1


gkmSVM Predicts Effect of mutations in melanocyte enhancers
Direct Validation

Lee—Beer Nat Gen 2015



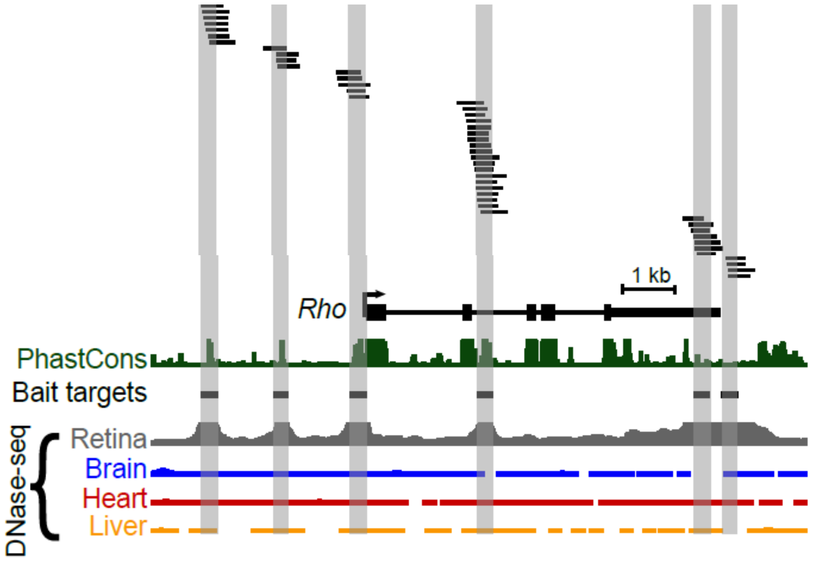
McCallion Lab

Similar results comparing to MPRAs:

Patwardhan et al., 2012

Kheradpour et al, 2013

gkmSVM Predicts MPRA Expression *in vivo* in Mouse Retina

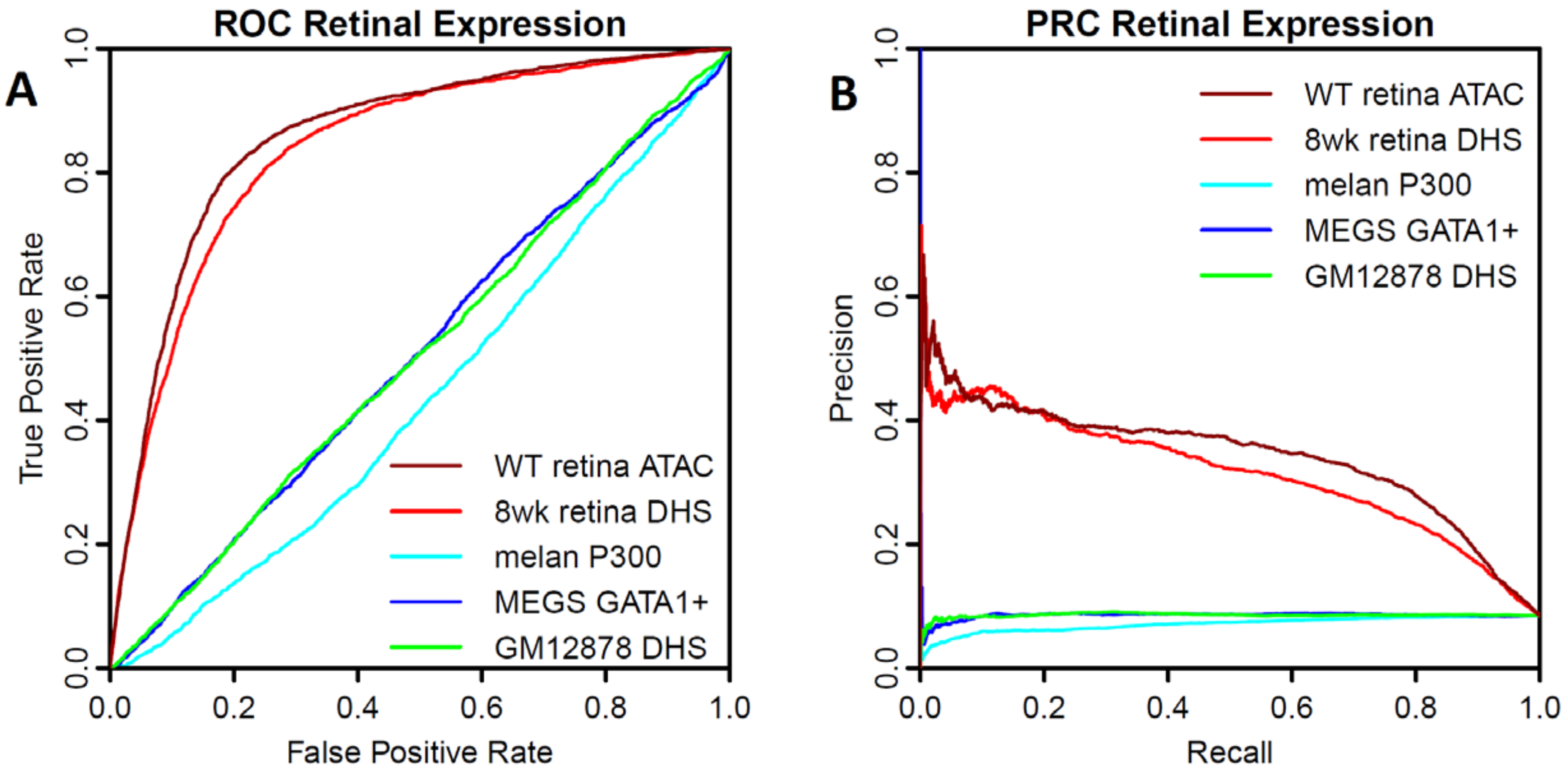


36005 DHS+ elements tested in mouse retina
(Shen—Corbo GR 2015)

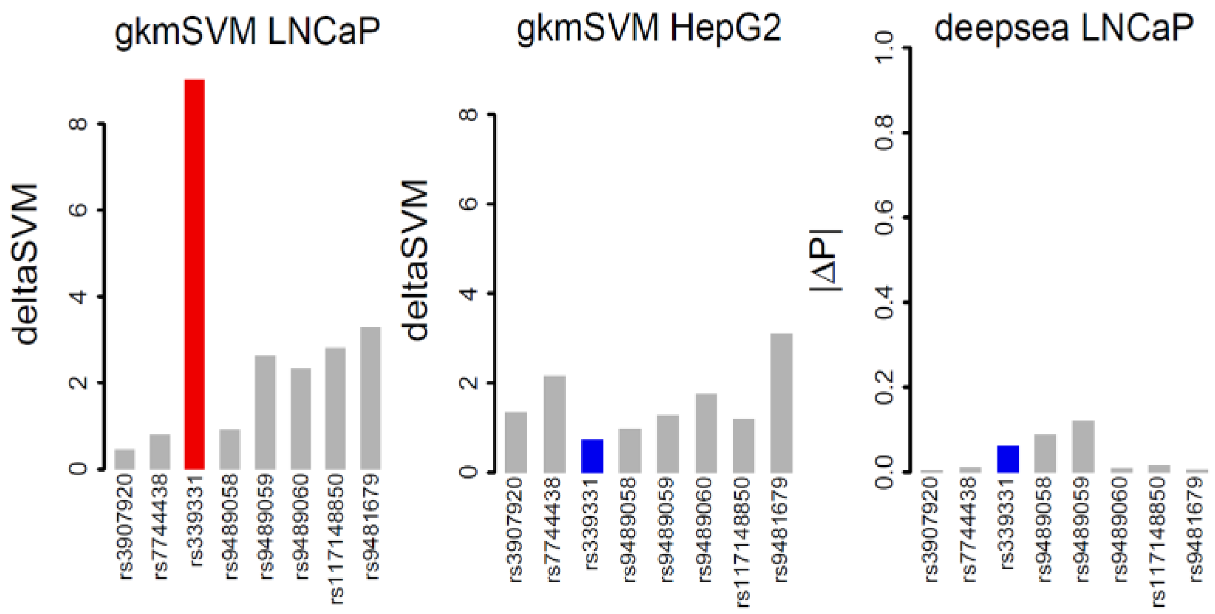
Pos: Strong expression 2156 seqs
Neg: Low expression 23147 seqs

gkmSVM trained on retina DHS or ATAC predicts expression, but not other cell types

Beer Human Mut 2017



Validated causal disease SNPs only score higher than flanking negative SNPs when gkmSVM trained on relevant cell type



some cell types (immune) are over-represented in ENCODE data set

Deepsea misses Rfx6 prostate cancer SNP, even though test set AUROC is high for LNCaP class

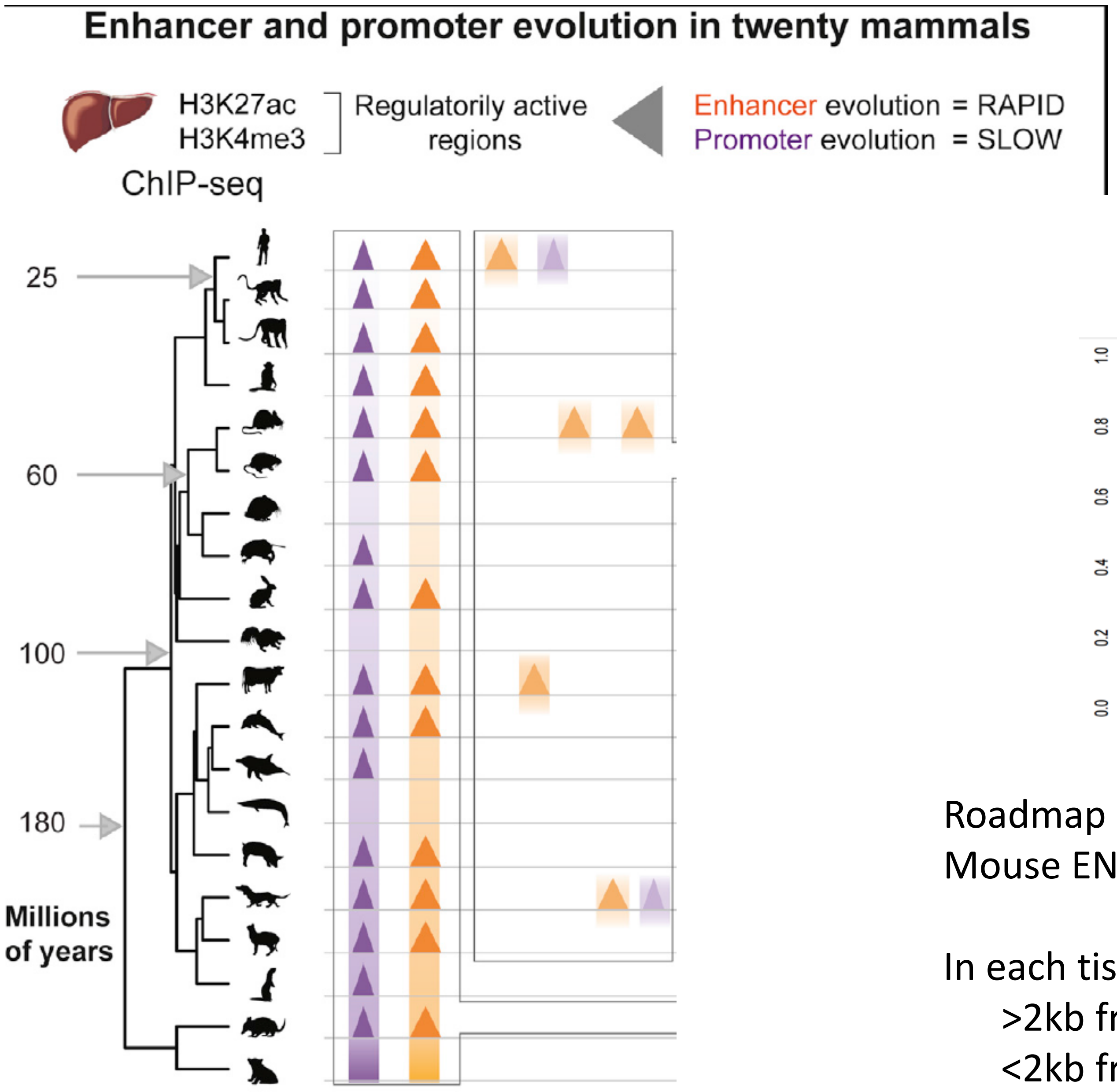
Beer Human Mutation 2017

gkm-SVM identifies *RFX6* prostate cancer SNP rs339331 (Huang NG 2014) (red) from among flanking SNPs when trained on:

- ENCODE LNCap DHS (prostate, red)
- but not HepG2 DHS (liver, blue)
- Deepsea predicts *RFX6* SNP has low ΔP for LNCaP DHS.

- Imbalance in training data can skew predictive accuracy across classes
- models with similar test-set CV AUROC can give very different predictions for variant impact

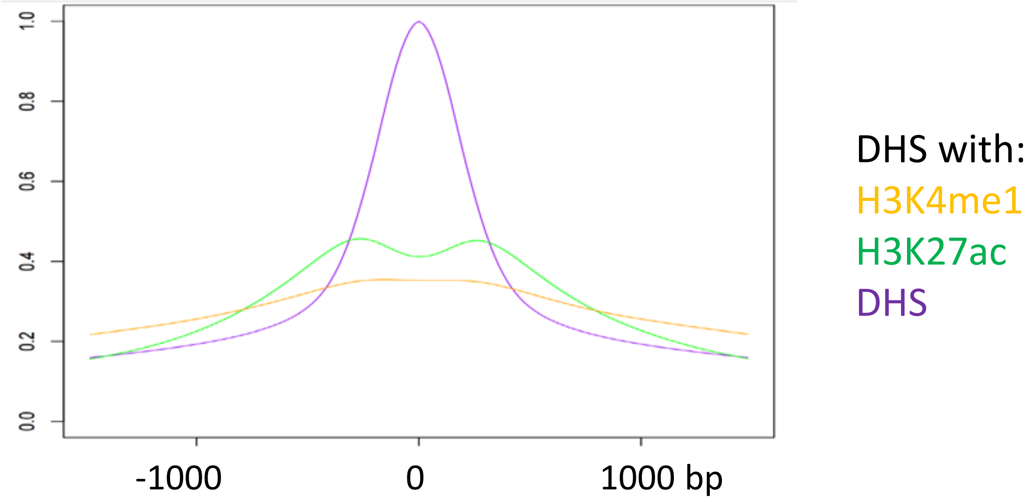
Villar, Berthelot,... Flicek, Odom   Cell 2015


Enhancer and promoter evolution in twenty mammals
H3K27ac
H3K4me3
Regulatorily active regions
Enhancer evolution = RAPID
Promoter evolution = SLOW
ChIP-seq
25
60
100
180
Millions of years

TFBS are conserved,
liver H3K27Ac activity is conserved
but enhancers are not alignable, why?

Genome-wide Cross-correlation of
DHS with chromatin marks

$$(f * g)(\Delta) = \sum_x f(x) g(x+\Delta)$$



Roadmap DHS (73 Human Tissues)
Mouse ENCODE DHS (53 Mouse Tissues)

In each tissue, select 10000 strongest cell-specific peaks
>2kb from TSS = enhancers
<2kb from TSS = promoters → train vs GC matched neg set

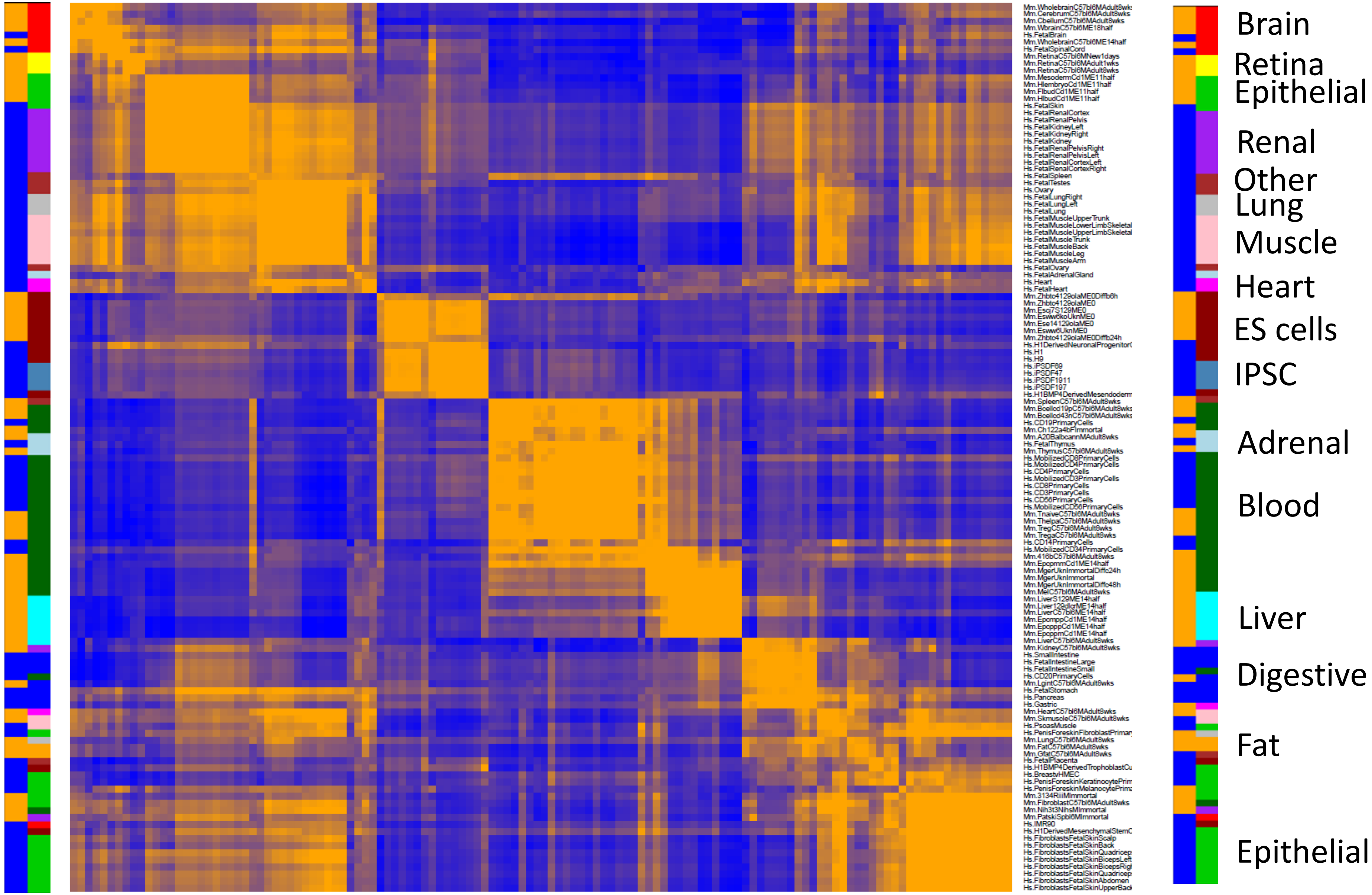
gkm-SVM Identifies Similar Sequence Features in Matched Human and Mouse Tissues

Train gkm-SVM on 10000 mouse ENCODE (53) and human Roadmap DHS (73, Stam)

Cluster by correlation across all gapped kmer weights $C(w_i,w_j)$: median AUC > 0.9

🟦 human low corr

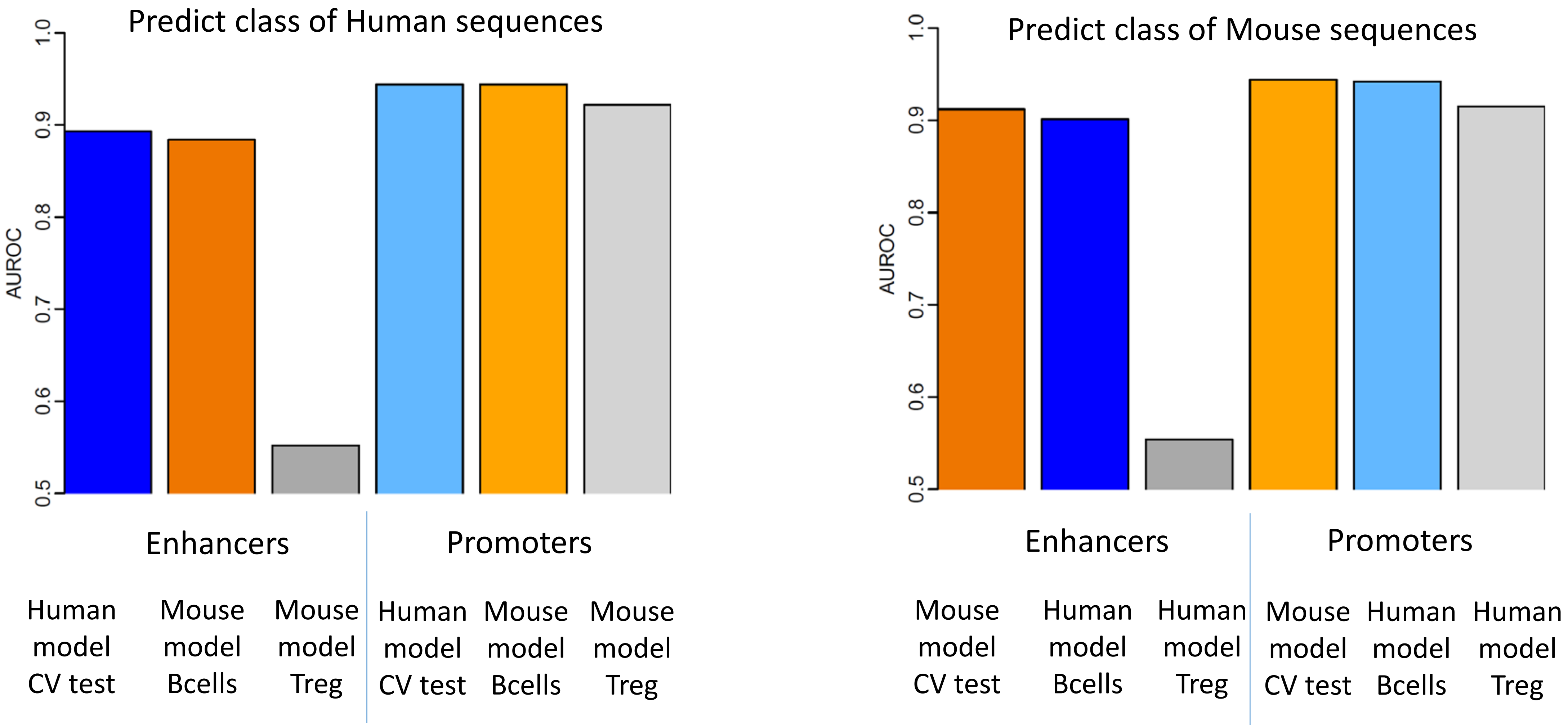
🟧 mouse high corr



Select reciprocal-best-hit: 11 matched Human-Mouse sample pairs

Human **ROADMAP** tissue/cell-type	Mouse **ENCODE** tissue/cell-type	Top wts → common TFBS
Small_Intestine	Lgint MAdult8wks	AP1, HNF4a, KLF
Fetal_Spinal_Cord	Wholebrain ME14half	Rfx2, HEB, Chx10
Heart	Heart MAdult8wks	Mef2, NF1/Tlx, AP1, Gata
Psoas_Muscle	Skmuscle MAdult8wks	Mef2, AP1, MyoG
CD19_Primary_Cells ⟷	**Bcell CD19+ MAdult8wks**	SpiB, Runx1, NFkB, PU.1
iPS_DF_19_7	ES-CJ7 S129 ME0	Oct, Sox, Klf, TEAD1
Fetal_Brain	Wbrain ME18half	Rfx2, Chx10, Olig2
Penis_Foreskin_Fibroblast_Primary_Cells	Nih3t3 Nihs MImmortal	AP1, HEB, Runx1, TEAD1
Mobilized_CD34_Primary_Cells	EpcpmmCd1ME14half	SpiB, Runx1, Gata, AP1
Fetal_Thymus	Thymus MAdult8wks	Runx1, Tcf, ETS1
Mobilized_CD3_Primary_Cells	Treg MAdult8wks	BATF, ETS1, NFkB

Train gkm-SVM on B Cell regions in one species, score B Cell sequences in other species


Predict class of Human sequences

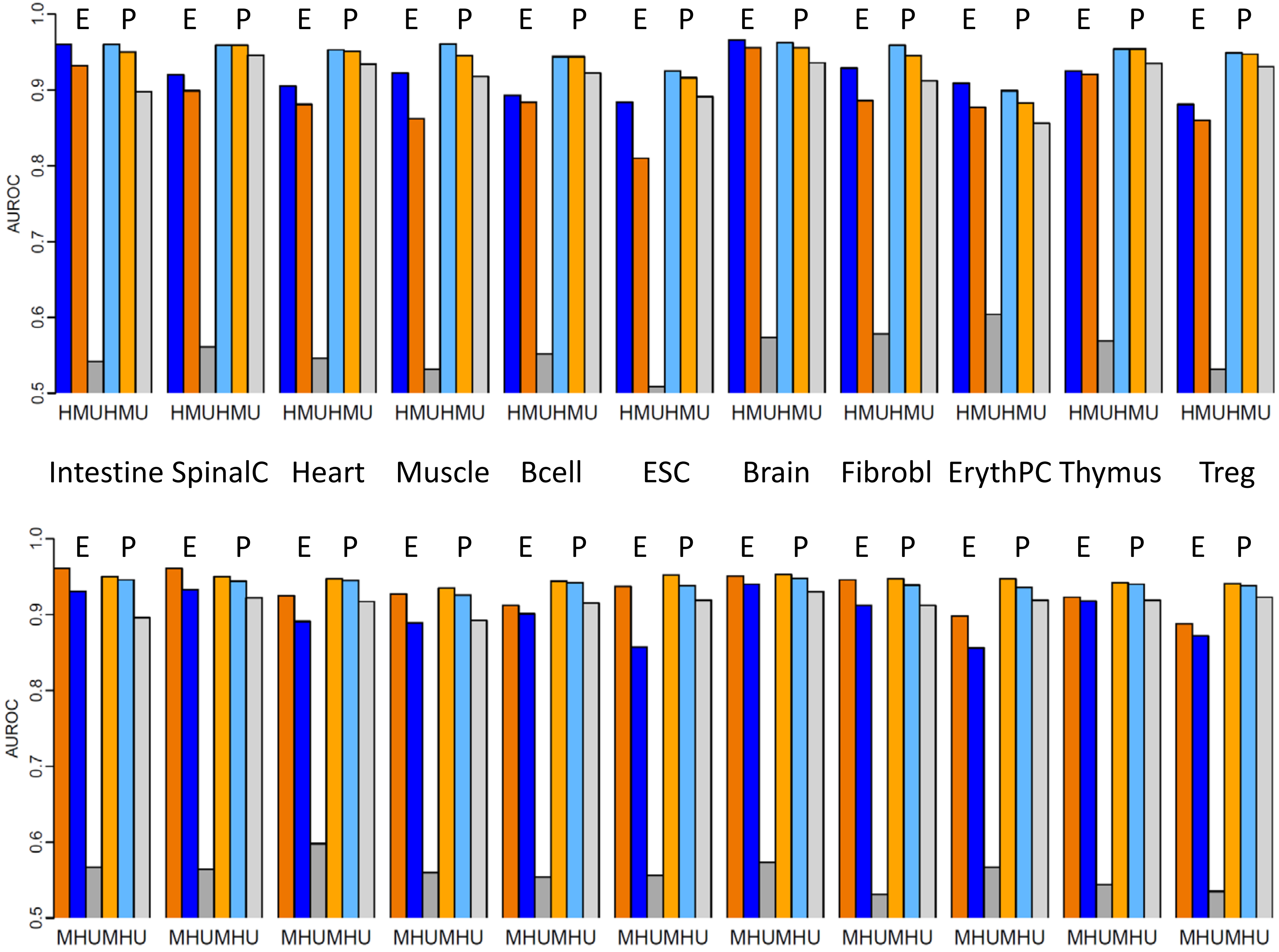
Predict class of Mouse sequences

- Both cell-type specific enhancers and promoters are predictable from gapped kmer sequence features
- Regulatory vocabulary of cell types is conserved between human and mouse for **both enhancers and promoters**
- **Enhancer** regulatory vocabulary is cell-specific
- **Promoter** regulatory vocabulary is more cell-type independent

Score Human regions

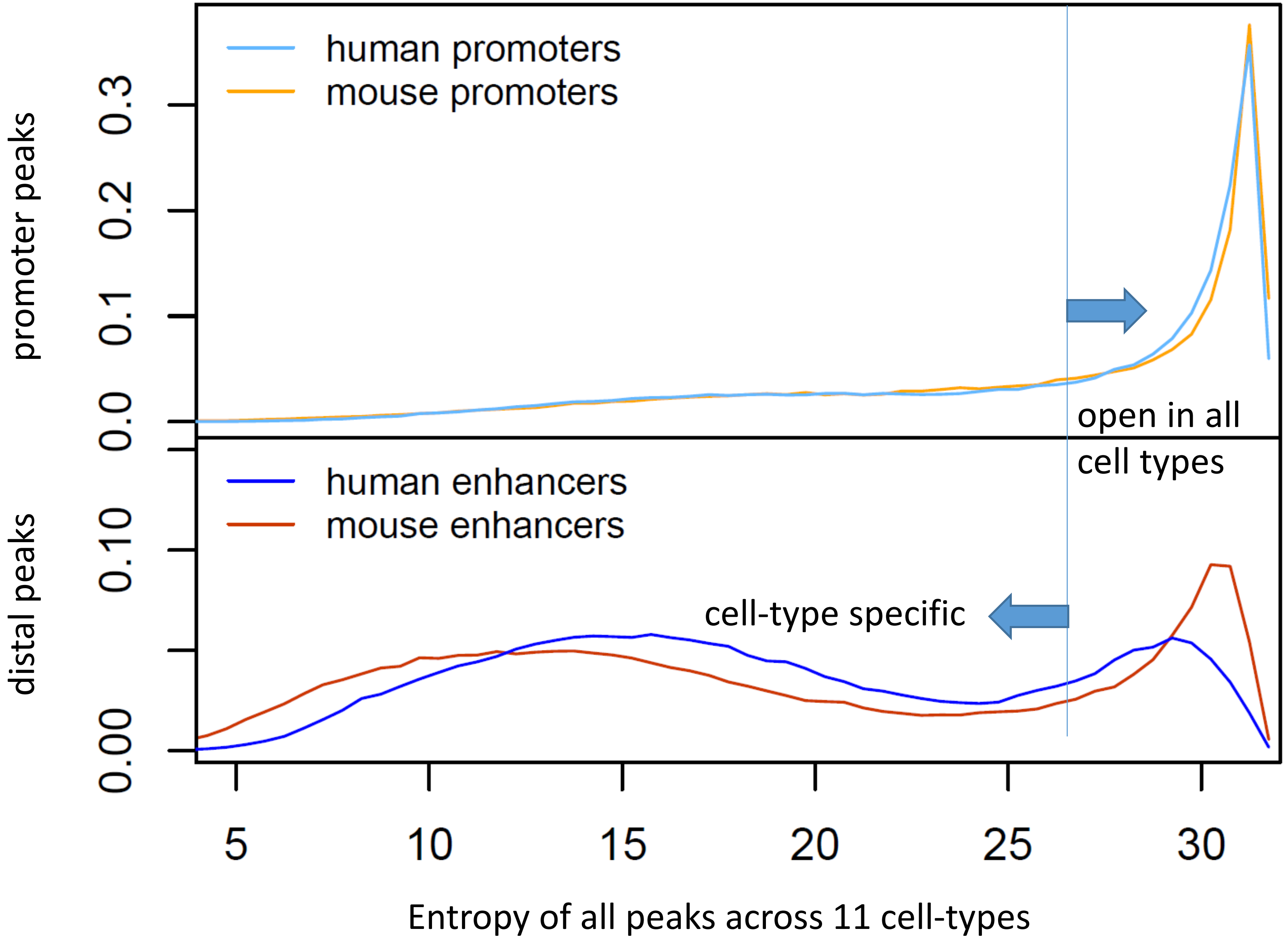
seq model:

- 🟦 Human Enh
- 🟧 Mouse Enh
- ▨ Unmatched Enh
- 🟦 Human Prom
- 🟧 Mouse Prom
- □ Unmatched Prom



Score Mouse regions

Entropy Distribution Confirms That Promoters Are Generally Open Across All 11 Cell-types

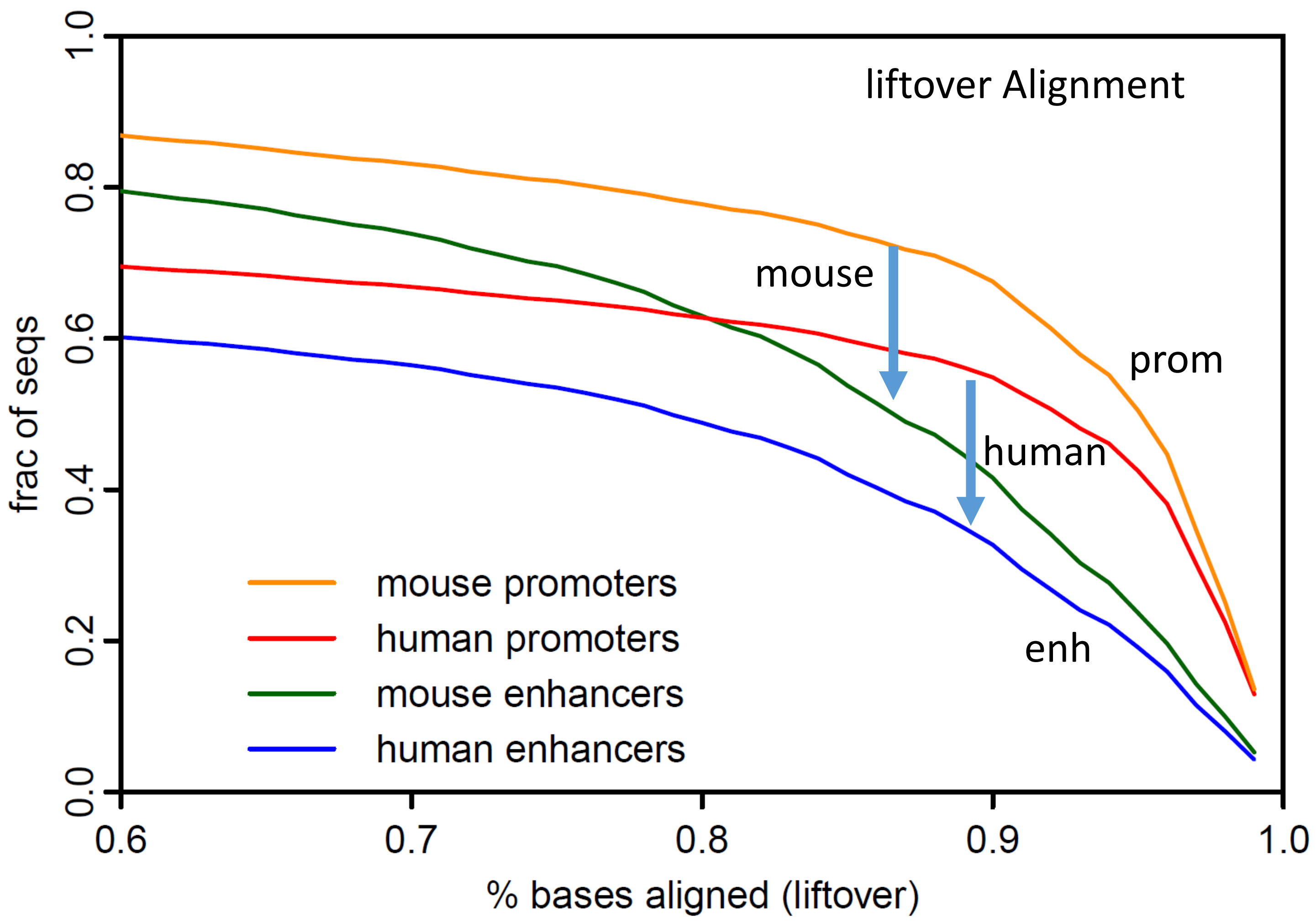


Consistent with summary statistics of:

Cheng 2014

Vierstra 2014

Although Regulatory Vocabulary of Enh and Prom are Both Conserved, Enhancers are much less Alignable than Promoters



similar conclusions using bnMapper or PhyloP

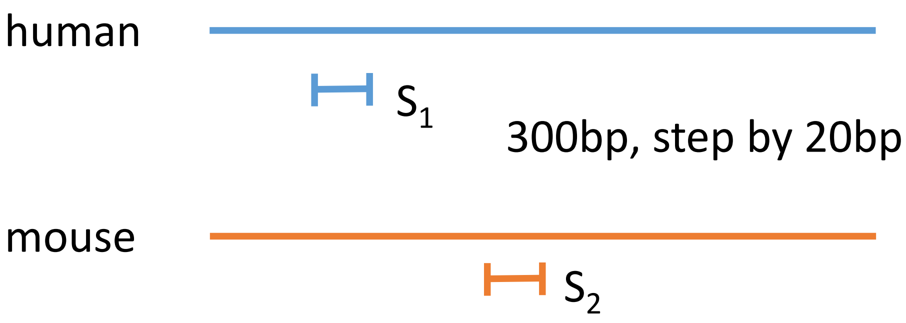
Gapped kmer word composition can detect conservation of enhancers between human and mouse

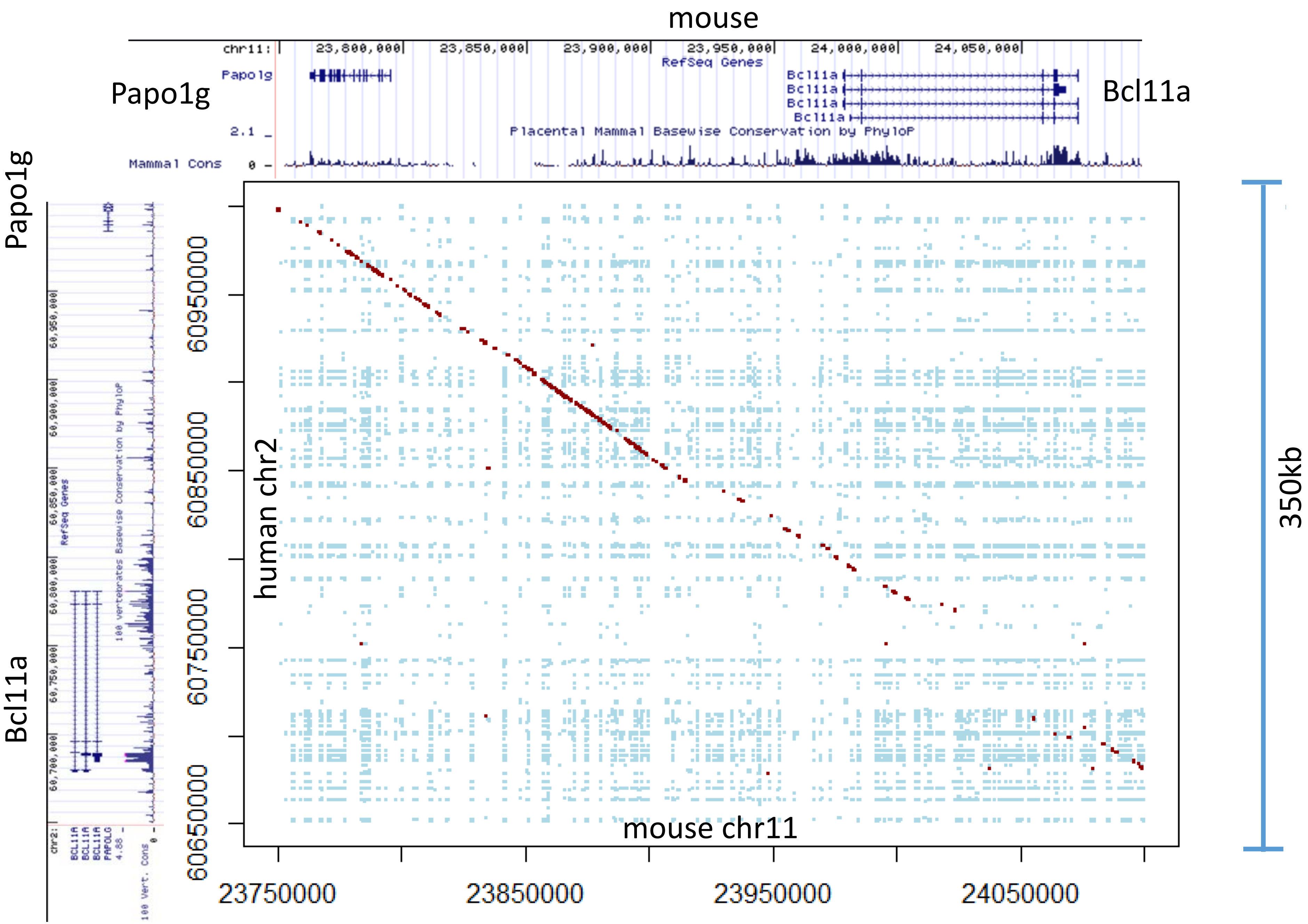
But sequence alignment cannot

Use gapped kmer kernel to detect conservation:

$$K(S_1,S_2)=\frac{\langle f^{S_1},f^{S_2}\rangle}{\|f^{S_1}\|\|f^{S_2}\|}$$

dot product of normalized gapped kmer count vectors for two sequences S_1 S_2





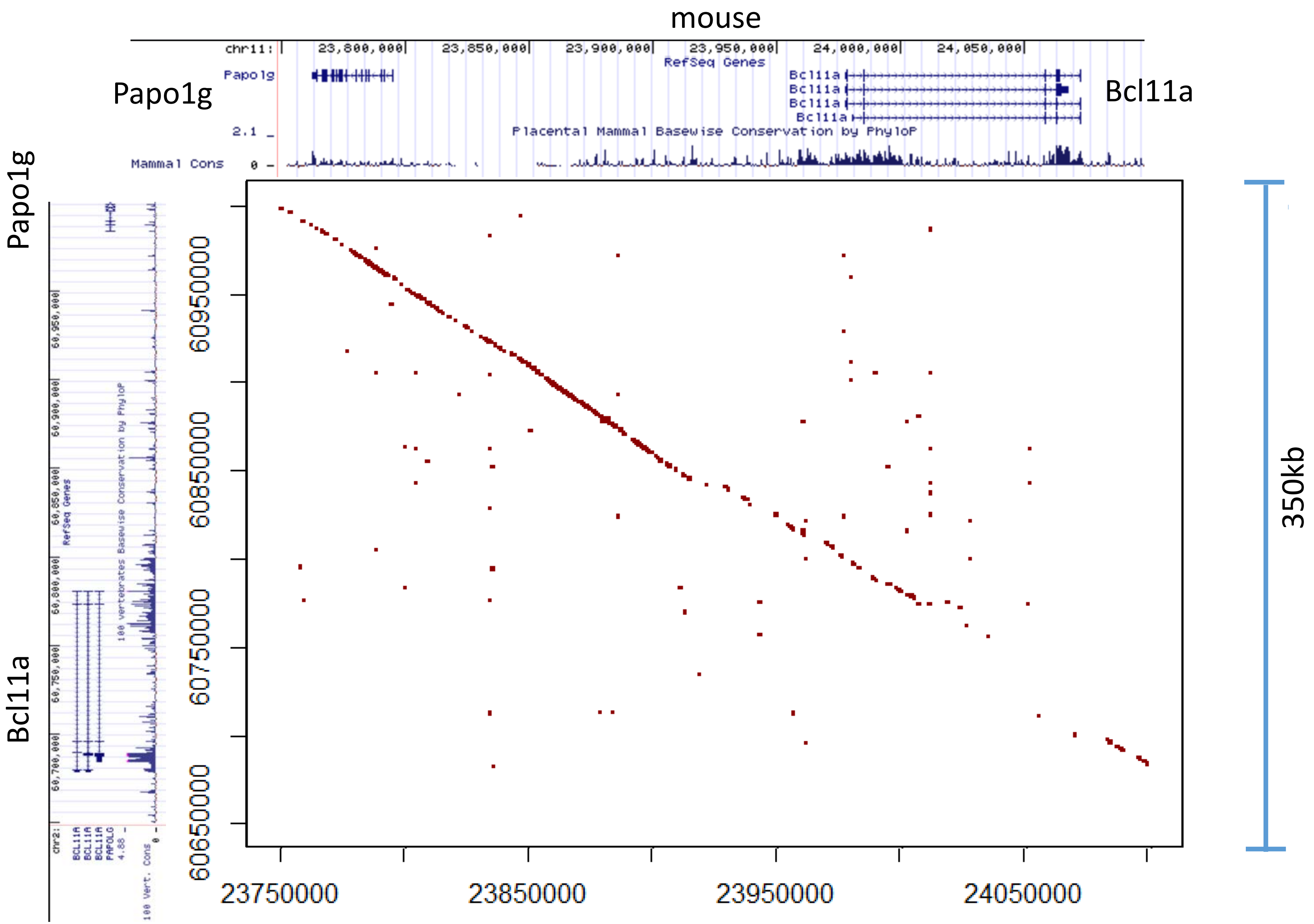
K>0.2

🟥 <=20 hits

🟦 >20 hits:
SINE
LINE

>300,000,000 comparisons of two 300bp seqs

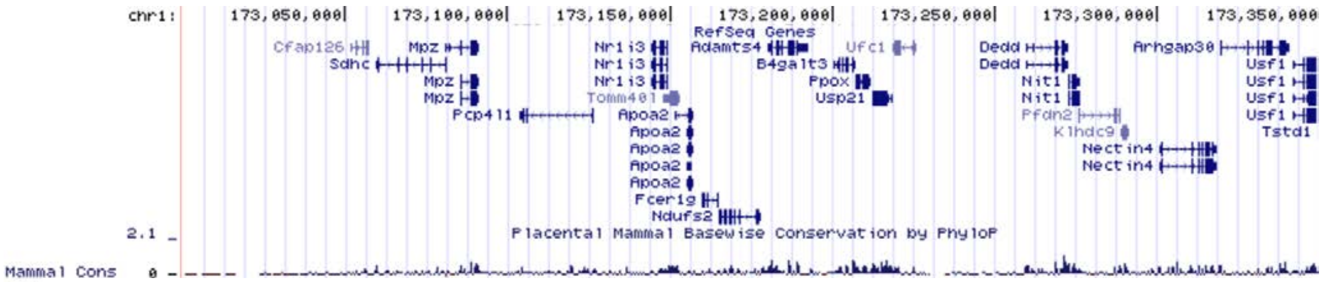
diagonal = syntenic blocks of similar gapped kmer composition



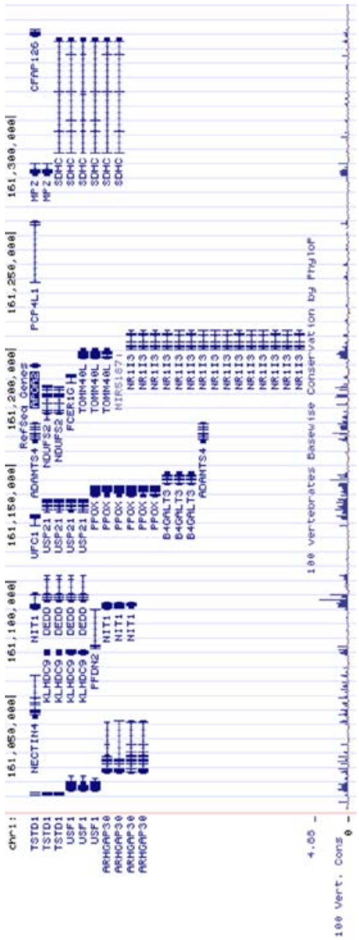
Better to remove repeats with PCA

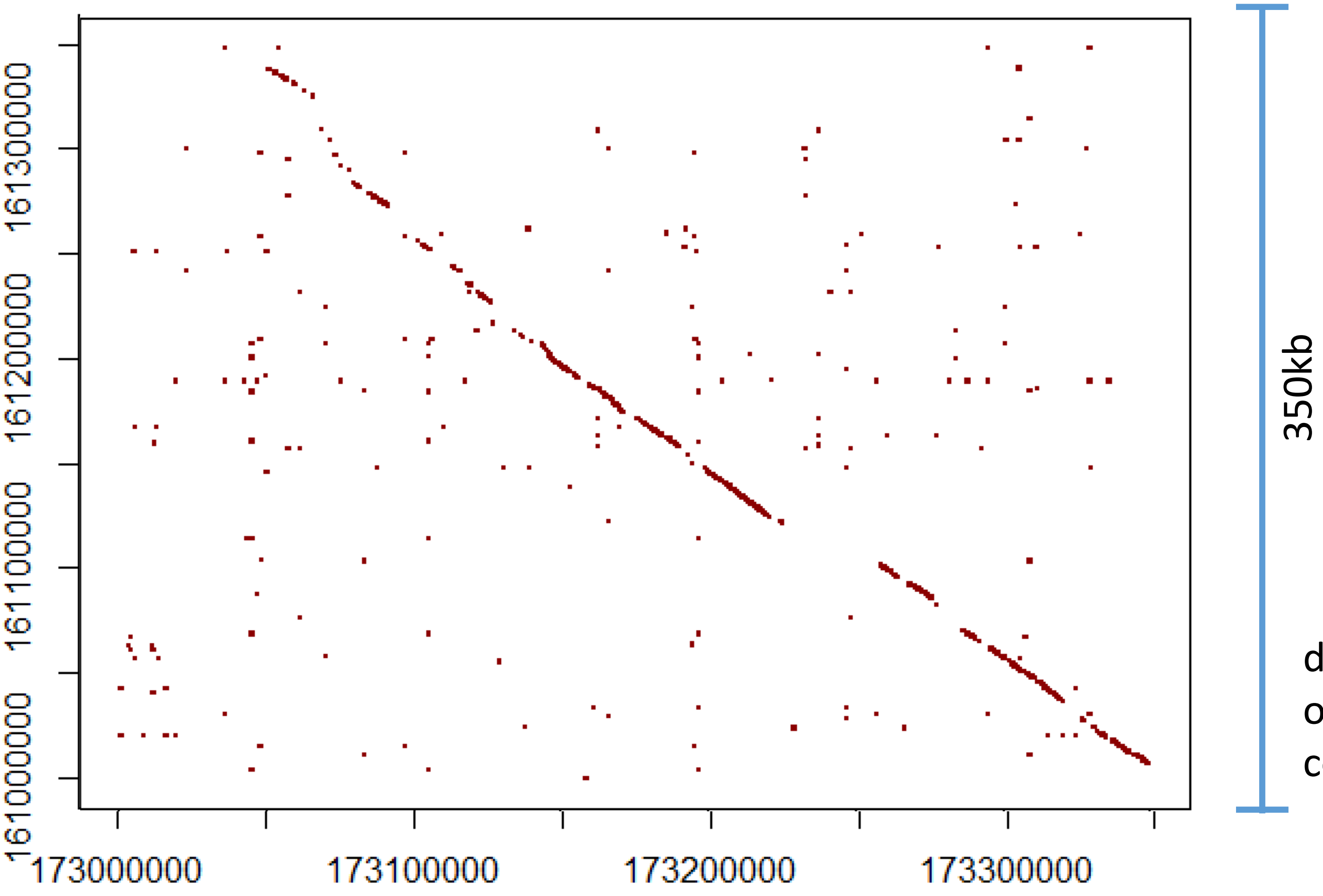
diagonal = syntenic blocks of similar gapped kmer composition

mouse Apoa2



human Apoa2





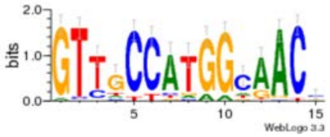
diagonal = syntenic blocks of similar gapped kmer composition

Mouse gkmSVM regulatory vocabulary interprets human GWAS Schizophrenia SNP

Schizophrenia SNP has large deltaSVM when gkm-SVM trained on

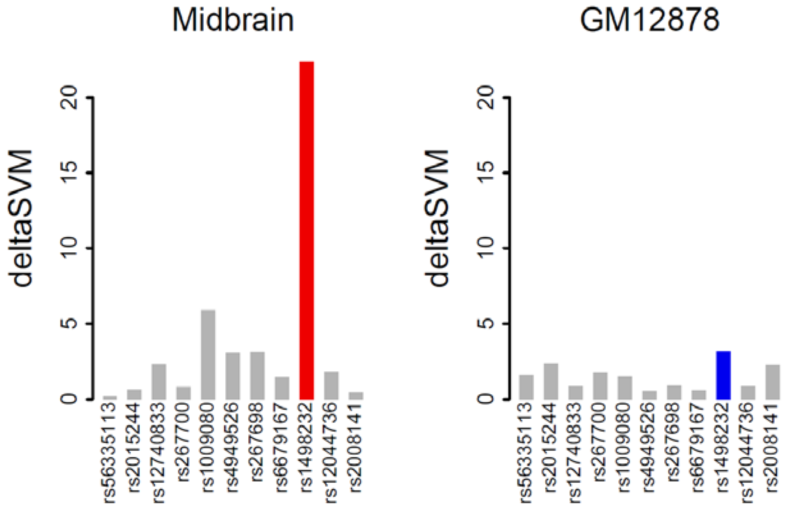
- adult mouse dentate gyrus ATAC-seq (Song Nature 2017)
- or midbrain DA neurons ATAC-seq (McCallion Lab)

clearly disrupts RFX BS:

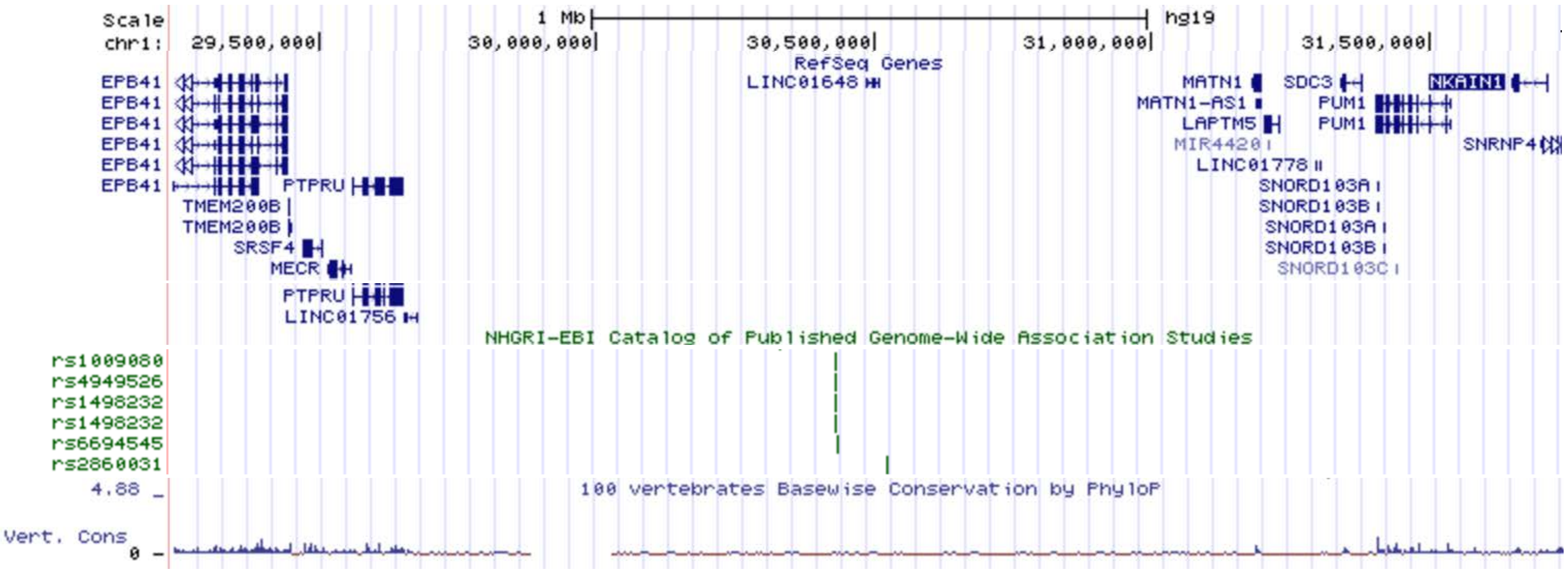


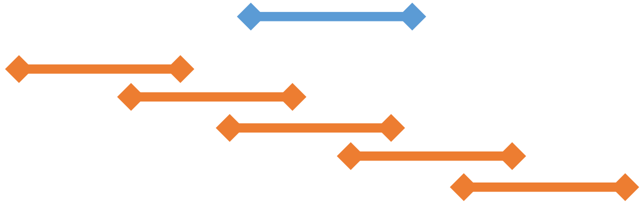
```
rs1498232
Ref allele: CCGTTTCCATGGCAACCAG   0.53
Alt allele: CCGTTTCCACGGCAACCAG   0.47
Ref 10-mer    wt   Alt 10-mer    wt      Diff
CCGTTTCCAT    1.27 CCGTTTCCAC    0.25    1.02
CGTTTCCATG    2.07 CGTTTCCACG    0.75    1.32
GTTTCCATGG    6.08 GTTTCCACGG    2.20    3.89
TTTCCATGGC    2.86 TTTCCACGGC    1.01    1.85
TTCCATGGCA    2.15 TTCCACGGCA    0.76    1.39
TCCATGGCAA    3.66 TCCACGGCAA    1.32    2.35
CCATGGCAAC    7.93 CCACGGCAAC    2.84    5.09
CATGGCAACC    4.31 CACGGCAACC    1.61    2.70
ATGGCAACCA    2.45 ACGGCAACCA    0.90    1.55
TGGCAACCAG    1.96 CGGCAACCAG    0.77    1.19
                                deltaSVM=-22.35
```

CCATGGCAAC	7.927474
CCATGGAAAC	6.0849012
CCATAGCAAC	5.8148892
CCATGGTAAC	5.515136
CATGGCAACC	4.3127716
CATGGCAACA	4.298156
CTATGGCAAC	4.1949162
CCTTGGCAAC	3.8884124
CCCTGGCAAC	3.7884638
CCATGGGAAC	3.7854512
CCATGACAAC	3.7819106
CTATGGAAAC	3.729401
TCCATGGCAA	3.6644538
...	

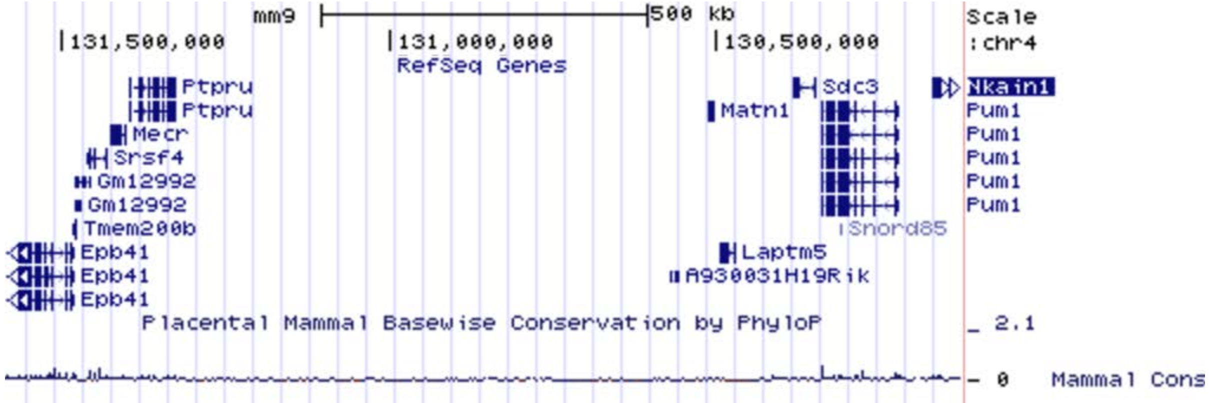


human locus

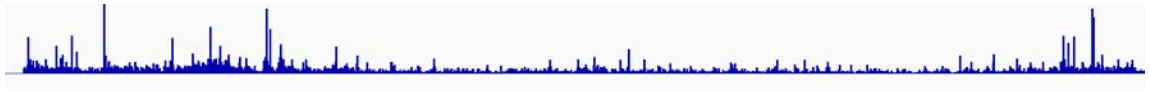


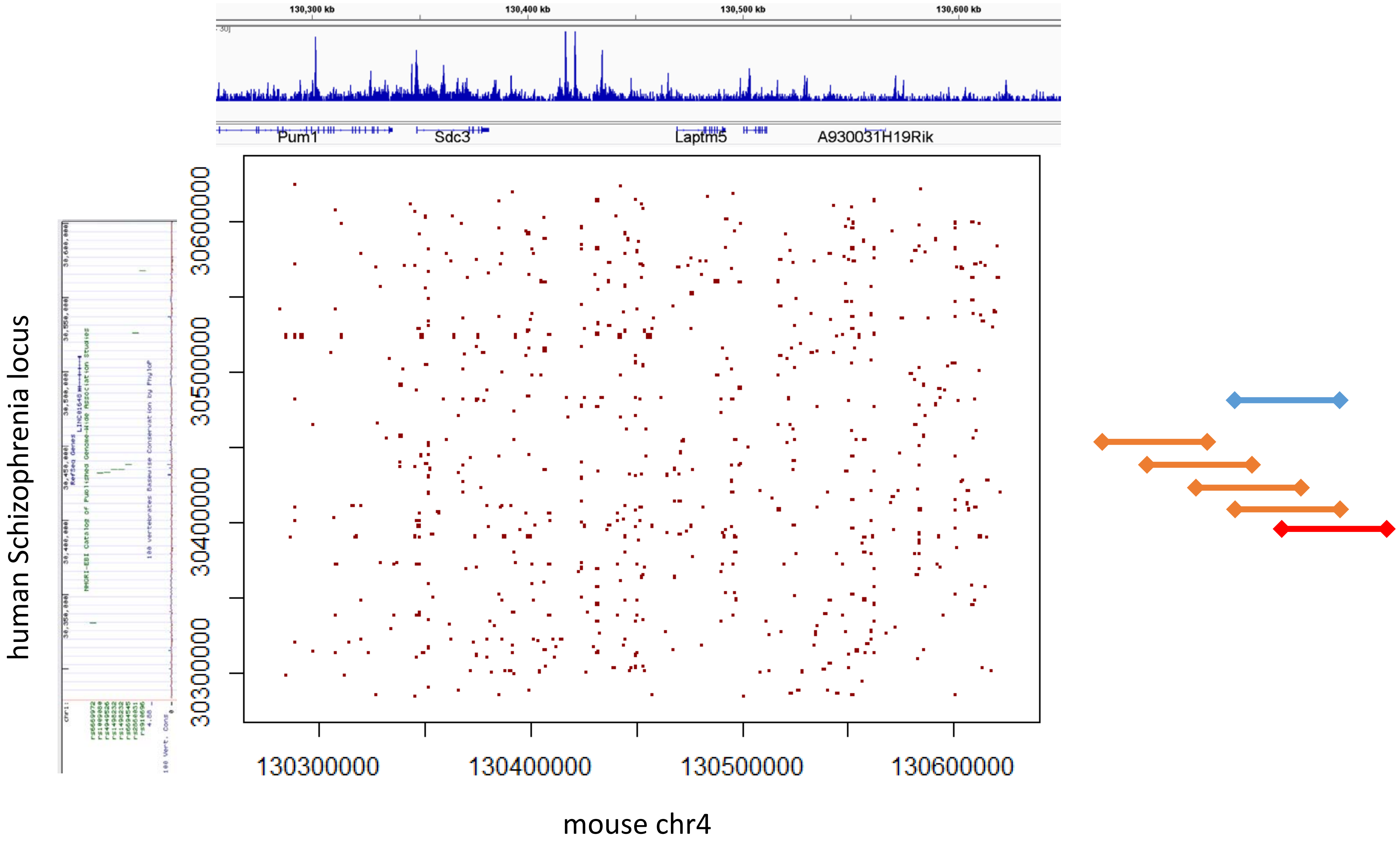


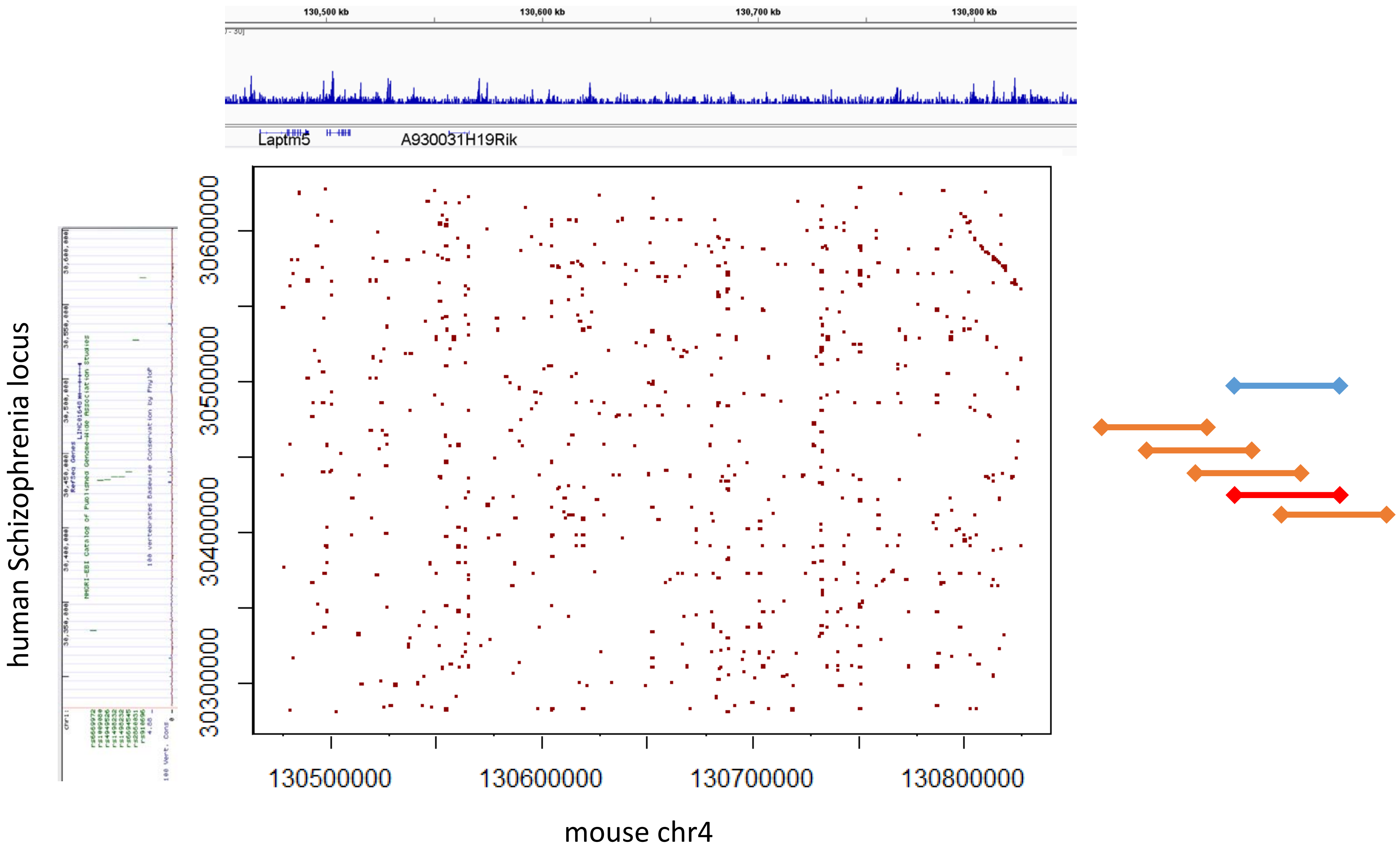
mouse locus

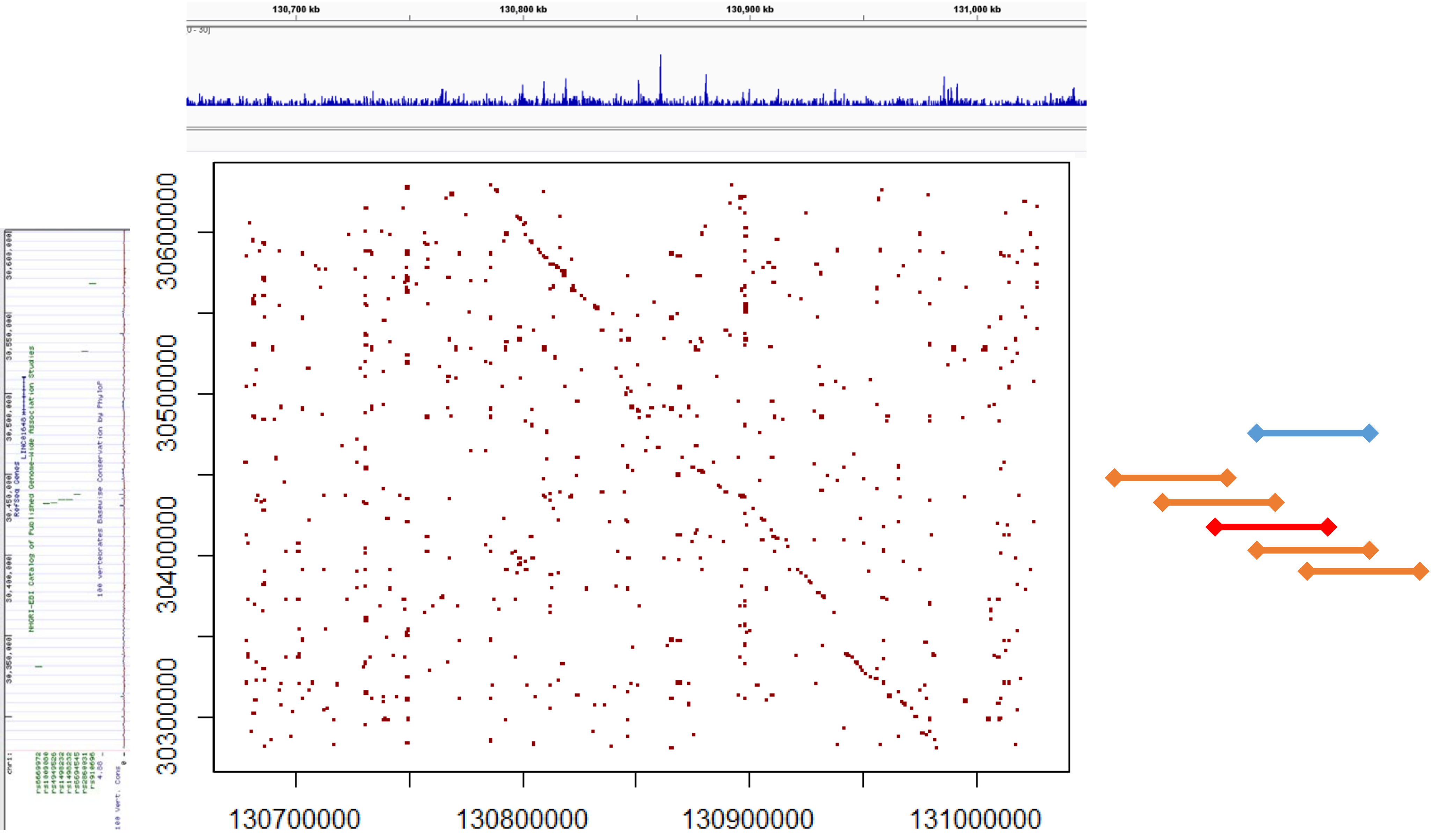


adult mouse
dentate gyrus ATAC-seq



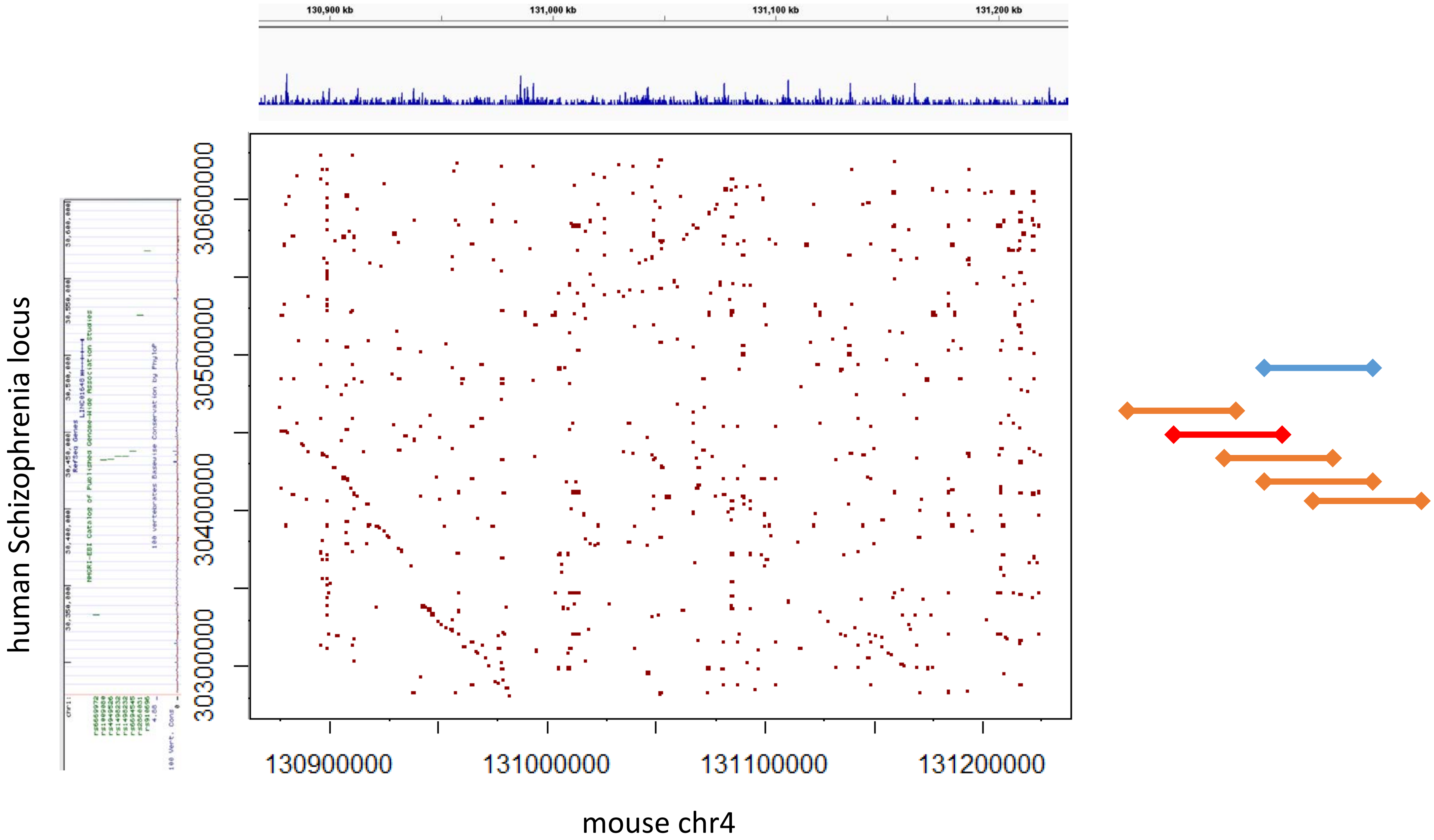

130,300 kb
130,400 kb
130,500 kb
130,600 kb
Pum1
Sdc3
Laptm5
A930031H19Rik
human Schizophrenia locus
30600000
30500000
30400000
30300000
130300000
130400000
130500000
130600000
mouse chr4


130,500 kb
130,600 kb
130,700 kb
130,800 kb
Laptm5
A930031H19Rik
human Schizophrenia locus
chr1:
30,350,000
30,400,000
30,450,000
30,500,000
30,550,000
30,600,000
RefSeq Genes
LINC01648
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
rs6669972
rs1009080
rs4949526
rs1498232
rs1498232
rs6694545
rs2858831
rs910696
100 vertebrates Basewise Conservation by PhyloP
4.88
0
100 Vert. Cons
30600000
30500000
30400000
30300000
130500000
130600000
130700000
130800000
mouse chr4

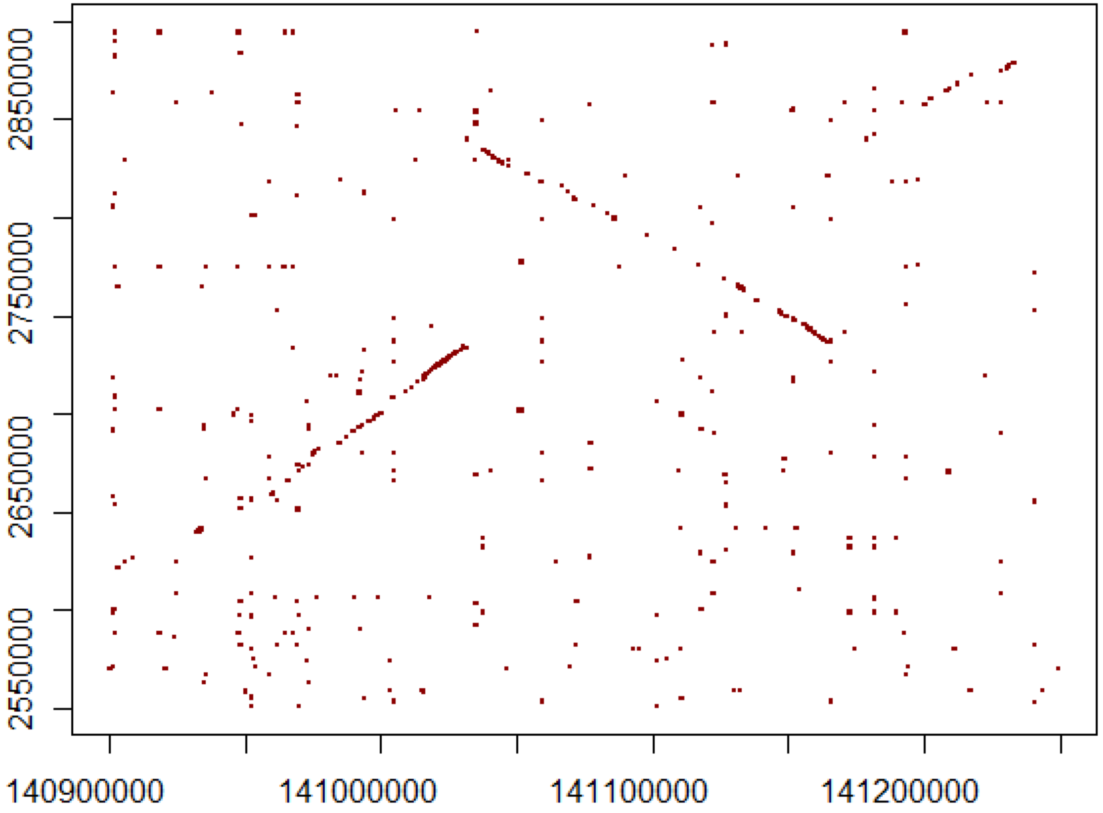


human Schizophrenia locus

mouse chr4


130,900 kb
131,000 kb
131,100 kb
131,200 kb
human Schizophrenia locus
chr1:
30,350,000
30,400,000
30,450,000
30,500,000
30,550,000
30,600,000
RefSeq Genes
LINC01648
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
rs6669972
rs1009080
rs4949526
rs1498232
rs1498232
rs6694545
rs2856831
rs910696
4.88 _
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0 _
30300000
30400000
30500000
30600000
130900000
131000000
131100000
131200000
mouse chr4

gkm-SVM detects inversion in gna12 locus



Summary:

- Cell-specific enhancers and promoters are predictable from gapped kmer sequence features
- Regulatory vocabulary of cell types is conserved between human and mouse for **both enhancers and promoters**
- **Enhancer** regulatory vocabulary is cell-specific
- **Promoter** regulatory vocabulary more cell-type independent
- **gkm-SVM** kernel can detect syntenic blocks of conserved gapped-kmer composition, independent of cell type

Future directions:

- Use segment detection to map conserved chains
- Generalize across multiple species
- Develop heuristic algorithm to apply genome wide