

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

_____ *Mike Beer* _____

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Regulation Saturation Assessment

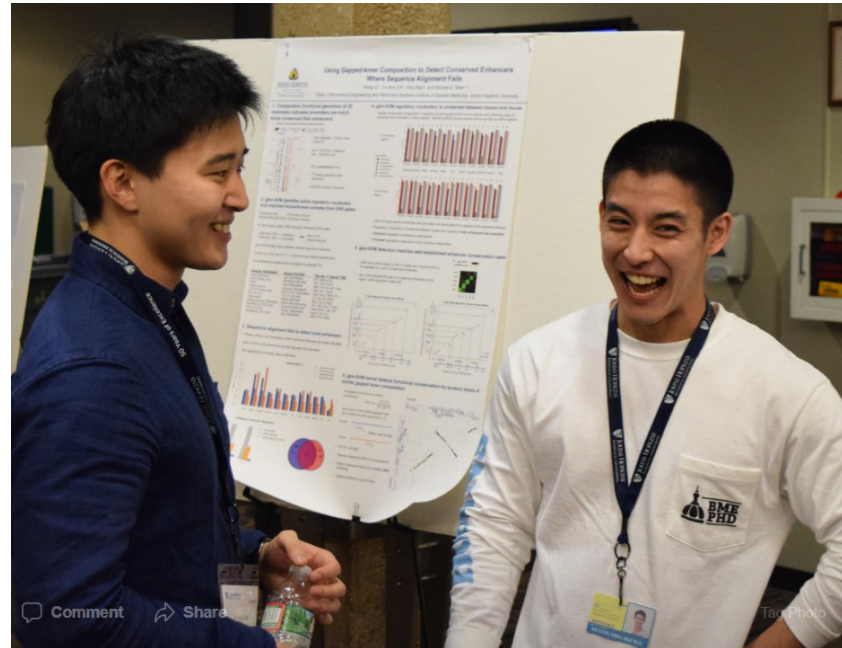
Mike Beer

Dustin Shigaki

Dept of Biomedical Engineering
Institute of Genetic Medicine
July 6, 2018



JOHNS HOPKINS

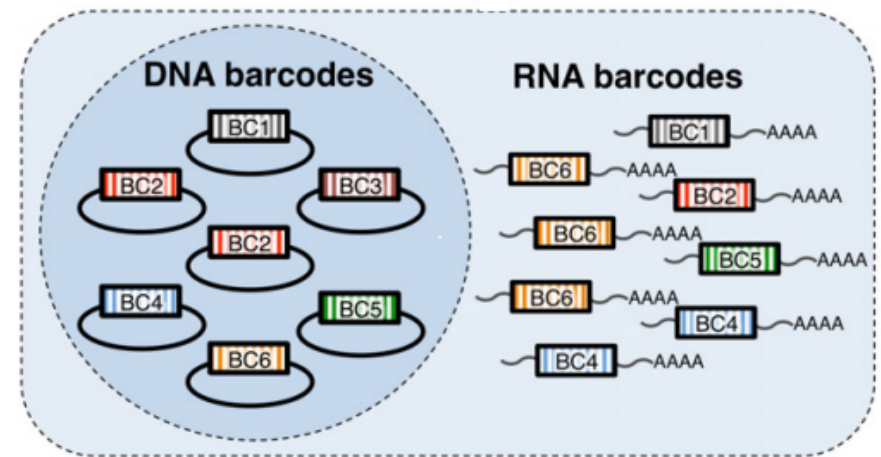
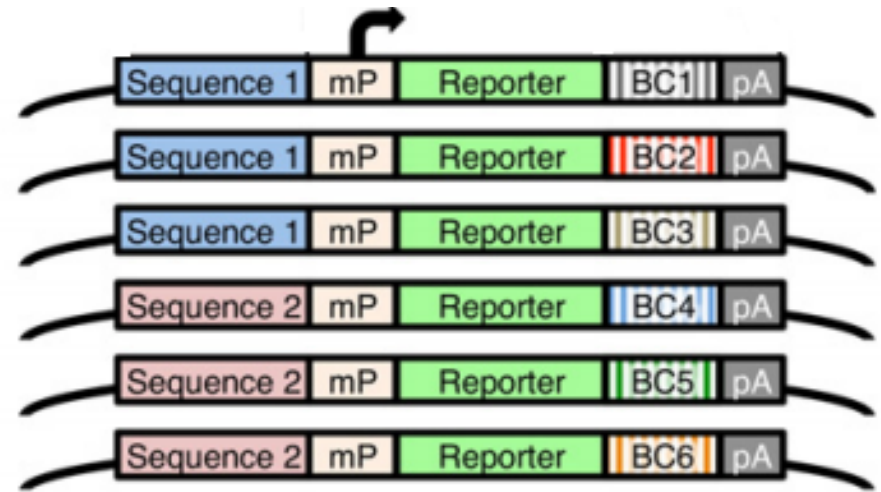
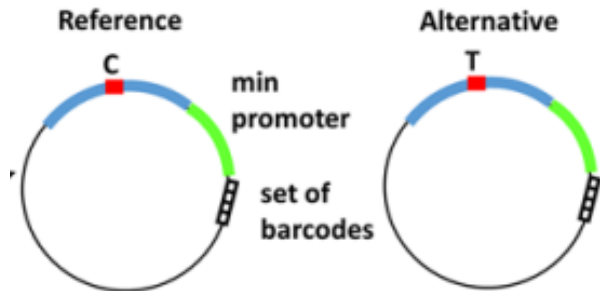


Experimental Design

Promoters: Cells transfected with plasmid containing promoter candidate upstream of a reporter gene (has a unique “barcode” sequence tag).

Enhancers: Plasmids contained enhancer region + minimal promoter.

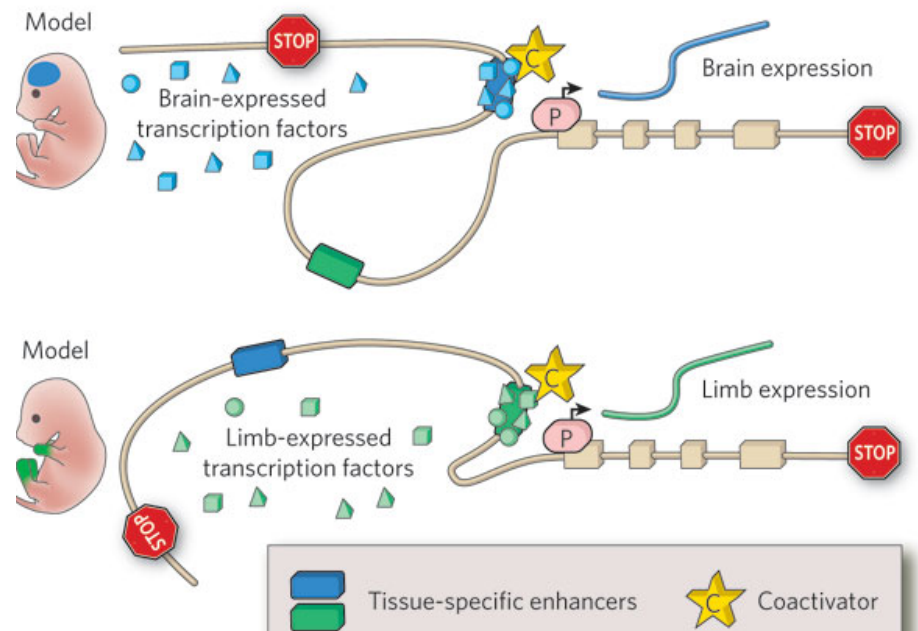
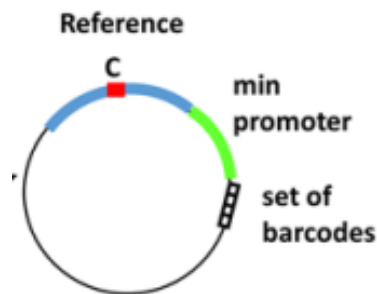
Regulatory regions inserted were reference or alternates generated by random PCR mutagenesis ~1 mutation per 100bp



Inoue et al. *Genomics* 2015.

$(\text{Log}(\text{RNA/plasmid}))_{alt} - (\text{Log}(\text{RNA/plasmid}))_{ref}$
gives an estimate of the impact of a variant

Mechanisms of Promoter and Enhancer Action

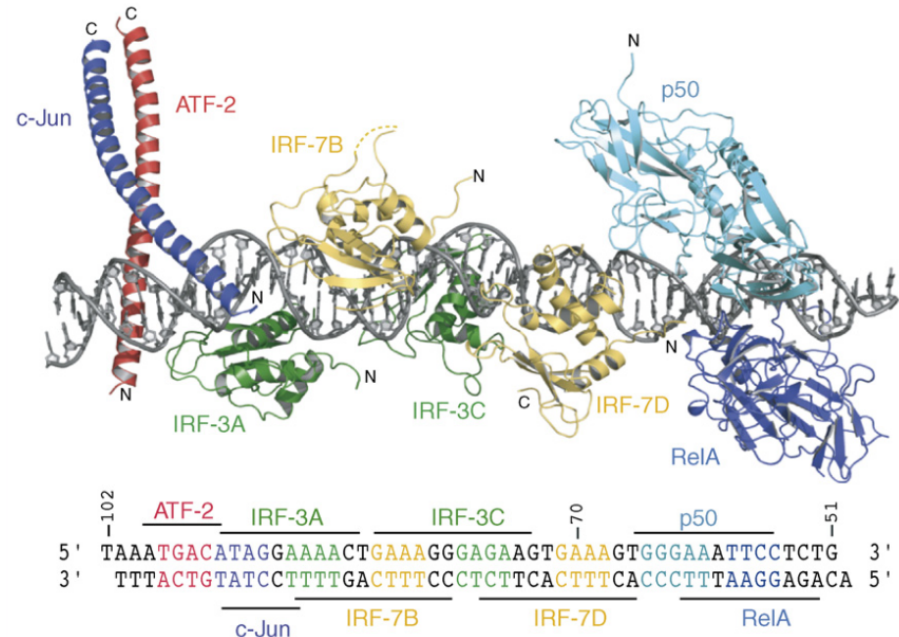


Panne 2008

Tested inserts 200-600bp

Episomal vs. genomic function

Determining cell specific TFs is critical



Disease Associations and Relevant Cells for Tested Regions

Promoters

region	disease association	cell line assayed	size (bp)
F9	Hemophilia B	HepG2	300
GP1BB	Bernard-Soulier Syndrome	HEL 92.1.7	384
HBB	Thalassemia	HEL 92.1.7	187
HBG	Hereditary fetal hemoglobin	HEL 92.1.7	274
HNF4A	Maturity-onset diabetes	HEK293T	285
LDLR	Familial hypercholesterolemia	HepG2	318
MSMB	Prostate cancer	HEK293T	591
PKLR	Pyruvate kinase deficiency	K562	469
TERT	Various types of cancer	HEK293T, GBM	258

Enhancers

region	disease association	cell line assayed	size (bp)
IRF4	Human pigmentation	SK-MEL-28	451
IRF6	Cleft lip	HaCaT	600
MYC	Various types of cancer	HEK293T	600
SORT1	cholesterol & myocardial infraction	HepG2	600
ZFAND3	Type 2 diabetes	MIN6	578

Cell Types, cont'd

cell line	cell type origin	regions tested
HepG2	hepatocellular carcinoma	F9,LDLR,SORT1
K562, HEL 92.1.7	erythroleukemia	GP1BB,HBB,HBG,PKLR
HEK293T	embryonic kidney + SV40	HNF4A,MSMB,TERT,MYC
SK-MEL-29	melanoma	IRF4
HaCaT	keratinocyte	IRF6
MIN6	mouse pancreatic beta cells	ZANDF3

Well
represented
in ENCODE

* Rare in
ENCODE

To compare models I used:

- ENCODE DHS for HepG2, K562, HEK293, NHEK, Melano
- and MIN6 ATAC-seq data from Stitzel lab AJHG 2018

Training and Prediction Data

$(\text{Log}(\text{RNA/plasmid}))_{alt} - (\text{Log}(\text{RNA/plasmid}))_{ref}$ estimates impact of variant

For each promoter/enhancer, about 25% of the data was released for training.

#Chrom	Pos	Ref	Alt	Value	Confidence
X	138612669	T	A	-0.17	0.07
X	138612669	T	C	-0.26	0.24
X	138612669	T	G	0.34	0.05
X	138612670	A	C	0	0
X	138612670	A	G	0.22	0.2
X	138612670	A	T	0.12	0.03

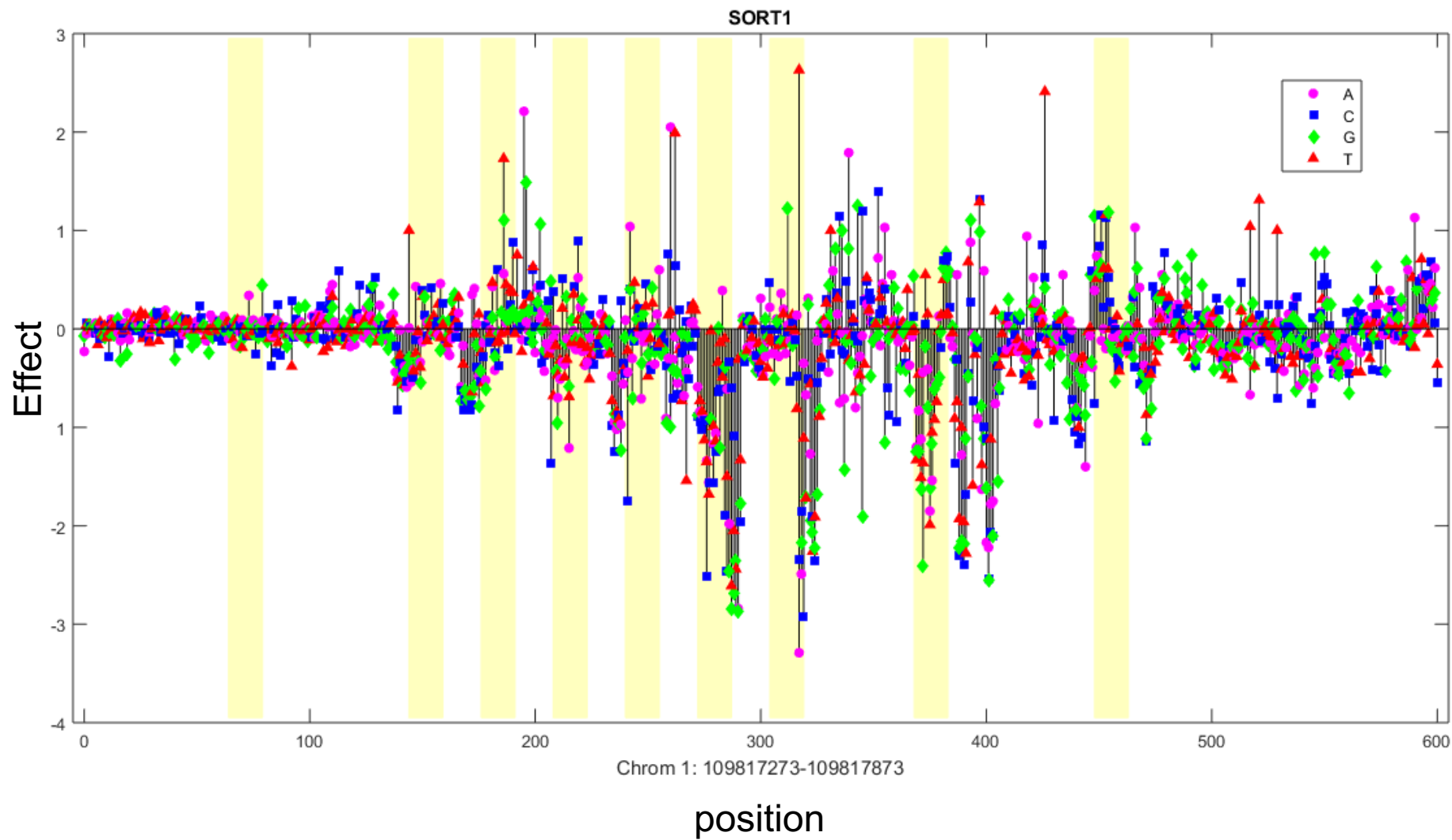
For remaining 75%, participants asked to predict direction of change in expression:

-1 for repressive effect 0 for very little or no effect 1 for activating effect

Chrom	Pos	Ref	Alt	Promoter_Enhancer	Direction	P_Direction	Confidence	SE
1	109817274	G	A	SORT1	1	0.797	0.75	1
1	109817274	G	C	SORT1	0	0.797	0.25	1
1	109817274	G	T	SORT1	0	0.777	0.25	1

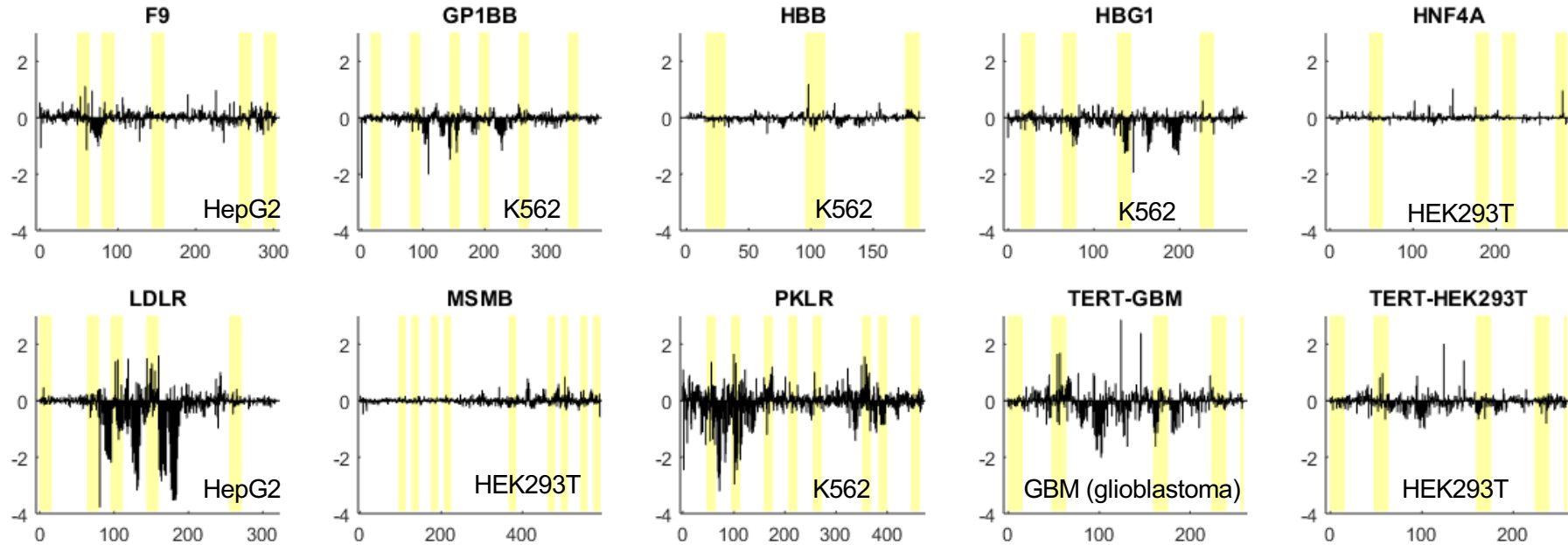
(Example from Group 1)

Experimental Results SORT1

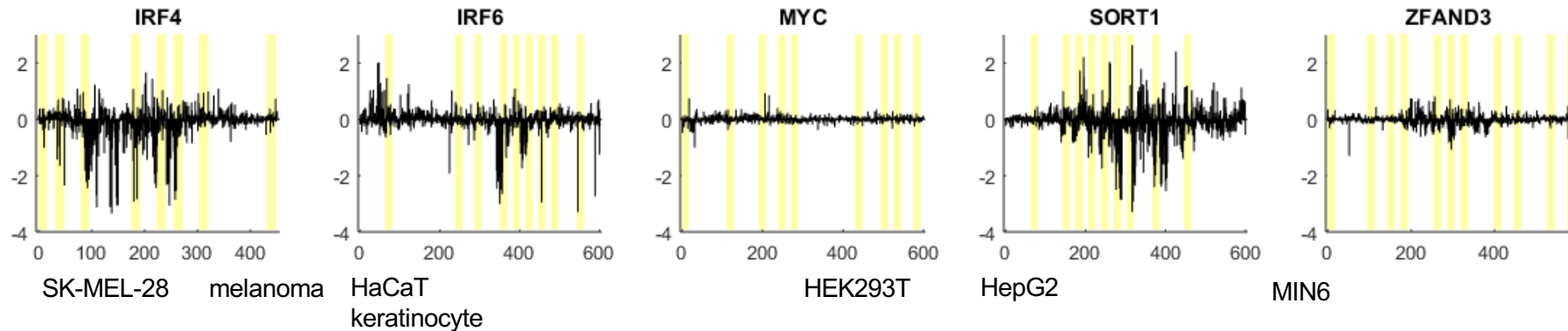


Training and Test Data

Promoters

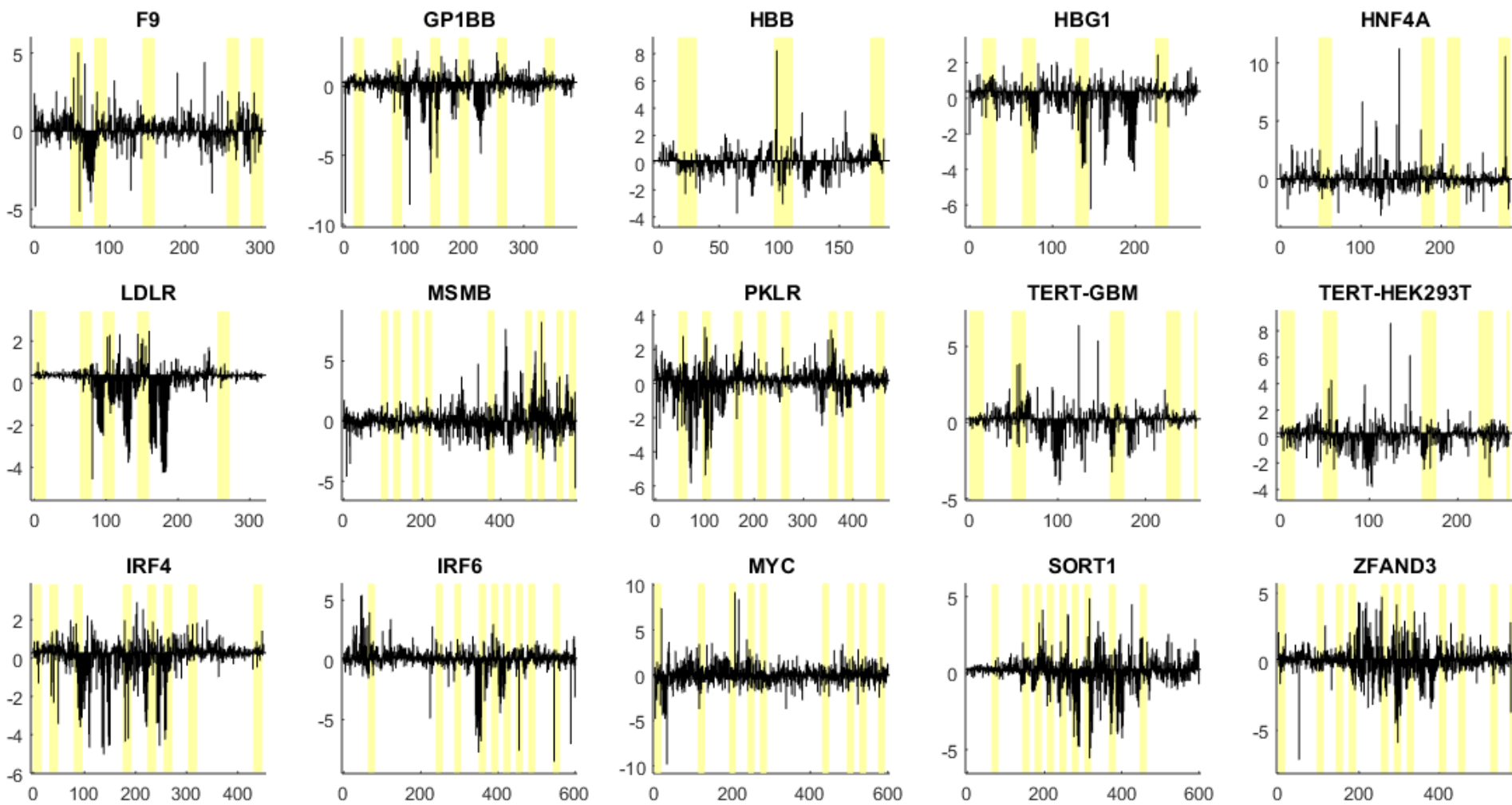


Enhancers

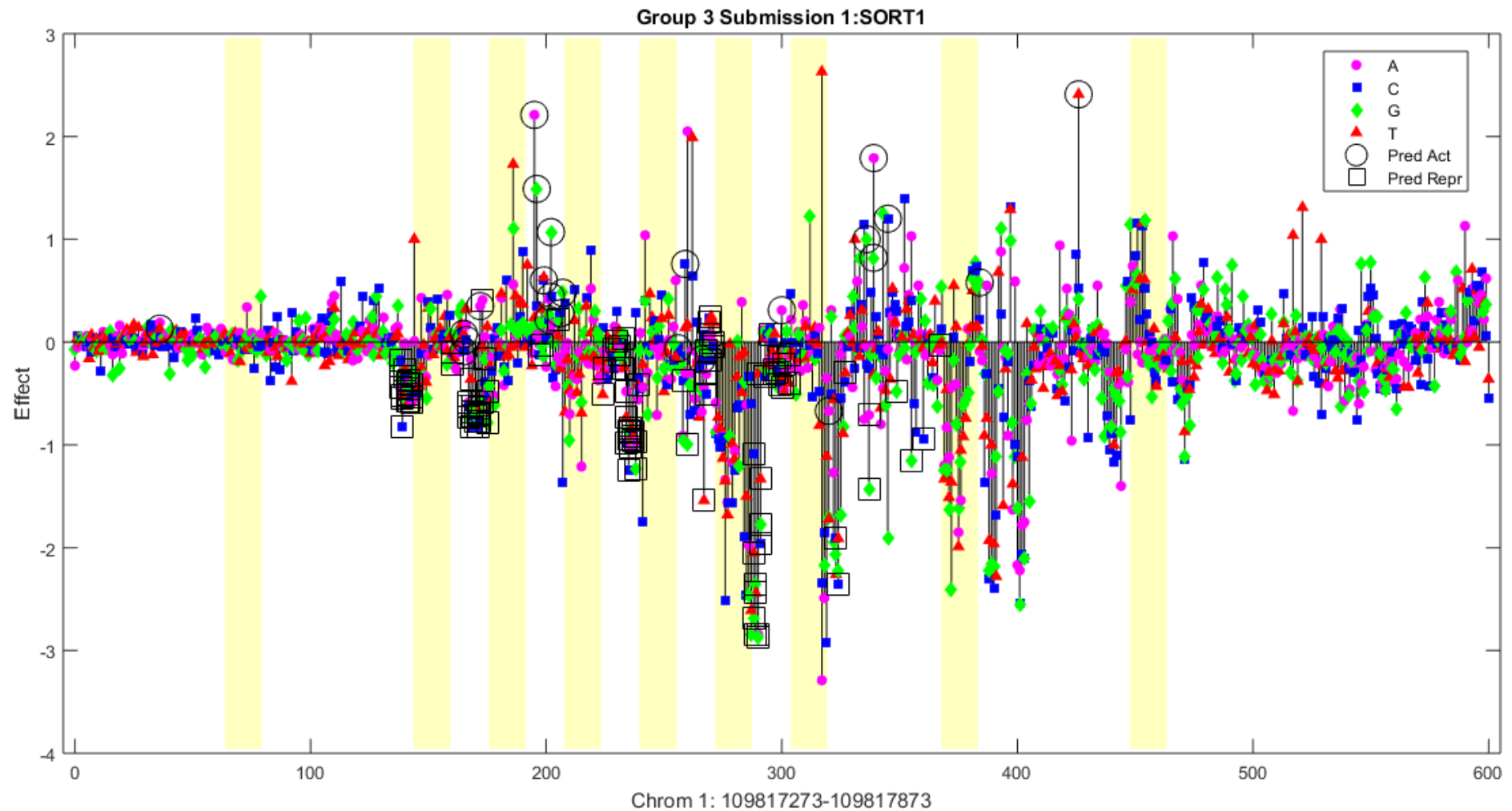


Some variable sensitivity, low signal for HNF4A, MSMB, MYC

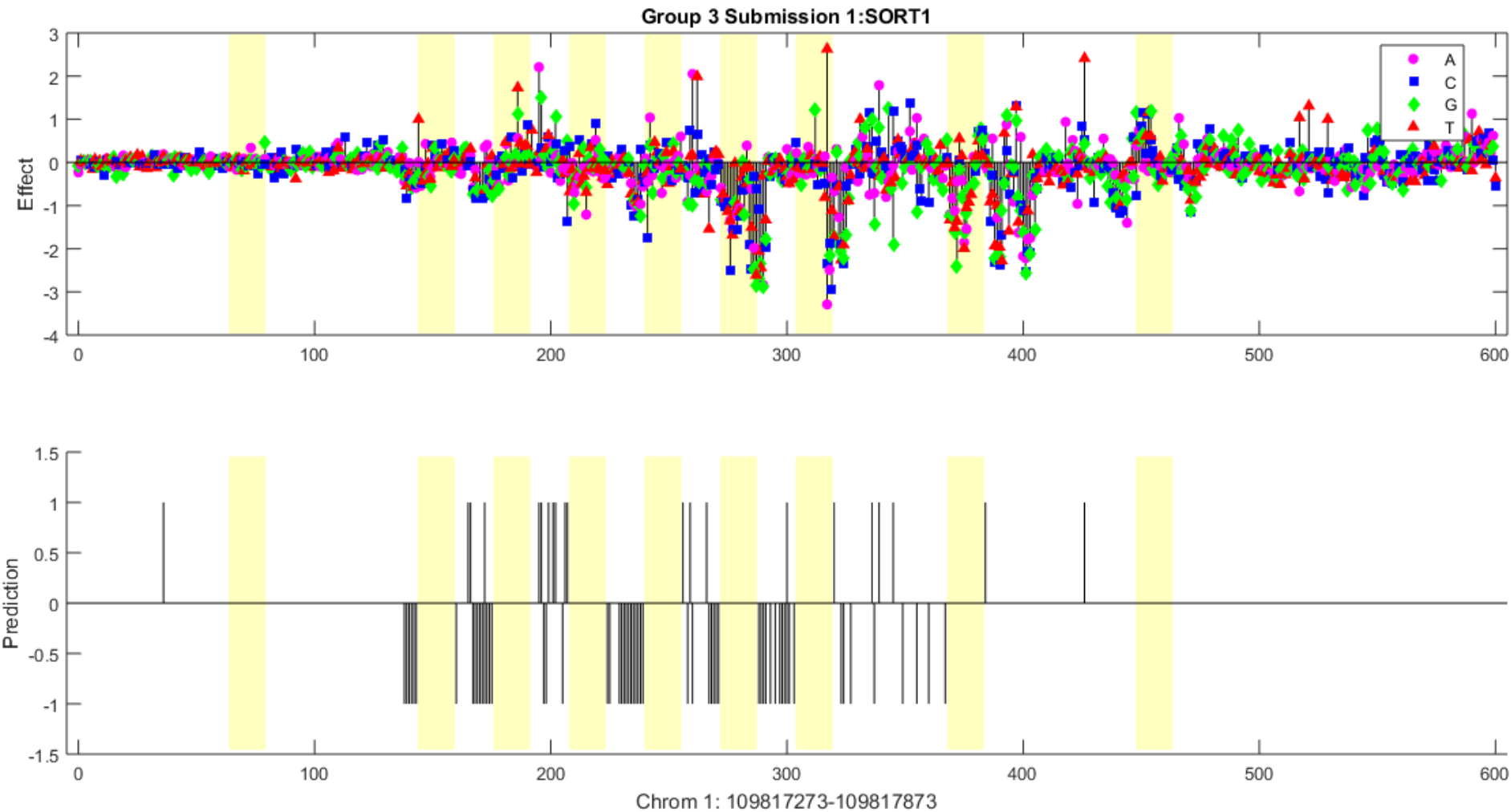
Training and Prediction Data normalized to unitary variance



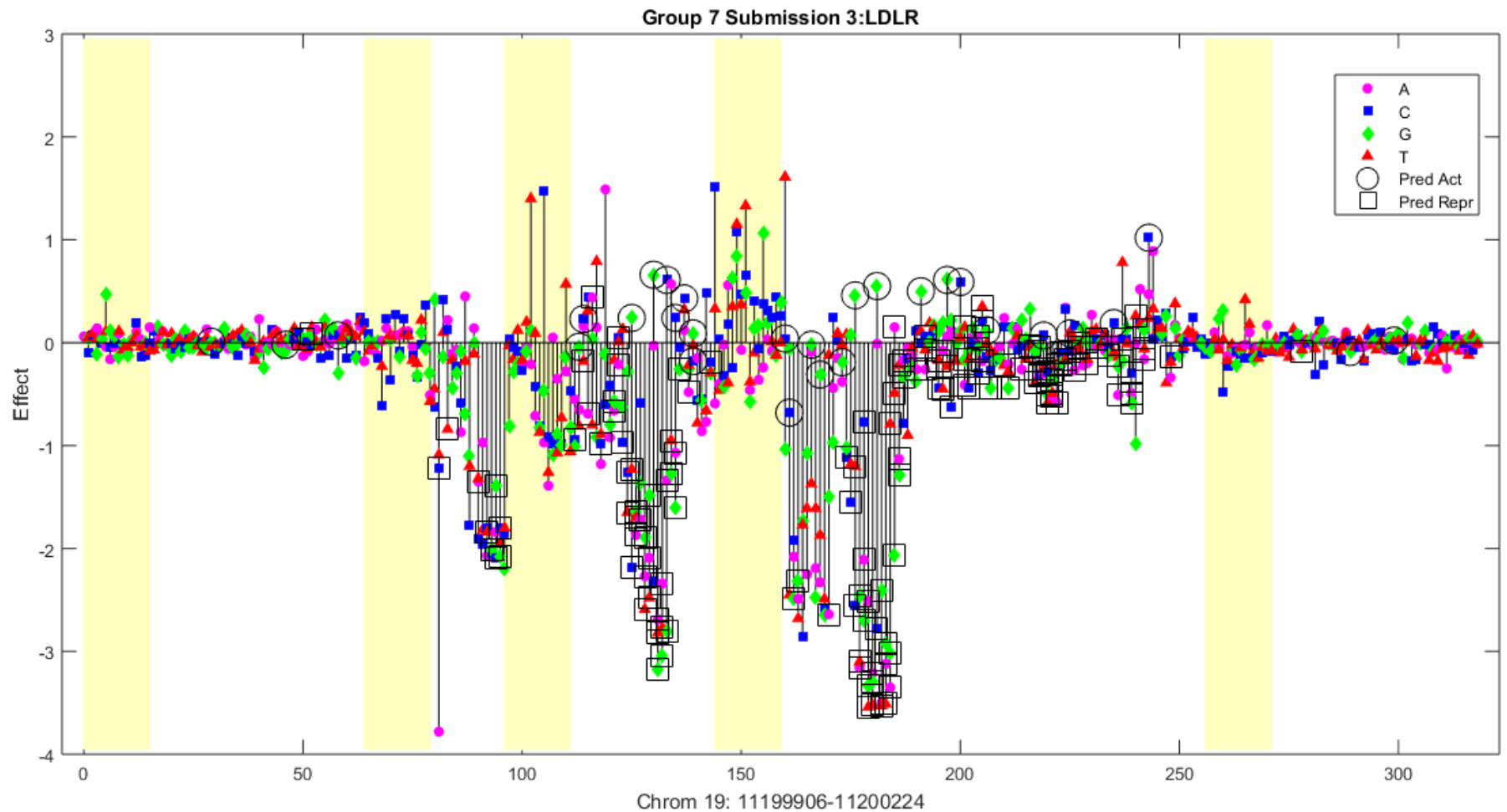
Sample Prediction Data Group3 SORT1



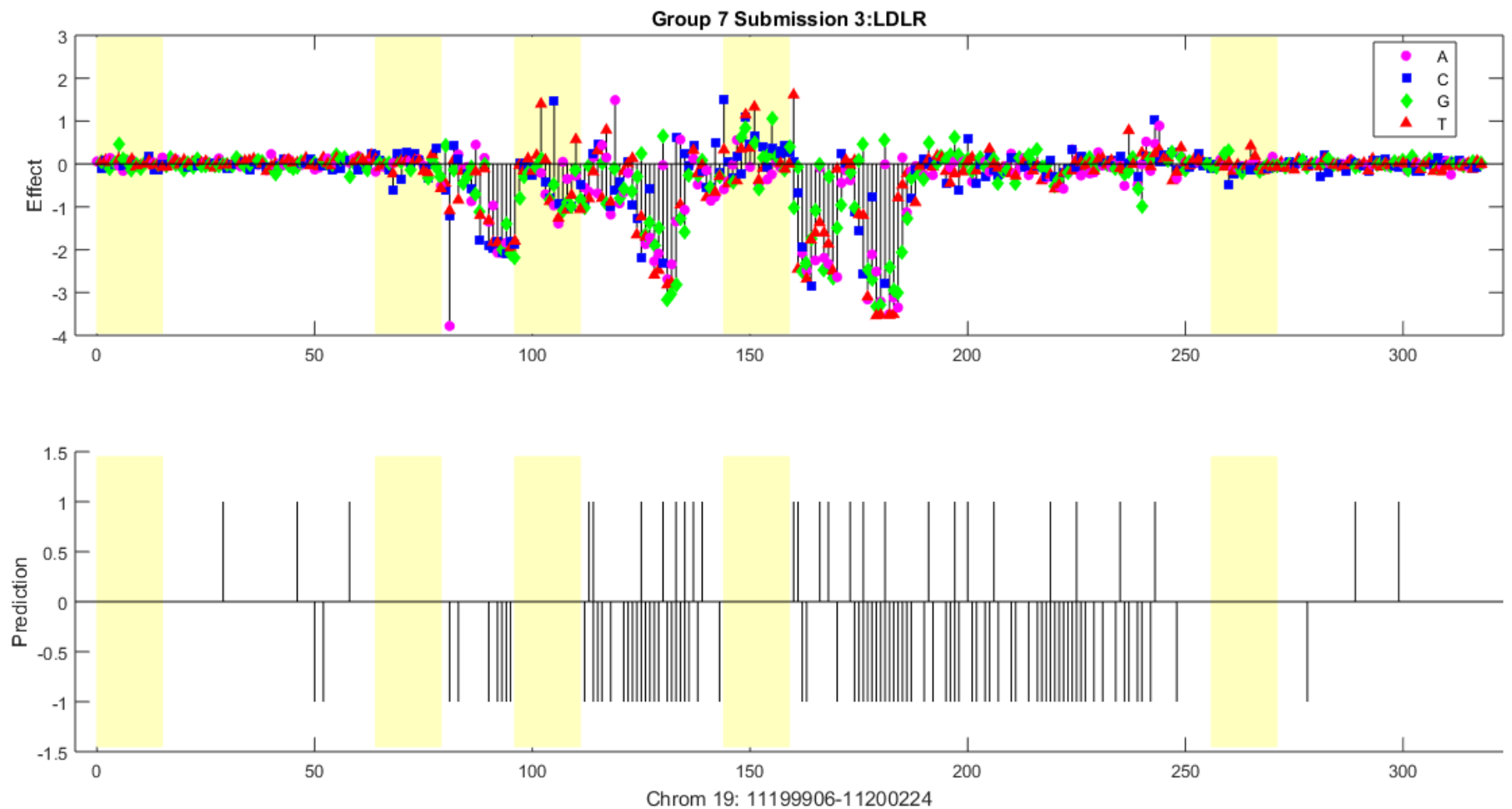
Sample Prediction Data Group 3 SORT1



Sample Prediction Data Group 7 LDLR



Sample Prediction Data Group 7 LDLR



Prediction Groups

	name	submitter	summary rank
Group_1	Biu	Ron Unger, Bar-Ilan Univ, Israel	
Group_2	ANA	Aashish Adhikari UC Berkeley	
Group_3	BoyleLab	Alan Boyle, Univ Michigan	1
Group_4	Gao Group	Yuan Shi, Peking University	
Group_5	Epimetheans	John Reid, Cambridge University	3
Group_6	autosome.ru	Ivan Kulakovskiy, Inst. Math. Probs Biol., Russia	
Group_7	Zhongxia	Zhongxia Yan, UC Berkeley	2

Prediction Methods Summary

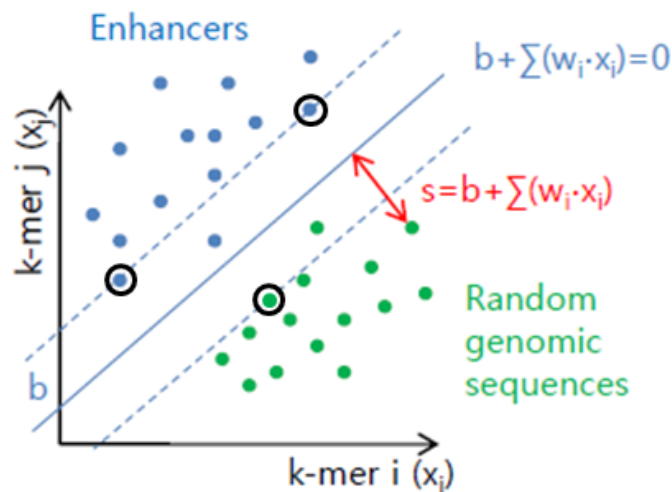
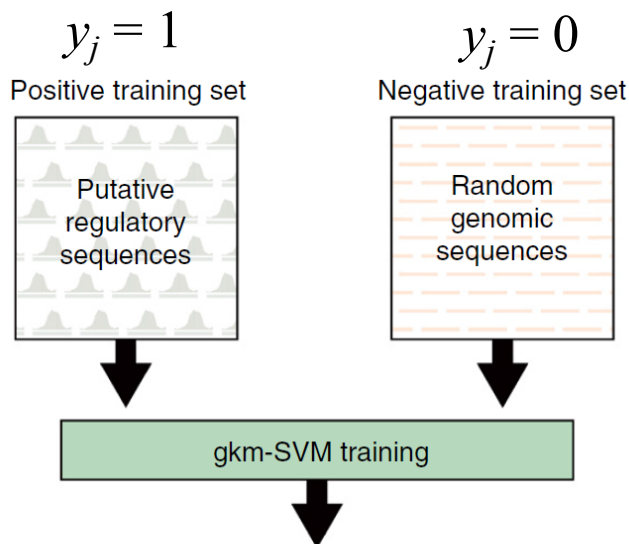
Group1	1-1	Random Forest	CADD, FATHMM,						
	1-2								
Group2	2-1/4	XGBoost	ENCODE +						
			Roadmap features						
Group3	3-1	MultiClass RF	Chip-Seq peaks, DHS				+DeepSea features		
	3-2		peaks, PWMs				+DeepBind features		
	3-3					+PWM features			
	3-4								
	3-5								
	3-6								
Group4	4-1	DNN	Conv filters	ENCODE feats					
	4-2								
Group5	5-1	XGBoost	Conservation, DHS, DeepSea Neural Networks, Region Identity,						
	5-2		Substitution						
	5-3								
	5-4								
	5-5								
	5-6								
	5-7								
Group6	6-1	XGBoost	Conservation, DHS, FANTOM5,						
			ATAC-Seq, PWMs						
Group7	7-1	MultiClass RF	Chlp-seq, DHS, PWMs, DeepSea scores,						
	7-2		DeepBind scores						
	7-3								
	7-4								

Comparison to Previous Published Methods

CADD scores: Kircher Nat Gen 2014

neither deltaSVM or CADD use training data

gkm-SVM: 2014 Ghandi, Lee, Beer, PLOS Comp Bio **deltaSVM**: 2015 Nat Gen CRAN R-package: 2016 Bioinformatics



$$S = \sum_i w_i x_{ij} + b$$

$$K(S_1, S_2) = \frac{\langle x_{S_1} \cdot x_{S_2} \rangle}{\|x_{S_1}\| \|x_{S_2}\|}$$

Regulatory sequence vocabulary

All unique gapped k-mers	SVM weights	
TGA-TCA---	3.275	Positive regulatory activity
-TGA-TCA--	3.147	
--ATGA-T-A	3.091	
-ATGA-T-A-	2.992	
...	...	
A--AAAT-G-	0.0002	Neutral
TTT-TC-T--	0.0001	
-AAA-A-TG	-0.0001	
-TTTT-CC-	-0.0001	
...	...	
--CACCTG--	-1.171	Negative regulatory activity
----CACCTG	-1.183	
-CACCTG---	-1.253	
CACCTG----	-1.282	

Ref: TCTCTGCAAC**C**AAAGACAGA
Alt: TCTCTGCAAA**A**AAAGACAGA

WT 10-mer	Weight	Variant 10-mer	Weight	Difference
TCTCTGCAAC	0.012	TCTCTGCAAA	0.298	0.286
CTCTGCAACA	0.082	CTCTGCAAAA	0.104	0.021
TCTGCAACAA	0.280	TCTGCAAAAA	0.114	-0.166
CTGCAACAAA	0.330	CTGCAAAAAA	-0.029	-0.359
TGCAACAAAG	0.441	TGCAAAAAAG	-0.025	-0.466
GCAACAAAGA	0.784	GCAAAAAAGA	0.008	-0.776
CAACAAAGAC	1.031	CAAAAAAGAC	0.109	-0.922
AACAAAGACA	0.545	AAAAAAGACA	-0.453	-0.998
ACAAAGACAG	0.671	AAAAAGACAG	-0.516	-1.187
CAAAGACAGA	-0.036	AAAAGACAGA	-0.478	-0.442
▲		▲		

deltaSVM = -5.007

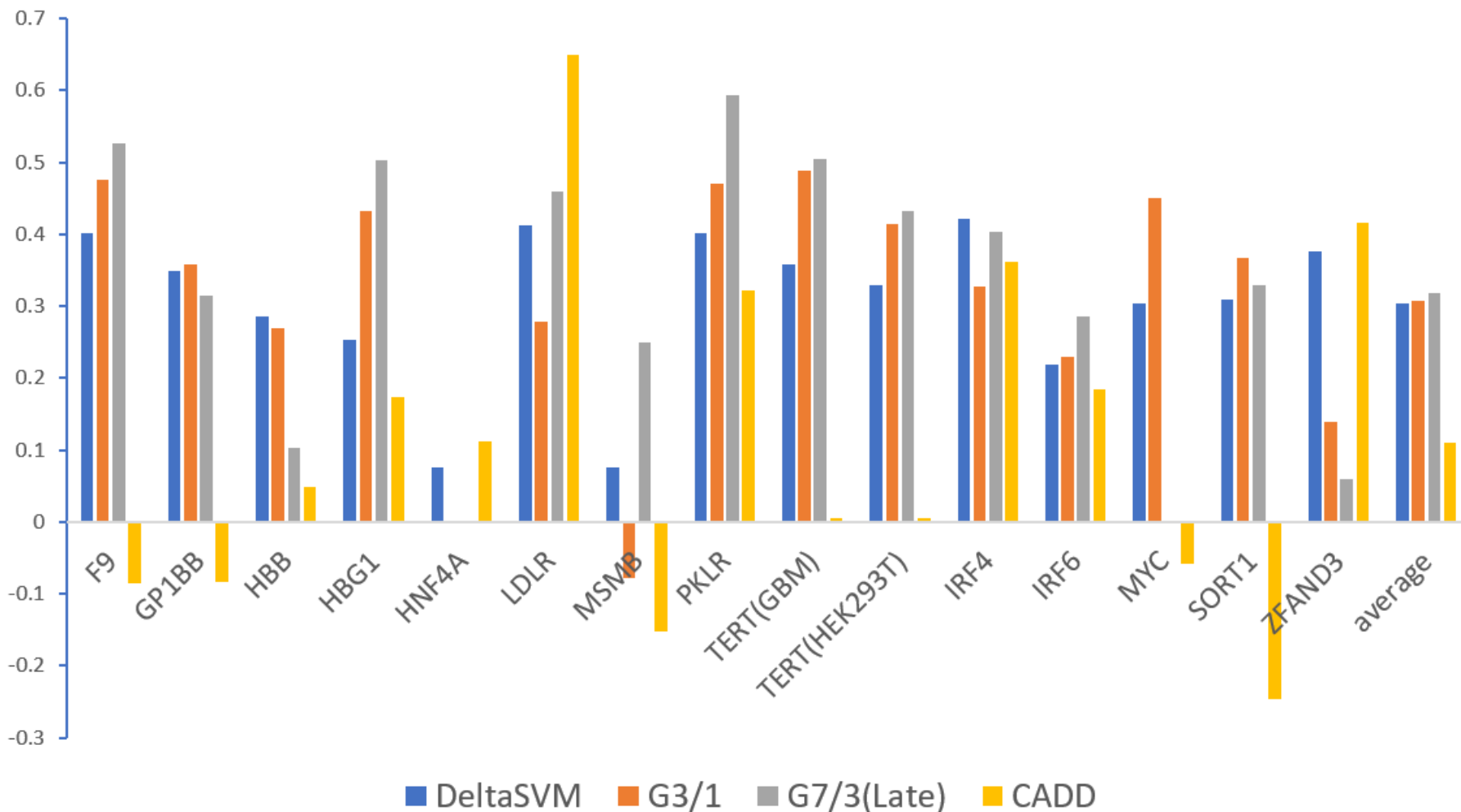
Performance Comparisons Test Data

AUROC

		Pearson	rank	Spearman	rank	Pos v Neg	Pos v Rest	Neg v Rest	Rank
Group1	1-1	0.060				0.505			
	1-2	0.069				0.544			
Group2	2-1/4								
Group3	3-1	0.318	1	0.2379	tie	0.825	0.708	0.805	1
	3-2	0.157				0.730			
	3-3	0.131				0.524			
	3-4	0.301				0.842			
	3-5	0.151				0.835			
	3-6	0.121				0.518			
Group4	4-1	0.047				0.530			
	4-2	0.041				0.556			
Group5	5-1	0.252		0.2391	tie	0.711			
	5-2	0.230				0.707			
	5-3	0.228				0.709			
	5-4	0.228				0.709			
	5-5	0.255				0.714			
	5-6	0.255	3	0.2347		0.714	0.608	0.691	3
	5-7	0.255				0.714			
Group6	6-1	0.103				0.537			
Group7	7-1	0.283				0.699			
	7-2	0.283				0.699			
	7-3	0.318	1	0.2492	tie	0.762	0.706	0.776	2
	7-4	0.203				0.643			
DeltaSVM		0.305				0.840	0.649	0.779	
CADD		0.110							

Pearson Correlation on Test Data by Region

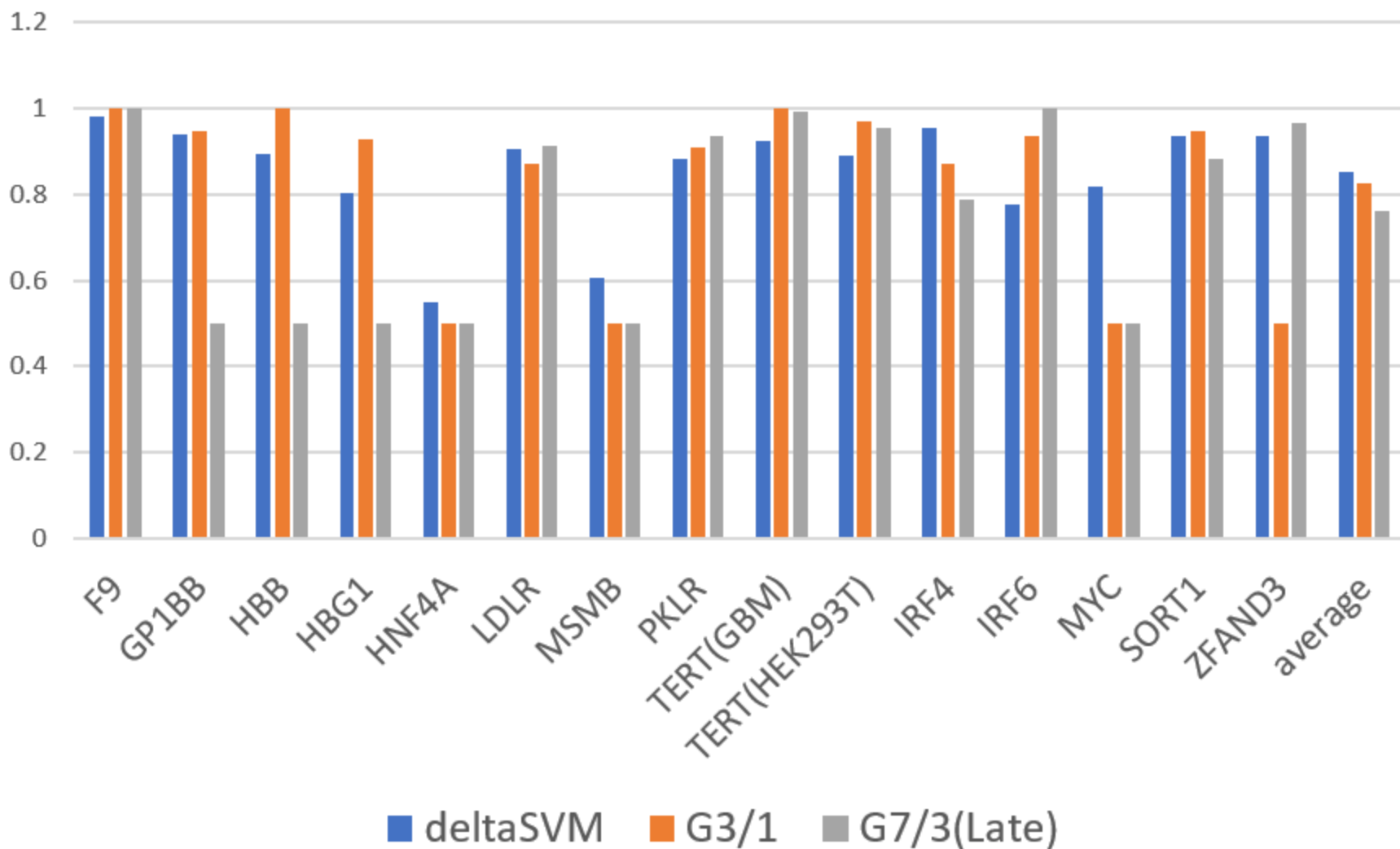
Test Set Pearson Correlation



HNF4A, MSMB, difficult for all methods, ZFAND3 difficult for 3-1, 7-3, MYC mixed

AUROC on Test Data by Region

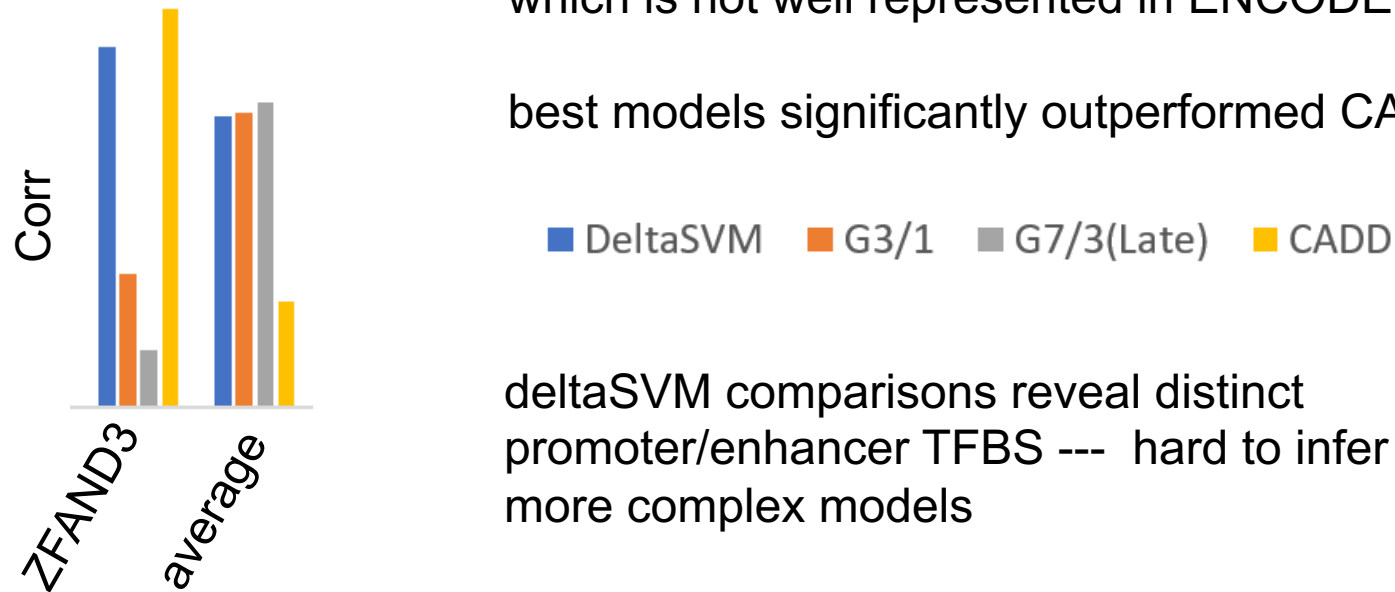
Pos v Neg AUROC



Performance Comparison Summary

All best models did well (3-1, 5-6, 7-4) using features including TF filters learned from DeepSea DNN trained on ENCODE data

Comparable performance to deltaSVM except for MIN6 ZFAND3 enhancer which is not well represented in ENCODE set



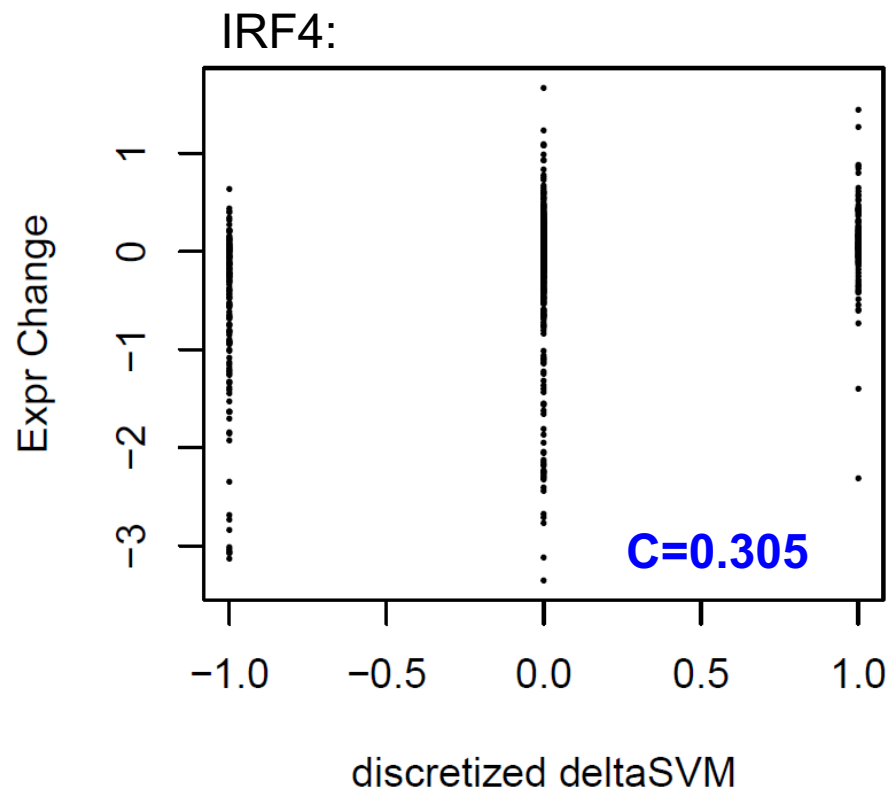
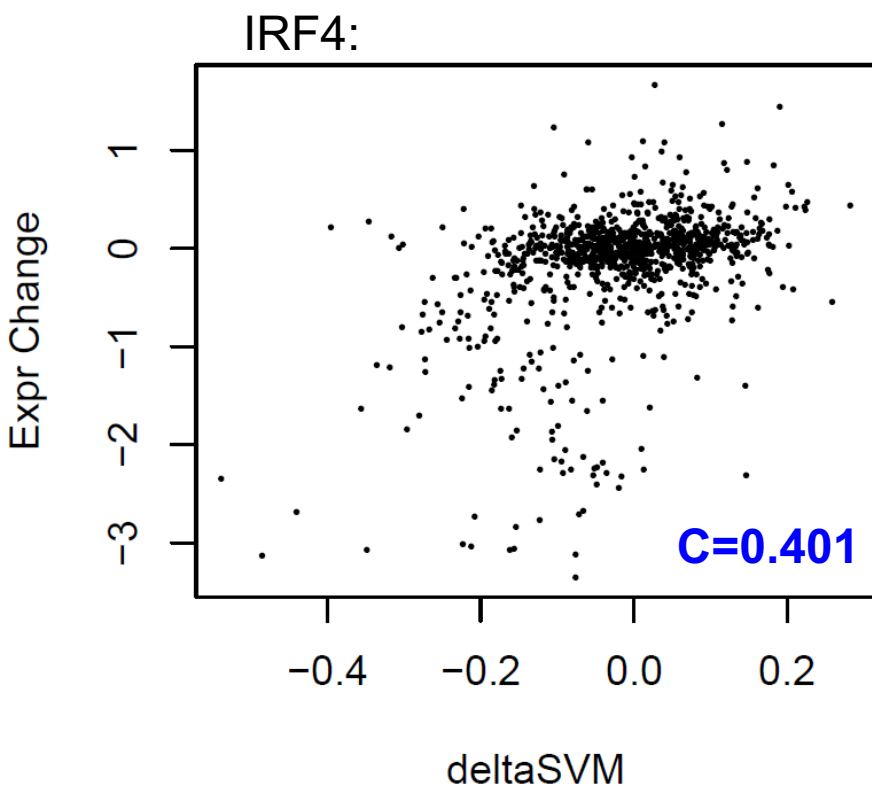
best models significantly outperformed CADD

deltaSVM comparisons reveal distinct promoter/enhancer TFBS --- hard to infer from more complex models

Discretization of Predictions Loses Information and Limits Predictive Power

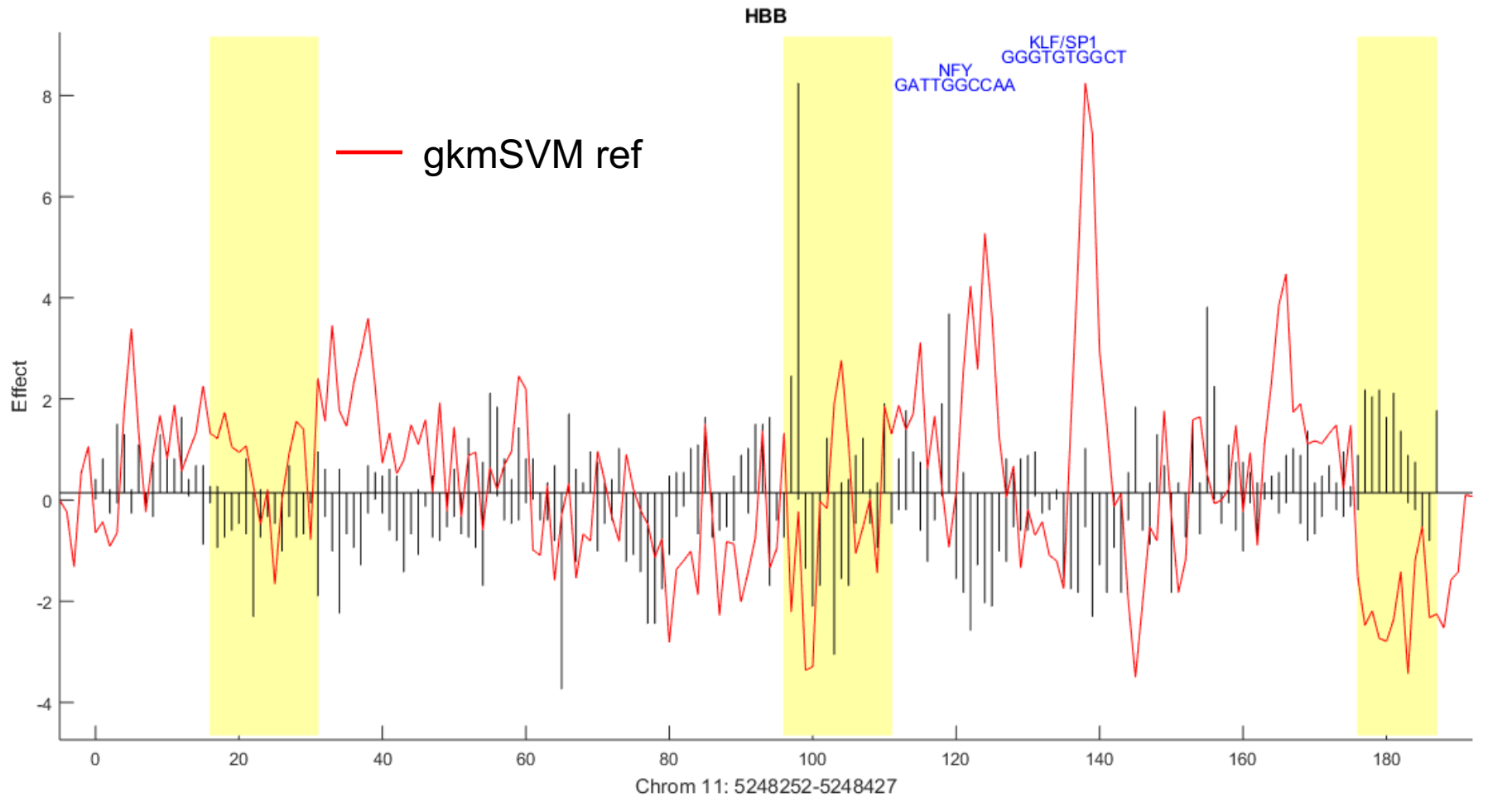
Although most models produce continuous scores, predictors were asked to predict a discrete class label (-1, 0, 1)

All models were compared this way, but results could improve without discretization



Some methods predicted no +1 or -1 for some regions → AUC = 0.5

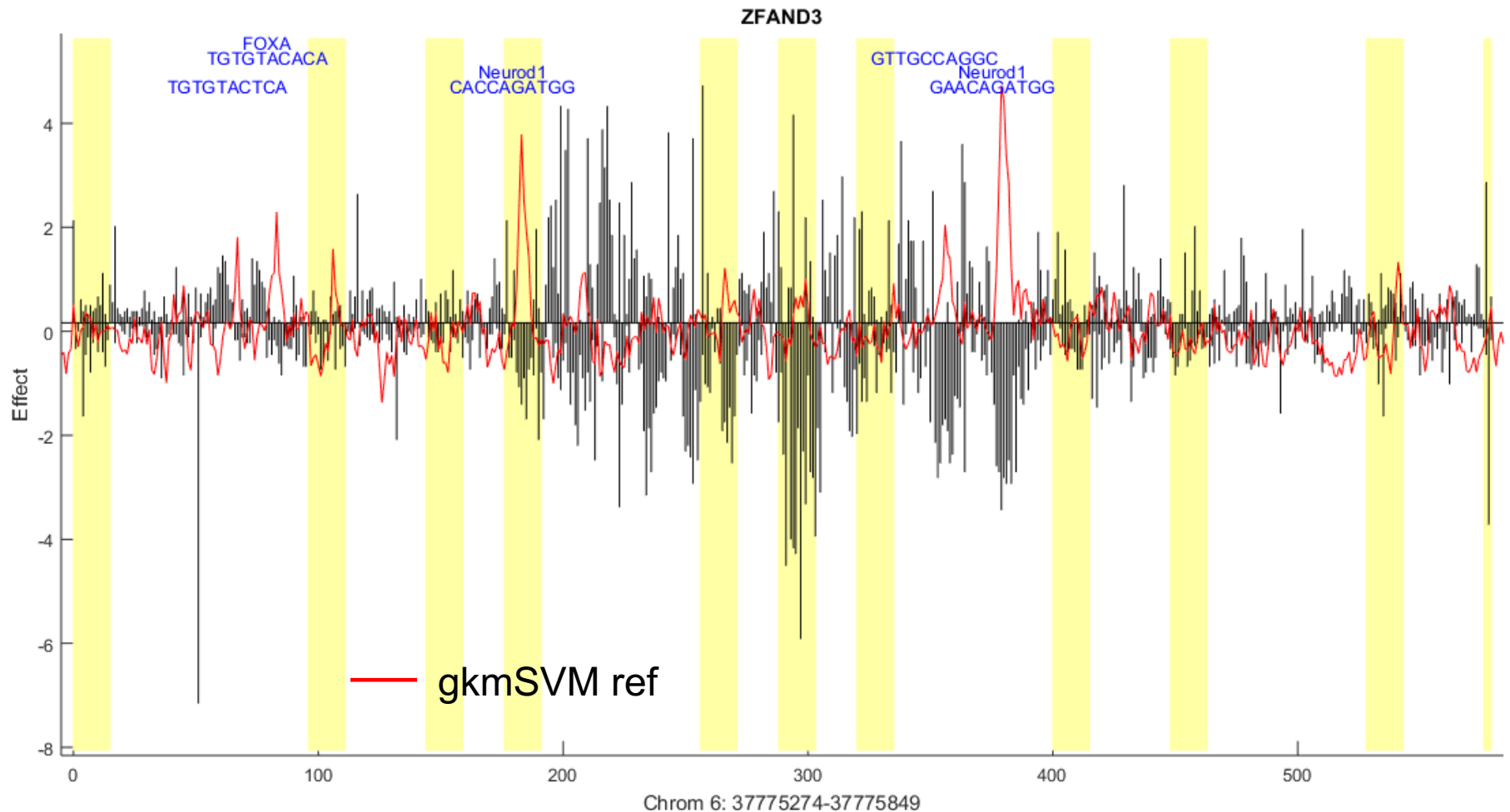
Simpler Models Perform Comparably and are Interpretable



Common promoter features: NFY, KLF/SP1, ELK4

repressive
TFBS

Simpler Models Perform Comparably and are Interpretable



Conclusions and Lessons Learned

- Groups 3-1 and 7-3 performed almost equally well, 5-6 was third
- Accurate TF models more important than classifier method used, multiclass RF beat DNN, but used features learned from DNN trained on all ENCODE cell types
- Performance comparable to deltaSVM trained on chromatin accessibility data from one relevant cell type
- Recommendation: Better if predictors submit continuous scores, I'm requesting this from Groups 3, 5, 7.
- Thanks to providers for interesting data set!