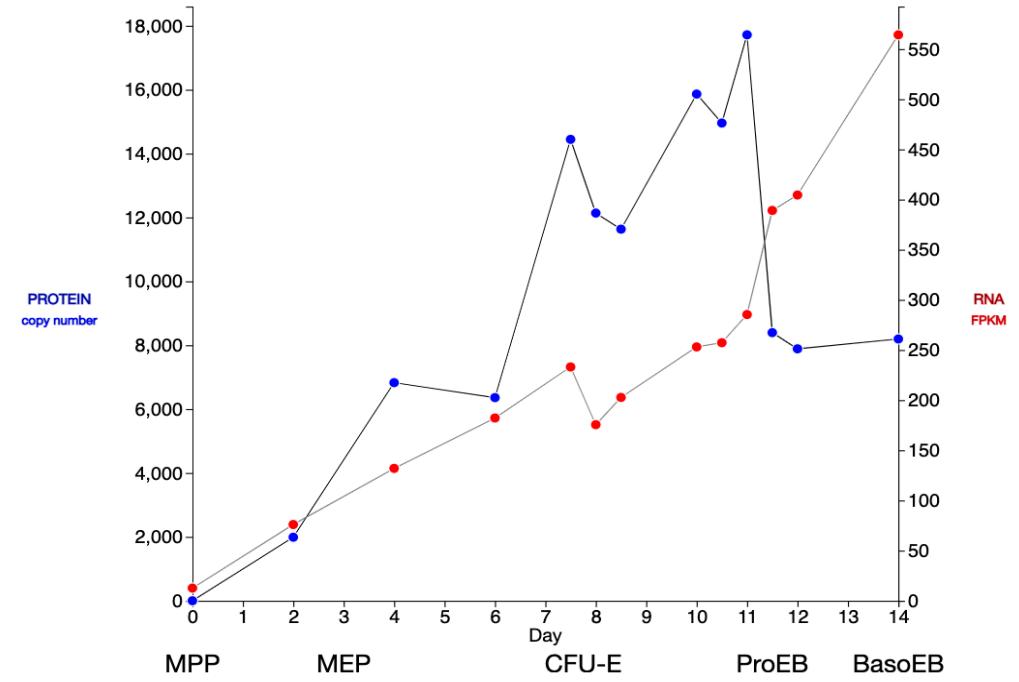


Exploring DDX3X as a possible regulator of late stage erythropoiesis

```
hagfish:~/github/TrenaProjectErythropoiesis/explore/slides/ddx3x.pptx
```

Does the helicase RNA-binding protein DDX3 play a repressive role in late-stage erythropoiesis? (KLF1, for example)



A trena model predicting late time series KLF1 protein levels from tf and rbp mRNA. DDX3X exhibits strong anti-correlation.

gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
TCF7	7987.480	933.939		0.786	0.933	462112.9	0.000	tf	1 KLF1p
DDX3X	-37.274	-2.751	-0.750	-0.931	1272151.2	0.000	rbp	2	KLF1p
BUD13	327.039	48.303		0.893	0.921	885696.0	0.000	rbp	3 KLF1p
FOXJ3	0.000	45.994		0.857	0.903	215373.5	0.000	tf	4 KLF1p
MLXIPL	0.000	658.374		0.786	0.893	379064.8	0.000	tf	5 KLF1p
FXR2	0.000	-16.033		-0.857	-0.892	1205324.6	0.000	rbp	6 KLF1p
TAF1	0.000	-52.400		-0.893	-0.890	268943.3	0.000	tf	7 KLF1p
CREB1	0.000	-30.324		-0.857	-0.884	401021.6	0.146	tf	8 KLF1p
CREB3L1	0.000	-14.480		-0.643	-0.879	0.0	0.000	tf	9 KLF1p
BHLHE41	0.000	-32.751		-0.857	-0.876	437511.0	0.000	tf	10 KLF1p

Traditional mRNA-only model of KLF1 mRNA

DDX3X not in this model

gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	target
NFE2	0.09884844	0.04443212	0.9912088	0.9521618	13649.742	0.014626449	tf	KLF1
TFDP1	0.00000000	0.07349797	0.9824176	0.9255354	12271.343	0.000000000	tf	KLF1
NFIX	0.00000000	0.24739791	0.9780220	0.8963371	10138.555	0.000000000	tf	KLF1
E2F4	0.97629969	0.10899634	0.9604396	0.9826948	14133.772	0.105645738	tf	KLF1
TROVE2	0.00000000	0.97831683	0.9516484	0.9253720	5398.729	0.000000000	rbp	KLF1
SP1	0.95177748	0.77836063	0.9472527	0.9602376	14783.940	0.000000000	tf	KLF1
CREB3L1	0.00000000	2.22935516	0.9428571	0.8851608	9153.254	0.000000000	tf	KLF1
SP2	3.46471606	0.93729805	0.9252747	0.9699924	10429.620	0.180998366	tf	KLF1
SP4	0.00000000	3.27724321	0.9208791	0.8463264	3345.139	0.004377712	tf	KLF1
MXII1	0.00000000	0.10077246	0.8417582	0.8820152	12439.674	0.000000000	tf	KLF1

Occurrence of RBPs across 91 trena late-time, protein models rank <= 10 - the origin of our interest in DDX3X

```
head(as.data.frame(sort(table(tbl.rbp10$gene), decreasing=TRUE)), n=12)
  Var1 Freq
1  DDX3X   36
2  RBM47   12
3  BUD13   11
4    FXR2    9
5  NCBP3    8
6 HNRNPC    7
7  DGCR8    6
8 IGF2BP2    5
9  GTF2F1    4
10 EIF3D    3
11 HNRNPA1   3
12 YTHDF1    3
```

First suggestion that RBPs play an important regulatory role

nature machine intelligence

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [nature machine intelligence](#) > [articles](#) > [article](#)

Article | [Published: 06 July 2020](#)

Deep learning decodes the principles of differential gene expression

[Shinya Tasaki](#)✉, [Chris Gaiteri](#), [Sara Mostafavi](#) & [Yanling Wang](#)

[Nature Machine Intelligence](#) 2, 376–386 (2020) | [Cite this article](#)

2719 Accesses | 5 Citations | 4 Altmetric | [Metrics](#)

ⓘ A [preprint version](#) of the article is available at bioRxiv.

Abstract

Identifying the molecular mechanisms that control differential gene expression (DE) is a major goal of basic and disease biology. Here, we develop a systems biology model to predict DE and mine the biological basis of the factors that influence predicted gene expression to understand how it may be generated. This model, called DEcode, utilizes deep learning to predict DE based on genome-wide binding sites on RNAs and promoters.

Ranking predictive factors from DEcode indicates that clinically relevant expression changes between thousands of individuals can be predicted mainly through the joint action of post-transcriptional RNA-binding factors. We also show the broad potential applications of DEcode to generate biological insights, by predicting DE between tissues, differential transcript usage, and drivers of ageing throughout the human lifespan, of gene co-expression relationships on a genome-wide scale, and of frequently differentially expressed genes across diverse conditions. DEcode is freely available to researchers to identify influential molecular mechanisms for any human expression data.

Many trena models based on Brand Lab ATAC-seq & rna-seq
have high ranking RBP regulators, of which these are a few

gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	target	rank
DICER1	0.00000000	0.0062452937	0.9032967	0.9120385	0.9713989	0.000000e+00	rbp	SETDB1	1
WDR33	0.00000000	0.0078875570	0.8901099	0.8997973	0.7223530	0.000000e+00	rbp	SETDB1	2
CELF2	0.00000000	0.0028067171	0.9032967	0.8646410	1.3961041	0.000000e+00	rbp	SIN3A	1
ELAVL1	0.00000000	0.0116977349	0.9560440	0.9468186	12.1966384	0.000000e+00	rbp	HDAC1	1
NUDT21	0.12451657	0.0074525831	0.9516484	0.9724939	2.7952158	0.000000e+00	rbp	HDAC1	2
ELAVL1	0.15170004	0.0086121576	0.9780220	0.9687025	11.4359292	1.048655e-01	rbp	TTF2	1
STAU1	0.00000000	0.0101752059	0.8769231	0.8574263	6.7501646	0.000000e+00	rbp	BACH1	1
RBM47	0.00000000	0.5405639499	0.8021978	0.6185401	249.7586023	3.596359e-01	rbp	EGR1	2
YTHDF3	0.00000000	0.0287455602	0.8153846	0.9034688	4.8673295	0.000000e+00	rbp	CHD1	1
ZC3H7B	0.00000000	0.0230283429	0.8109890	0.3752218	7.9507915	1.429642e-02	rbp	CTCF	2
ZNF622	0.00000000	0.0088633373	0.6439560	0.5722783	0.7220828	0.000000e+00	rbp	HCFC1	1
UPF1	0.00000000	0.0108760556	0.6219780	0.8797393	0.3070512	9.121209e-04	rbp	HCFC1	2
HNRRNP D	0.00000000	0.0010512699	0.8373626	0.8978015	9.4437056	0.000000e+00	rbp	HDAC3	1
CPSF6	0.29989361	0.0223494818	0.7846154	0.9380618	9.0419695	0.000000e+00	rbp	HDAC3	2
GNL3	0.00000000	0.0036452591	0.9868132	0.9698988	17.1581988	0.000000e+00	rbp	WDR5	1
XPO5	0.08859308	0.0107686309	0.9868132	0.9849281	5.0860417	0.000000e+00	rbp	WDR5	2
YTHDC2	0.50052393	0.1330242787	0.9648352	0.9414947	136.6402884	0.000000e+00	rbp	HTLF	1
ZC3H7B	0.24920261	0.0495009007	0.8901099	0.8143120	45.8075127	1.791433e-01	rbp	IKZF1	1
GRWD1	0.00000000	0.0088811055	0.9648352	0.8869665	11.7401567	0.000000e+00	rbp	KDM1A	1
HNRRNP C	0.01342793	0.0017316454	0.9648352	0.9672496	5.3479374	1.245692e-03	rbp	KDM1A	2
ALKBH5	0.00000000	0.0024654729	0.8461538	0.8820103	0.4456919	0.000000e+00	rbp	KLF13	2
AKAP8L	0.28092341	0.0247987225	0.9648352	0.9888217	669.5751053	0.000000e+00	rbp	KLF3	1
NCBP3	0.21008816	0.0358613235	0.7846154	0.9223705	1.1634607	6.171183e-02	rbp	NR2C2	1
PPIG	0.00000000	0.0192974061	0.8725275	0.8883840	15.7410171	0.000000e+00	rbp	MTA1	1
CPSF1	0.00000000	0.0176644924	0.8417582	0.9090611	15.1944702	0.000000e+00	rbp	MTA1	2
AARS	0.00000000	0.0004400508	0.8417582	0.6760254	0.3126562	0.000000e+00	rbp	NRF1	1
STAU1	0.00000000	0.0170639014	0.8505495	0.7158587	26.7849950	0.000000e+00	rbp	OGT	2
GPKOW	0.00000000	0.1405171457	0.7934066	0.8999727	244.8837285	0.000000e+00	rbp	RAD21	2
GPKOW	0.00000000	0.1900721375	0.9868132	0.8136937	3699.1040391	0.000000e+00	rbp	NFE2	2
RBM47	0.00000000	0.0438308837	0.8329670	0.6730729	0.3976047	1.697814e-02	rbp	SAP130	1
METAP2	0.01745403	0.0088293723	0.7978022	0.7864391	10.3890050	5.398406e-03	rbp	SMC3	1
HTLF	0.00000000	0.0195494482	0.7626374	0.7745289	1.0360071	8.199274e-04	rbp	SMC3	2
NCBP3	1.04298111	0.0881554065	0.8065934	0.9008382	12.6288794	4.733430e-02	rbp	SUZ12	1
EFTUD2	0.00000000	0.0217600078	0.9692308	0.9088222	170.9882037	0.000000e+00	rbp	PARP1	2
TIAL1	0.07711095	0.0277386896	0.8945055	0.9267908	4.7950508	0.000000e+00	rbp	HDAC2	2

RBP binding across the genome

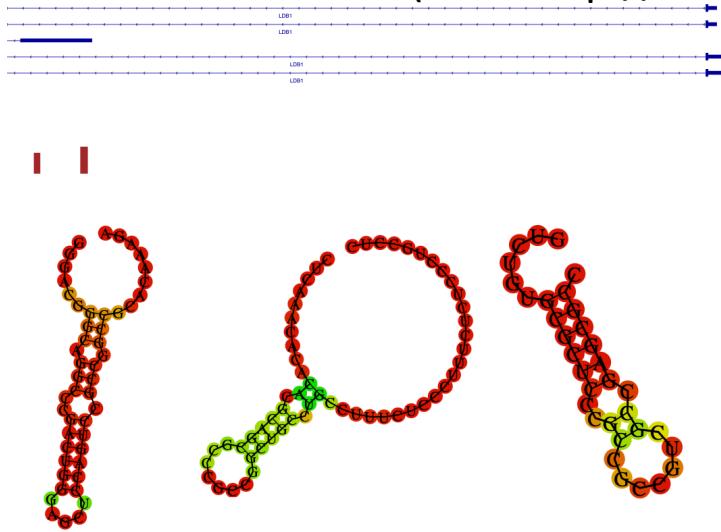
	rbp	model.count	genome per million	total binding sites
1	DDX3X	11	1684009	6.53
2	FXR2	7	69694	100.44
3	RBM47	4	45759	87.41
4	BUD13	3	96181	31.19
5	DGCR8	1	86482	11.56
6	DICER1	1	18449	54.20
7	GPKOW	1	13999	71.43
8	HNRNPA1	1	1275055	0.78
9	HNRNPC	1	7995744	0.13
10	IGF2BP1	1	164484	6.08
11	IGF2BP3	1	155409	6.43
12	LIN28B	1	564575	1.77
13	NONO	1	48197	20.75
14	NPM1	1	897	1114.83
15	PUM2	1	44530	22.46
16	SF3A3	1	33000	30.30
17	TROVE2	1	6945	143.99

Tasks

It appears that DDX3X is associated with translational repression in a number of systems.

~700 genes in the paper that Paul cited in heart, and this paper PMID: 17667941 presents data supporting a general translational repression role in a few cell lines

Predicted structure of the rna binding sites of DDX3X
(from <http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAlign.cgi>)



From Marjorie: I think it could suggest coregulation and this is something that could be tested experimentally. Looking into DDX3X, I found this

<https://app.oxfordabstracts.com/events/1385/program-app/submission/186037>
that says DDX3X binds to highly structured 5'UTR. So the question becomes if our DDX3X targets have a highly structured 5'UTR.

This is a first attempt at examining mRNA structures in eCLIP K562 DDX3X binding sites in 7/100 of your erythropoiesis genes - the seven which showed the most interesting srm/rna-seq discordance:

CTCF E2F4 KLF1 LDB1 SMC3 TAL1 ZBTB7A

Only ZBTB7A had no such binding sites.

For each gene, in the slides attached below, you will find:

- 1) a table of the binding sites, with scores. All seem to be more or less in a transcripts 5' UTR.
- 2) the genome view, showing gene structure and the binding sites (binding sites in dark red)
- 3) predicted mRNA structure of each binding site from <http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RFNAfold.cgi>
 - a) they provide two prediction methods (MFE: maximum free energy, and Centroid). Where they disagree, I show both
 - b) if the two predictions agree, I show just the MFE.

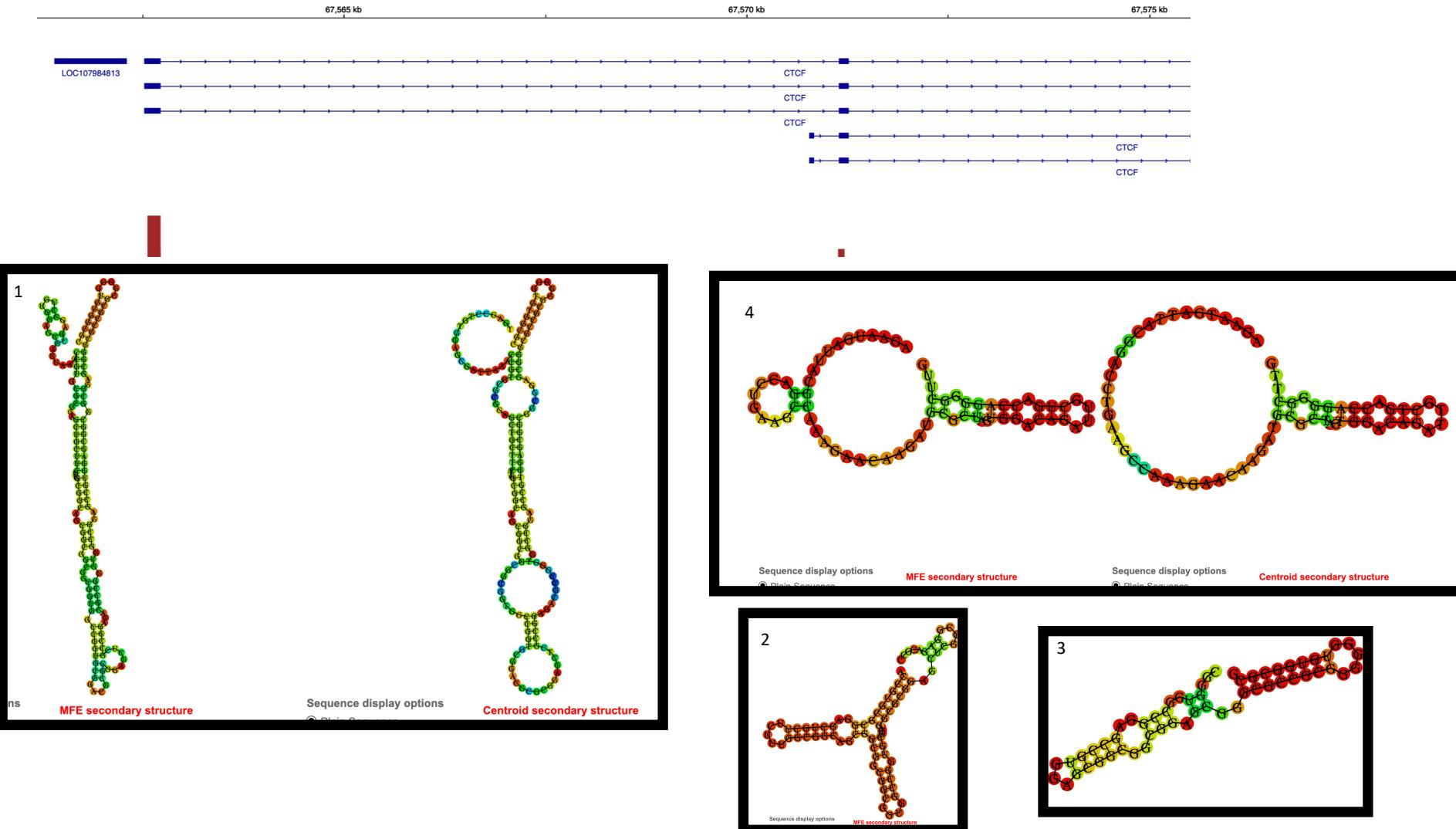
I am eager to hear what you think of all this, and what I should do next.

(I will do a quick survey to find out how common eCLIP K562 DDX3X binding sites are across the entire genome. If 6/7 of all genes have such sites, then these results will not be so interesting.)

CTCF

chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr16	67562565	67562720	156	+	DDX3X	eCLIP	K562	ENCODE 23.994476	CTCF	hg38_genes_promoters
2	chr16	67562586	67562665	80	+	DDX3X	eCLIP	K562	ENCODE 7.839952	CTCF	hg38_genes_promoters
3	chr16	67562665	67562721	57	+	DDX3X	eCLIP	K562	ENCODE 5.006577	CTCF	hg38_genes_5UTRs
4	chr16	67571145	67571213	69	+	DDX3X	eCLIP	K562	ENCODE 4.1668016	CTCF	hg38_genes_introns

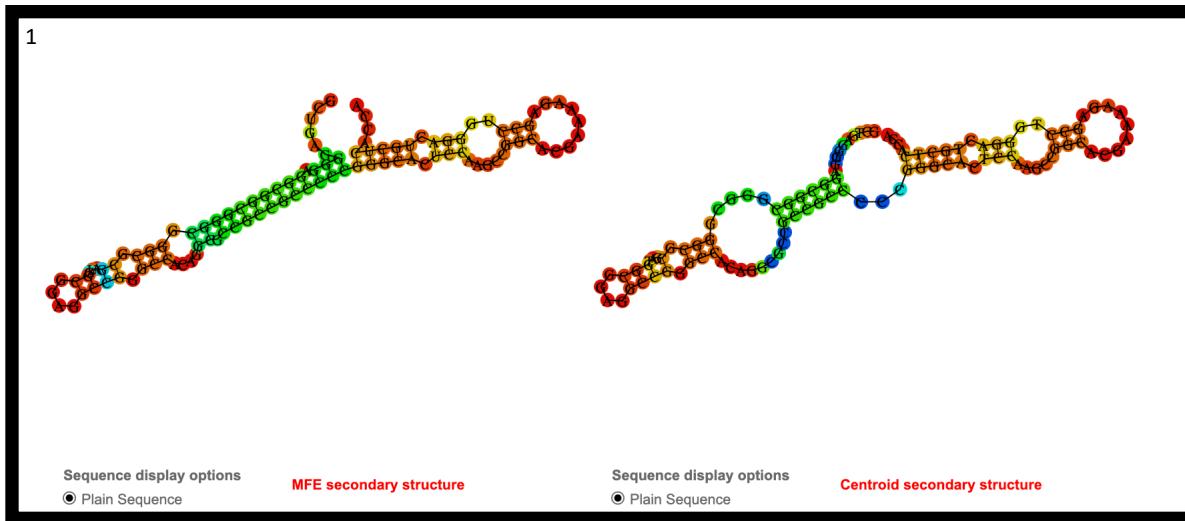
Binding sites 2 and 3 overlap with long higher-scoring binding site 1



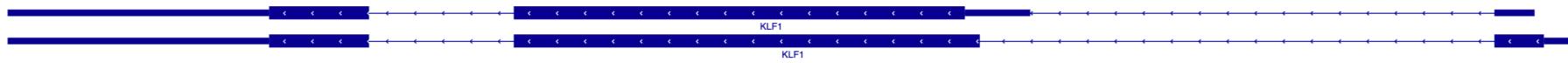
E2F4



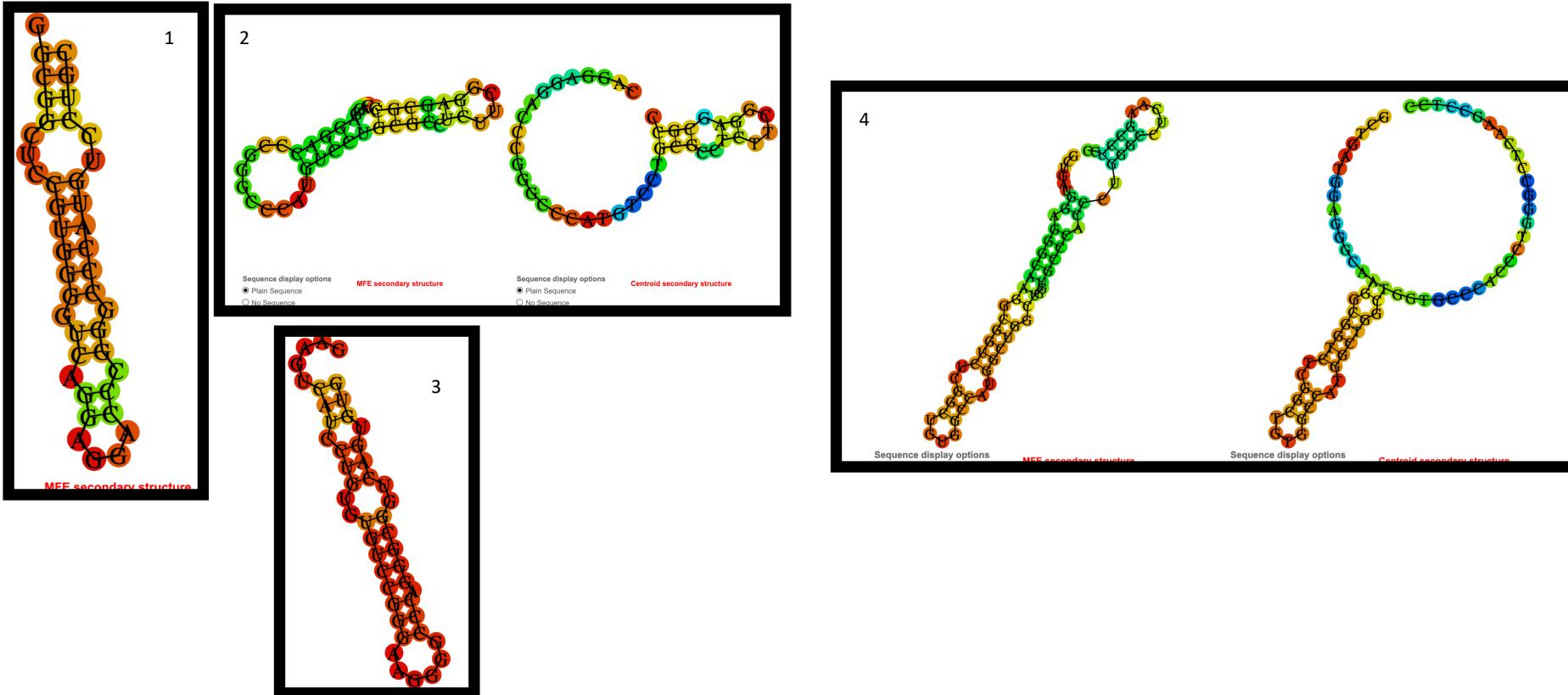
chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr16	67192201	67192306	106	+ DDX3X	eCLIP	K562	ENCODE 31.31813	E2F4	hg38_genes_1to5kb	



KLF1

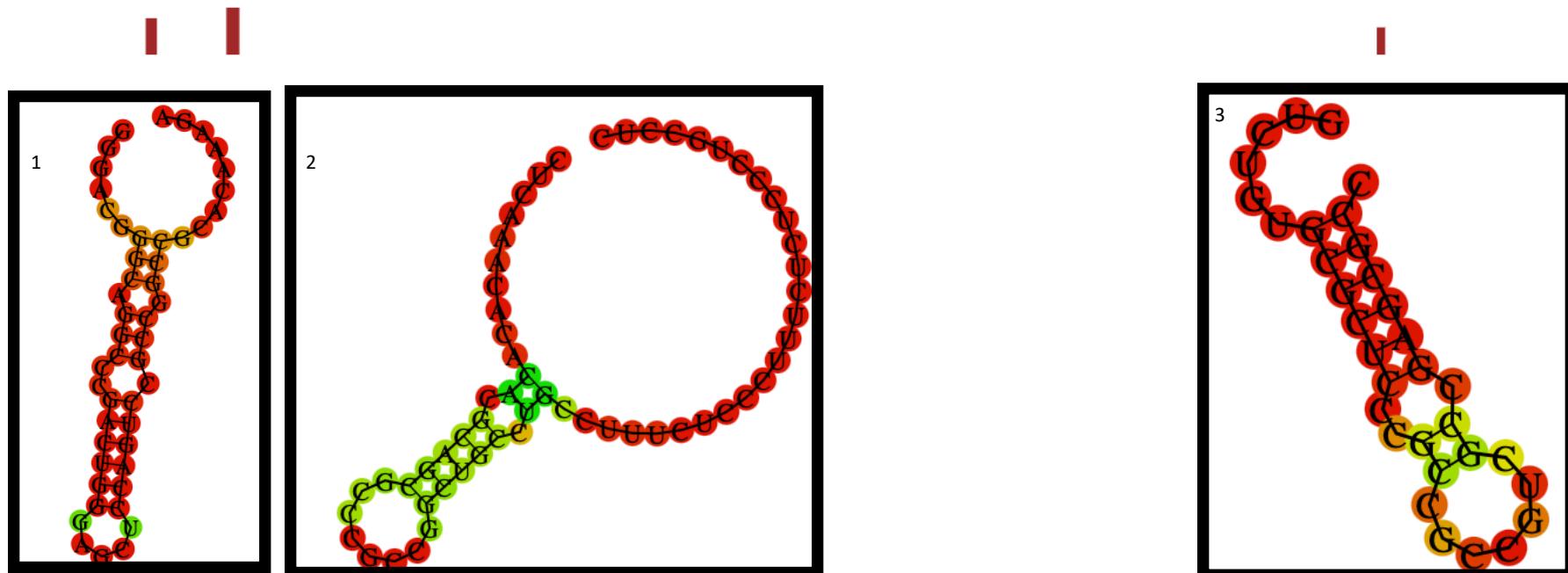
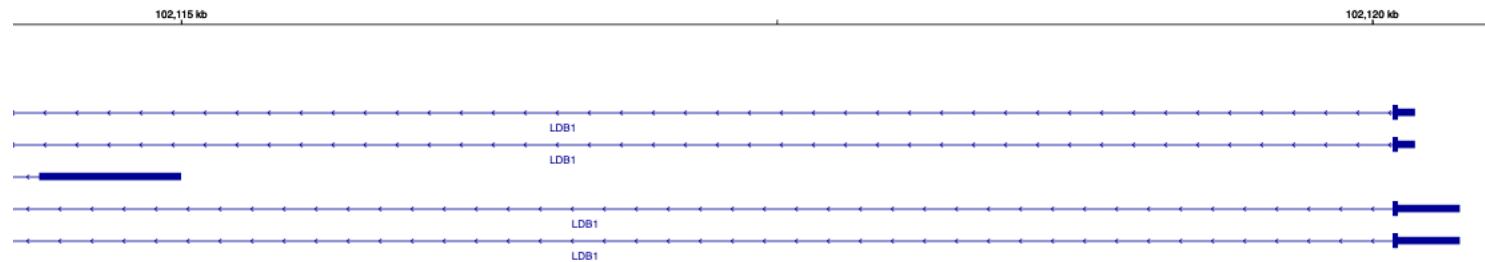


chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature	
1	chr19	12886081	12886122	42	-	DDX3X	eCLIP	K562	ENCODE	4.610510	KLF1	hg38_genes_exons
2	chr19	12886097	12886138	42	-	DDX3X	eCLIP	K562	ENCODE	10.145122	KLF1	hg38_genes_exons
3	chr19	12887060	12887105	46	-	DDX3X	eCLIP	K562	ENCODE	25.962056	KLF1	hg38_genes_exons
4	chr19	12887105	12887175	71	-	DDX3X	eCLIP	K562	ENCODE	61.814090	KLF1	hg38_genes_exons



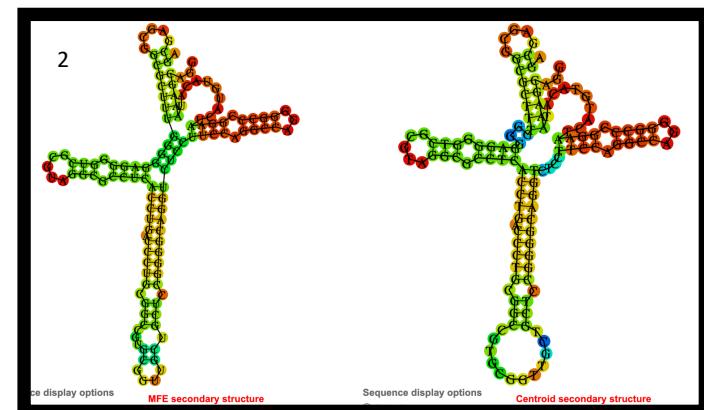
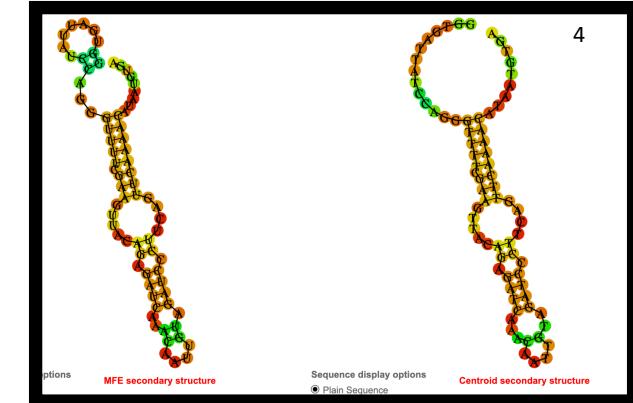
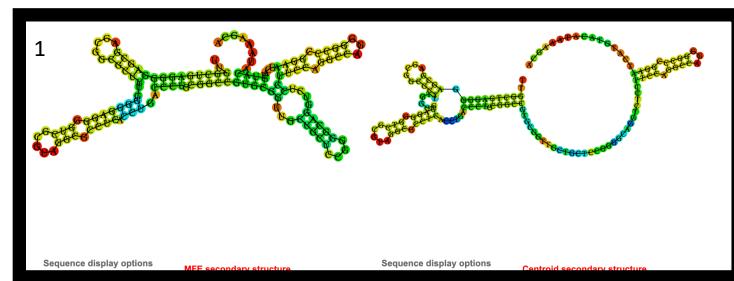
LDB1

chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr10	102114520	102114567	48	-	DDX3X	eCLIP	K562	ENCODE	6.003685	LDB1 hg38_genes_introns
2	chr10	102120403	102120438	36	-	DDX3X	eCLIP	K562	ENCODE	4.927915	LDB1 hg38_genes_exons
3	chr10	102114902	102114963	62	-	DDX3X	eCLIP	K562	ENCODE	15.796315	LDB1 hg38_genes_introns



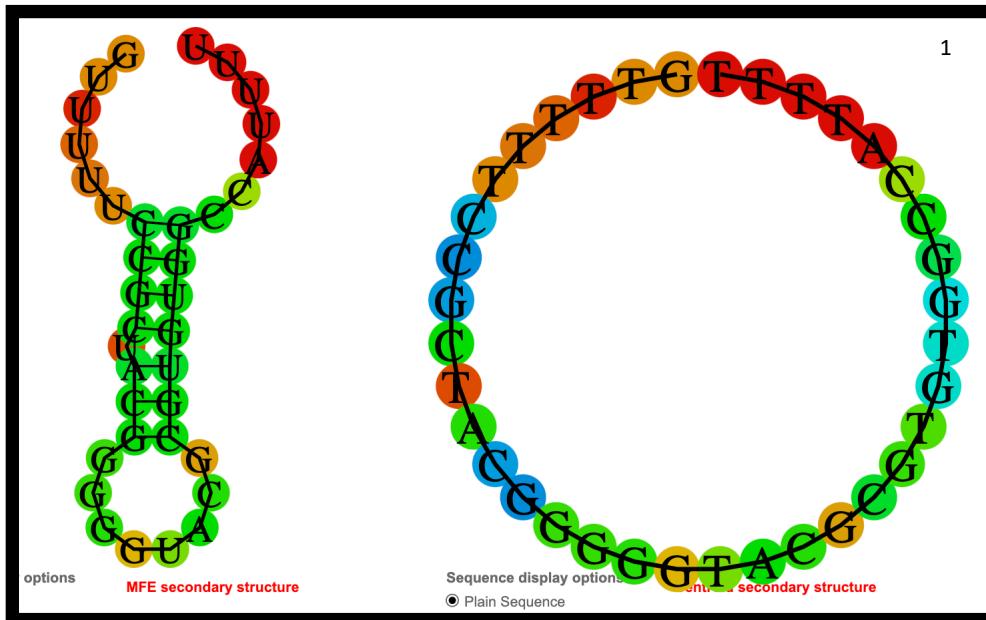
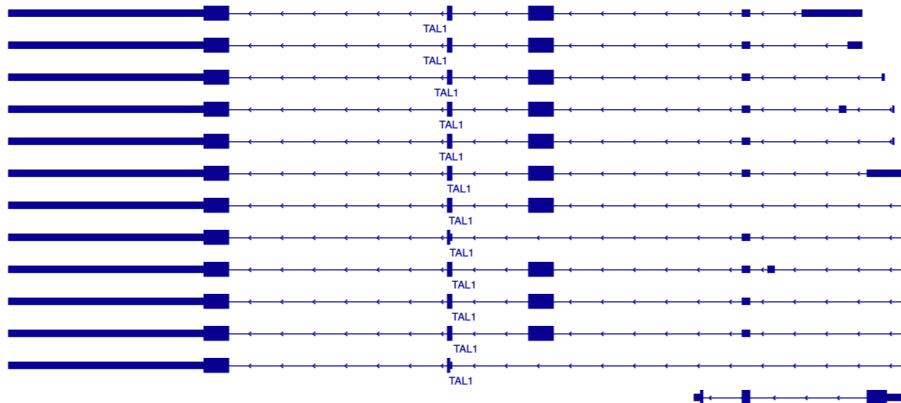
SMC3

chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr10	110567697	110567830	134	+	DDX3X	eCLIP	K562	ENCODE 44.68689	SMC3	hg38_genes_promoters
2	chr10	110567709	110567832	124	+	DDX3X	eCLIP	K562	ENCODE 29.39116	SMC3	hg38_genes_promoters
3	chr10	110568937	110569013	77	+	DDX3X	eCLIP	K562	ENCODE 10.61689	SMC3	hg38_genes_introns
4	chr10	110568937	110569010	74	+	DDX3X	eCLIP	K562	ENCODE 8.09189	SMC3	hg38_genes_introns



TAL1

chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr1	47232167	47232202	36	-	DDX3X	eCLIP	K562	ENCODE 6.323707	TAL1	hg38_genes_1to5kb

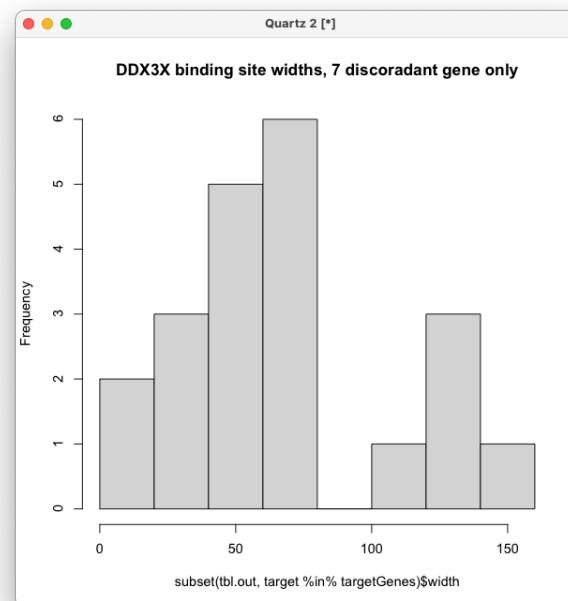
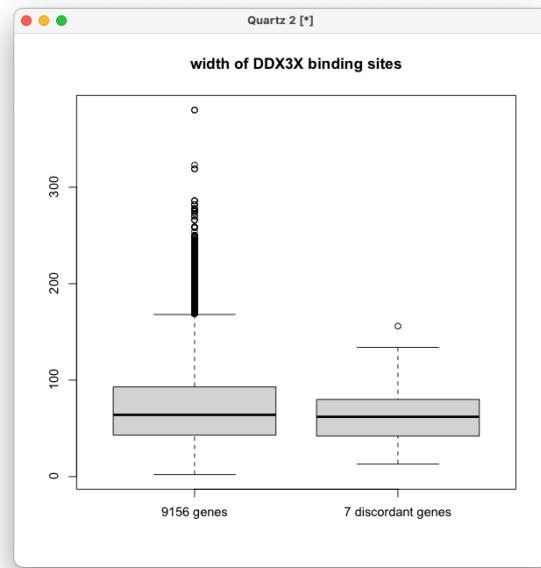


ZBTB7A

No K562 DDX3X binding sites

Jeff asked: It would be interesting to know the frequency of long DDX3X binding sites in the K562 eCLIP data. And the frequency of long DDX3X sites in 5'UTRs in the K562 eCLIP data

One combined box plot, then two histograms of the same data. I'll have to do a bit more work to answer the second question, regarding 5'UTRs, but FWIW, most of the sites I have seen are, I thin, in (or very near) the 5'UTR. Isn't that what we seen in the genome views I sent earlier today? I might be confused on this point however.



DDX3X has binding sites on 9125 genes out of 16168 reported in K562 cells: 56%
The top 7 discordant rna/srm genes had K562 binding for 6, or 86%. A more rigorous enrichment analysis is surely needed.

Some background: DDX3X (you may know all this and more) is a DEAD-box helicase which regulates RNA processing and metabolism by unwinding short double-stranded (ds) RNAs. Sharing a helicase core composed of two RecA-like domains (D1D2), DDXs function in an ATP-dependent, non-processive manner. As an attractive target for cancer and AIDS treatment, DDX3X and its orthologs are extensively studied, yielding a wealth of biochemical and biophysical data, including structures of apo-D1D2 and post-unwound D1D2:single-stranded RNA complex, and the structure of a D2:dsRNA complex that is thought to represent a pre-unwound state [song & ji, nature 2019] But probably more relevant here is recently recognized DDX3X pleiotropy: Multifunctional RNA-binding proteins influence mRNA abundance and translational efficiency of distinct sets of target genes: <https://www.biorxiv.org/content/10.1101/2021.04.13.439465v2>

...This led us to denote these RBPs as “multifunctional RBPs” – context-specific RBPs whose functional outcome depends on the set of mRNAs it targets. A key example appears to be the multifunctional RBP DDX3X (Mo et al., 2021; Soto-Rifo et al., 2012), whose abundance correlates significantly with the mRNA levels of 339 target genes ($\text{padj} = 2.83 \times 10^{-5}$; Glass' $\Delta = 6.9$) and the translational efficiency of 730 target genes ($\text{padj} = 5.25 \times 10^{-5}$; Glass' $\Delta = 11.89$), of which only 43 targets overlap between both sets (Figure 3A and 3C). The consequences of DDX3X binding for mRNA abundance (positive correlation) or TE (negative correlation) are opposite, though this is not the case for all multifunctional RBPs (Figure S3B). [more from this paper on the next slide]

I am not sure if we can find a signature of non-canonical function (that is, other than dsDNA unwinding) from the eCLIP data. But for you to critique, and probably discard, the results I show only include long binding sites for DDX3X. This may turn out to be a poor idea. FWIW, most of the reported target genes have both short (2bp) and long ($\geq 20\text{bp}$) binding sites, or none at all.

More notes on the multiple functions of DDX3X, mention of 5'UTR structure

From Multifunctional RNA-binding proteins influence mRNA abundance and translational efficiency of distinct sets of target genes (bioRxiv, July 2021)

Of these, 21 impacted both RNA expression and translation, albeit for virtually independent sets of target genes. We highlight a subset of these, including G3BP1, PUM1, UCHL5, and DDX3X, where dual regulation is achieved through differential affinity for target length, by which separate biological processes are controlled. Similar to the RNA helicase DDX3X, the known splicing factors EFTUD2 and PRPF8 - all identified as multifunctional RBPs by our analysis - selectively influence target translation rates depending on 5' UTR structure.

More notes on the multiple functions of DDX3X, mention of 5'UTR structure, email discussion with Marjorie and Jeff (2 nov 21)

marjorie:

I think it is very possible that DDX3X would repress translation through binding structured 5'UTR.

For example, GATA1 has been proposed to be translated more efficiently because it has a highly structured 5'UTR.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4087046/>

Specifically, the paper says:

The 5' end of human *GATA1* mRNA is predicted to be highly structured ([Supplementary Fig. 13c](#)) and thus translation may be more readily impaired in settings where the translation initiation potential was reduced, such as with ribosomal protein haploinsufficiency.[23,29,30](#) In support of this, we found that the *GATA1* 5' UTR restricted translation to a greater extent than other endogenous 5' UTRs of similar length using a reporter assay ([Supplementary Fig. 13d](#)). Additionally, other transcripts with highly structured 5' UTRs showed reduced association with larger polysomes ([Supplementary Fig. 14](#)), while 5' UTRs that were unstructured were found in the genes that did not change or were upregulated in these polysomes

GATA1 shows the expected discrepancy (RNA up, protein down). So it could be regulated by DDX3X too.

I think the highly structured 5'UTR of *GATA1* mRNA could explain both its increased translation (increased binding of ribosomes through an unknown mechanism) and decreased translation (if bound by DDX3X).

jeff:

Biochemical evidence that DDX3Xs can function as a translational repressor comes from the Shio, 2008 Oncogene paper. In this paper an interaction between DDX3X and eIF4E was required for repression. eIF4E interacts with the mRNA cap so it is reasonable to think that DDX3X would also localize to the 5' end of mRNA when it interacts with eIF4E.

Another possibility is that DDX3X induces formation of stress granules that sequester translation factors. In this case DDX3X interaction with RNA might not be necessary for repression.

paul:

I have annotated and have scores for all the DDX3X binding sites in K562 and HepG2 cells. Filtering by width > 5, and eCLIPs score > 5, we have:

K562: 5710 binding sites in the 5'UTRs of 3327 genes

HepG2: 11945 binding sites in the 5'UTRs of 5711 genes

These numbers do not change much if the minimum binding site width is increased to 20.

Next up? Find motifs & structures shared across these binding sites?

These tasks emerged from our Monday (11 October 2021) meeting.

Any suggestions, corrections, other tasks?

- 1) DDX3X has binding sites on 9125/16168 genes reported in K562 cells: 56%.
For the top discordant genes, 6/7 have DDX3X binding sites: 86%

Jeff suggests I survey all 100 of the srm/rna-seq genes. Does binding site frequency fall off as discordance does?

- 2) Is there a clearcut DDX3X 5' motif?

- 3) interpolate the iTRAQ data to give it the same timepoints as srm & rna-seq. Compare itraq and srm.
Is the itraq useful? - by which we mean: does it correlate well with the srm?

- 4) What other cell types do we have for DDX3X?

For DDX3X and (for instance) CTCF as a target, we have these counts:

HEK293	HEK293T	HepG2	K562
151	1	5	4
93.8%	0.6%	3.1%	2.5%

For DDX3X with any of 18098 genes:

HEK293	HEK293T	HepG2	K562
1596651	2697	48671	35990
94.8%	0.2%	2.9%	2.1%

--- jeff replies to email

for 1) I think it would be interesting to see how the number of targets (as opposed to binding sites, perhaps this what you meant) varies with discordance.

Also for 1) it may be that there is not a strong relationship between discordance and targets. For example it is possible that DDX3X expression correlates with target protein expression but not so well with target RNA expression. Can we also look at target numbers in relation to the strength of DDX3x correlation with target protein expression? I think you mentioned that you already divided our genes into 3 classes based on discordance.

I'm not sure about using HEK293 as the number of targets seems quite different compared to HepG2 and K562. Also in the paper that you referenced

<https://www.biorxiv.org/content/10.1101/2021.04.13.439465v2.full>

they only used HepG2 and K562 for some reason.

From Jeff: I noticed that they processed the eCLIP data as described here:

Processed eCLIP data of 150 RBPs were obtained from ENCODE (Davis et al., 2018) for HepG2 ($n = 103$) and K562 ($n = 120$) cell lines. Datasets consisted of BED files containing eCLIP peaks and BAM files containing reads mapped to the human genome (GRCh38.p10/hg38). The identification of robust eCLIP peaks across replicates and cell lines was performed as suggested by Van Nostrand and colleagues (Van Nostrand et al., 2020). First, we used BEDTools (Quinlan & Hall, 2010) to quantify the coverage of each predicted peak using input (mock) and immunoprecipitation (IP, antibody against RBP) BAM files. Next, for each peak, the relative information content was defined as $\pi_i \times \log_2(\pi_i/\pi_q)$, where π and q are the sums of reads mapping to the peak in IP and negative control respectively. The information content was used to calculate the Irreproducible Discovery Rate (IDR) (Li et al., 2011), a parameter indicating reproducible peaks across biological replicates. A significant and reproducible peak was defined meeting an IDR cut-off < 0.01 , $p\text{-value} \leq 10^{-5}$ and fold-enrichment (FC) > 8 . In case two or more peaks overlapped the same genomic region, the most significant one was included in the peak table. Additionally, non-overlapping peaks were pooled into a single table, in order to get a complete set in both cell lines. While CLIP data was produced in a non-cardiac setting, CLIP signals are usually preserved among similarly expressed genes of the same RBP independent of the cell line, with peak differences instead reflecting cell type-specific expression rather than binding affinity (Van Nostrand et al., 2020). Additionally, for the muscle-specific splicing repressor RBM20, which was not part of the ENCODE dataset but included for its importance for cardiac splicing and heart disease (Maatz et al., 2014; Guo et al., 2012), significant rat RBM20 HITS-CLIP targets were obtained from Maatz et al. (Maatz et al., 2014) and converted to GRCh38.p10/hg38 genomic coordinates. Only 143 RBPs with expression in human heart tissue were kept (mean FPKM across samples > 1 ; 142 ENCODE RBPs and RBM20). Overall, we retrieved an average of 4,300 eCLIP-seq peaks per experiment. Finally, we mapped these peaks to the annotated transcriptome (Ensembl v.87) and, for each RBP experiment, all the genes supported by at least one CLIP-seq peak were defined as putative target genes.

Would it be worthwhile following this methodology?

Does DDX3X binding site frequency fall off as discordance does?

```
hagfish: ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants/ddx3x.vs.discordance.100.genes.R
function build.discordanceAndHitsTable
> tbl.discordance <- calculate.discordance()
> head(tbl.discordance)
  gene protein cor.early cor.late cor.all corDelta
1 TCF3   TCF3p    0.83   -0.86   0.09   -1.69
2 HLTF   HLTFp    0.89   -0.68   0.07   -1.56
3 E2F4   E2F4p    0.89   -0.64   0.25   -1.53
4 CTCF   CTCFp    0.71   -0.79   0.03   -1.50
5 KLF1   KLF1p    0.83   -0.64   0.49   -1.47
6 E2F8   E2F8p    0.77   -0.68   0.24   -1.45
> tbl.hits <- get.ddx3x.binding.sites(tbl.discordances)
> head(tbl.hits)
  chrom start endpos width strand gene method celltype accession score target targetfeature
1 chr21 29298873 29298955 83 + DDX3X eCLIP K562 ENCODE 13.414056 BACH1 hg38_genes_introns
2 chr21 29321220 29321310 91 + DDX3X eCLIP K562 ENCODE 13.700035 BACH1 hg38_genes_introns
3 chr21 29298920 29298954 35 + DDX3X eCLIP K562 ENCODE 3.801192 BACH1 hg38_genes_introns
4 chr20 50190841 50190940 100 + DDX3X eCLIP K562 ENCODE 13.768770 CEBPB hg38_genes_exons
5 chr20 50190847 50190902 56 + DDX3X eCLIP K562 ENCODE 5.641245 CEBPB hg38_genes_exons
6 chr5 98928732 98928793 62 - DDX3X eCLIP K562 ENCODE 7.700012 CHD1 hg38_genes_1to5kb
combined:
head(tbl.discordance)
  gene protein cor.early cor.late cor.all corDelta ddx3x.hits
1 TCF3   TCF3p    0.83   -0.86   0.09   -1.69      5
2 HLTF   HLTFp    0.89   -0.68   0.07   -1.56      4
3 E2F4   E2F4p    0.89   -0.64   0.25   -1.53      2
4 CTCF   CTCFp    0.71   -0.79   0.03   -1.50      4
5 KLF1   KLF1p    0.83   -0.64   0.49   -1.47      7
6 E2F8   E2F8p    0.77   -0.68   0.24   -1.45      3
# flip the sign of corDelta so that loss of correlation is positive. does this measure then correlate with DDX3X eCLIP hits?
with(tbl.discordance, cor((-1 * corDelta), ddx3x.hits, method="spearman", use="pairwise.complete"))
[1] 0.08152603
```

Jeff suggests: use the late time protein-RNA correlation for the DDX3X binding site correlation. It could be that there are different regulatory mechanisms at play during early and late time points and DDX3X becomes important only at later times.

```
with(tbl.discordance, cor(cor.late, ddx3x.hits, use="pairwise.complete"), method="spearman") # [1] 0.01495144
with(tbl.discordance, cor(cor.early, ddx3x.hits, use="pairwise.complete"), method="spearman") # [1] 0.1151263
with(tbl.discordance, cor(cor.all, ddx3x.hits, use="pairwise.complete"), method="spearman") # [1] 0.0443076
with(tbl.discordance, cor(cor.late, hit.01, use="pairwise.complete")) # [1] -0.01905391
with(tbl.discordance, cor(corDelta, hit.01, use="pairwise.complete")) # [1] -0.0124624 "hit.01" is no/any hits
```

Jeff's suggestions upon seeing results on previous slide (10 oct 2021, 1051a)

For the question about whether late time course falling mRNA/srm correlation, for each gene/protein pair, corresponds to the number of eCLIP binding sites on the gene, it looks like you've used a corDelta between early and late RNA-protein correlation for the DDX3X binding suite correlation. I think it would be interesting to just use the late time protein-RNA correlation for the DDX3x binding site correlation. It could be that there are different regulatory mechanisms at play during early and late time points and DDX3X becomes important only at later times.

Also maybe the number of DDX3X sets is not the key metric. How about a metric that considers 1) if an eCLIP site is present and 2) the correlation between DDX3X expression and target gene protein expression at late time points? Perhaps weighted equally?

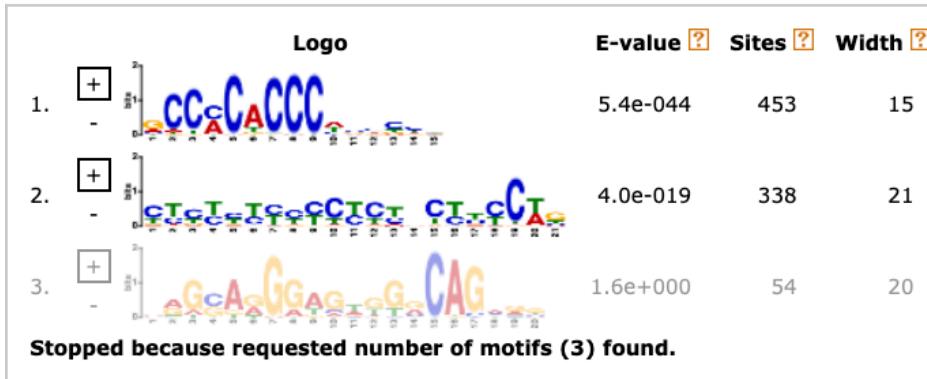
Regarding the eCLIP data processing presented in <https://www.biorxiv.org/content/10.1101/2021.04.13.439465v2.full> do you think this processing step is worth trying?

Is there a clearcut DDX3X 5' motif?

By example, copied off the meme website, reproduced on hagfish:

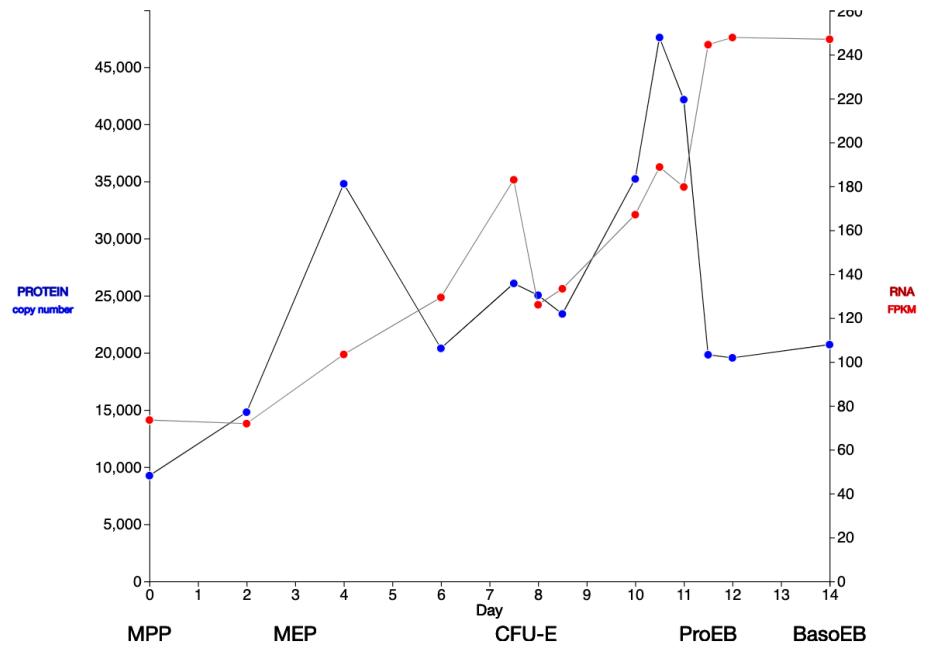
```
meme tests/common/Klf1.fa -dna -oc . -nostatus -time 14400 -mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun classic -revcomp -markov order 0
```

DISCOVERED MOTIFS



Take a look at the input file which produced these results – a standard fasta file

One cherry-picked example: LDB1



gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
MLXIPL	39244.672	2580.296	0.857	0.982	20007835	0	tf	1	LDB1p
VENTX	2222.523	1706.431	0.857	0.947	5456316	0	tf	2	LDB1p
DDX3X	0.000	-11.515	-0.893	-0.888	4633037	0	rbp	3	LDB1p
ETV2	0.000	422.499	0.964	0.885	0	0	tf	4	LDB1p
ELK4	0.000	154.268	0.857	0.877	15124506	0	tf	5	LDB1p
FXR2	0.000	-53.384	-0.929	-0.859	6780227	0	rbp	6	LDB1p

gene	protein	cor.early	cor.late	cor.all	corDelta
LDB1	LDB1p	0.54	-0.61	0.19	-1.15

Notes: hagfish:~/github/TrenaProjectErythropoiesis/explore/rbp/discordants/findMotifs.R

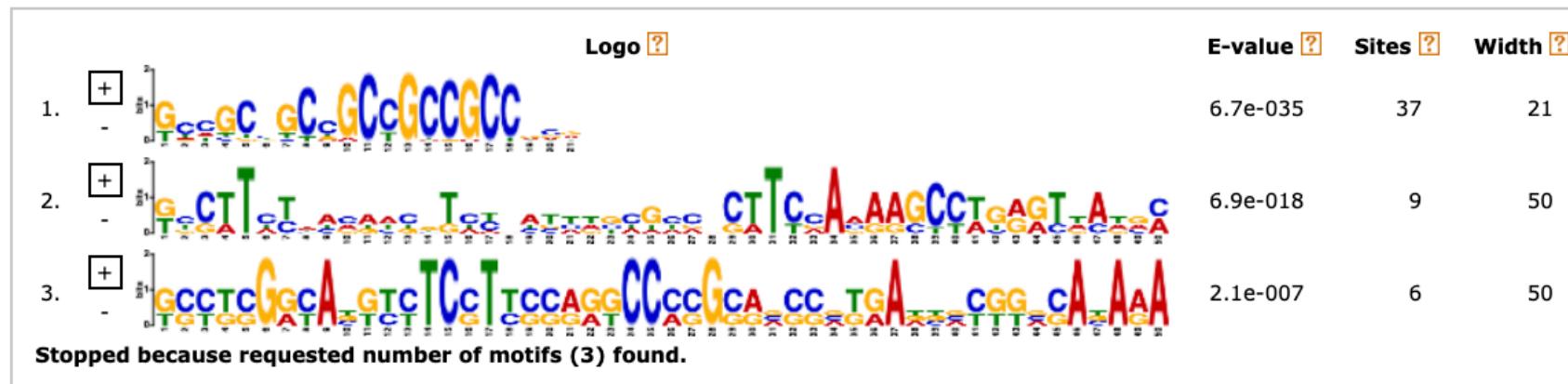
Motif discovery in DDX3X eCLIPs sequences

A few sequences from
targetGenes-97-ddx3x-k562.fa
(220 in all)

```
>KLF1-chr19:12887105-12887175-DDX3X-K562
GCTGATGGAGGCCAAGCCGCTCGGCTGTGGCATGGCTGGTCCCCACCCCTGGGCTCAAGCTCC
>KLF1-chr19:12891165-12891202-DDX3X-K562
CACTGTAGCCTCGGCAGTGAACCGGGAGGTACTACAG
>KLF1-chr19:12891270-12891394-DDX3X-K562
GGTCGCCGCTGGCTCGCTCTGAGAGAGCATGGCCCTGAGAGGCGTCTCCGTGCGGCTGCTGAGCCGCGACCC
GGCTGCACGTCCTCGCACGTGGGCTCGTCCGGCGCAGACC
>KLF1-chr19:12886081-12886122-DDX3X-K562
GGCGGCTCCGTGGGTCAAGGAGGACCCGGGCCATGTCCCTGC
>KLF3-chr4:38664217-38664307-DDX3X-K562
GTCATGTGACTGCCCGAGTTGGTGCAGGAGCCAGAGGGAGGCCAGGGAGCCAGGCCGGAGCCGGGGCCGAGCCG
GACCGCACCGA
>LDB1-chr10:102114520-102114567-DDX3X-K562
GGGACGGGCAGGCCGACTGGGAGCTCCAGTCCGCCGCCGACAAG
>LDB1-chr10:102120403-102120438-DDX3X-K562
GTGTCTGTGCGCTCCCGCCGCGTCGCCGAGCGCCC
>LDB1-chr10:102114902-102114963-DDX3X-K562
ACTCAAACACACACGCGCCCGCCGCTGCCTCTCCCTTCTCCCTGCCTCC
>MAFG-chr17:81927562-81927685-DDX3X-K562
GGGCGGGCCGCCGCGCTCCGAGGGCCGCGGGAGGGACCGCGCGAGGGTCCGGGGCCGGGCTGGAGGACTCG
CCGCCTGCGCGGGCCGGCCGAGCGCACTGGAAAGGCCGGGCCG
```

Todo: rerun meme asking for up to 6 motifs: next 3 are also scored at ~2e-7)

DISCOVERED MOTIFS



pval	leftFlank	sequence	rightFlank	motif	gene	score
1.91e-06	GTGTCTGTGC	GCTCCCGCCGCCGTGCGCGAG		CGCCC motif.01	LDB1	5.718967

Notes from A large-scale binding and functional map of human RNA-binding proteins

Van Nostrand & Yeo, Nature 29 jul 2020

and from Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins.

Genome Biology, Apr6 2020

Both refer to ENCSR456FVU, 2670 files.

From published paper to best DDX3X peaks

"Data Availability" section of the July 2020 Nature paper:

Raw and processed data sets are accessible using accession identifiers provided in [Supplementary Data 2](#) or can be found using the following publication file set accession identifiers at the ENCODE Data Coordination Center (<https://www.encodeproject.org>): eCLIP (ENCSR456FVU), knockdown RNA-seq (HepG2: ENCSR369TWP; K562: ENCSR795JHH; secondary analysis files including DEseq, rMATS, MISO, and CUFFDIFF output: ENCSR413YAF; batch corrected gene expression and splicing analysis: ENCSR870OLK), RBNS (ENCSR876DCD), and ChIP-seq (ENCSR999WIC).

Search ENCODE portal for ENCSR456FVU at [encodeproject.org](https://www.encodeproject.org):

ENCODE: Encyclopedia of DNA Elements

/documents/cd785330-9240-4913-a1ac-bd765b41b223/
Metadata table for the files included in publication data set ENCSR456FVU

/documents/cc918b99-faf8-491d-90fe-1c520dcbb7d9c/
List of file download links for files included in publication data set ENCSR456FVU

Publication file set: A Large-Scale Binding and Functional Map of Human RNA Binding Proteins: eCLIP data, including fastq, bam, peak files and reproducible peak files.
Lab: Gene Yeo, UCSD
Project: ENCODE

107 pages

FileSet
ENCSR456FVU
● released

```
grep -i ddx3x ~/drop/ENCSR456FVU_metadata.tsv | grep peaks | awk '{print $2}' | uniq  
ENCSR930BZL  
ENCSR648LAH  
Searched both of these ids from the encode home page  
Added all files in each to the cart  
Downloaded files.txt to  
~/github/TrenaProjectErythropoiesis/inst/extdata/ddx3x/fromEncode/ENCSR456FVU/  
xargs -L 1 curl -O -J -L < files.txt
```

Reproducible peak files:
DDX3X K562 ENCF667MXK
DDX3X HepG2 ENCF8850WJ

Notes from **A large-scale binding and functional map of human RNA-binding proteins**

Van Nostrand & Yeo, Nature 29 jul 2020

and from Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins.

Genome Biology, Apr6 2020

Both refer to ENCSR456FU, 2670 files.

For about half of the RBPs assayed (37 of 78), we were able to identify highly enriched kmers of five nucleotides (nt) ($k = 5$) that could be clustered into a single motif. The remaining RBPs had more complex patterns of binding, best described by two motifs (32 of 78), or even three or more motifs (9 RBPs).

MEME finds at least 2, maybe 3-6 motifs for DDX3X in sequence for unfiltered POSTAR2 hits for late time discordant genes. Is this the same thing being discussed in the paper?

<https://www.encodeproject.org/publication-data/ENCSR456FVU/>

A Large-Scale Binding and Functional Map of Human RNA Binding Proteins: eCLIP data, including fastq, bam, peak files and reproducible peak files.

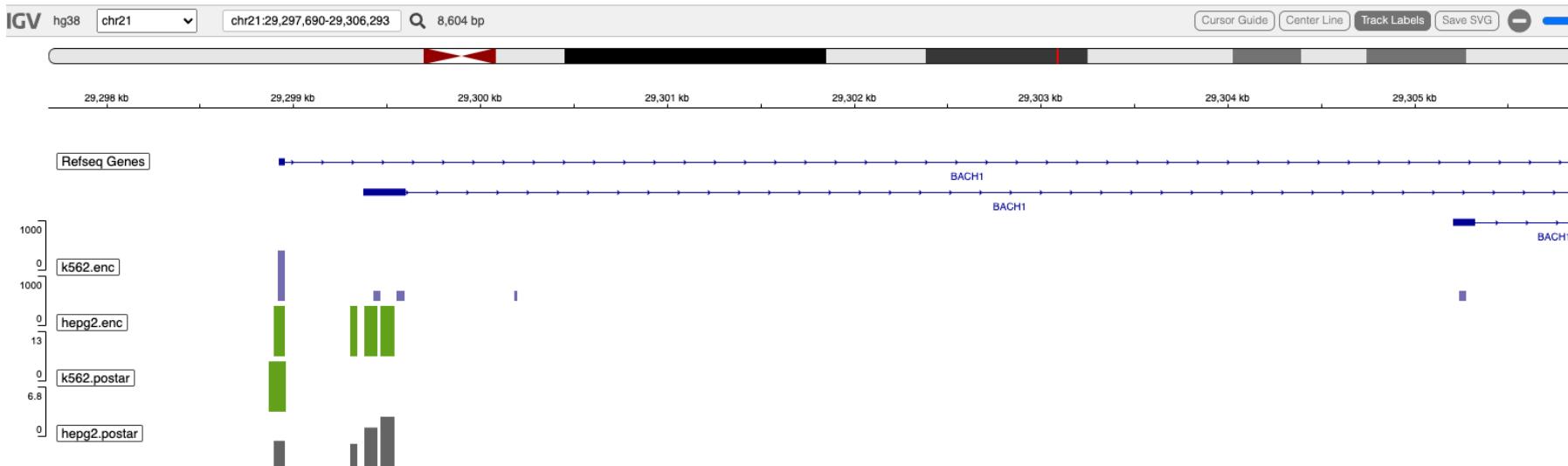
Better link: https://www.encodeproject.org/encore-matrix/?type=Experiment&status=released&internal_tags=ENCORE
shared cart with 4 ddx3x files, eclips and sirna, k562 and HepG2

<https://www.encodeproject.org/carts/ad08e2ec-68e2-4583-bc81-2de0d3125326/>

```
cd ~/github/TrenaProjectErythropoiesis/inst/extdata/ddx3x/fromEncode/  
xargs -L 1 curl -O -J -L < files.txt  
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92175
```

```
12676904 Oct 26 07:30 ENCFF017FYD.bigWig  
13104665 Oct 26 07:30 ENCFF294VBN.bigWig  
2434733 Oct 26 07:30 ENCFF565FNW.bigBed  
1151051 Oct 26 07:30 ENCFF901BYH.bed.gz  
      538 Oct 26 07:25 files.txt  
    3411 Oct 26 07:30 metadata.tsv
```

Trying to figure out the relative merits of POSTAR2 and



Email to Jeff, Marjorie, cory (26 oct 2021) contrasting POSTAR2 with YEO direct from ENCODE, DDX3X binding sites around BACH1

Here is a result for you to study, then give me guidance if you will.

Two sources of data.

First, from Gene Yeo, published in two similar papers:

2019: A large-scale binding and functional map of human RNA-binding proteins, Nature
2020: Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins.
Genome Biology, Apr6 2020

These both point to the same ENCODE data set: [ENCSR456FU](#), which contains 2670 files, 4 for every celltype/RBP pair, with just two cell types K562 and HepG2.

I extracted narrowPeaks from the ENCODE archives from two files (K562 and HepG2)

Second, from POSTAR2, by region (in the vicinity of BACH1's TSS) for the same two cell types.

It looks to me like POSTAR2 used YEO, then filtered out the lower scoring binding sites. And it appears that POSTAR put the scores on a similar but not identical scale: K562 POSTAR scales from 0-14, YEO K562 from 0-7.

From this one example, and from



Consider this also, where a higher-scoring intronic hit in YEO K562 is missing from POSTAR2:



Jeff replies, asking

Thanks Paul. I agree that the sites probably didn't survive because of their quality score or length.

From the POSTAR2 paper:

For the newly collected CLIP-seq datasets, we used the uniform preprocessing pipeline from CLIPdb (7) to preprocess the raw data. Briefly, we first trimmed the adaptor sequences from the raw reads using FASTX-Toolkit package (http://hannonlab.cshl.edu/fastx_toolkit). We only retained reads with quality score above 20 in 80% of their nucleotides. The reads shorter than 13 nt after adaptor trimming were discarded. Finally, we collapsed identical reads to minimize polymerase chain reaction duplicates.

They don't mention whether the reproducibility of the data was used in the evaluation.
Could you check to see if the sites are reproducibly found in the replicate measurements?

My earlier notes include this:

Reproducible peak files:

DDX3X K562 ENCFF667MXK

DDX3X HepG2 ENCFF885OWJ

Where <https://www.encodeproject.org/files/ENCFF667MXK/> has two hg19 peaks files. Examine them perhaps after liftover, in the same BACH1 Region, to see what they suggest.

<https://www.encodeproject.org/search/?searchTerm=ENCFF601TGO>

The IDR (Irreproducible Discovery Rate) framework is a unified approach to measure the reproducibility of findings identified from replicate experiments and provide highly stable thresholds based on reproducibility. Unlike the usual scalar measures of reproducibility, the IDR approach creates a curve, which quantitatively assesses when the findings are no longer consistent across replicates. In layman's terms, the IDR method compares a pair of ranked lists of identifications (such as ChIP-seq peaks). These ranked lists should not be pre-thresholded i.e. they should provide identifications across the entire spectrum of high confidence/enrichment (signal) and low confidence/enrichment (noise). The IDR method then fits the bivariate rank distributions over the replicates in order to separate signal from noise based on a defined confidence of rank consistency and reproducibility of identifications i.e the IDR threshold.

The method was developed by Qunhua Li and Peter Bickel's group and is extensively used by the ENCODE and modENCODE projects and is part of their ChIP-seq guidelines and standards.

Yeo, Nature 2020, A large-scale binding and functional map of human rna-binding proteins - **replicability**

To work towards developing a comprehensive understanding of the binding and function of the human RBP repertoire, we used **five assays** to produce 1,223 replicated data sets for 356 RBPs that participate in diverse aspects of RNA biology and encompass diverse sequence and structural characteristics (Fig. 1a, b, Supplementary Data 1, 2). see https://hbctraining.github.io/Intro-to-ChIPseq/lessons/07_handling-replicates-idr.html

To identify reproducible and significantly enriched peaks across biological replicates, a modified IDR method was used (Supplementary Text, Supplementary Fig. 10). Unless otherwise noted, the final set of reproducible and significant peaks was identified by requiring that the replicate-merged peak meet an IDR cutoff of 0.01 as well as $P \leq 0.001$ and fold enrichment ≥ 8 (using the geometric mean of $\log_2(\text{fold enrichment})$ and $-\log_{10}(P)$ between the two biological replicates). Finally, 57 'blacklist' regions were identified that were common artefacts across multiple data sets and lacked normal peak shapes (manual inspection indicated these often contain either adaptor sequences or tRNA fragments; Supplementary Data 11). IDR peaks that overlapped these blacklist regions were removed to yield the final set of reproducible peaks used in all analyses in this manuscript (unless otherwise indicated) (Supplementary Data 4).

Five assays

eCLIP

RNA Bind-N-Seq

Immunofluorescence

Knockdown rna-seq

RBP ChIP-seq

Sequence, Structure, and Context Preferences of Human RNA Binding Proteins, Mol Cell 2018

Uses High-throughput RNA Bind-n-Seq Assay. **DDX3X** not included in the 78 RBP assessed

RNA-binding proteins (RBPs) orchestrate the production, processing, and function of mRNAs. Here we present the affinity landscapes of 78 human RBPs using an unbiased assay that determines the sequence, structure, and context preferences of these proteins *in vitro* by deep sequencing of bound RNAs. These data enable construction of "RNA maps" of RBP activity without requiring crosslinking-based assays. We found an unexpectedly low diversity of RNA motifs, implying frequent convergence of binding specificity toward a relatively small set of RNA motifs, many with low compositional complexity. Offsetting this trend, however, we find extensive preferences for contextual features distinct from short linear RNA motifs, including spaced "bipartite" motifs, biased flanking nucleotide composition, and bias away from or towards RNA structure. Our results emphasize the importance of contextual features in RNA recognition, which likely enable targeting of distinct subsets of transcripts by different RBPs that recognize the same linear motif.

yeo nature 2020 continued, Assessment and analysis of eCLIP data sets, replicability:

in this study we required peaks to meet stringent criteria of enrichment relative to input (fold enrichment ≥ 8 and $P \leq 0.001$). We further required that significant peaks be reproducibly identified across both biological replicates using an irreproducible discovery rate (IDR) approach (Methods, Supplementary Fig. 10c). Finally, we removed peaks that overlapped with 57 ‘blacklist’ regions (many of which contain either adaptor sequences or tRNA fragments) that show consistent artefactual signal (Supplementary Data 11). Downsampling analysis indicated that peaks were robustly detected at standard sequencing depth even in genes with low expression (transcripts per million (TPM) near or even below 1)

2 nov 2021

Quick test: annotate all DDX3X hits associated with BACH1
K562 and HepG2 cell lines, Yeo data from ENCODE

I have annotated and have scores for all the DDX3X binding sites in K562 and HepG2 cells.
Filtering by width > 5, and eCLIPs score > 5, we have:

K562: 5710 binding sites in the 5'UTRs of 3327 genes

HepG2: 11945 binding sites in the 5'UTRs of 5711 genes

These numbers do not change much if the minimum binding site width is increased to 20.

Next up? Find motifs & structures shared across these binding sites?

1/3 questions from Jeff (4 nov 2021)

- I am thinking about a model where DDX3X represses translation in a way that doesn't necessarily involve binding to 5'UTRs. Trena RNA based models predict 36 DDX3X targets amongst our 100 genes. What are these genes? I wonder if they are enriched for discordant genes?

tbl.trena.top.3

gene	protein	cor.early	cor.late	cor.all	corDelta
CTCF	CTCFp	0.71	-0.79	0.03	-1.50
SMC3	SMC3p	0.49	-0.71	-0.13	-1.20
TAL1	TAL1p	0.49	-0.71	0.14	-1.20
E2F4	E2F4p	0.89	-0.64	0.25	-1.53
KLF1	KLF1p	0.83	-0.64	0.49	-1.47
TRIM33	TRIM33p	0.37	-0.64	0.05	-1.01
LDB1	LDB1p	0.54	-0.61	0.19	-1.15
ZBTB7A	ZBTB7Ap	0.71	-0.61	0.34	-1.32
KLF3	KLF3p	0.09	-0.54	0.26	-0.62
CREBBP	CREBBPp	-0.14	-0.43	-0.31	-0.29
KMT2D	KMT2Dp	-0.71	-0.43	-0.54	0.29
WDHD1	WDHD1p	0.26	-0.36	0.35	-0.61
JUND	JUNDp	-0.43	0.32	0.11	0.75
GATAD2A	GATAD2Ap	0.26	0.54	0.40	0.28
POLR2A	POLR2Ap	0.03	0.61	0.30	0.58
SMARCC1	SMARCC1p	0.37	0.64	0.55	0.27
HMGB3	HMGB3p	0.26	0.75	0.53	0.49
SETDB1	SETDB1p	-0.14	0.75	0.43	0.89
SIRT6	SIRT6p	0.83	0.75	0.80	-0.08
GTF3C2	GTF3C2p	0.14	0.75	0.53	0.61

tbl.trena.top.5

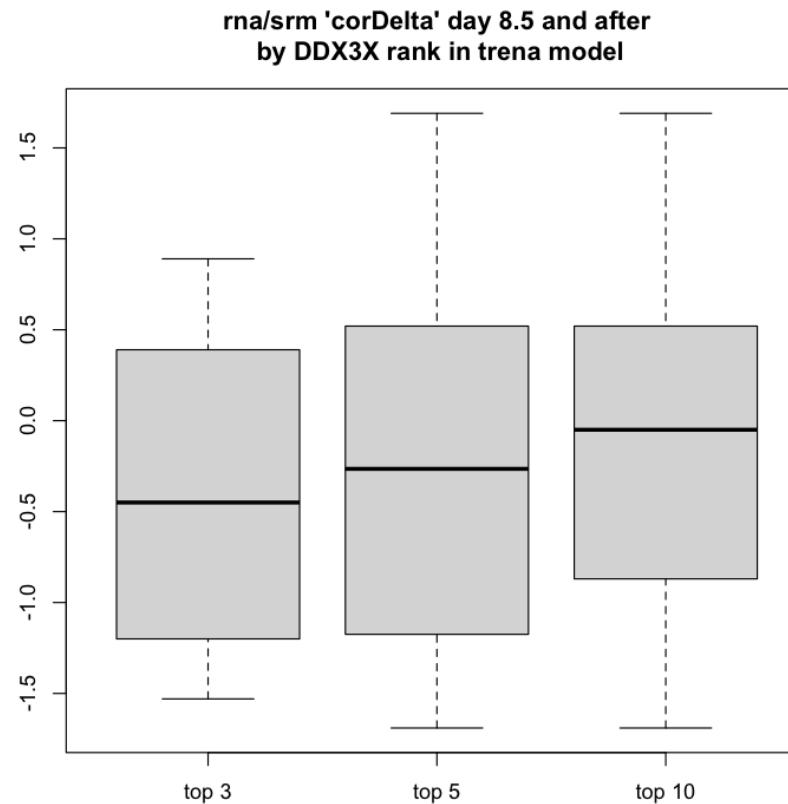
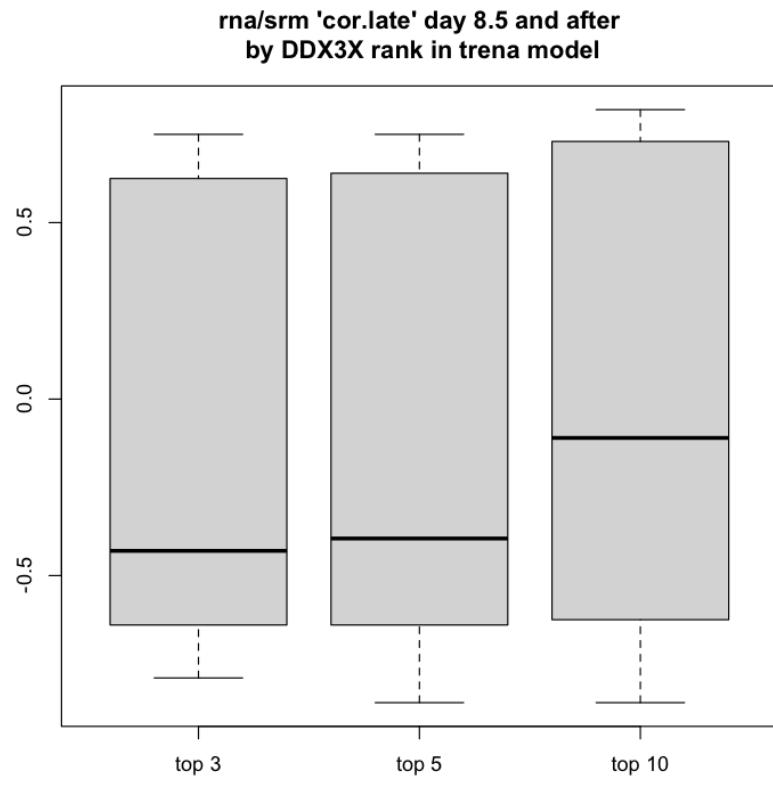
gene	protein	cor.early	cor.late	cor.all	corDelta
TCF3	TCF3p	0.83	-0.86	0.09	-1.69
CTCF	CTCFp	0.71	-0.79	0.03	-1.50
USF1	USF1p	-0.14	-0.75	-0.32	-0.61
SMC3	SMC3p	0.49	-0.71	-0.13	-1.20
TAL1	TAL1p	0.49	-0.71	0.14	-1.20
E2F4	E2F4p	0.89	-0.64	0.25	-1.53
KLF1	KLF1p	0.83	-0.64	0.49	-1.47
TRIM33	TRIM33p	0.37	-0.64	0.05	-1.01
LDB1	LDB1p	0.54	-0.61	0.19	-1.15
KDM6A	KDM6Ap	-0.37	-0.61	-0.38	-0.24
ZBTB7A	ZBTB7Ap	0.71	-0.61	0.34	-1.32
KLF3	KLF3p	0.09	-0.54	0.26	-0.62
CREBBP	CREBBPp	-0.14	-0.43	-0.31	-0.29
KMT2D	KMT2Dp	-0.71	-0.43	-0.54	0.29
WDHD1	WDHD1p	0.26	-0.36	0.35	-0.61
RCOR1	RCOR1p	0.14	0.14	0.25	0.00
POU2F1	POU2F1p	0.77	0.18	0.11	-0.59
JUND	JUNDp	-0.43	0.32	0.11	0.75
GATAD2A	GATAD2Ap	0.26	0.54	0.40	0.28
POLR2A	POLR2Ap	0.03	0.61	0.30	0.58
SAP130	SAP130p	-0.60	0.64	0.08	1.24
SMARCC1	SMARCC1p	0.37	0.64	0.55	0.27
HMGB3	HMGB3p	0.26	0.75	0.53	0.49
NELFE	NELFEP	-0.94	0.75	0.27	1.69
SETDB1	SETDB1p	-0.14	0.75	0.43	0.89
SIN3A	SIN3Ap	0.20	0.75	0.51	0.55
SIRT6	SIRT6p	0.83	0.75	0.80	-0.08
GTF3C2	GTF3C2p	0.14	0.75	0.53	0.61

Some questions from Jeff (4 nov 2021)

- I am thinking about a model where DDX3X represses translation in a way that doesn't necessarily involve binding to 5'UTRs. Trena RNA based models predict 36 DDX3X targets amongst our 100 genes. What are these genes? I wonder if they are enriched for discordant genes?

tbl.trena.top.10					
gene	protein	cor.early	cor.late	cor.all	corDelta
TCF3	TCF3p	0.83	-0.86	0.09	-1.69
CTCF	CTCFp	0.71	-0.79	0.03	-1.50
USF1	USF1p	-0.14	-0.75	-0.32	-0.61
SMC3	SMC3p	0.49	-0.71	-0.13	-1.20
TAL1	TAL1p	0.49	-0.71	0.14	-1.20
E2F4	E2F4p	0.89	-0.64	0.25	-1.53
KLF1	KLF1p	0.83	-0.64	0.49	-1.47
SUZ12	SUZ12p	0.09	-0.64	-0.22	-0.73
TRIM33	TRIM33p	0.37	-0.64	0.05	-1.01
LDB1	LDB1p	0.54	-0.61	0.19	-1.15
NRF1	NRF1p	-0.60	-0.61	-0.37	-0.01
KDM6A	KDM6Ap	-0.37	-0.61	-0.38	-0.24
ZBTB7A	ZBTB7Ap	0.71	-0.61	0.34	-1.32
NR3C1	NR3C1p	0.09	-0.54	-0.23	-0.62
KLF3	KLF3p	0.09	-0.54	0.26	-0.62
CREBBP	CREBBPp	-0.14	-0.43	-0.31	-0.29
KMT2D	KMT2Dp	-0.71	-0.43	-0.54	0.29
WDHD1	WDHD1p	0.26	-0.36	0.35	-0.61
RCOR1	RCOR1p	0.14	0.14	0.25	0.00
POU2F1	POU2F1p	0.77	0.18	0.11	-0.59
JUND	JUNDp	-0.43	0.32	0.11	0.75
GATAD2A	GATAD2Ap	0.26	0.54	0.40	0.28
POLR2A	POLR2Ap	0.03	0.61	0.30	0.58
SAP130	SAP130p	-0.60	0.64	0.08	1.24
SMARCC1	SMARCC1p	0.37	0.64	0.55	0.27
WDR5	WDR5p	0.03	0.68	0.21	0.65
CTBP2	CTBP2p	0.49	0.71	0.62	0.23
CHD4	CHD4p	0.77	0.75	0.70	-0.02
HMGB3	HMGB3p	0.26	0.75	0.53	0.49
NELFE	NELFEP	-0.94	0.75	0.27	1.69
SETDB1	SETDB1p	-0.14	0.75	0.43	0.89
SIN3A	SIN3Ap	0.20	0.75	0.51	0.55
SIRT6	SIRT6p	0.83	0.75	0.80	-0.08
GTF3C2	GTF3C2p	0.14	0.75	0.53	0.61
PARP1	PARP1p	0.66	0.79	0.75	0.13
ELF1	ELF1p	-0.31	0.82	0.41	1.14

Median anti-correlation stronger when DDX3X is in top
3 or top 5 predictors in the trena model

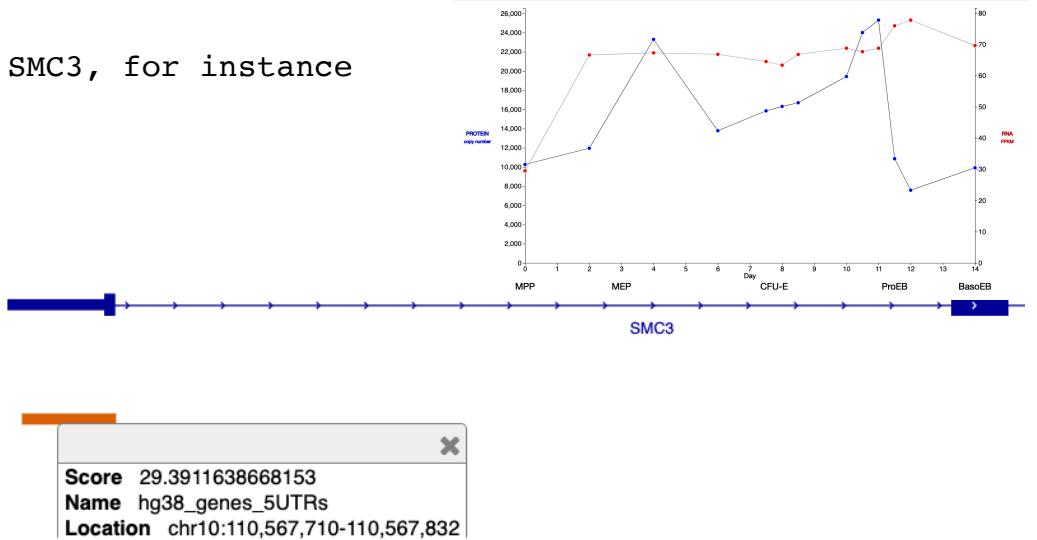


suggesting that higher trena rank for DDX3X (where mRNA predicts target protein level), leads to stronger median negative correlation.

2/3 questions from Jeff (4 nov 2021)

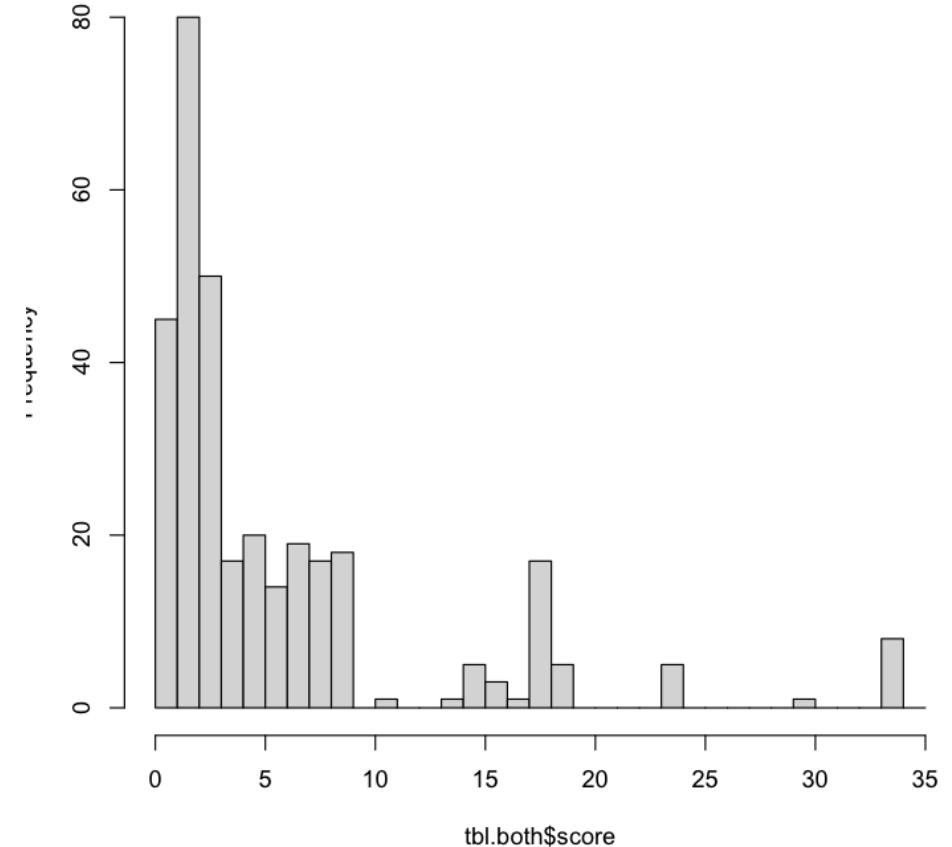
Among the 36 Trena DDX3X targets, how many are DDX3X targets based on eCLIPS data? If some are DDX3X targets by eCLIPs, it would be interesting to know where DDX3X binds along the RNA.

SMC3, for instance



If eCLIPS score is not used to filter, then 34/36 DDX3X-regulated genes identified by trena have 5'UTR binding, in one or another associated transcript.

**DDX3X hits in k562 in 5'UTR, eCLIP score
34/36 genes with DDX3X trena predictors**



On seeing these results, Jeff asks (email, 513p, 4 nov 2021)

Amongst the remaining ~64 SRM genes (for which DDX3X is not a high ranking regulator) how many have 5'UTR DDX3X binding?

The score distribution would be interesting to see as well.
see my answer in the slide

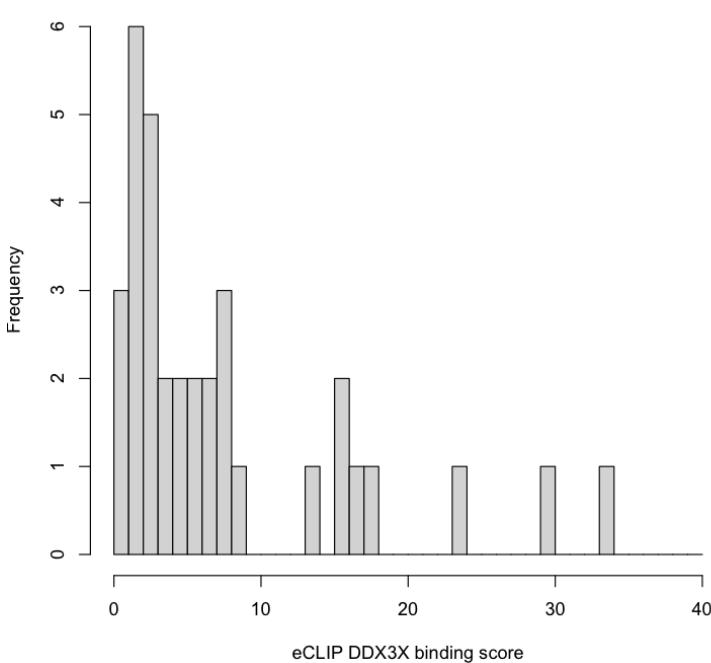
On seeing the results on the previous slide, Jeff asks (email, 513p, 4 nov 2021)

Amongst the remaining ~64 SRM genes (for which DDX3X is not a high ranking regulator) how many have 5'UTR DDX3X binding?

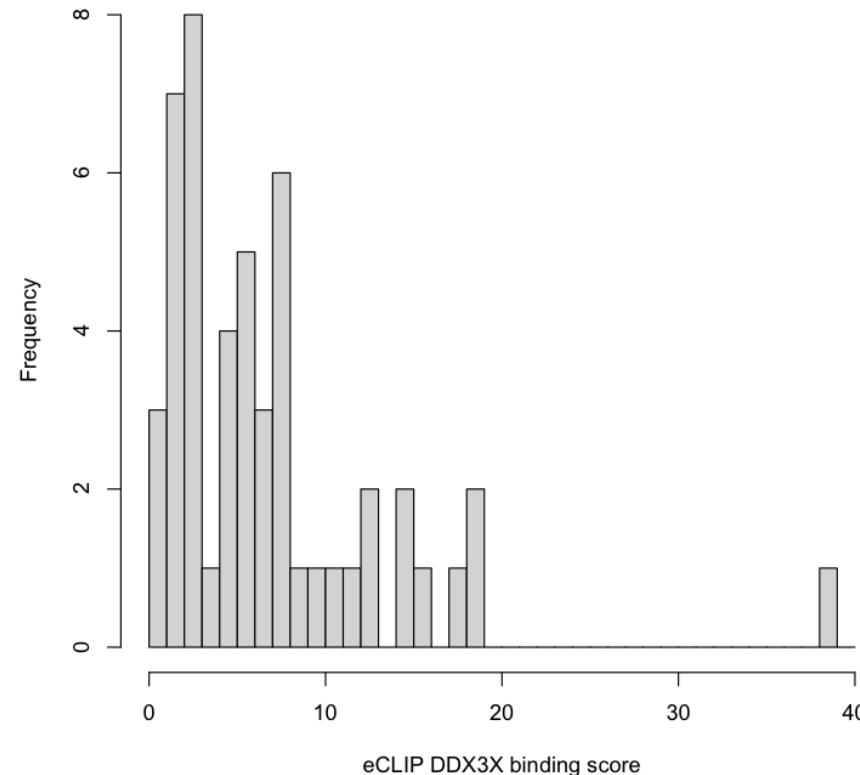
The score distribution would be interesting to see as well.

Conclusion: no significant difference when DDX3X is anywhere in the top 10 trena predictors.

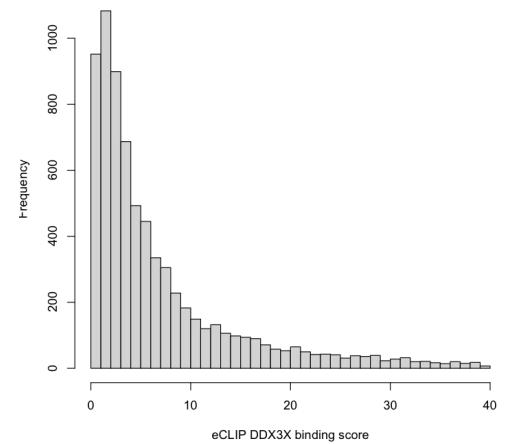
DDX3X hits in k562 in 5'UTR, eCLIP score
34 of 36 genes with DDX3X in top 10 trena predictors



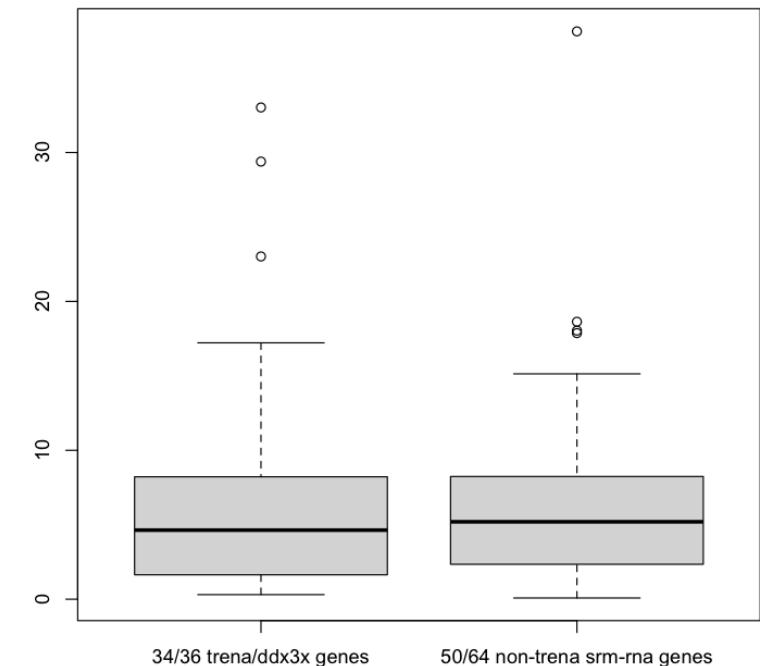
DDX3X hits in k562 in 5'UTR, eCLIP score
50 of 64 genes without trena/DDX3X predictors



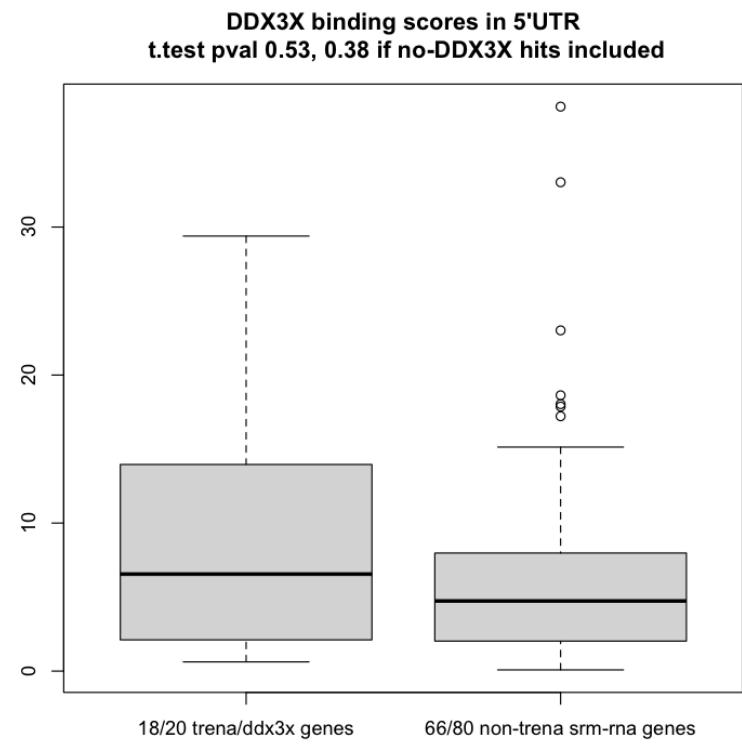
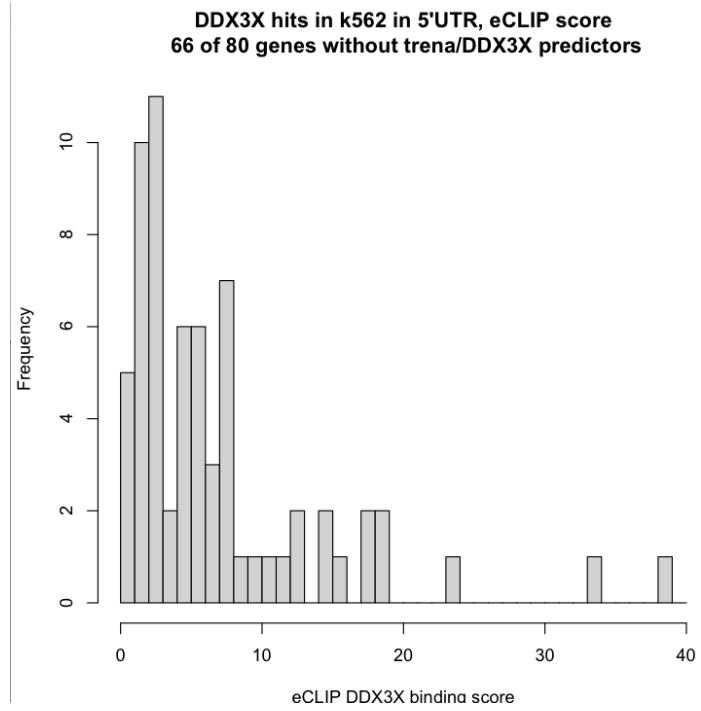
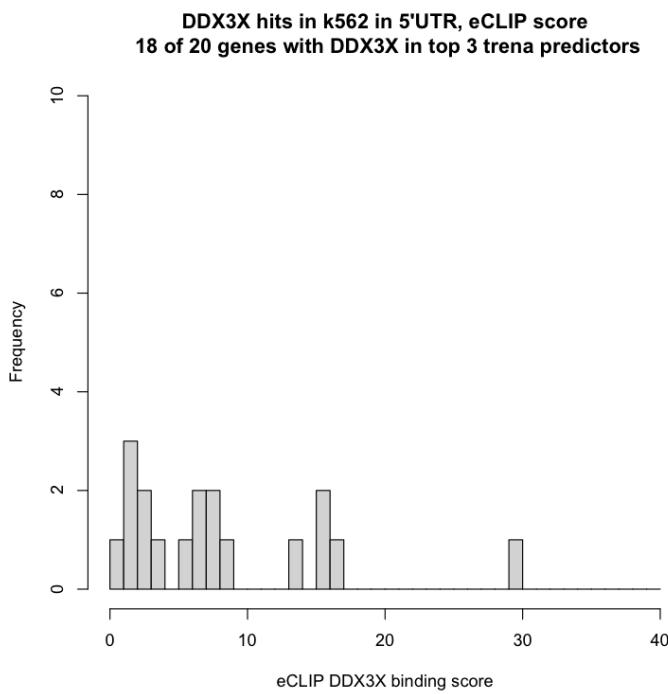
DDX3X hits in k562 in 5'UTR, eCLIP score
best hit in 7540 genes (359 scores > 40 removed)



DDX3X binding scores in 5'UTR
t.test pval 0.689, 0.2 if no-DDX3X hits included



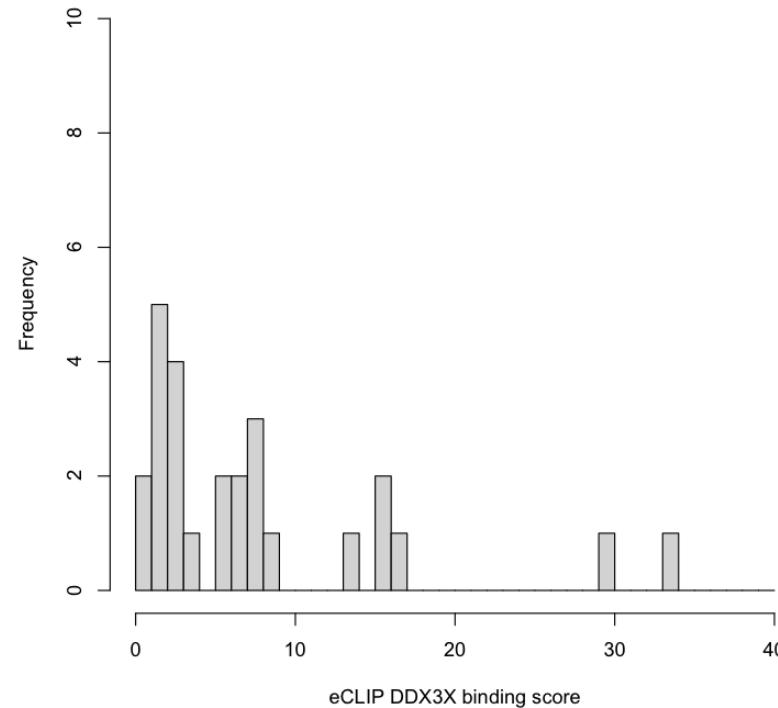
Compare DDX3X 5'UTR binding site scores in genes with and without
DDX3X trena rank <= 3:
no signal



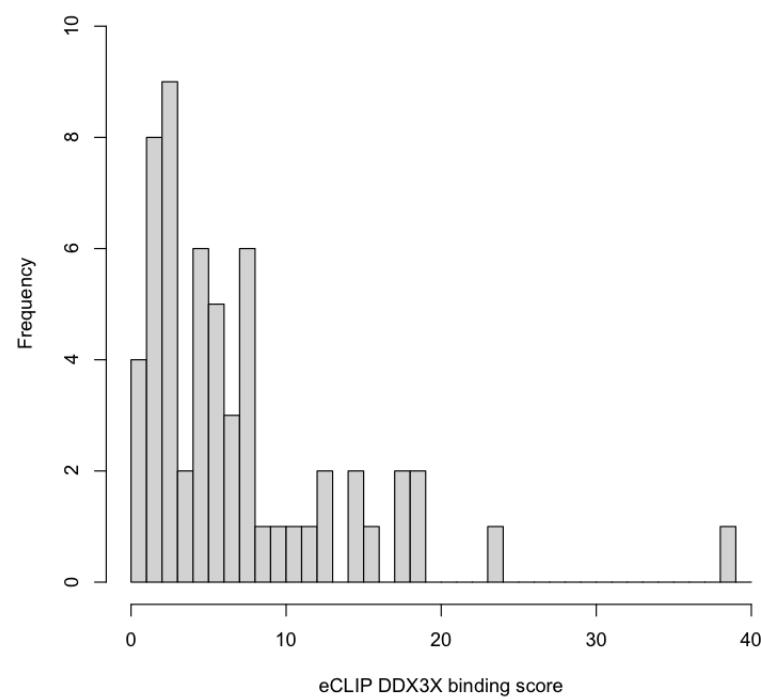
Compare DDX3X 5'UTR binding site scores in genes with and without
 DDX3X trena rank ≤ 5 :
 again no signal

No signal in any of these comparisons but perhaps noteworthy is that 18/20, 26/28 and 34/36 trena-predicted target genes had DDX3X binding sites, but significantly larger percentages of non-trena genes lacked binding sites: 50/64, 58/72 and 66/80 – where all genes are from the 100 rna-seq, srm set. The two trena targets lacking binding sites are **ZBTB7A** and **JUND**.

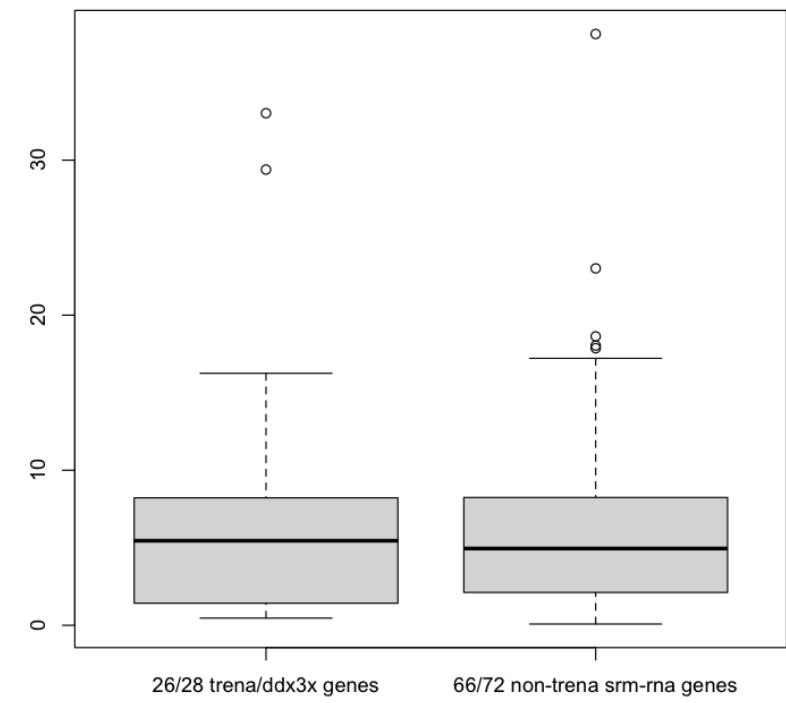
DDX3X hits in k562 in 5'UTR, eCLIP score
 26 of 28 genes with DDX3X in top 5 trena predictors



DDX3X hits in k562 in 5'UTR, eCLIP score
 58 of 72 genes without trena/DDX3X predictors



DDX3X binding scores in 5'UTR
 t.test pval 0.69, 0.39 if no-DDX3X hits included

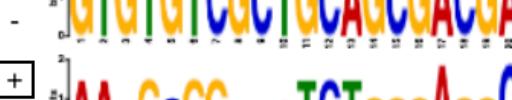


3/3 questions from Jeff (4 nov 2021)

- Can you tell me how DDX3x protein and RNA expression profiles compare during the time course. Or should I ask, if one used DDX3X protein data for modeling, would Trena still predict DDX3X as a top regulator of the current DDX3X targets? Here is the DDX3X protein data:

Gene Symbol	Day0	Day2	Day4	Day6	Day8	Day10	Day11	Day12
DDX3X	44.7	88.4	109	132.5	149.5	182.6	174.8	118.5

meme discovers motifs in K562 5'UTRs in
 7822 DDX3X binding sites in 2105 genes
 where eCLIP score in top quartile of 31k hits in 7524 genes

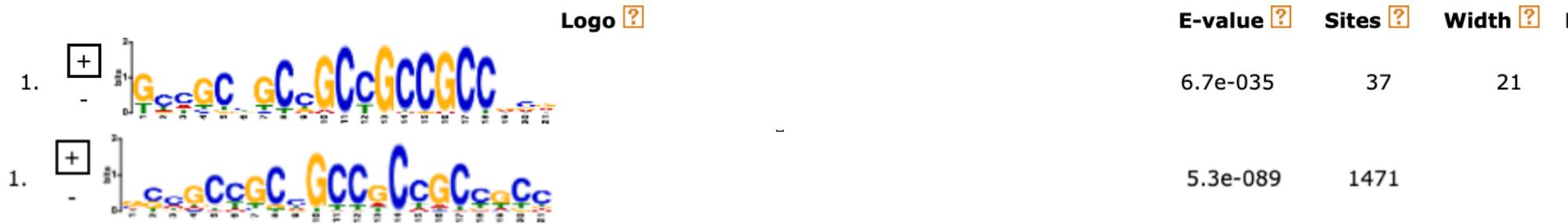
		E-value ?	Sites ?	Width ?
1.	+ - 	5.3e-089	1471	21
2.	+ - 	7.5e-046	13	49
3.	+ - 	1.8e-022	9	50
4.	+ - 	8.7e-014	840	21
5.	+ - 	1.9e-007	10	50
6.	+ - 	9.3e-016	23	50
7.	+ - 	7.3e-002	2	32
8.	+ - 	7.1e-004	8	41
9.	+ - 	2.8e-003	3	48
10.	+ - 	3.2e-005	4	50

todo list from 8 nov 21 zoom meeting

1. using Jeff's latest TMT dataset, which has 1000's of proteins:
 - A. identify those which have day 11-14 drops and those which don't
 - B. count DDX3X binding sites in the corresponding gene 5'UTR
 - C. also count binding sites for BUD13, RBM47, FXR2, probably others, all of which showed up in earlier trena models
2. check DDX3X dominant motif & full eCLIP binding site with Dfam, maybe Arian
3. check DDX3X motif/eCLIP binding sites for eQTLs
4. run new trena models with TMT proteins, predicting rna-seq from proteins, then protein abundance from protein abundance, via tf and RBP DNA binding sites

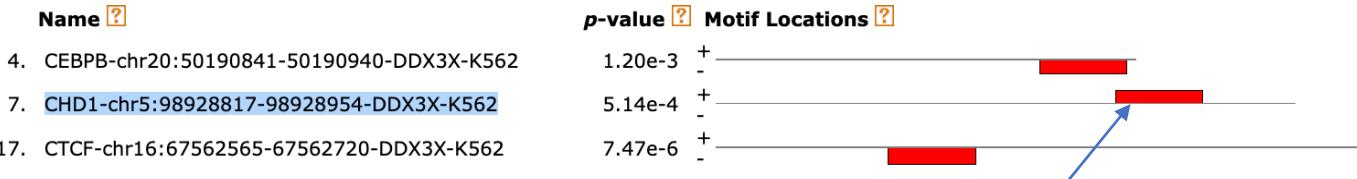
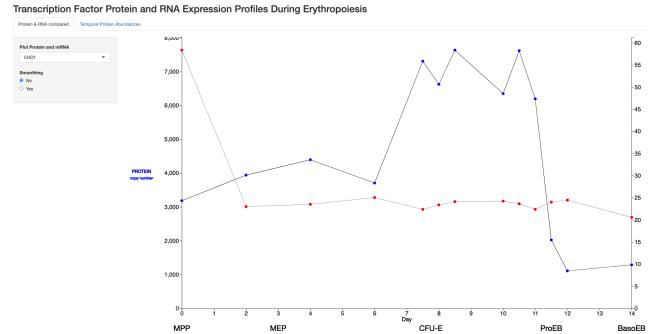
check DDX3X dominant motif & full eCLIP binding site with Dfam, maybe Arian

Returning to the eCLIP DDX3 binding sites - not filtered for 5'UTR intersection, 37/259 were matched by this motif, roughly the same as the motif found for DDX3X hits in 5'UTR for ~3k genes shown just below.



The sequences for the first set are easier to retrieve, so randomly selecting among those 37, I submitted 5 of them to Dfam, which found no transposable elements. I checked with Robert to make sure I was using Dfam properly.

The eQTL hypothesis: if 5'UTR DDX3X binding is functional, then any mutation in the binding site might show up in changed gene expression. To explore this possibility, I examine CHD1.



```
tbl.blood.eqtls <- eQTL_Catalogue.fetch(unique_id="Lepik_2017.blood",
                                           nThread = 4,
                                           chrom = 5,
                                           bp_lower=match.start,
                                           bp_upper=match.end)
```

qtl_id	Lepik_2017.blood
molecular_trait_id.QTL	ENSG00000153922
chromosome.QTL	5
position.QTL	98928917
ref.QTL	A
alt.QTL	AGCC
variant.QTL	chr5_98928917_A_AGCC
ma_samples.QTL	48
maf.QTL	0.0509554
pvalue.QTL	0.0123618
beta.QTL	0.0495628
se.QTL	0.0197334
type.QTL	INDEL
ac.QTL	48
an.QTL	942
median_tpm.QTL	21.363
rsid.QTL	rs1172668401
gene.QTL	CHD1

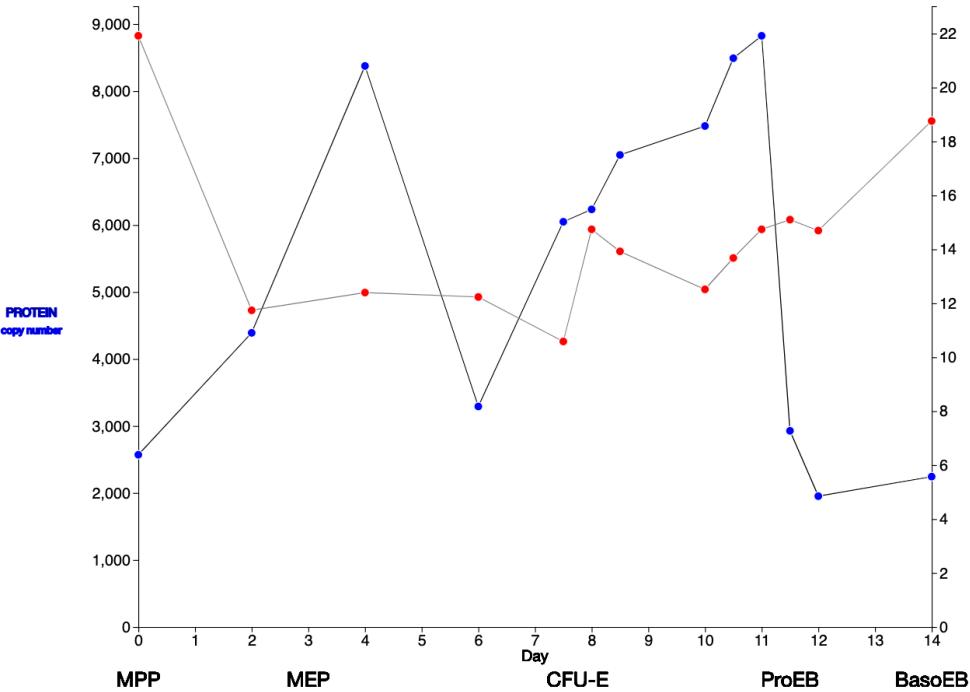
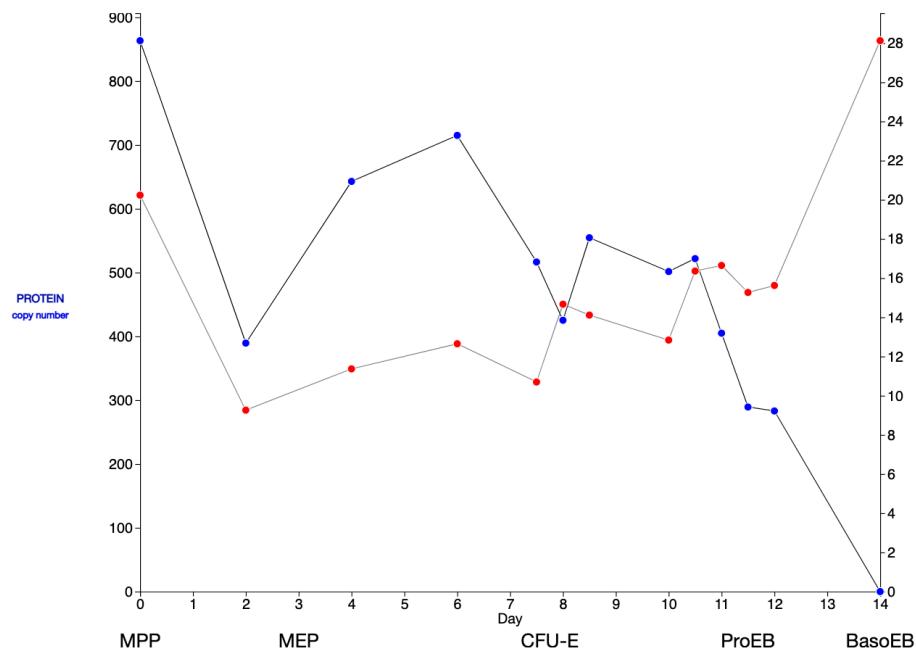
Question from Jeff (11 nov 2021)

I'm curious about your analysis of DDX3X binding sites in 5'UTRs of the SRM genes. It seems like most of the genes have at least one binding site in their 5'UTR. Interestingly most of the proteins also decrease in abundance starting around day 10 or 11.

Could it be that most of the proteins that decrease in abundance starting at day 10 or 11 have a DDX3X binding site in the 5'UTR of their mRNA?

CBP (CREBBP)

BACH1, for instance:



email from Jeff (16 nov 2021)

I attach the curated list highlighted in blue. Criteria for inclusion on the list included:

- 1) Dropping protein levels at day 10.5 or 11 and RNA levels not dropping (usually increasing) across at least 3 time points. Typically days 11-11.5-12.
- 2) TMT expression profile agrees with SRM.

I think it would be interesting to look for RBP binding sites in this group of genes. Of course it would be helpful to also have a set of non-discordant genes to see if RBPs are enriched in one set vs the other set. We could use the genes highlighted in green plus STAT1 and 2 for this set or a randomly selected set of genes from the genome. Focusing on 5'UTRs and/or 3'UTRs maybe yield interesting results

In particular it would be of interest to see if DDx3X binding sites are enriched in the 5'UTRs set of discordant genes. Both the number of sites and just the presence or absence of a binding site for DDx3x would be interesting.

Would this list be useful for TRena modeling? Maybe the criteria that I choose here would help in guiding the correlation analysis?

1	Protein	clear drop at day 10_11	TMT1 agrees	RNA protein discordance	anticorrelated	
2	BACH1	Y	y	Y		drop at 10.5
3	BCL11A_XL_L	Y	y	Y		drop at 11 then correlated increase at end
4	CBP	Y	y	Y		drop at 11 then correlated increase at end
5	CHD1	Y	T not as strong but trending dc	Y		drop at 10.5, then correlated at end
6	coREST	Y	y	Y		drop at 11 then correlated decrease at end
7	CTCF*	Y	y	Y		drop at 11, increase at end
8	DOT1L	Y	y	Y		drop at 11
9	E12/E47	Y	y	Y		drop at 11
10	E2F8	Y	NA	Y		drop at 11 then correlated increase at end
11	GATA1	Y	y	Y		drop at 10 but anti-corr starts at 10.5
12	GFI1B	Y	y	Y		drop at 11
13	GR	Y	NA	Y		drop at 11 then correlated increase at end
14	HCFC1	Y	y	Y		drop at 10.5 then correlated at end
15	HDAC3	Y	y	Y		drop at 10.5
16	HLTF	Y	y	Y		drop at 10.5, but anti-corr starts at 11
17	IKZF1	Y	y	Y		drop at 11
18	KAT2A	Y	y	Y		drop at 11 then correlated decrease at end
19	KLF1*	Y	y	Y		drop at 11 then correlated increase at end
20	KLF3	Y	y	Y		drop at 11 then correlated increase at end
21	LDB1*	Y	y	Y		drop at 10.5, but anti-corr starts at 11
22	MAFF	Y	NA	Y		drop at 8.5
23	MAFG	Y	y	Y		drop at 11 then correlated at end
24	MED1	Y	y	Y		drop at 10.5 but anti correlation starts at 11
25	MLL3	Y	y	Y		drop at 11
26	MLL4 (KMT2D)	Y	y	Y		drop at 10, but anti-corr at 11
27	MTA1	Y	y	Y		drop at 10.5 but strong anti corr starts at 11
28	NFE2	Y	NA	Y		drop at 10.5, anti- corr at 11
29	NR2C2	Y	y	Y		drop at 11
30	OGT	Y	y	Y		drop at 10.5, but anti-corr at 11
31	PSIP1	Y	y/n	Y		drop at 11
32	RAD21	Y	y	Y		drop at 10.5, but anti-corr at 11
33	SCML2	Y	y	Y		drop at 10.5, but anti-corr at 11
34	SMC3*	Y	y	Y		drop at 11
35	SSRP1	Y	y	Y		drop at 11
36	SUZ12	Y	y	Y		drop at 10.5, but anti-corr at 11
37	TAL1*	Y	y	Y		drop at 11 increase at end
38	TFDP1	Y	y	Y		drop at 8 but anti-corr at 10.5
39	TRIM33	Y	y	Y		drop at 11
40	USF1	Y	y	Y		drop at 11
41	UTX	Y	y	Y		drop at 11
42	WDHD1	Y	y	Y		drop at 11
43	ZBTB7A*	Y	y	Y		drop at 11
44						
45	PO2F1	Y	n	Y		drop at 10.5, but anti-corr at 11
46	E2F4*	Y	n	Y		drop at 11 then correlated increase at end
47	KLF13	Y	n	Y		drop at 11
48						
49	STAT1	N	y	Y	opposite	increase at 11
50	STAT2	N	NA	Y	opposite	increase at 11

49	STAT1	N	Y	Y	opposite	increase at 11
50	STAT2	N	NA	Y	opposite	increase at 11
51						
52						
53	ETF1	?	slight drop SRM, increase TMT	RNA up then down		
54	EP300	?	TMT down SRM up	RNA up		
55	CEBPB	?	TMT up/SRM down	RNA slight increase		
56	ETO2	?	TMT decrease SRM unclear	RNA down		
57	STAT3	?	TMT down SRM up	RNA down		
58	CHD3	Y	TMT down more than SRM	N		
59	CHD4	Y		N		
60	CTBP2	Y		N		
61	DNMT1	Y		N		
62	EGR1	Y		N		
63	ELF1	Y		N		
64	ERG	N	continuous decrease	N		
65	ETV6	Y	continuous decrease	N		
66	FLI1	N	continuous decrease	N		
67	GATA2	N	drop begins at day 8.5	N		
68	GATAD2A	Y		N		
69	HDAC1	Y		N		
70	HMGB3	Y		N		
71	HXB4	Y		N		
72	KDM1A	Y		N		
73	MNDA	N	continous decrease	N		
74	NELFE	Y		N		
75	PARP1	Y		N		
76	RFX5	Y		N		
77	RPB1	Y		N		
78	RUNX1	Y		N		
79	SET1B	Y		N		
80	SETB1	Y		N		
81	SIN3A	Y		N		
82	SIR6	Y		N		
83	SMCA4	Y		N		
84	SMRC1	Y		N		
85	SNF5	Y		N		
86	SOX6	N	increase	N		
87	SPI1	N	continuous decrease	N		
88	SPT16	Y	TMT	N		
89	STA5A	Y		N		
90	T2FA	Y		N		
91	TF2B	Y		N		
92	TF3C2	Y		N		
93	TFCP2	Y		N		
94	WDR5	Y		N		

email to jeff (17 nov 2021)

I am starting on this now. I have two strong DDX3X motifs from 5'UTRs across about 1400 genes.
I picked these on these criteria:

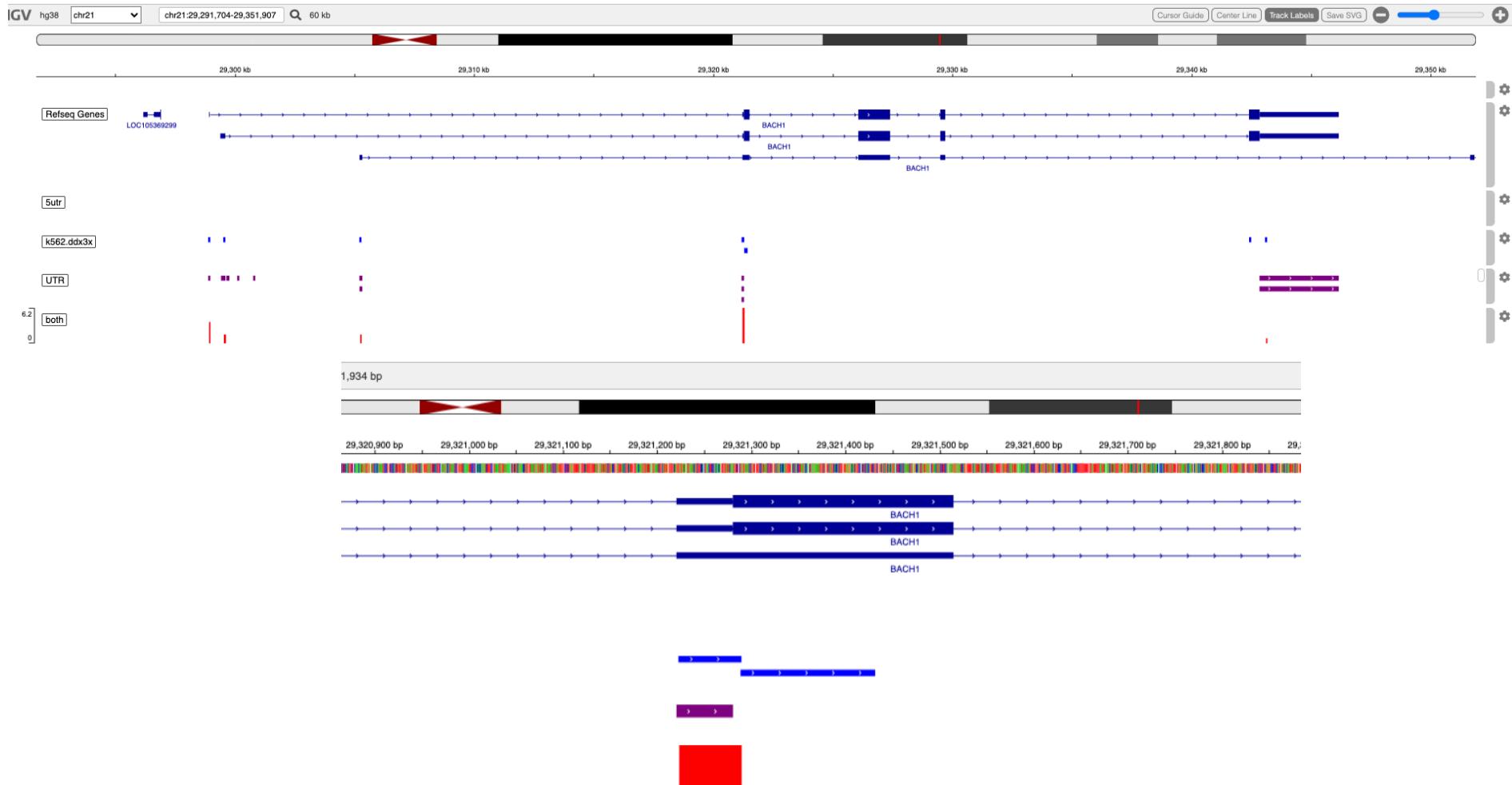
ENCODE K562 reported 82K 5'UTR DDX3X hits across the genome, with multiple hits in many genes
Each hit has a score, I kept only those hits in the top quartile.

This target set of ~1400 provides two strong motifs, the top 2 shown below. I propose to look for these, and score their match, in the 42 late-discordant genes you identified in blue, and the green genes + STAT1 and STAT2 for contrast.

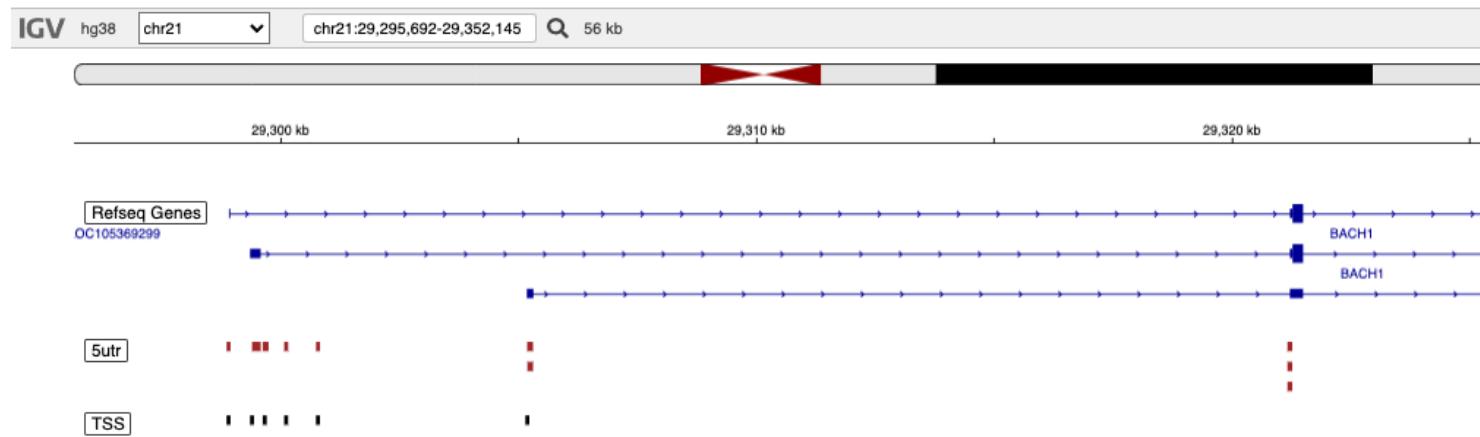
- Paul

		E-value ?	Sites ?	Width ?
1.		5.3e-089	1471	21
2.		7.5e-046	13	49
3.		1.8e-022	9	50
4.		8.7e-014	840	21
5.		1.9e-007	10	50
6.		9.3e-016	23	50

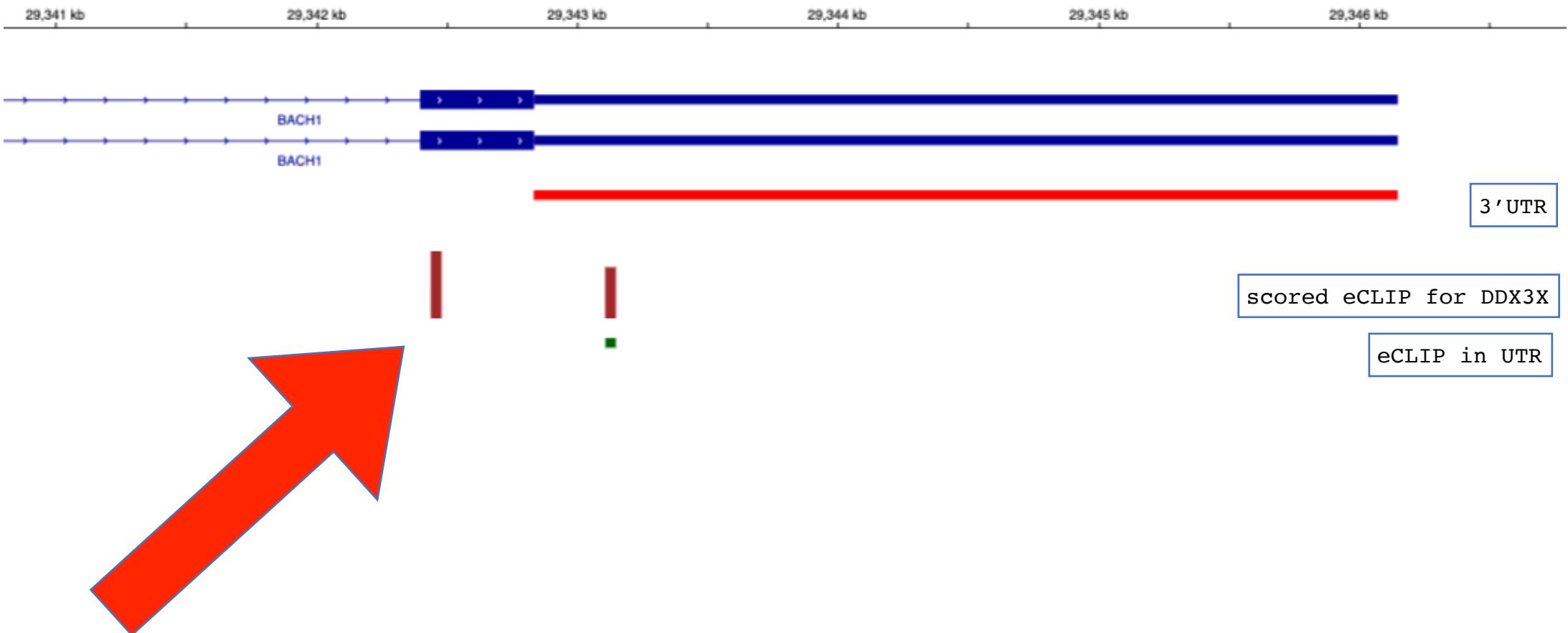
The method: find all binding sites for DDX3X (and soon, other RBPs also) which overlap 5' or 3' UTRs. Interesting that the first BACH1 exon is annotated as a 5'UTR. after discussing with Jeff, this resolution proposed: annotate exon intersections (T/F) for each UTR/DDX3x intersection. should be infrequent. the import or inconsequentiality of this can be figured out later. 4% of 5'UTRs overlap exons.



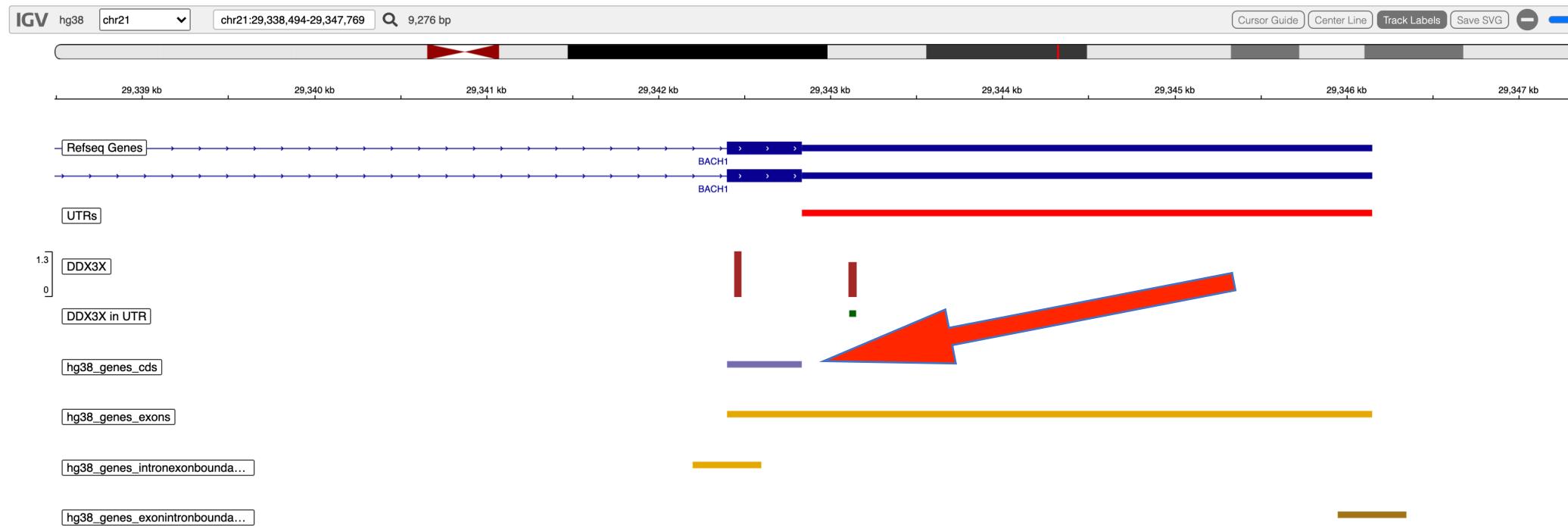
Jeff and me, email exchange (19 nov 330p); One more bit of information. I turned to Ensembl, queried them for 5'UTRs and for TSS for all BACH1 transcripts. Here's what I (Paul) got, pictured below.
We -could- drop 5'UTRs which have no TSS nearby. But I think you are saying, "keep them all, because after splicing, those downstream 5'UTRs are found at the front of the actual spliced transcript, just where they should be".



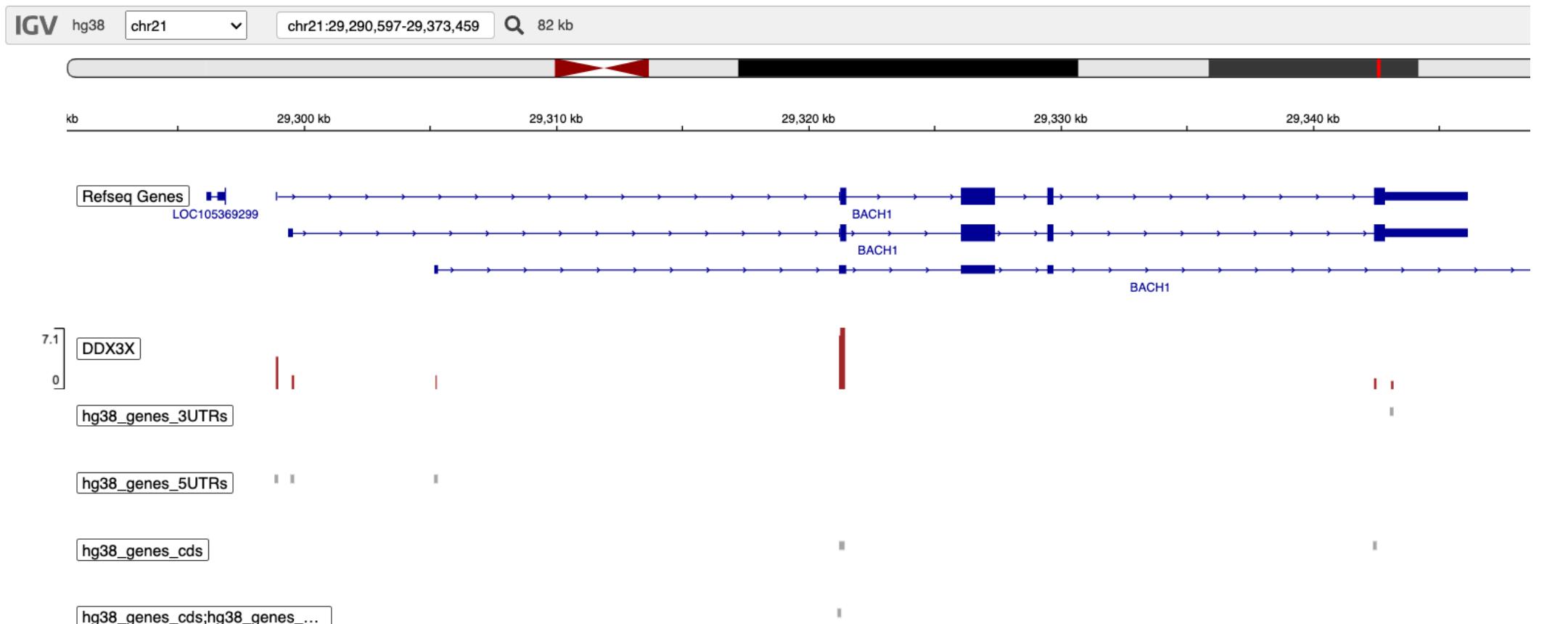
Question for Jeff: should this DDX3X binding site - pointed to by red arrow be included in our assessment of eCLIP in UTRs? not in a proper 3'UTR, but maybe interestingly functional nonetheless.



possibly interesting intersection of that non-utr DDX3X binding site in a 435 bp CDS



automated and visual test, DDX3X and BACH1, all binding sites fall in 5'UTR, 3'UTR and or CDS genic regions.



chrom	start	end	width	score	regionType	
1	chr21	29298920	29298954	35	3.801192	hg38_genes_5UTRs
2	chr21	29299553	29299594	42	1.659932	hg38_genes_5UTRs
3	chr21	29305242	29305275	34	1.659932	hg38_genes_5UTRs
4	chr21	29321223	29321289	67	6.215896	hg38_genes_cds;hg38_genes_5UTRs
5	chr21	29321289	29321431	143	7.116253	hg38_genes_cds
6	chr21	29342443	29342478	36	1.306399	hg38_genes_cds
7	chr21	29343109	29343149	41	1.002156	hg38_genes_3UTRs

**DDX3X binding in 3 gene sets: "clear drop", "no drop"
1000 genes picked at random (performed twice) and
day 10-11 drop, without and without rna/srm discord**

group	ddx3x.genes	ddx3x.3utr.sites	ddx3x.5utr.sites	ddx3x.cds.sites
clear drop genes	77/90	32	68	50
mean score	3.99	2.74	5.48	2.58
no drop genes	4/8	1	4	2
mean score	4.15	1.28	6.04	1.36
random set #1	423/1000	370	779	501
mean score	10.18	3.5	14.95	7.42
random set #2	394/1000	246	696	451
mean score	9.50	2.13	9.21	7.01
drop with discord	39/42	36	13	23
mean score	3.73	1.68	5.20	2.60
drop no discord	29/32	33	74	56
mean score	3.97	2.28	5.62	2.63

email from jeff: 6 dec 2021

I attach a file that has three tabs. 1) the original curated data, 2) called group 1 - genes that show protein/RNA discordance, 3) called group 2 - genes that do not show discordance.

In addition to testing for RBP enrichment amongst proteins that decrease, or not, at late time points, it would be interesting to see if RBPs are enriched amongst these two classes of genes - discordant and not discordant.

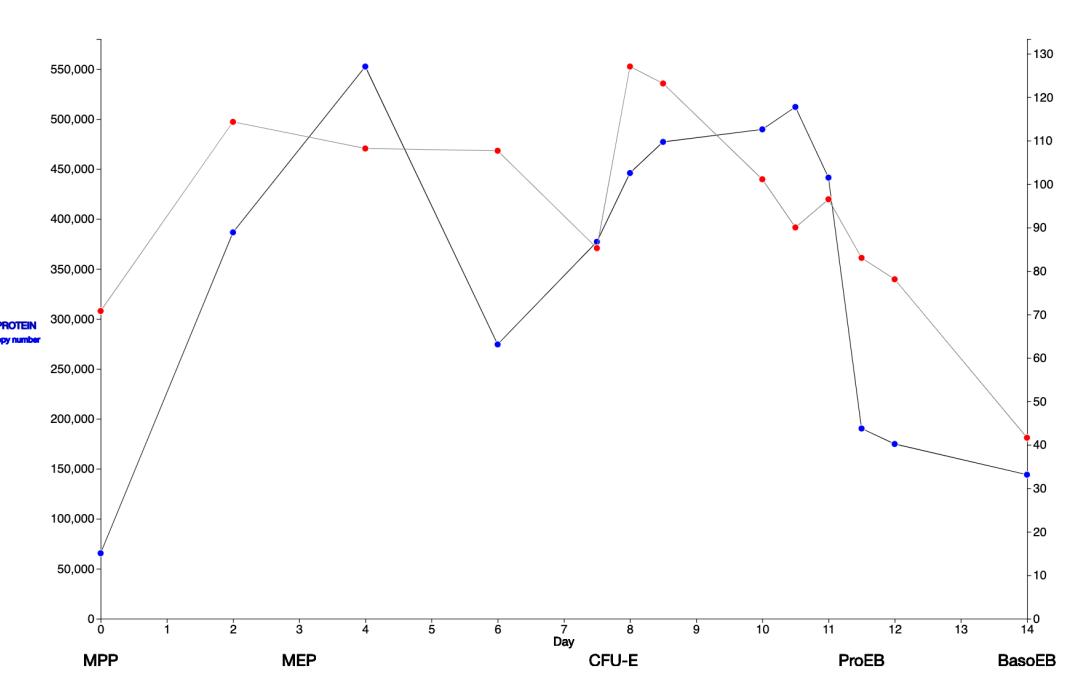
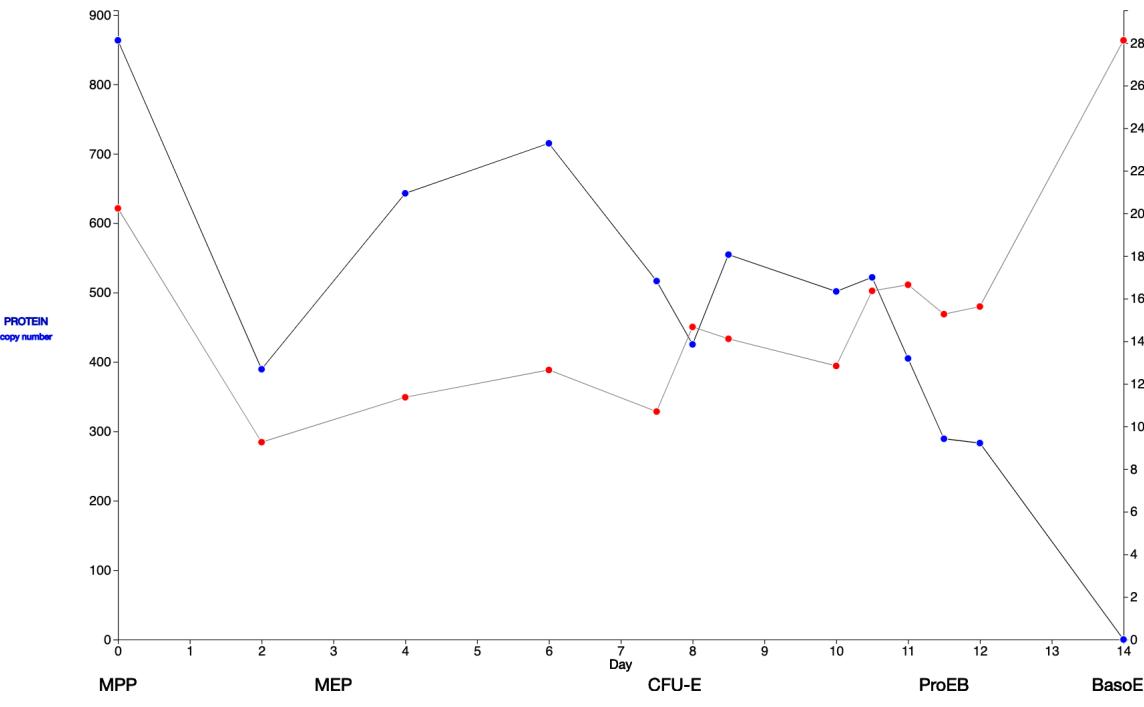
In our last meeting you reported that DDX3x could bind to 4/8 of the "no drop" genes. Could you tell which of these genes are DDX3x targets and which are not?

clear protein drop at day 10-11, with discordant rna (42): BACH1, BCL11A, CREBBP, CHD1, ...
same, but concordant rna (32): CHD3, CHD4, CTBP2, DNMT1, ...

Acting on Jeff's email

	rnaProteinDiscord	
clearProteinDrop	FALSE	TRUE
FALSE	6	0
TRUE	32	42

42 genes have clear (pronounced) protein abundance drop at day 10-11 AND rnaProteinDiscord (that is, opposite trends in rna-seq and SRM, **BACH1** for example. Protein abundance of this 42-member group may be enriched for DNA-binding protein regulation. **CHD4** shows the opposite effect. Is there contrasting RBP binding in this 32-member group?
the differences in quantity, max of srm copy number 900 for BACH1, 65000 for ZBTB7a



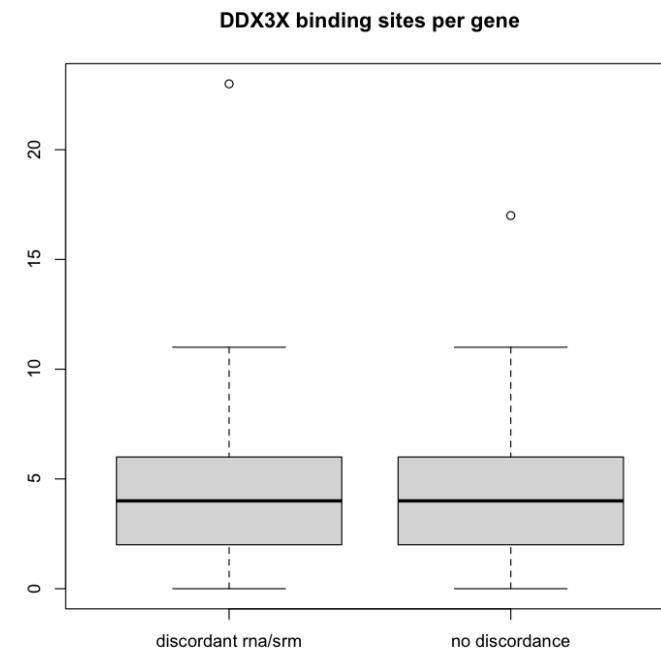
rna/srm discordance

	gene	DDX3X.bindingSites
1	BCL11A	0
2	ZBTB7A	0
3	KAT2A	1
4	MAFF	1
5	OGT	1
6	RCOR1	1
7	DOT1L	2
8	IKZF1	2
9	KDM6A	2
10	MED1	2
11	NR2C2	2
12	SCML2	2
13	SMC3	2
14	SUZ12	2
15	GFI1B	3
16	HDAC3	3
17	MTA1	3
18	NR3C1	3
19	USF1	3
20	CREBBP	4
21	E2F4	4
22	GATA1	4
23	HLTF	4
24	KLF1	4
25	MAFG	4
26	PSIP1	4
27	TFDP1	4
28	TRIM33	4
29	WDHD1	4
30	LDB1	5
31	CTCF	6
32	KMT2C	6
33	BACH1	7
34	RAD21	7
35	SSRP1	7
36	TAL1	7
37	TCF3	7
38	NFE2	8
39	CHD1	9
40	HCFC1	11
41	KMT2D	23

day 10-11 protein drop, with and without rna/srm discord
DDX3X binding sites do not predict discord.

rna/srm concordance

	gene	DDX3X.bindingSites
1	ETV6	0
2	RFX5	0
3	SIRT6	0
4	CHD3	1
5	CTBP2	1
6	HOXB4	1
7	GTF2F1	2
8	NELFE	2
9	SETD1B	2
10	SETDB1	3
11	SMARCA4	3
12	TFCP2	3
13	DNMT1	4
14	EGR1	4
15	ELF1	4
16	GTF2B	4
17	HDAC1	4
18	SMARCB1	4
19	KDM1A	5
20	SIN3A	5
21	WDR5	5
22	GATAD2A	6
23	GTF3C2	6
24	STAT5A	6
25	SUPT16H	6
26	HMGB3	7
27	RUNX1	7
28	PARP1	10
29	CHD4	11
30	POLR2A	11
31	SMARCC1	11
32	ZC3H11A	17



42 dropping proteins with rna/srm discord
32 dropping proteins without discord
6 proteins which do not drop:
ERG FLI1 GATA2 MNDA SOX6 SPI1
(GATA2 drops, but that starts at day 8.5)

Any absence of DDX3X binding sites in no-drop genes?

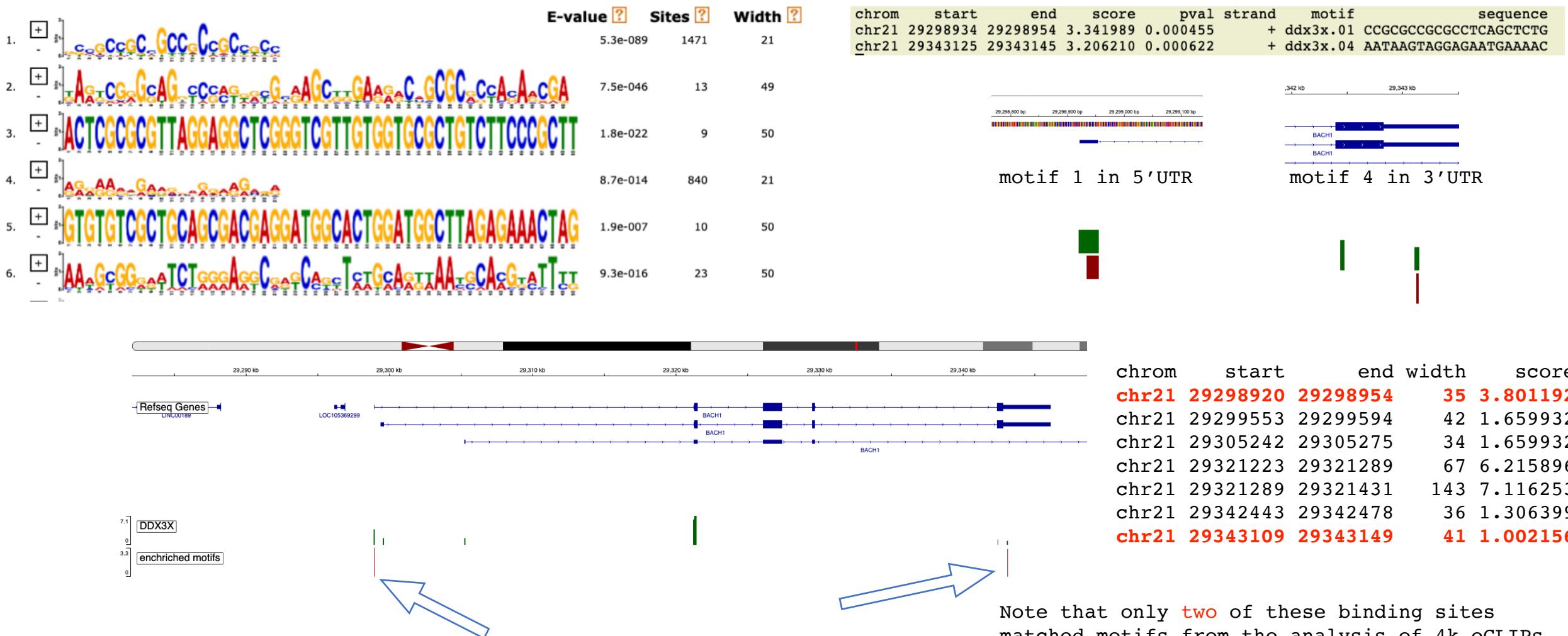
Jeff proposes these genes: ERG FLI1 GATA2 MNDA SOX6 SPI1 STAT1 STAT2 STAT3, of which ERG, FLI1, MNDA SOX6 have no binding sites.

me: If DDX3X plays a regulatory role, then like TFBS and functional TF activity, it may be controlled, not simple presence/absence, but by a complex set of interactions.

	chrom	start	end	width	score	regionType	targetGene	rbp
1	chr3	128479650	128479686	37	1.192377	hg38_genes_3UTRs	GATA2	DDX3X
2	chr3	128480413	128480438	26	1.360104	hg38_genes_3UTRs	GATA2	DDX3X
3	chr3	128485882	128485902	21	1.422952	hg38_genes_cds	GATA2	DDX3X
4	chr3	128487710	128487909	200	14.101637	hg38_genes_5UTRs	GATA2	DDX3X
5	chr11	47378364	47378416	53	2.113769	hg38_genes_5UTRs	SPI1	DDX3X
6	chr2	191009875	191009911	37	1.306399	hg38_genes_cds	STAT1	DDX3X
7	chr2	191014010	191014047	38	4.233213	hg38_genes_5UTRs	STAT1	DDX3X
8	chr2	191014062	191014166	105	12.168697	hg38_genes_5UTRs	STAT1	DDX3X
9	chr12	56356508	56356579	72	1.192377	hg38_genes_cds;hg38_genes_5UTRs	STAT2	DDX3X
10	chr12	56360057	56360099	43	2.404108	hg38_genes_5UTRs	STAT2	DDX3X
11	chr17	42315253	42315337	85	1.673357	hg38_genes_3UTRs	STAT3	DDX3X
12	chr17	42315525	42315601	77	1.324955	hg38_genes_3UTRs	STAT3	DDX3X
13	chr17	42315712	42315774	63	1.401330	hg38_genes_cds;hg38_genes_3UTRs	STAT3	DDX3X
14	chr17	42388311	42388368	58	2.463934	hg38_genes_5UTRs	STAT3	DDX3X

DDX3X enriched motifs in BACH1 eCLIPs binding sites

from slide 46: DDX3X binding motifs from 5'UTR in > 4k genes



Note that only **two** of these binding sites matched motifs from the analysis of 4k eCLIPs genes, 1st and 7th in the above table. zoomed-in views at top right. note that the highest scoring hit had no matches to any of the 6 motifs found by the meme suite.

