deltaSVM: Predicting the Impact of Regulatory Mutations from DNA Sequence

Mike Beer JHU, Biomedical Engineering
Dongwon Lee Inst. of Genetic Medicine
Dave Gorkin
Maggie Baker
Andy McCallion
Ben Strober
Ale Asoni
M. Ghandi

7/10/2015

JOHNS HOPKINS

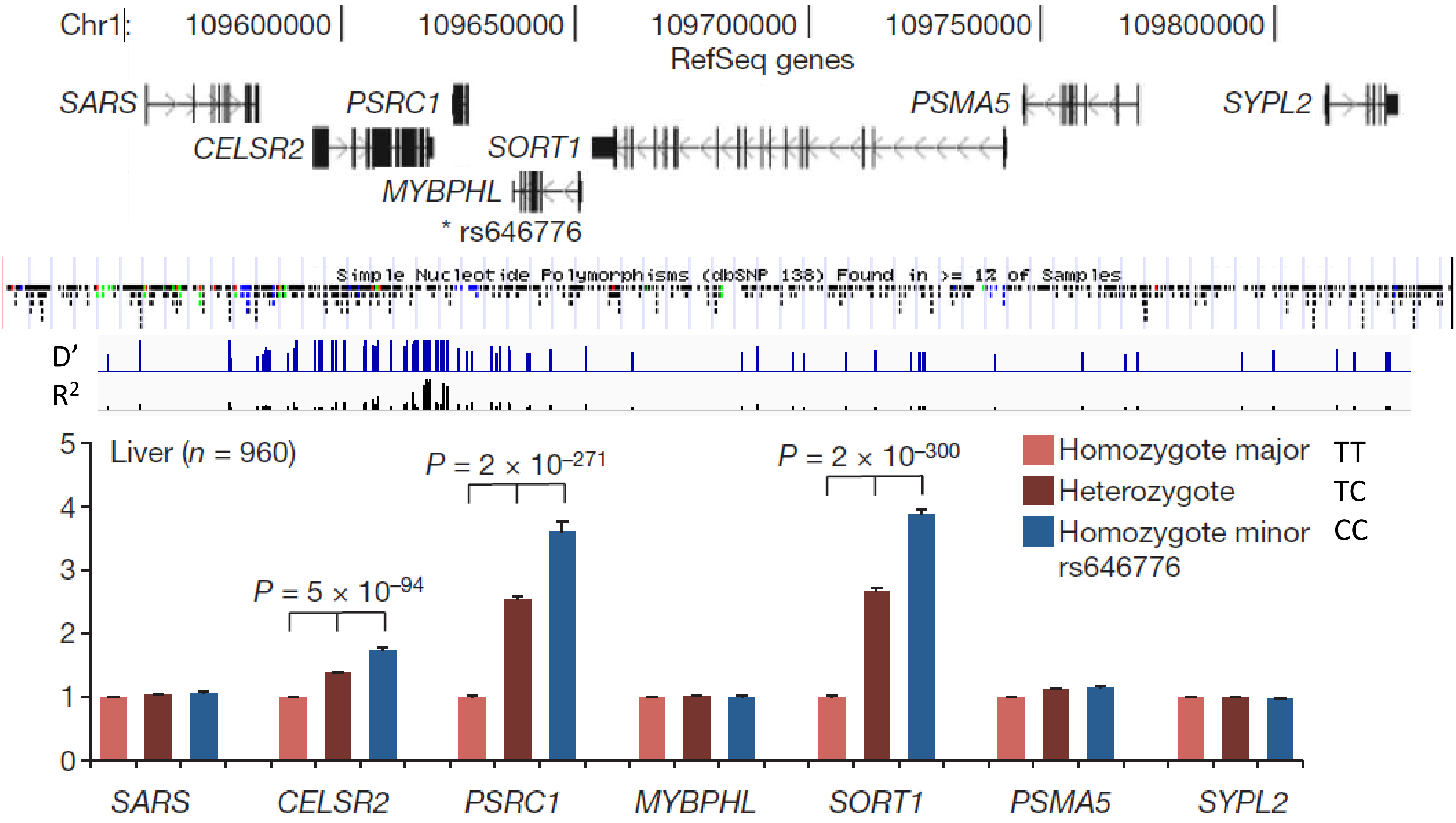
Support: NHGRI, ENCODE

- >5000 non-coding SNPs associated with common diseases
- SVM classifiers predict regulatory regions from DNA sequence features
  -- trained on Chip-seq, DHS, ATAC, histone active regions
  -- identify cofactors
- kmer based methods identify unbiased regulatory vocabulary
- Can these methods predict the quantitative impact of mutations?

From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus

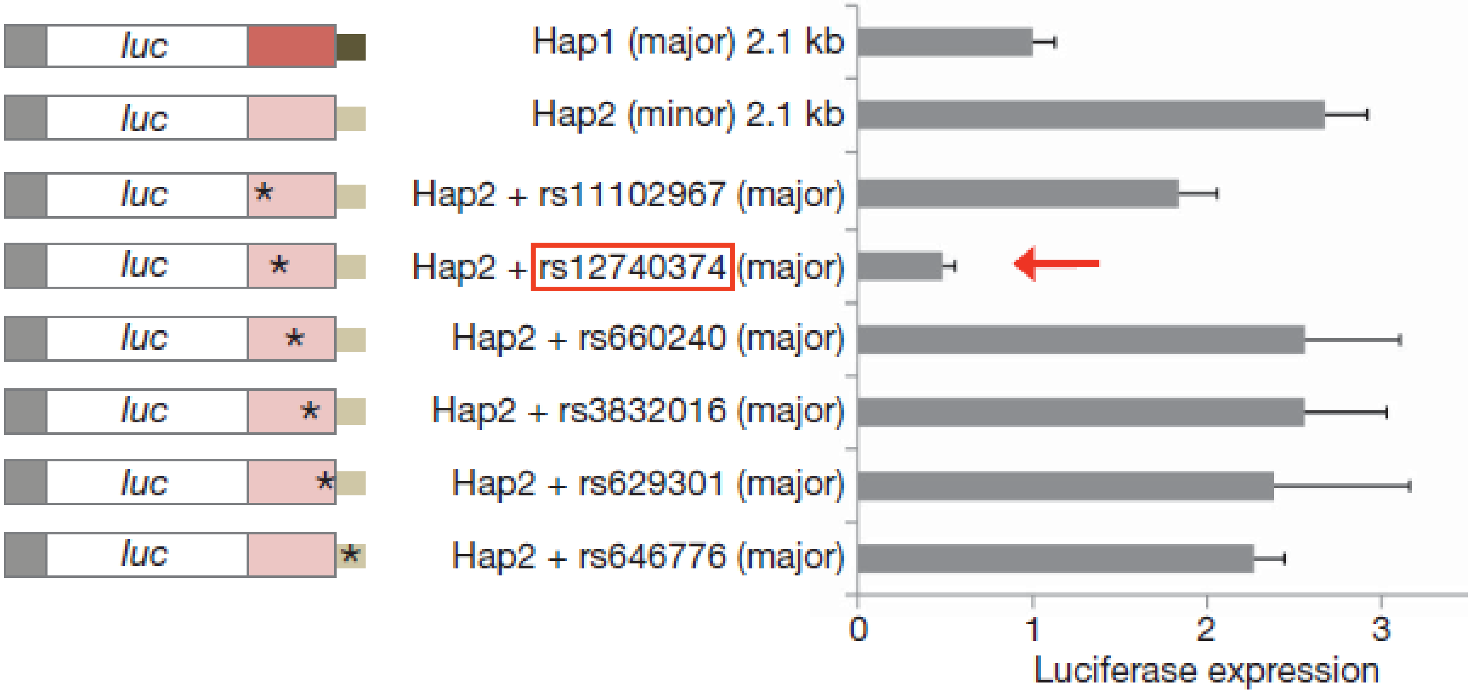
Musunuru et al Nature (2010)

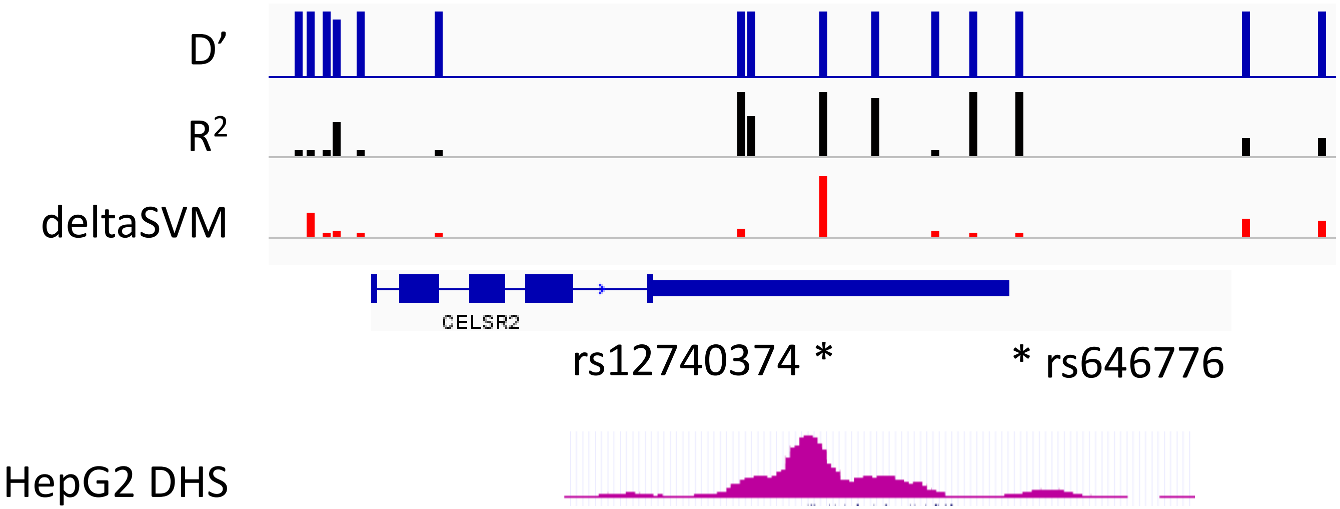
- rs646776 C vs T: LDL-C and MI $p<10^{-170}$ ~100,000 indiv, 24% increase LDL-VS



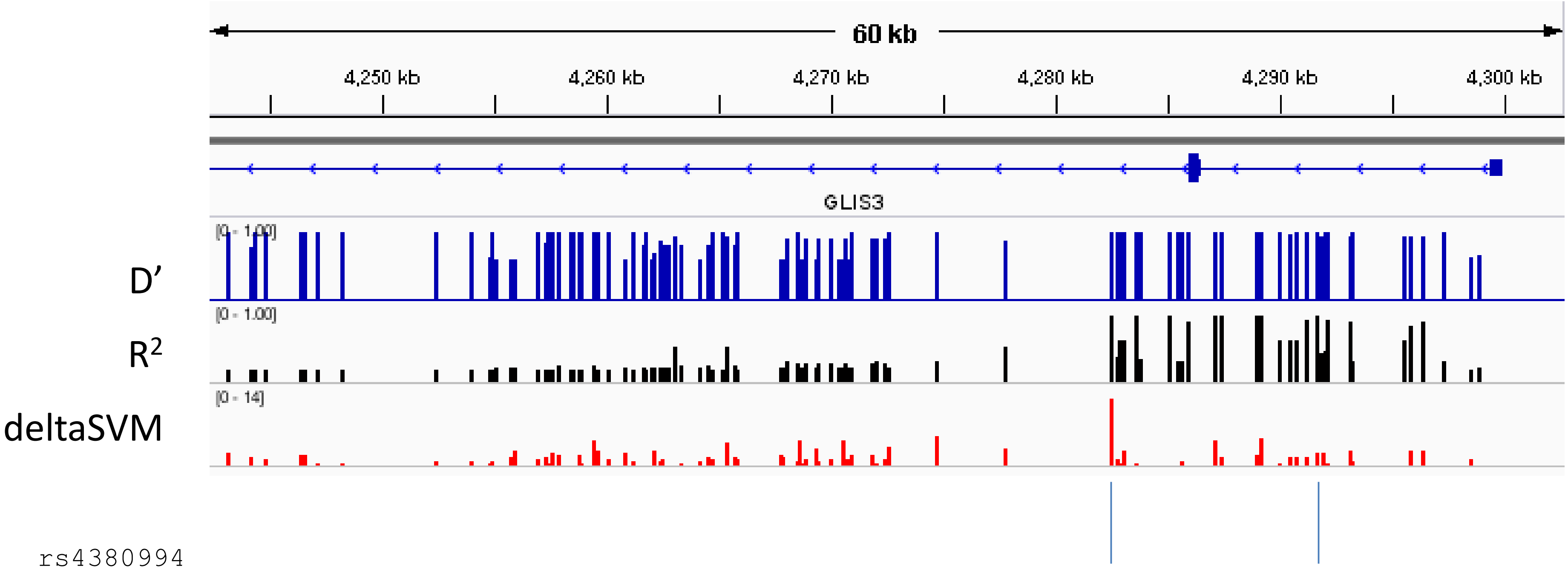
Narrowing in on causal SNP

Musunuru Nature (2010)





Novel prediction for causal SNP in GLIS3 type-1 diabetes risk locus



rs4380994 *
deltaSVM prediction

* rs7020673
type-1 diabetes

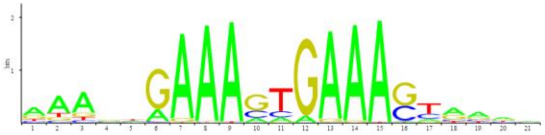
Barrett Nat Gen (2009)

```
rs4380994
Ref allele: TTCAAACTGAAACATCAGC
Alt allele: TTCAAACTGGAACATCAGC
```

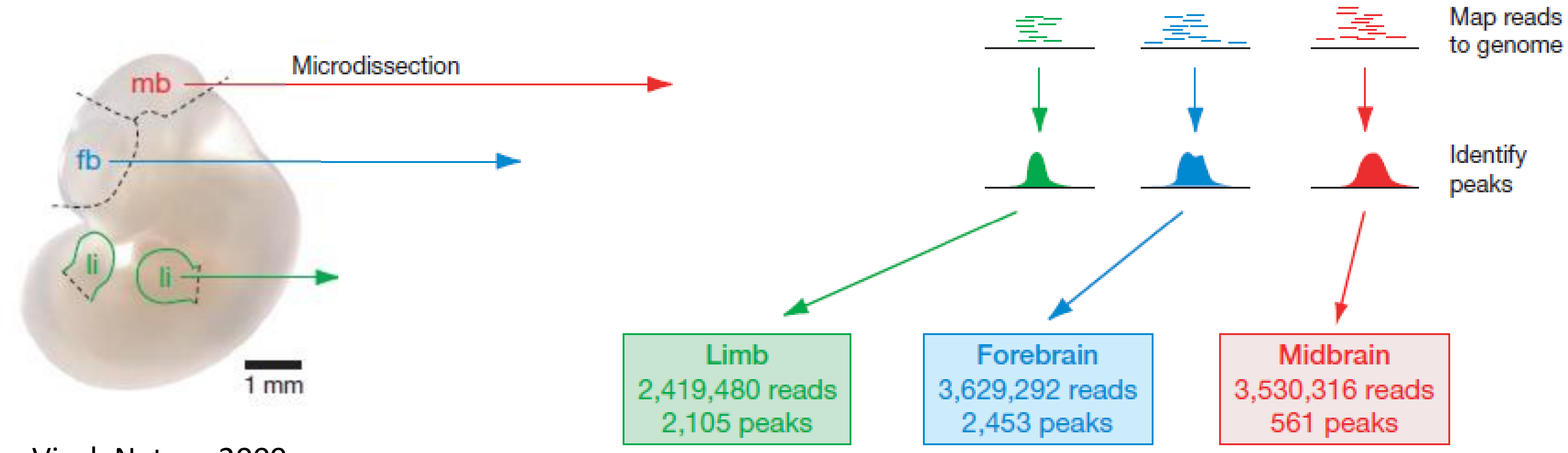
Ref 10-mer	weight	Alt 10-mer	weight	Diff
TTCAAACTGA	1.04	TTCAAACTGG	0.07	-0.97
TCAAACTGAA	1.55	TCAAACTGGA	-0.14	-1.69
CAAACTGAAA	2.12	CAAACTGGAA	0.38	-1.74
AAACTGAAAC	4.47	AAACTGGAAC	1.00	-3.47
AACTGAAACA	1.76	AACTGGAACA	0.06	-1.71
ACTGAAACAT	0.71	ACTGGAACAT	-0.20	-0.91
CTGAAACATC	0.65	CTGGAACATC	-0.48	-1.12
TGAAACATCA	0.73	TGGAACATCA	-0.46	-1.18
GAAACATCAG	0.11	GGAACATCAG	-0.59	-0.70
AAACATCAGC	-0.12	GAACATCAGC	-0.20	-0.09

deltaSVM = -13.59

Predicted causal SNP disrupts/creates IRF1/2 site:



P300 Chip-seq Bound Enhancers



Visel, Nature 2009

Lee, Karchin, Beer Genome Research 2011

DNA Sequences → k-mer frequency vectors

Leslie, Noble

- >chr10:6238300-6238926
- TTGTGGACGTCAGGGAGTGGGGATTGGAGGTTAGCCTTGTATCTCAGTATCTCCGATGCCT...
- >chr10:7757450-7758801
- GTATGTGCACAAAGCACACATTTTCTTTGTTAGCTGAAGCACGGCAGGGCAGGGTTTCACT...
- >chr10:8992150-8992551
- TGCCATGCTACTAACTCAGGACTTATTGTTACACGTACAAACATGTTTGGAATTCCAGTGC...

id	AAAAAA	AAAAAC	AAAAAG	AAAAAT	AAAACA	...	CACGTG	...	GATAAA	...	TTTAAA
seq1	3	2	0	1	2	1	1	1	1	1	0
seq2	2	1	1	1	2		1		0		0
seq3	1	0	0	1	1		0		0		1
seq4	2	1	1	0	0		0		1		1
...											

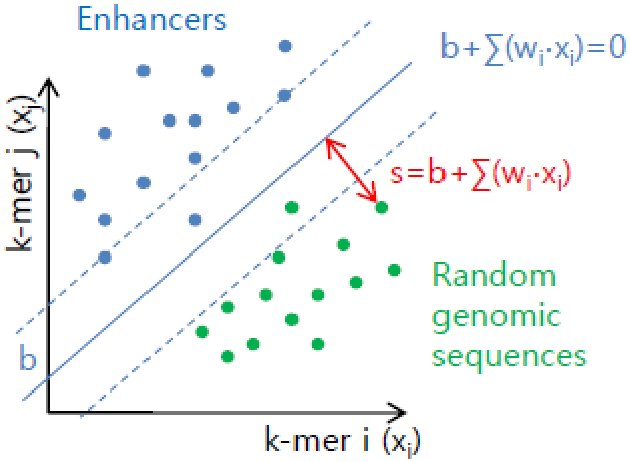
Generate GC matched negative set,   Train SVM to classify enhancers

Lee, Karchin, Beer   Genome Research Dec 2011

Sequences → k-mer frequencies →

k-mer	counts
5′-AAAAAA-3′ 3′-TTTTTT-5′	x_1
5′-AAAAAC-3′ 3′-TTTTTG-5′	x_2
5′-AAAAAG-3′ 3′-TTTTTC-5′	x_3
...	...
5′-TTTAAA-3′ 3′-AAATTT-5′	x_n

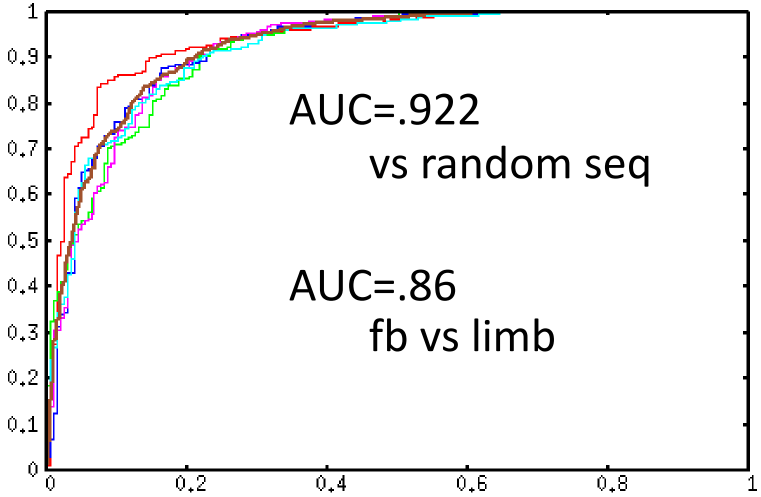
SVM training →



Predictive feature analysis

k-mer	weights (w_i)
5′-AATGAG-3′ 3′-TTACTC-5′	+3.94
5′-AATTAG-3′ 3′-TTAATC-5′	+3.84
5′-AGCTGC-3′ 3′-TCGACG-5′	+3.65
...	...
5′-CAGGTA-3′ 3′-GTCCAT-5′	-2.06

In 5-fold CV, can predict forebrain, mb, limb

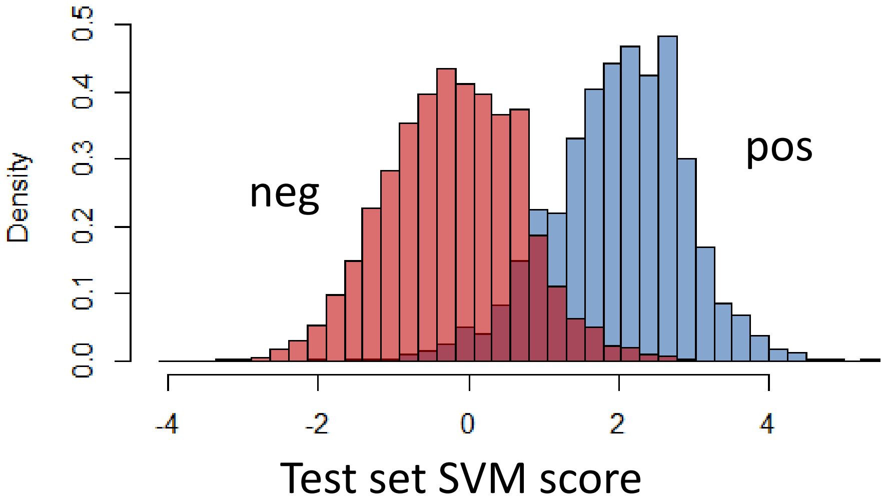


Large weight k-mers are relevant TFs (both **positive** and **negative**!)

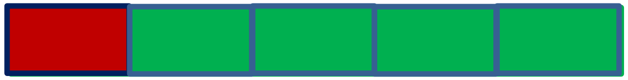
AATGAG	+3.94	Homeo
AATTAG	+3.95	Homeo
AGCTGC	+3.65	HLH
CAATTA	+3.62	Homeo
ACAAAG	+3.29	SOX
.		
.		
.		
AGGTGA	-1.97	ZEB1
ACCTGG	-2.03	ZEB1
CAGGTA	-2.06	ZEB1

Predictive Accuracy from SVM Score Distribution on Reserved Test Set

+ 2453 forebrain enhancers (~500-1000bp)
- 4-100x “random” seqs (GC,rpt,len matched)



5-fold Cross validation

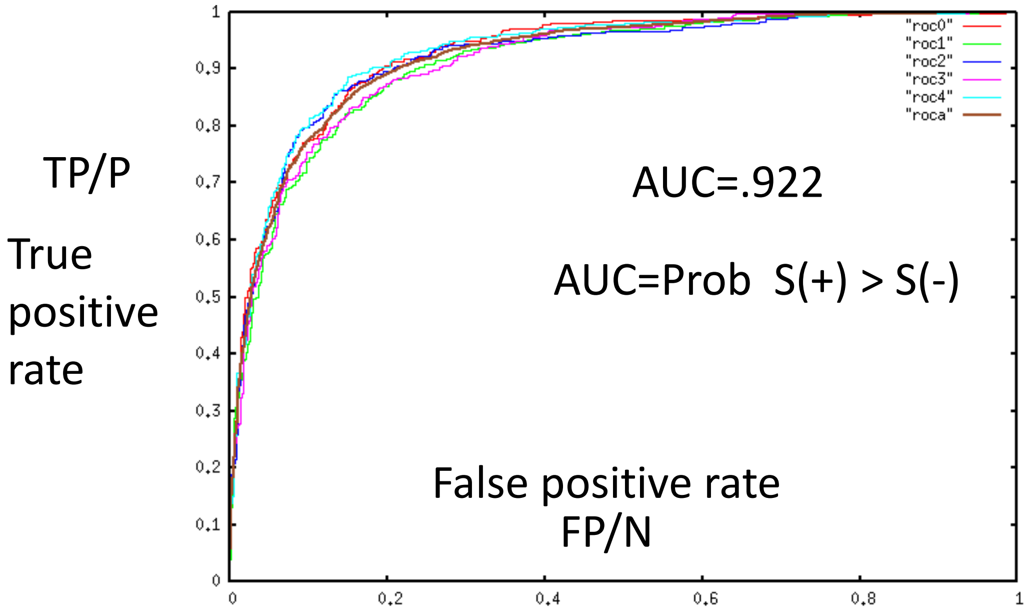


test training

SVM score: $S = \sum_i w_i x_i + b$

	Actual Value	
Predicted	**P** (pos)	**N** (neg)
PP (pos)	**TP**	**FP**
PN (neg)	**FN**	**TN**

region	score	class
2943	1.675	+1
2944	0.995	+1
2945	0.340	+1
2946	-0.237	+1
2947	0.704	+1
...		
3509	-4.305	-1
3510	-0.963	-1
3511	-1.106	-1
3512	-2.211	-1
3513	-2.544	-1

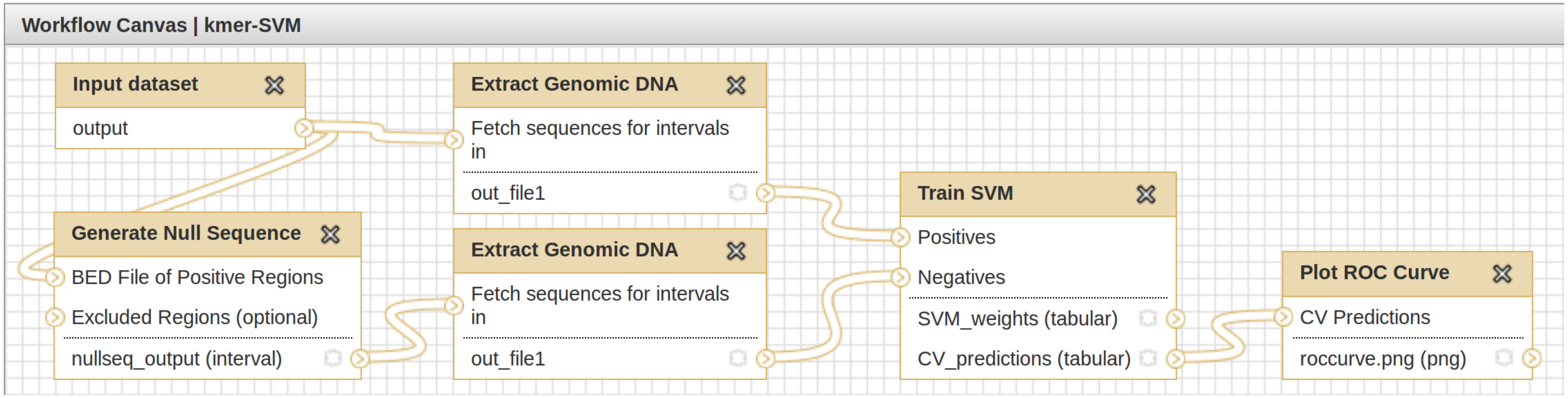


Kmer-SVM Can Accurately Predict Many Classes of Regulatory Elements

- Positive set of functional genomic elements (p300, DHS, H3K27Ac, in some tissue/cell-type)
- Generate GC/repeat matched negative set
- Train SVM, identify predictive kmers
- Median AUC > 0.9, (all mouse, human ENCODE) analyze weights to find DNA features (TFBS)

kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets http://kmersvm.beerlab.org

Fletez-Brant, Lee, McCallion, Beer NAR 2013



Workflow Canvas | kmer-SVM

Input dataset
output

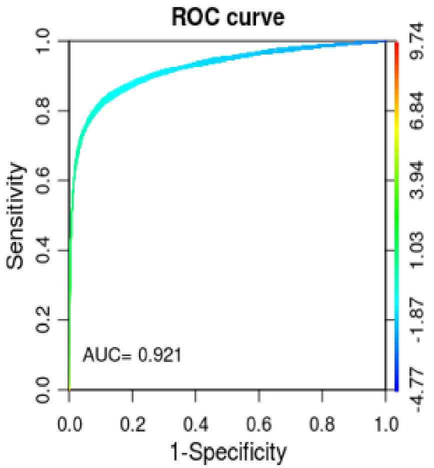
Extract Genomic DNA
Fetch sequences for intervals in
out_file1

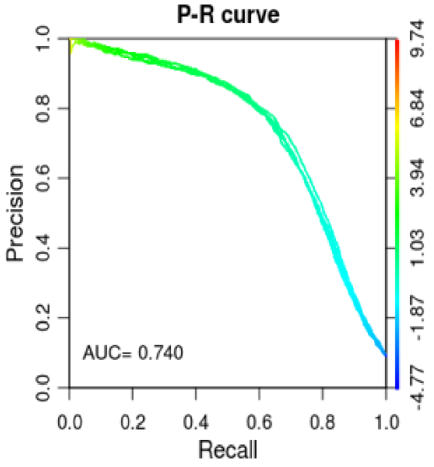
Generate Null Sequence
BED File of Positive Regions
Excluded Regions (optional)
nullseq_output (interval)

Extract Genomic DNA
Fetch sequences for intervals in
out_file1

Train SVM
Positives
Negatives
SVM_weights (tabular)
CV_predictions (tabular)

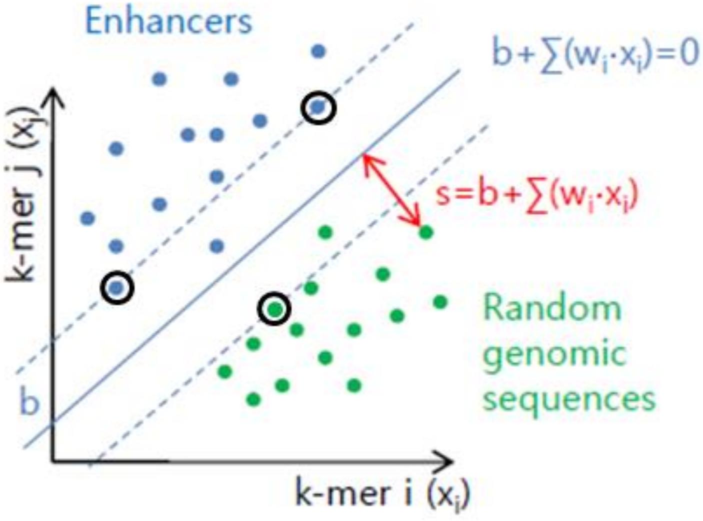
Plot ROC Curve
CV Predictions
roccurve.png (png)





6-mers	Revcomp	SVM Scores
	Positive 6-mers	
AAGGTC	GACCTT	10.05
AGGTCA	TGACCT	8.47
ACCTTG	CAAGGT	5.33
AGGTCG	CGACCT	5.17
GGTCAA	TTGACC	4.01
	Negative 6-mers	
GCAATA	TATTGC	-2.05
TGACCA	TGGTCA	-3.33
AAGGTA	TACCTT	-4.23
AGACCT	AGGTCT	-4.55
AGGTCC	GGACCT	-4.98

SVM can predict accurately, but individual weights can be noisy



Solution: Use **gapped kmer** distribution to find most likely estimate of kmer counts in training sequences

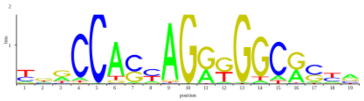
Ghandi, Beer J. Math. Biol. 2013

Elements of matrix mapping gapped kmers to estimated kmers take simple form:

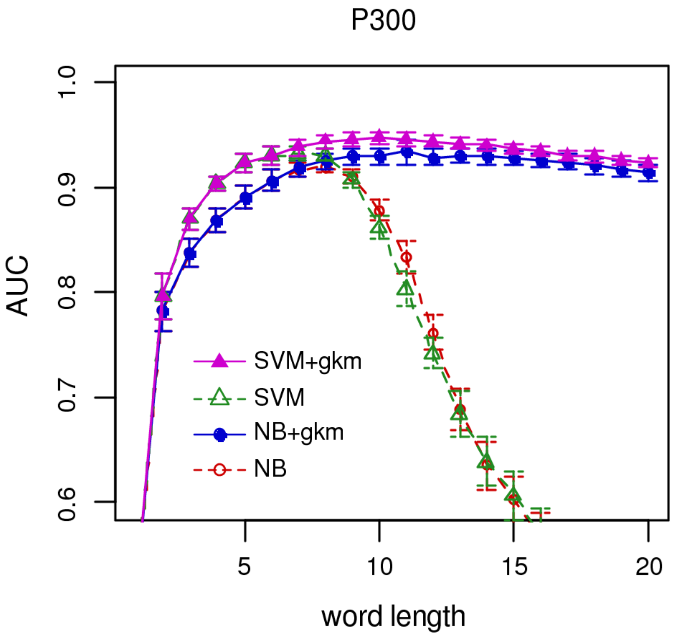
$$w(m) = \frac{(-1)^m}{4^k \binom{k}{u-m}} \frac{k-u}{k-u+m} \sum_{t=0}^{u-m} \binom{k}{t} 3^t$$

Or simply use **gapped kmers** as features

Ghandi, Lee, Beer PLOS Comp. Biol. 2014



actual counts:	**pos**	**neg**
CACCAGGGGG	42	0
CCACCAGGGG	44	3
CCACCTGGTG	26	2
CCACCAGGTG	33	0
gapped kmer counts:		
CA--AG--GG	748	252
CC-C--G-GG	702	212
CA-C-G--GG	670	205
CCA---GG-G	693	232
estimated counts:		
CACCAGGGGG	22.02	1.60
CCACCAGGGG	22.18	2.27
CCACCTGGTG	20.67	0.91
CCACCAGGTG	20.66	1.81



Kernel evaluation challenging in very large gapped kmer feature space

```
CA--AG--GG
CC-C--G-GG
CA-C-G--GG
CCA---GG-G
```

Typically we use (10,6): 10-mers with 6 informative posns

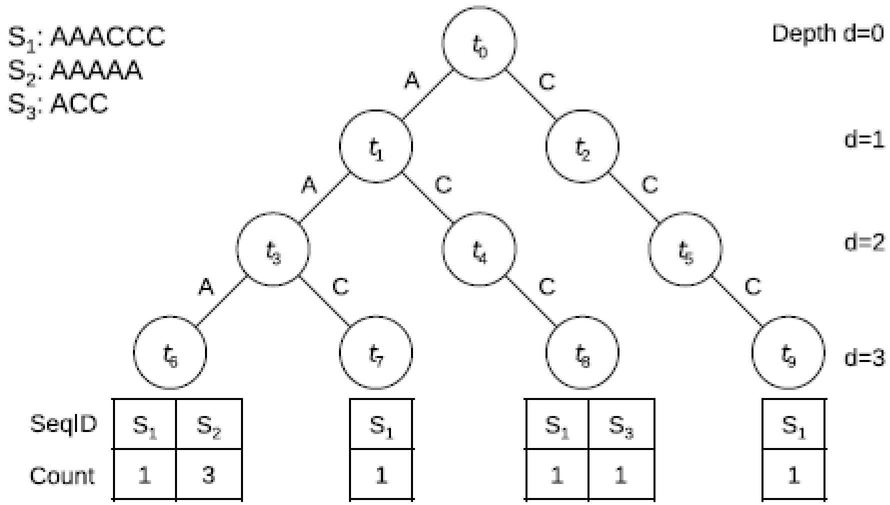
$4^6 \binom{10}{6}$= 860160

4^{10}=1048576

And 20,000 sequence elements:
5000-10000 positives
10000-50000 negatives

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

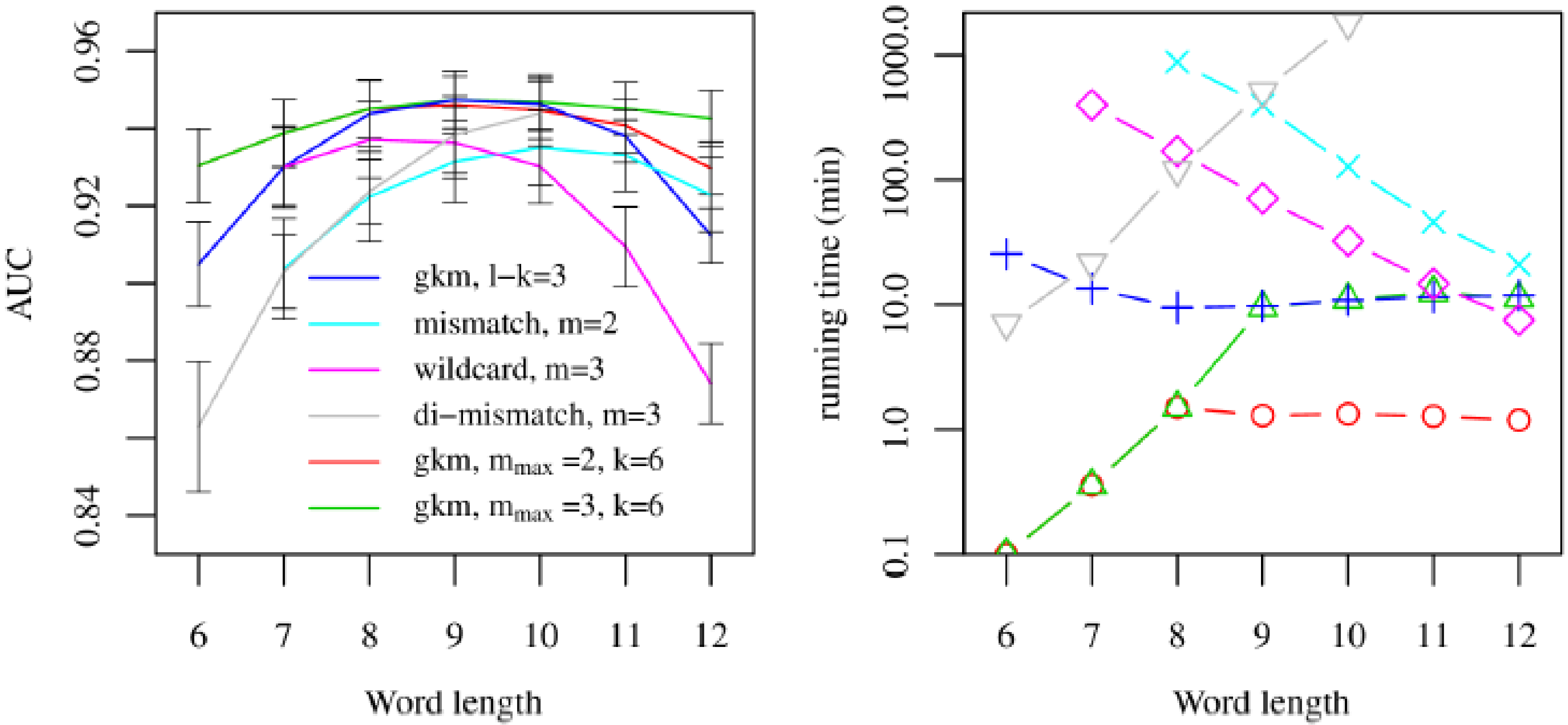
$$f^{S_1} = (x_i^{S_1}), i = 1..M$$



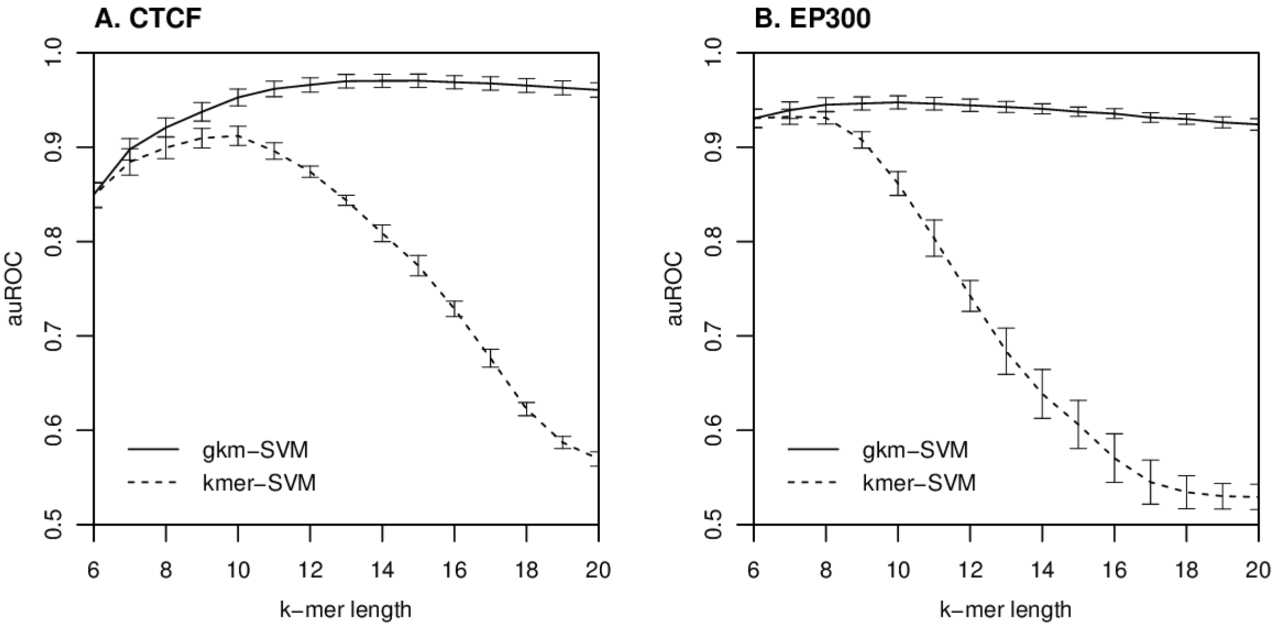
M. Ghandi 2014

Kernel evaluation challenging in very large gapped kmer feature space

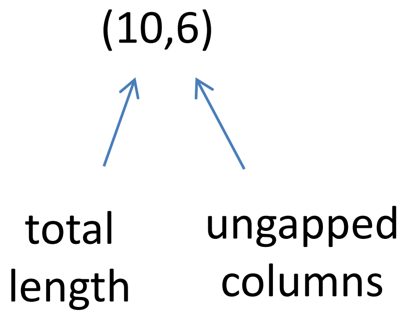
Our algorithm is optimized for densely populated trees, where statistical support for feature weights is high

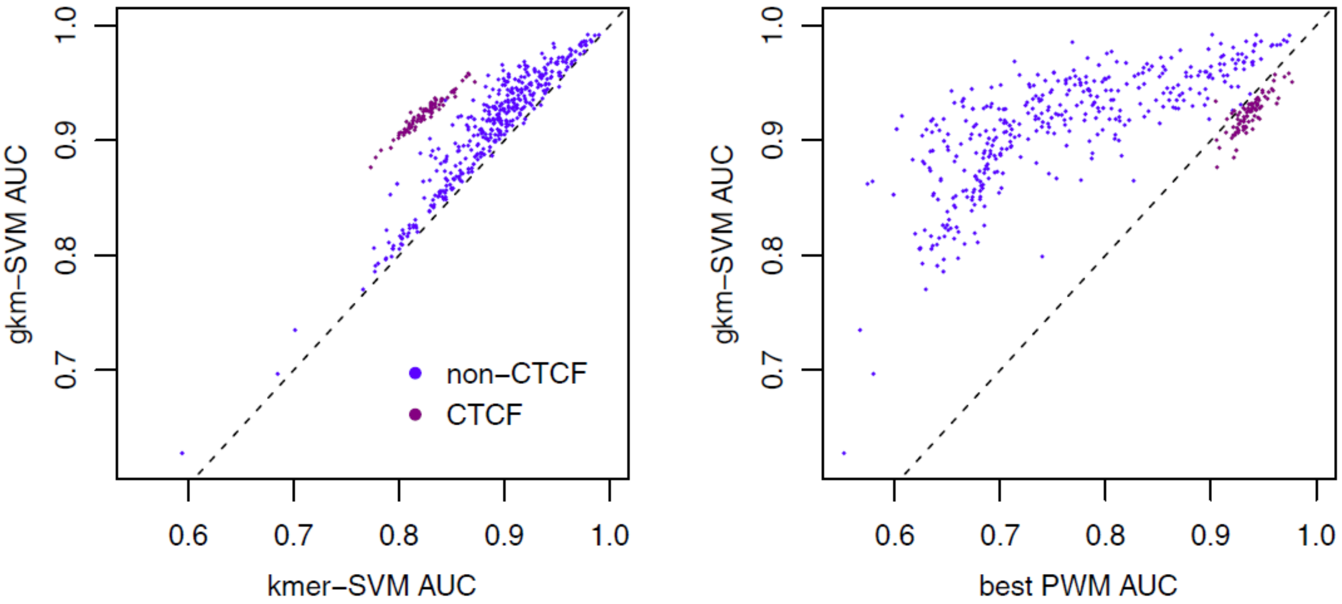


gkm-SVM predicts all human ENCODE TF ChIP-seq data accurately



Ghandi, Lee, Beer PLOS Comp. Biol. 2014

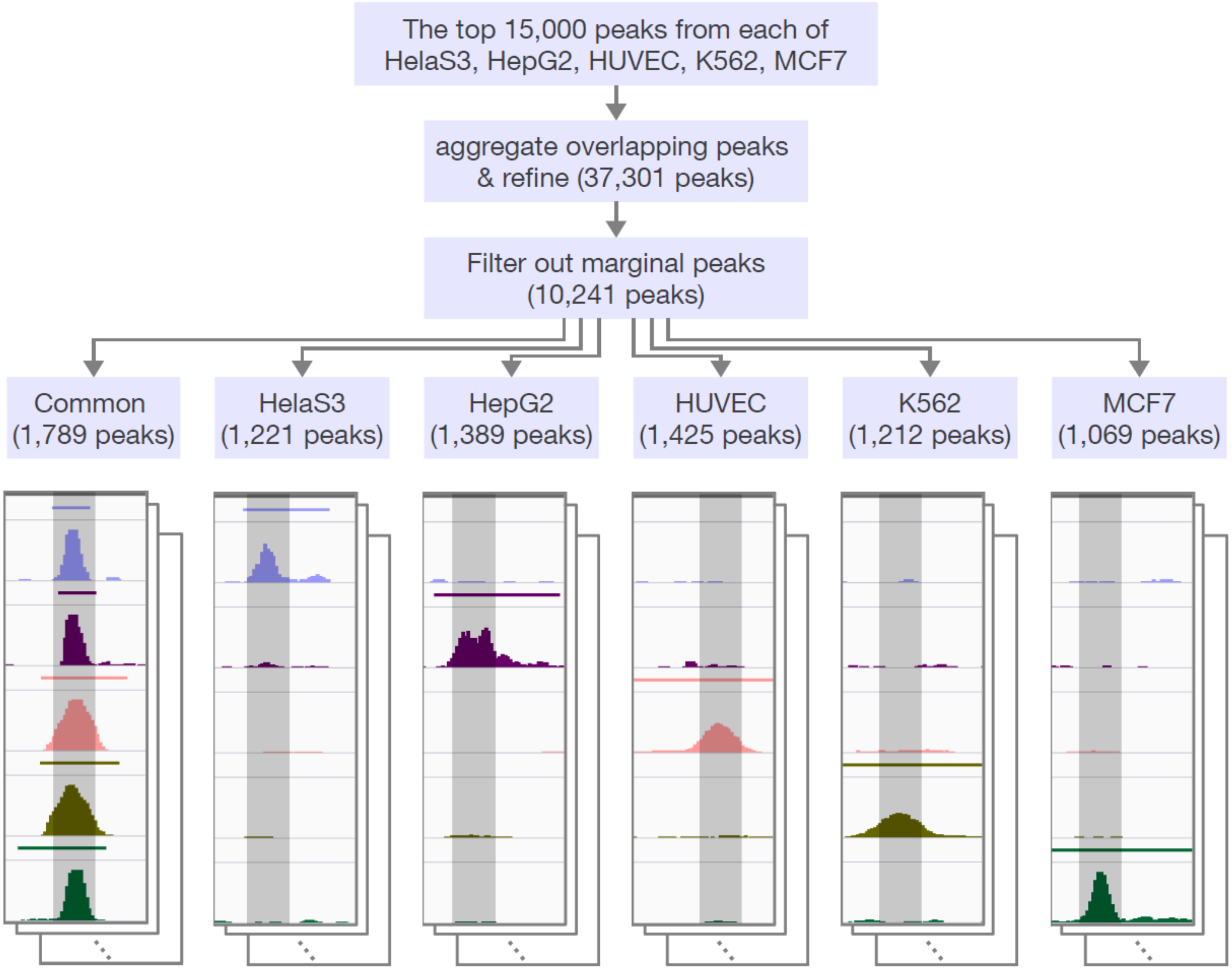




Wang , J Genome Res 2012

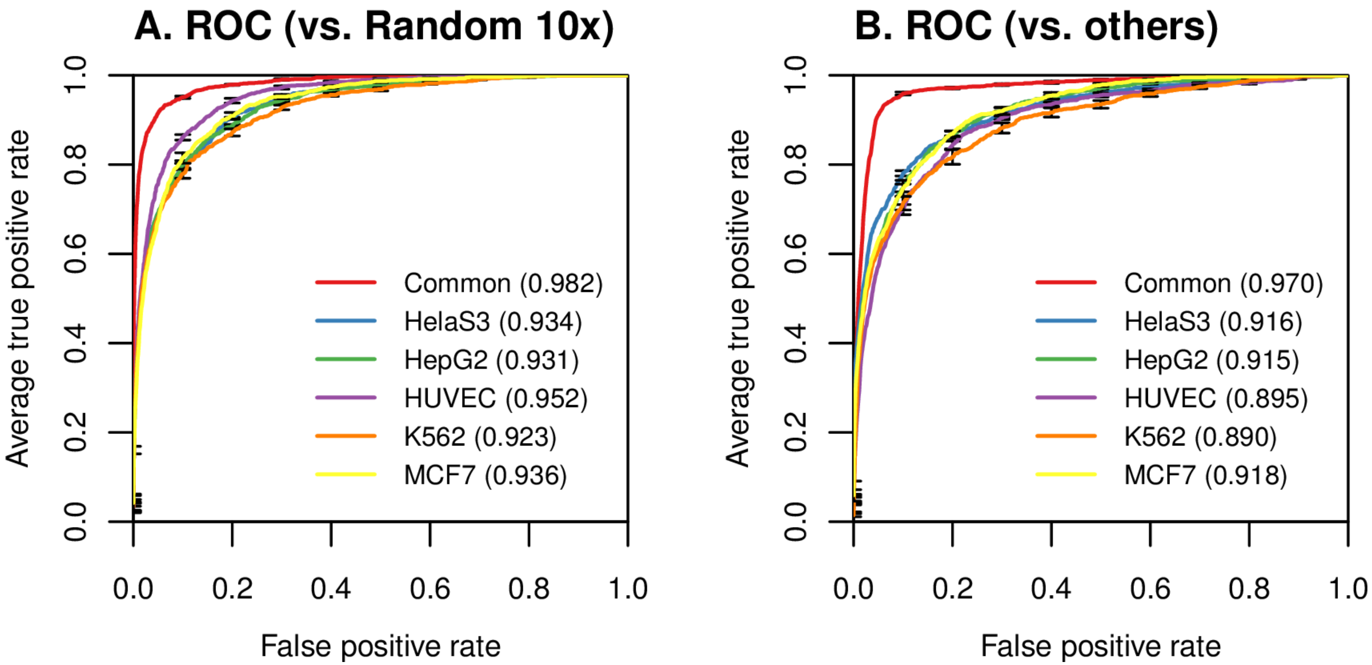
467 datasets

Application to Cell-type specific Binding of a Sequence Specific TF (MYC)



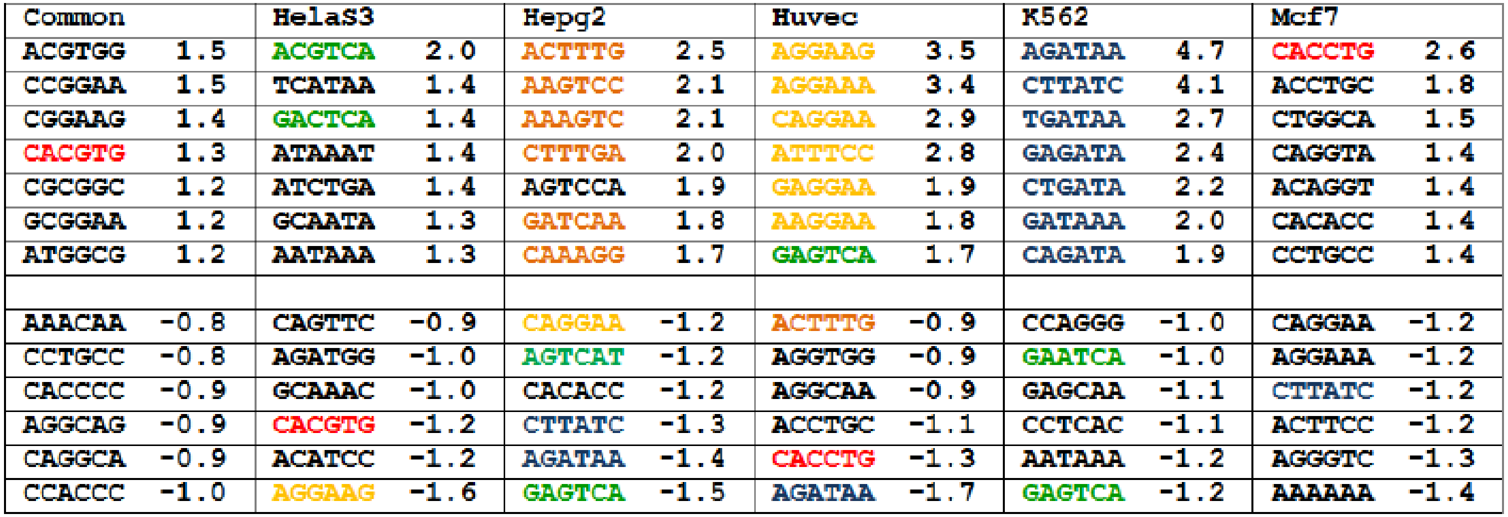
The top 15,000 peaks from each of HelaS3, HepG2, HUVEC, K562, MCF7
aggregate overlapping peaks & refine (37,301 peaks)
Filter out marginal peaks (10,241 peaks)
Common (1,789 peaks)
HelaS3 (1,221 peaks)
HepG2 (1,389 peaks)
HUVEC (1,425 peaks)
K562 (1,212 peaks)
MCF7 (1,069 peaks)

gkm-SVM Predicts Cell Specific Binding of Myc by Identifying Cofactors



Myc: CACGTG   AP1: TGAGTCA   HNF4a: TCAAAGG   Ets: AGGAAG   GATAA

kmer weights vs. other exclusive peaks:



Common		HelaS3		Hepg2		Huvec		K562		Mcf7	
ACGTGG	1.5	ACGTCA	2.0	ACTTTG	2.5	AGGAAG	3.5	AGATAA	4.7	CACCTG	2.6
CCGGAA	1.5	TCATAA	1.4	AAGTCC	2.1	AGGAAA	3.4	CTTATC	4.1	ACCTGC	1.8
CGGAAG	1.4	GACTCA	1.4	AAAGTC	2.1	CAGGAA	2.9	TGATAA	2.7	CTGGCA	1.5
CACGTG	1.3	ATAAAT	1.4	CTTTGA	2.0	ATTTCC	2.8	GAGATA	2.4	CAGGTA	1.4
CGCGGC	1.2	ATCTGA	1.4	AGTCCA	1.9	GAGGAA	1.9	CTGATA	2.2	ACAGGT	1.4
GCGGAA	1.2	GCAATA	1.3	GATCAA	1.8	AAGGAA	1.8	GATAAA	2.0	CACACC	1.4
ATGGCG	1.2	AATAAA	1.3	CAAAGG	1.7	GAGTCA	1.7	CAGATA	1.9	CCTGCC	1.4
AAACAA	-0.8	CAGTTC	-0.9	CAGGAA	-1.2	ACTTTG	-0.9	CCAGGG	-1.0	CAGGAA	-1.2
CCTGCC	-0.8	AGATGG	-1.0	AGTCAT	-1.2	AGGTGG	-0.9	GAATCA	-1.0	AGGAAA	-1.2
CACCCC	-0.9	GCAAAC	-1.0	CACACC	-1.2	AGGCAA	-0.9	GAGCAA	-1.1	CTTATC	-1.2
AGGCAG	-0.9	CACGTG	-1.2	CTTATC	-1.3	ACCTGC	-1.1	CCTCAC	-1.1	ACTTCC	-1.2
CAGGCA	-0.9	ACATCC	-1.2	AGATAA	-1.4	CACCTG	-1.3	AATAAA	-1.2	AGGGTC	-1.3
CCACCC	-1.0	AGGAAG	-1.6	GAGTCA	-1.5	AGATAA	-1.7	GAGTCA	-1.2	AAAAAA	-1.4

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

Train gkm-SVM on mouse ENCODE DHS or H3K27Ac signal

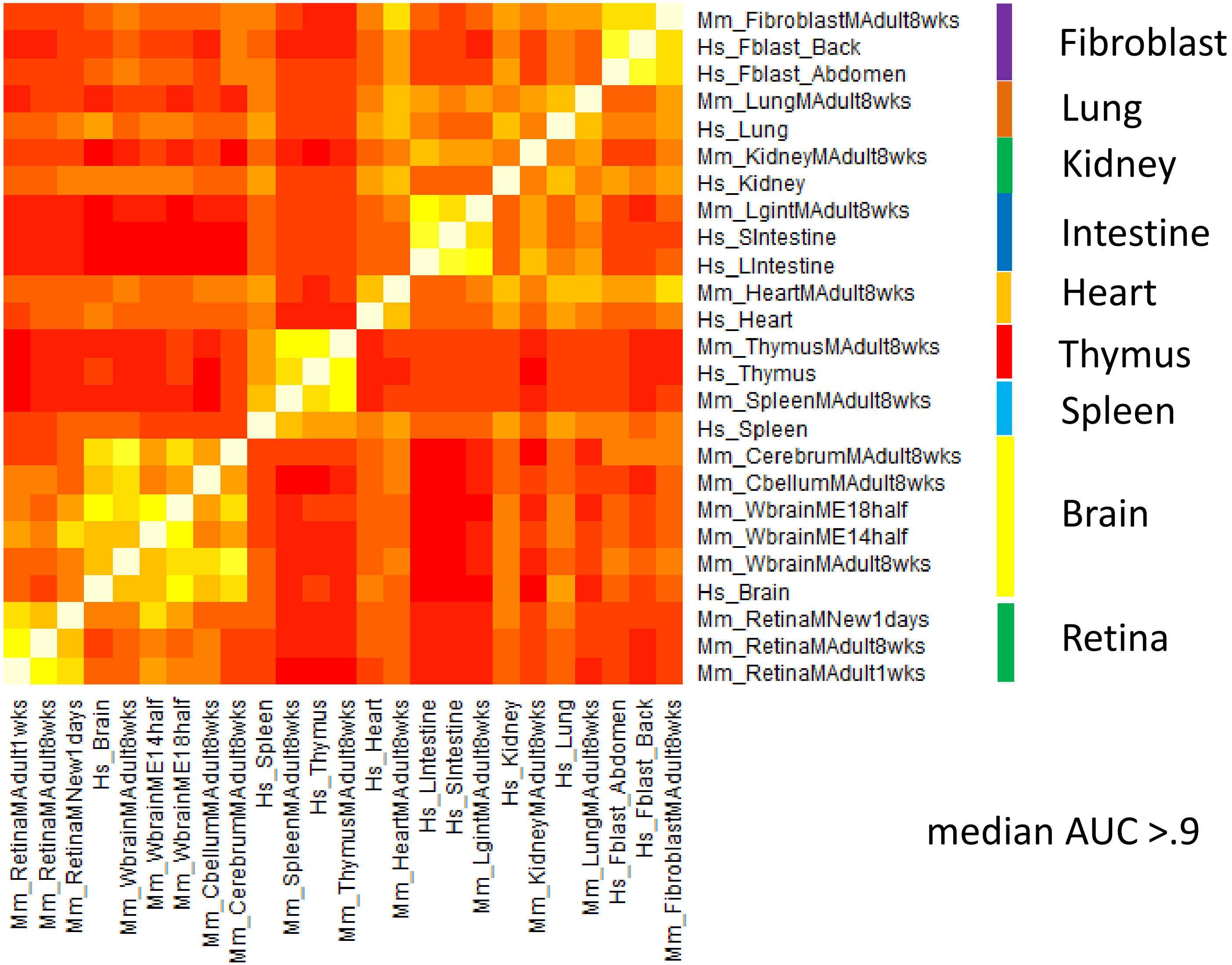
Train gkm-SVM on human Fetal Roadmap signal

- remove constituitively open regions

Compare weight vectors using correlation across all kmer weights

mouse ENCODE
Roadmap
Stam Lab

$C(w_i,w_j)$

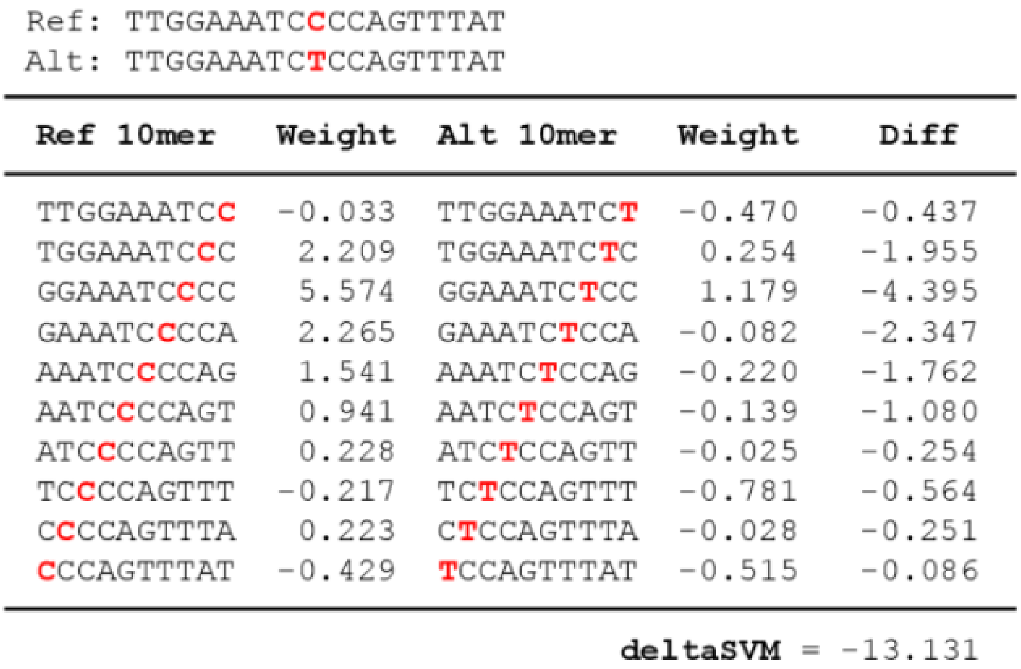


median AUC >.9

deltaSVM: gkm-SVM Weights can Predict Impact of Mutations

Lymphoblast regulatory regions:

10000 DHS regions in GM12878 (LCL) cells



```
Ref: TTGGAAATCCCCAGTTTAT
Alt: TTGGAAATCTCCAGTTTAT
```

Ref 10mer	Weight	Alt 10mer	Weight	Diff
TTGGAAATCC	-0.033	TTGGAAATCT	-0.470	-0.437
TGGAAATCCC	2.209	TGGAAATCTC	0.254	-1.955
GGAAATCCCC	5.574	GGAAATCTCC	1.179	-4.395
GAAATCCCCA	2.265	GAAATCTCCA	-0.082	-2.347
AAATCCCCAG	1.541	AAATCTCCAG	-0.220	-1.762
AATCCCCAGT	0.941	AATCTCCAGT	-0.139	-1.080
ATCCCCAGTT	0.228	ATCTCCAGTT	-0.025	-0.254
TCCCCAGTTT	-0.217	TCTCCAGTTT	-0.781	-0.564
CCCCAGTTTA	0.223	CTCCAGTTTA	-0.028	-0.251
CCCAGTTTAT	-0.429	TCCAGTTTAT	-0.515	-0.086

deltaSVM = -13.131

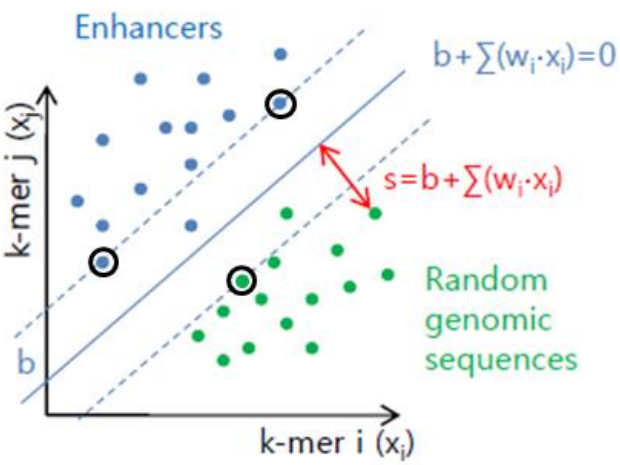
deltaSVM = change in gkm-SVM score

Lee Gorkin Smith Strober Asoni McCallion Beer Nat Gen 2015

k-mer	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
	...	
ATTCCAGGGA	TCCCTGGAAT	-1.502
GCGGGGGTCC	GGACCCCCGC	-1.507
GAGGGGGTCC	GGACCCCCTC	-1.514
CGGGACCCCC	GGGGGTCCCG	-1.533
GGACCCCCAC	GTGGGGGTCC	-1.545
CGGGGGGTCC	GGACCCCCCG	-1.546
GGGGGTCTCA	TGAGACCCCC	-1.588
CGGGGGTCCC	GGGACCCCCG	-1.680
CCCCCTCCCC	GGGGAGGGGG	-1.787
CCCCCTGCCC	GGGCAGGGGG	-1.893

deltaSVM Method Overview

- Define a set of cell type specific enhancers using DHS, ATAC-seq, or chromatin
- Train gkm-SVM to learn regulatory TF binding site vocabulary for given cell-type
- Calculate how each SNP changes SVM score to predict impact (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)

Leslie, Noble (2004, 2012)

k-mer counts:

CACCAGGGGG
CCACCAGGGG
CCACCTGGTG
CCACCAGGTG

gapped kmer counts:

CA--AG--GG
CC-C--G-GG
CA-C-G--GG
CCA---GG-G

```
Ref: TTGGAAATCCCCAGTTTAT
Alt: TTGGAAATCTCCAGTTTAT
```

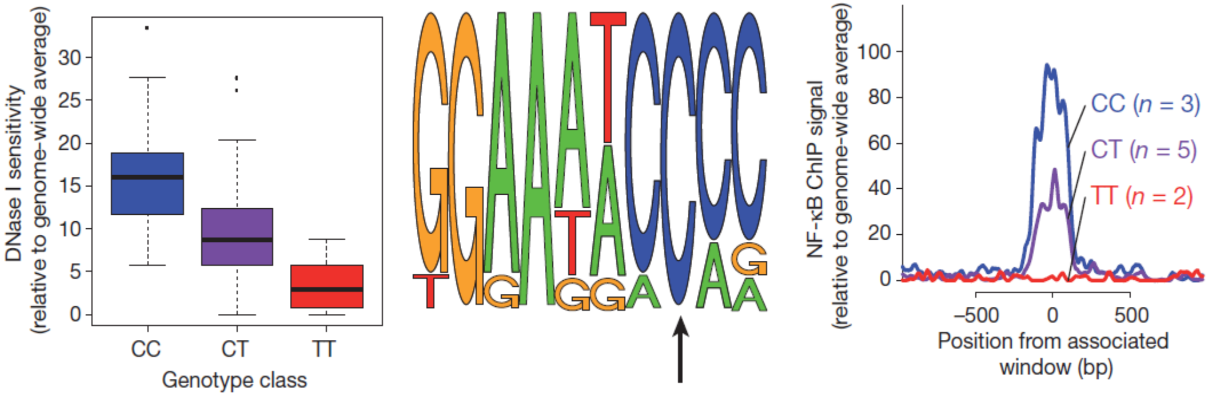
Ref 10mer	Weight	Alt 10mer	Weight	Diff
TTGGAAATCC	-0.033	TTGGAAATCT	-0.470	-0.437
TGGAAATCCC	2.209	TGGAAATCTC	0.254	-1.955
GGAAATCCCC	5.574	GGAAATCTCC	1.179	-4.395
GAAATCCCCA	2.265	GAAATCTCCA	-0.082	-2.347
AAATCCCCAG	1.541	AAATCTCCAG	-0.220	-1.762
AATCCCCAGT	0.941	AATCTCCAGT	-0.139	-1.080
ATCCCCAGTT	0.228	ATCTCCAGTT	-0.025	-0.254
TCCCCAGTTT	-0.217	TCTCCAGTTT	-0.781	-0.564
CCCCAGTTTA	0.223	CTCCAGTTTA	-0.028	-0.251
CCCAGTTTAT	-0.429	TCCAGTTTAT	-0.515	-0.086

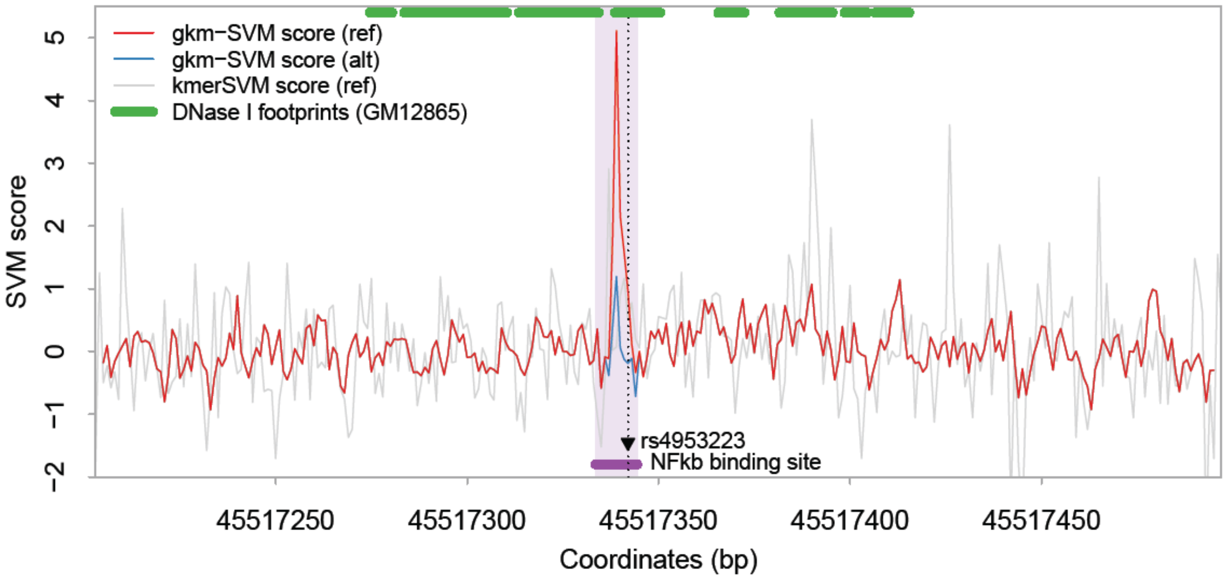
deltaSVM = -13.131

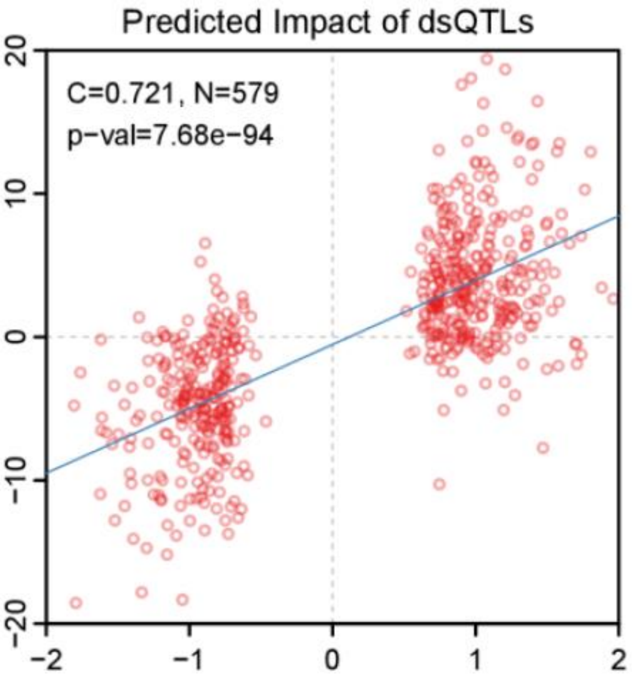
deltaSVM Predicts Functional SNPs

dsQTLs: DNaseI signal correlated with genotype at SNPs

Degner Nature 2012





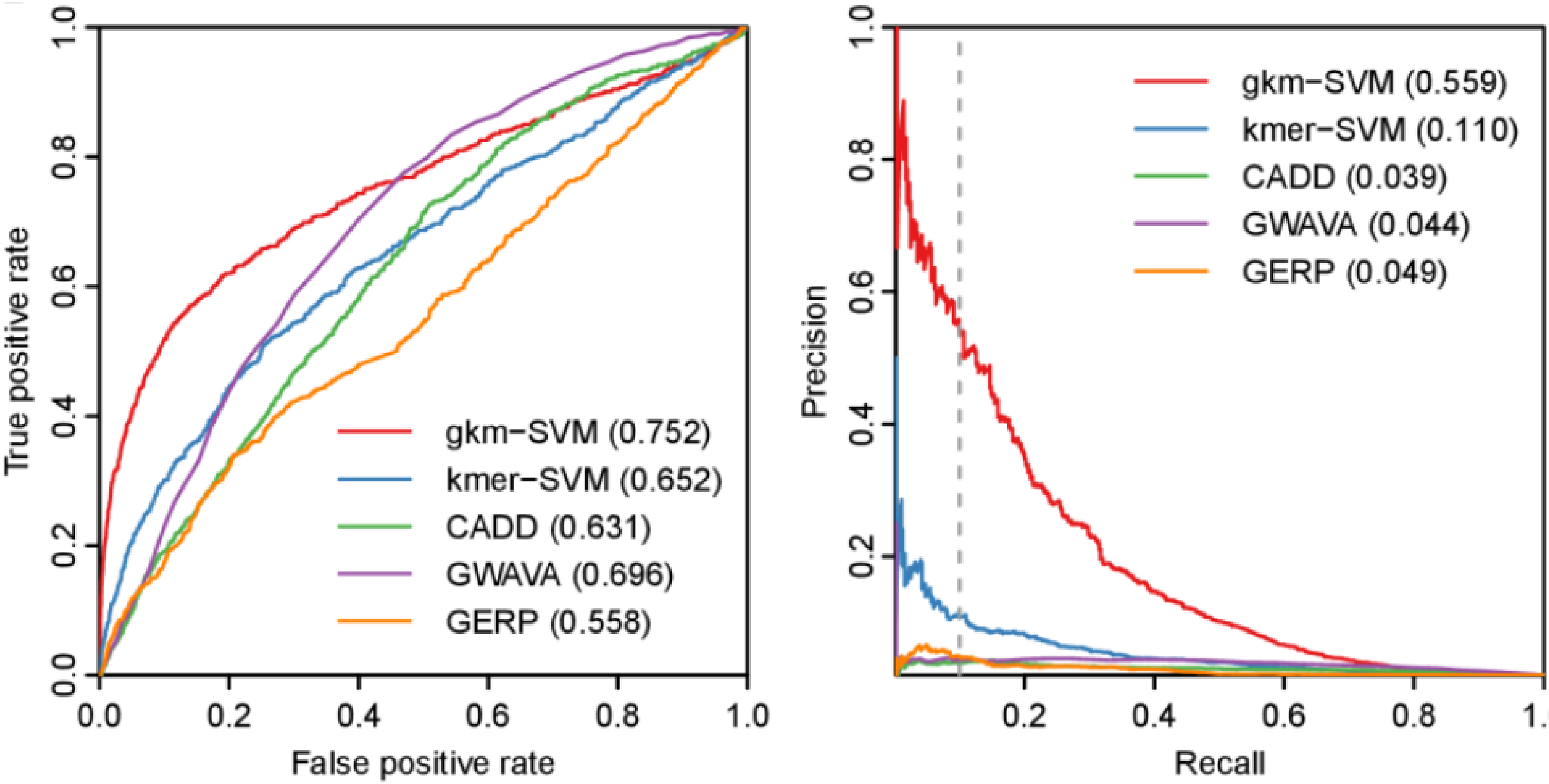


gkm-SVM deltaSVM

fold change in Dnase signal at dsQTLs

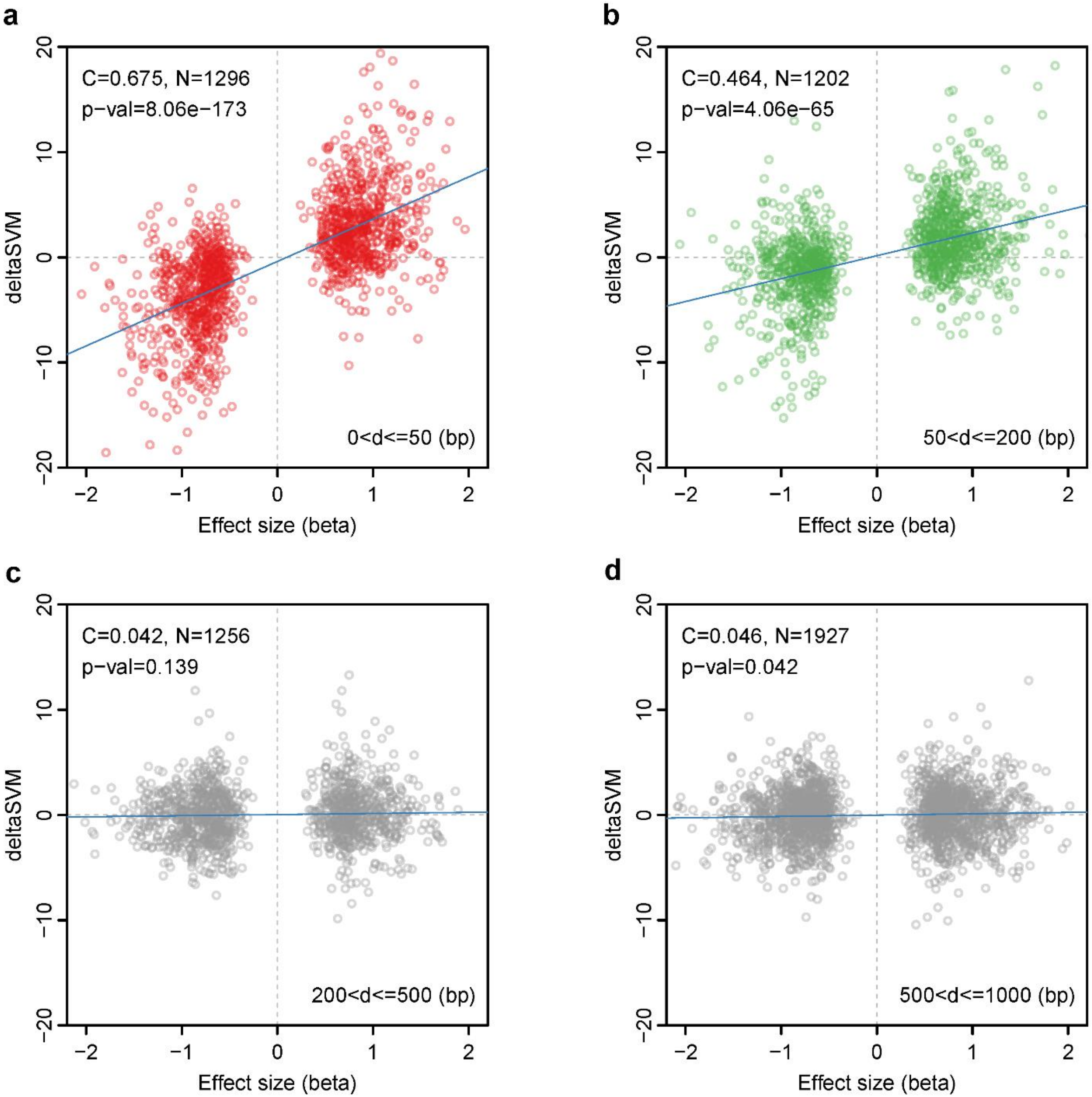
deltaSVM Predicts Functional Impact More Accurately Than Alternative Methods

Positive set:	579 dsQTLs
Negative set:	28,950 SNPs (50x)



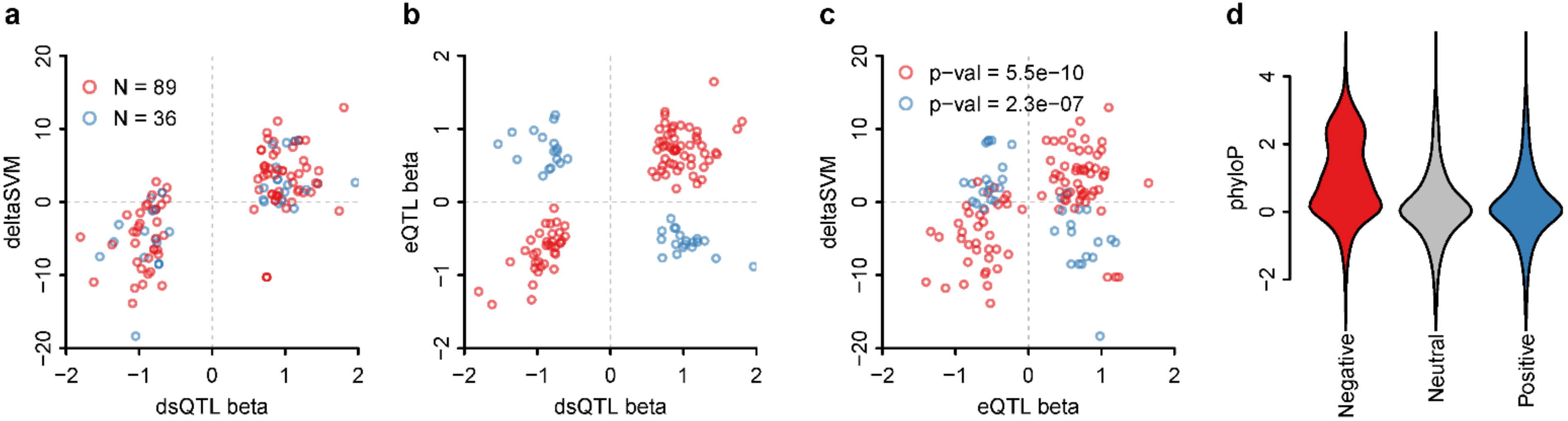
CADD: Combined Annotation Dependent Depletion	Kircher, Shendure, Nat Gen 2014
GWAVA: Genome-Wide Annotation of Variants	Ritchie, Flicek, Nat Meth 2014
GERP: Genomic Evolutionary Rate Profiling	Davydov, Sidow, Batzoglou PLOS CB 2010

dsQTLs act locally



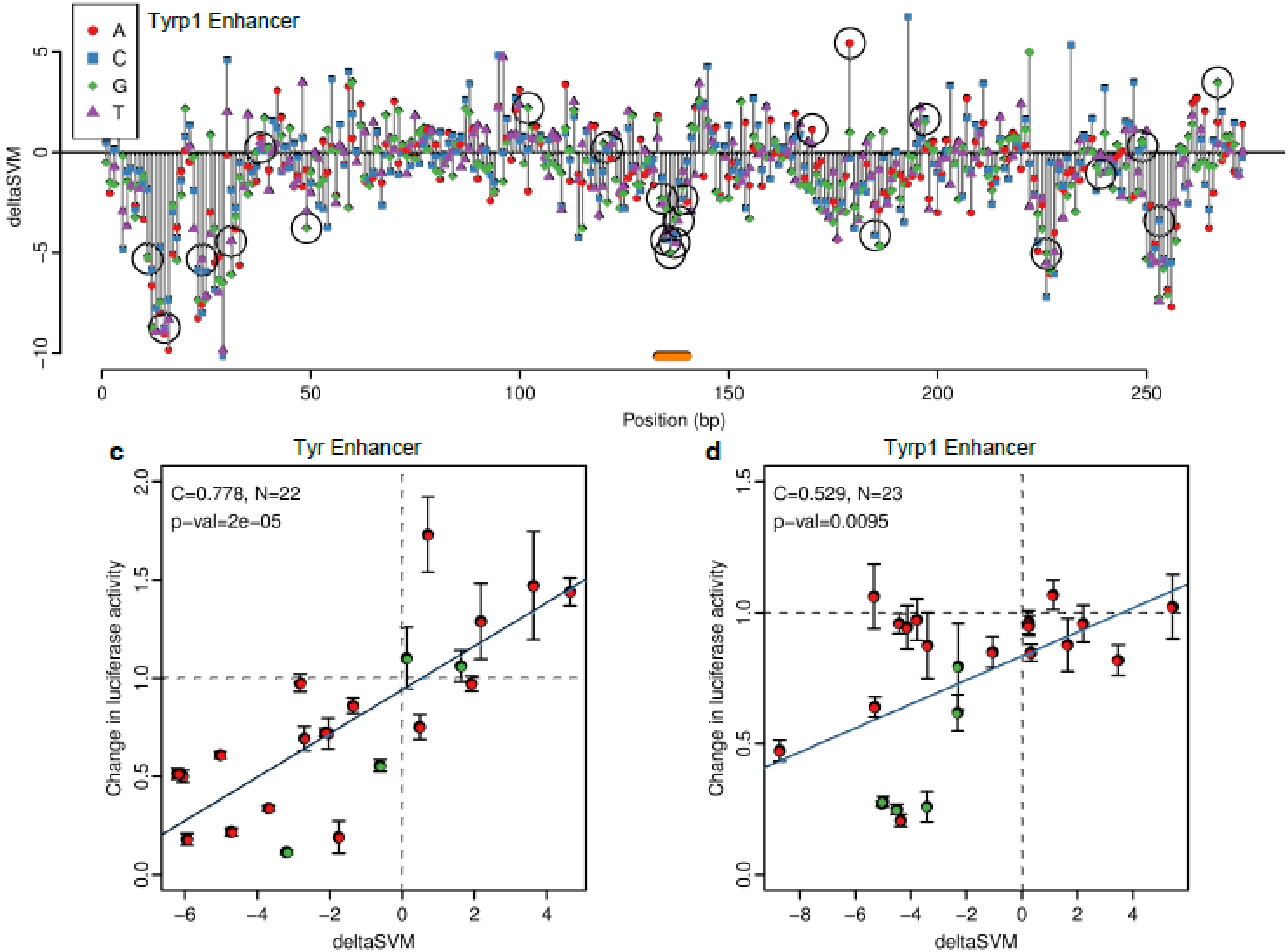
deltaSVM correlated with eQTLs
when dsQTL beta positively correlated with eQTL beta

~30% of dsQTLs are repressive



strongly negative deltaSVM bases
are evolutionarily constrained

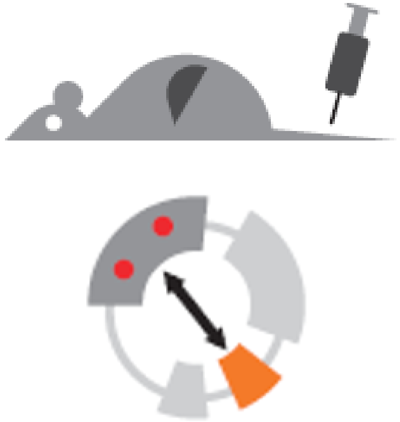
SVM Predicts Effect of SNPs in melanocyte enhancers

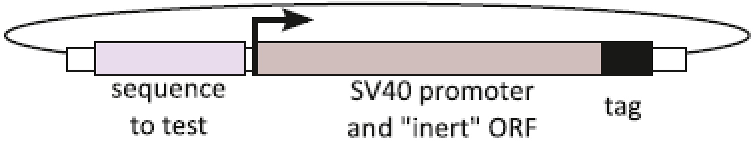


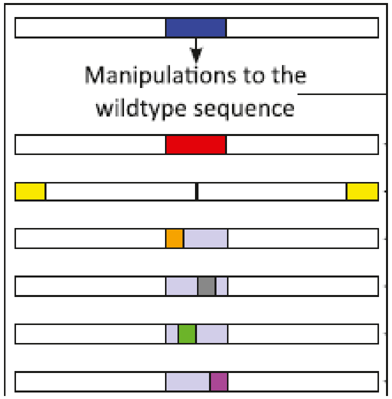
delta SVM Predicts Effect of SNPs in massively parallel enhancer assays

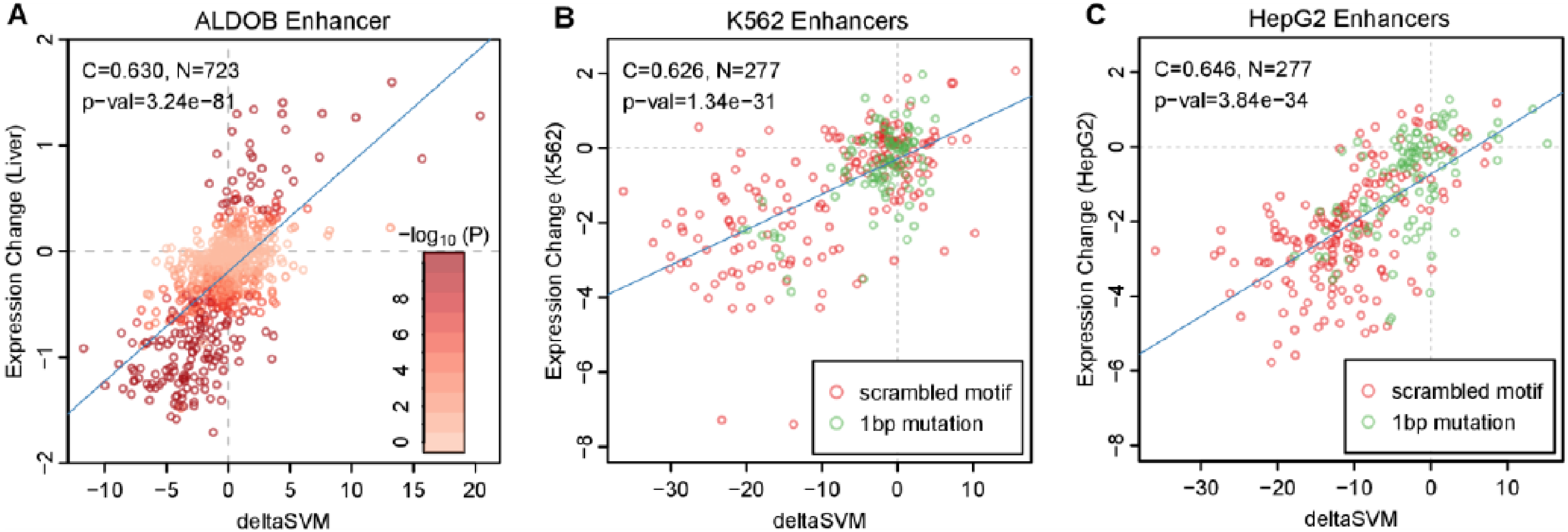
Patwardhan et al., 2012

Kheradpour et al, 2013







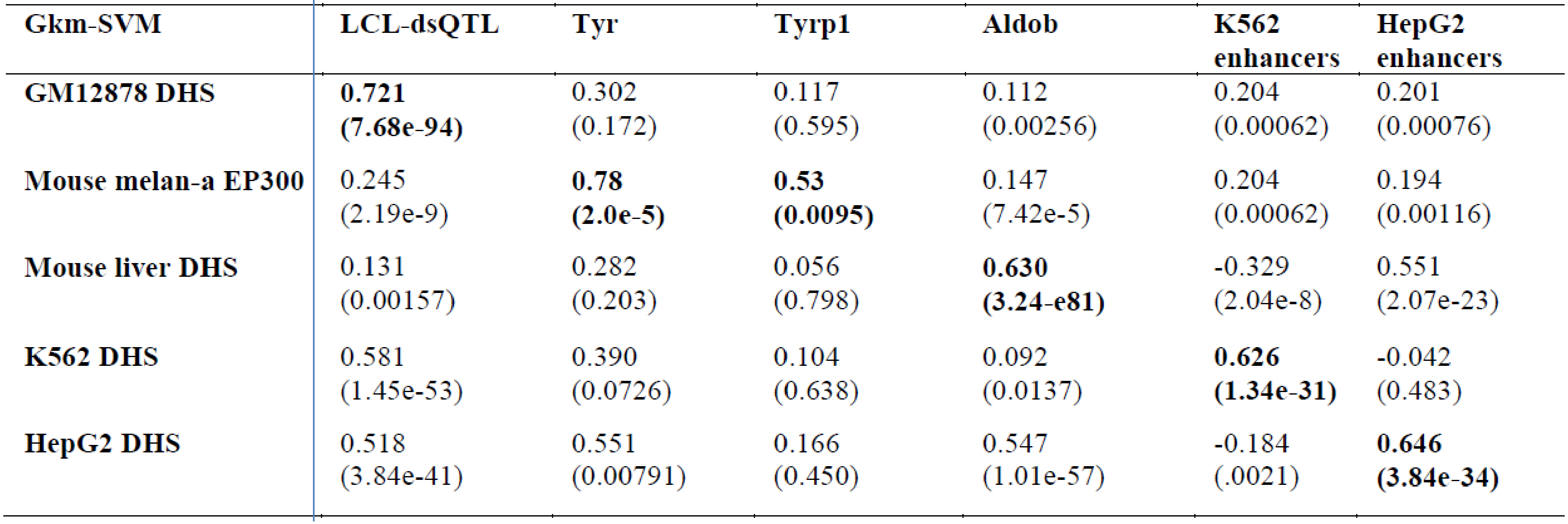


deltaSVM prediction is cell-type specific

Correlation between predictions and observed expression only when using related cell types

training
cell-type

Validation cell-type



Gkm-SVM	LCL-dsQTL	Tyr	Tyrp1	Aldob	K562 enhancers	HepG2 enhancers
GM12878 DHS	**0.721** **(7.68e-94)**	0.302 (0.172)	0.117 (0.595)	0.112 (0.00256)	0.204 (0.00062)	0.201 (0.00076)
Mouse melan-a EP300	0.245 (2.19e-9)	**0.78** **(2.0e-5)**	**0.53** **(0.0095)**	0.147 (7.42e-5)	0.204 (0.00062)	0.194 (0.00116)
Mouse liver DHS	0.131 (0.00157)	0.282 (0.203)	0.056 (0.798)	**0.630** **(3.24-e81)**	-0.329 (2.04e-8)	0.551 (2.07e-23)
K562 DHS	0.581 (1.45e-53)	0.390 (0.0726)	0.104 (0.638)	0.092 (0.0137)	**0.626** **(1.34e-31)**	-0.042 (0.483)
HepG2 DHS	0.518 (3.84e-41)	0.551 (0.00791)	0.166 (0.450)	0.547 (1.01e-57)	-0.184 (.0021)	**0.646** **(3.84e-34)**

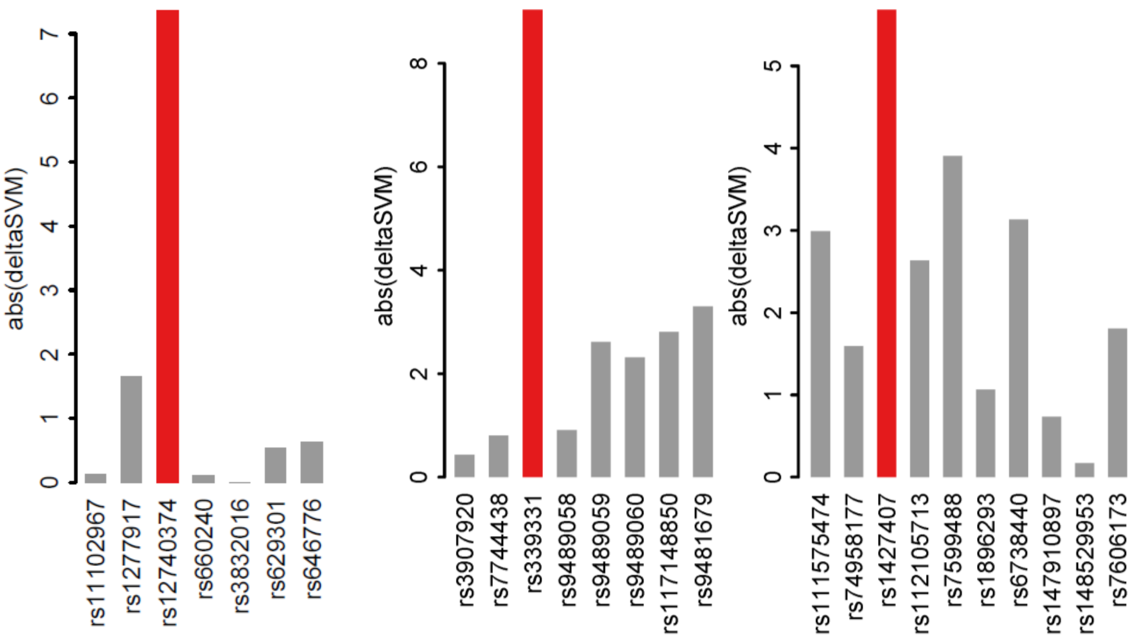
Underscores importance of training on appropriate cell type, developmental time, and biological state to identify disease relevant regulatory vocabulary

deltaSVM identifies previously validated GWAS disease associated SNPs

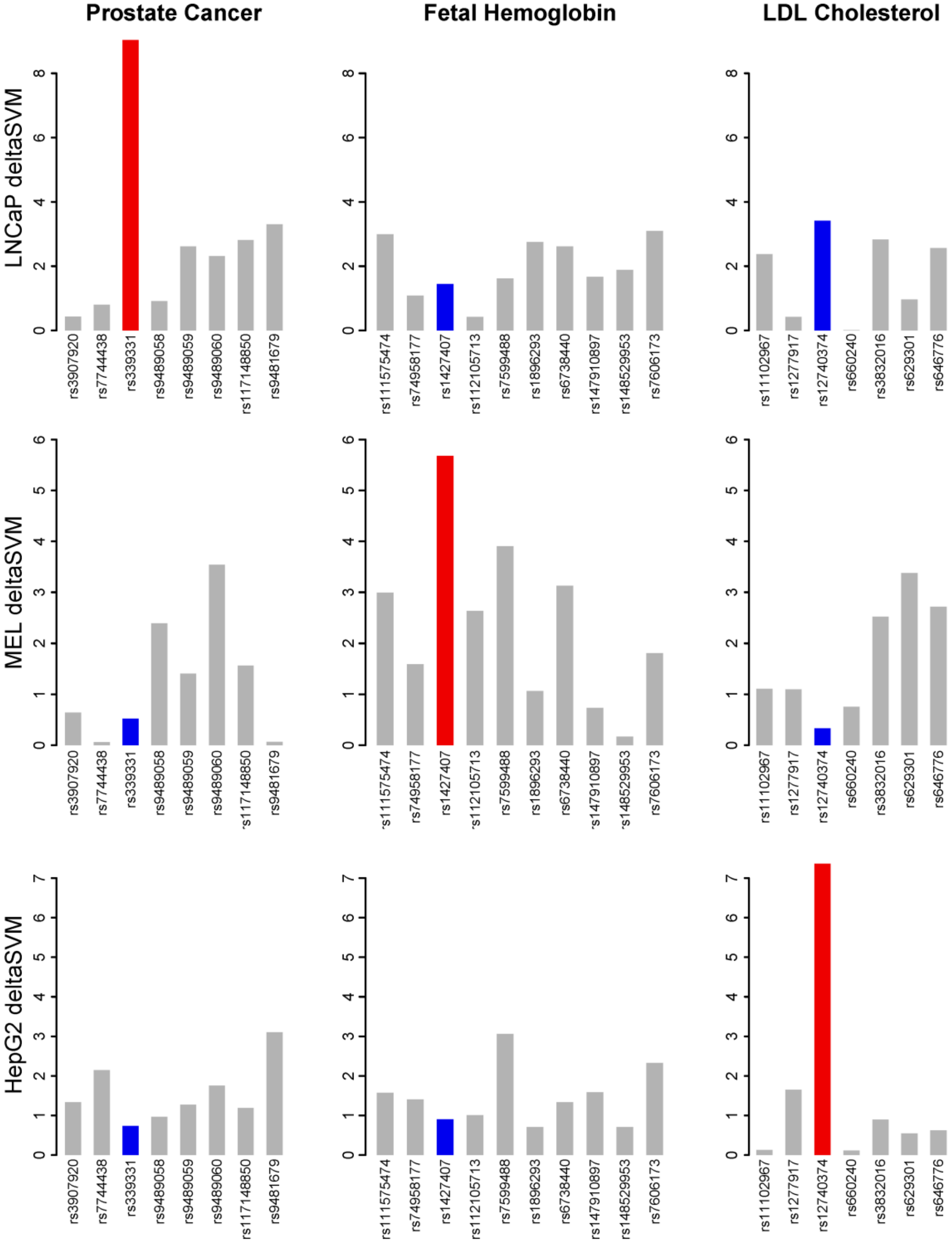
deltaSVM for:

validated causal SNP

flanking negative SNPs



Disease/trait:	LDL cholesterol	Prostate Cancer	Fetal Hemoglobin
gkm-SVM trained on DHS:	HepG2	LNCaP	MEL
gene:	*Sort1*	*Rfx6*	*Bcl11a*
Ref:	Musunuru Nature 2010	Huang Nat Gen 2014	Bauer, Orkin Science 2013

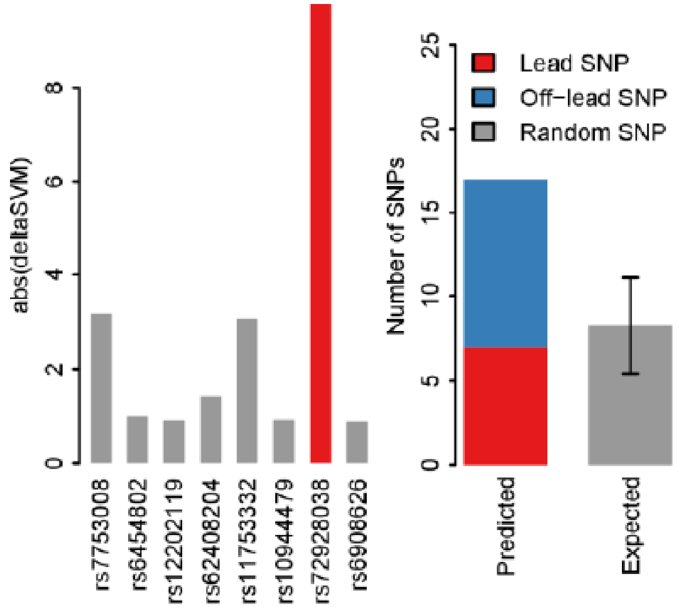


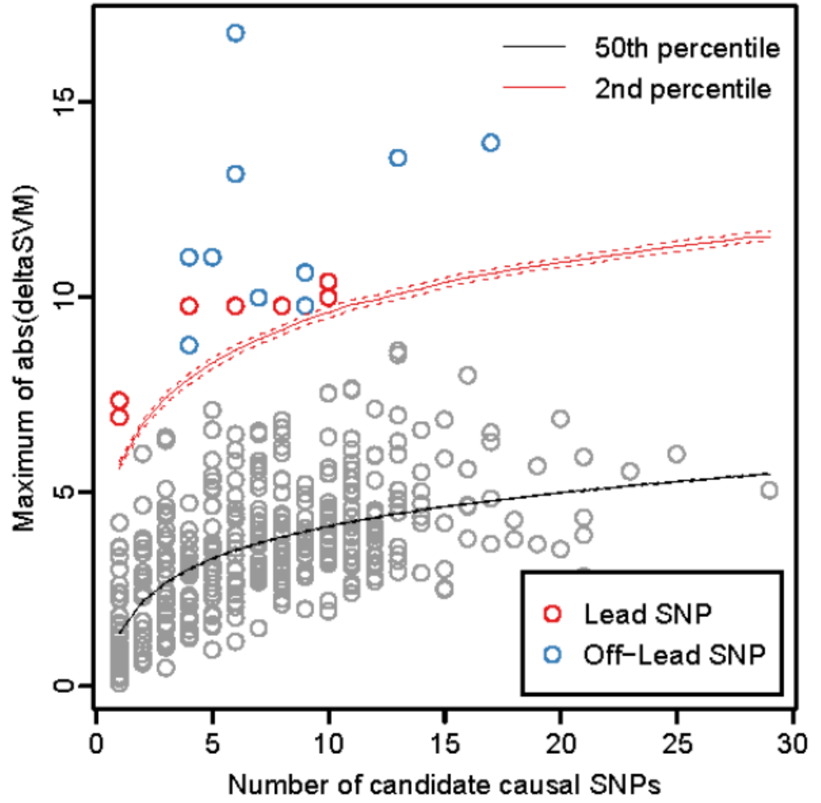
Validated causal SNPs only score higher than flanking negative SNPs when deltaSVM is trained on the relevant cell type

deltaSVM at GWAS associated SNPs predicts novel autoimmune SNPs

Train on Th1 DHS:

Type 1 Diabetes, Crohn’s Disease
Multiple Sclerosis, Celiac Disease
Primary Biliary Cirrhosis,
Rheumatoid Arthritis, Allergy
Autoimmune Thyroid Disease
Ulcerative Colitis, Vitiligo
Systemic Lupus Erythematosus





deltaSVM at GWAS LD SNPs score systematically higher than random SNP control sets

413 GWAS lead SNPs
use PICS [Fahr Nature 2014], 22822 R^2>.2 → 3114 PICS

11 very high confidence predictions

But, very challenging: not biggest impact, but right place and time

Now predicting and testing disease associated SNPs