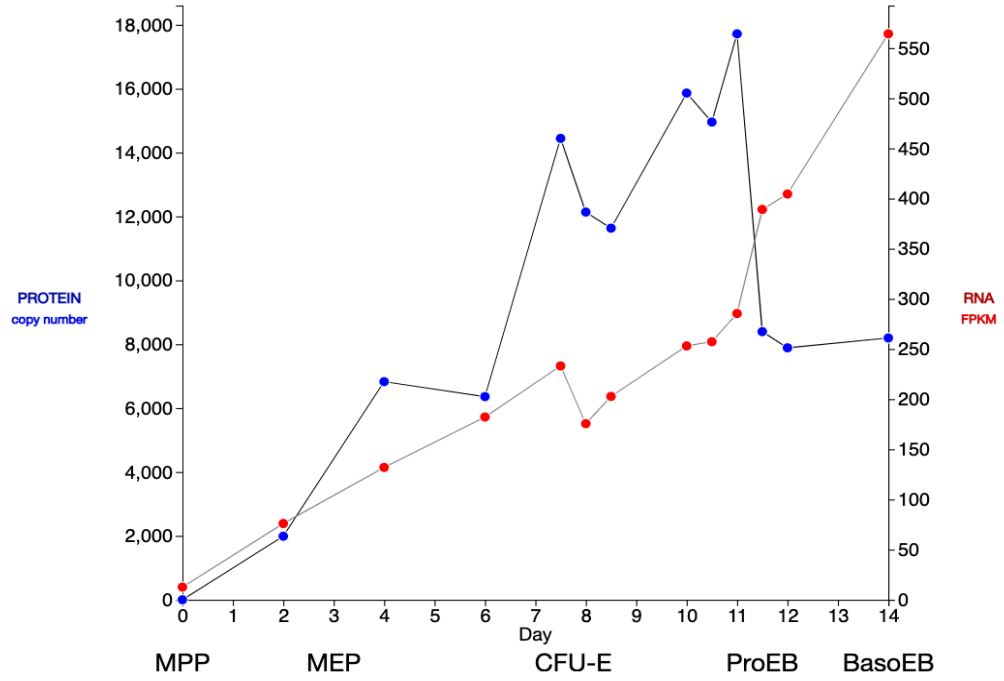
Exploring DDX3X as a possible regulator of late stage erythropoiesis

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
MXI1	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
HNRNPD	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	HLTF	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
HNRNPC	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
HLTF	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

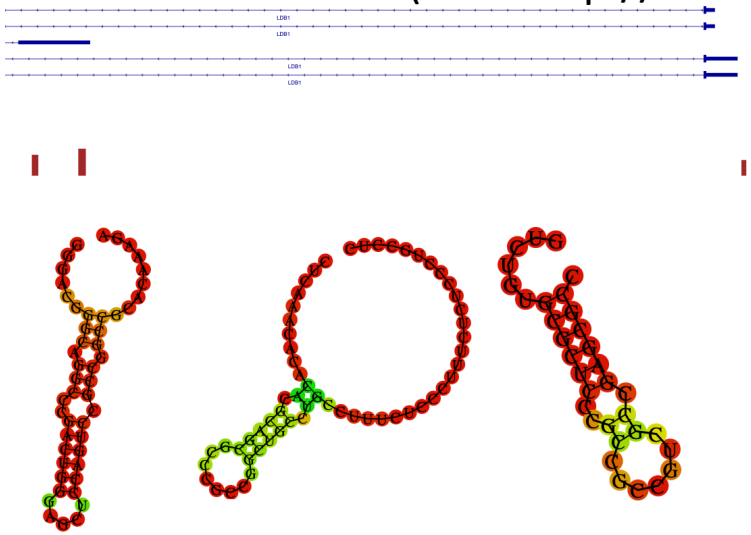
RBP binding across the genome

	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/RNAfold.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/RNAfold.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.

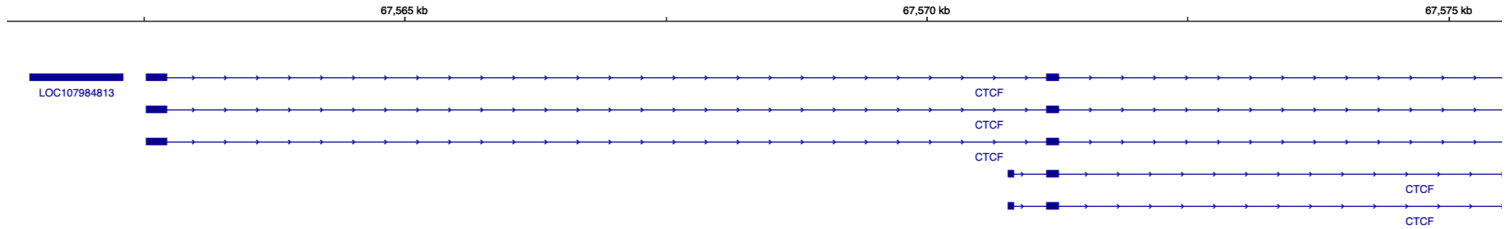
I am eager to hear what you think of all this, and what I should do next.

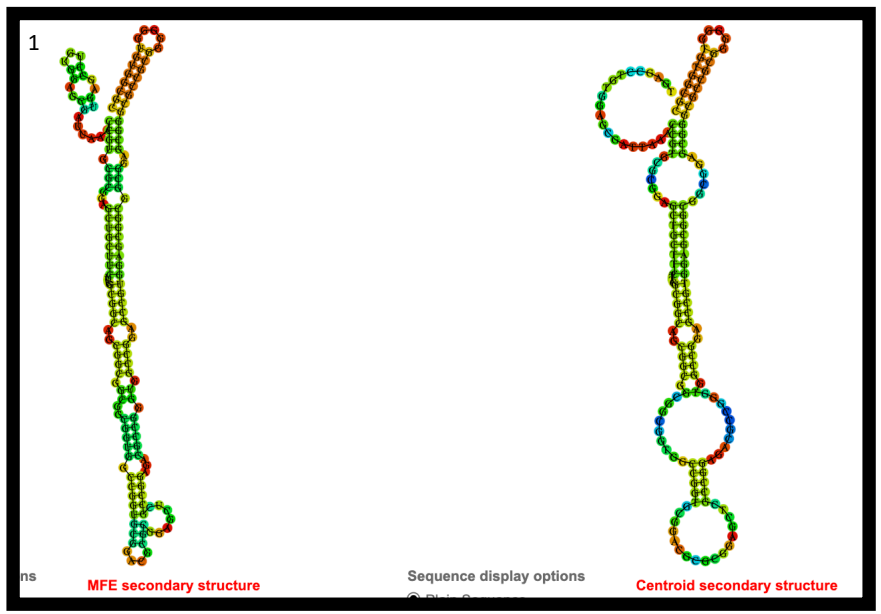
(I will do a quick survey to find out how common eCLIP K562 DDX3X binding sites are across the entire genome. If 6/7 of all genes have such sites, then these results will not be so interesting.)

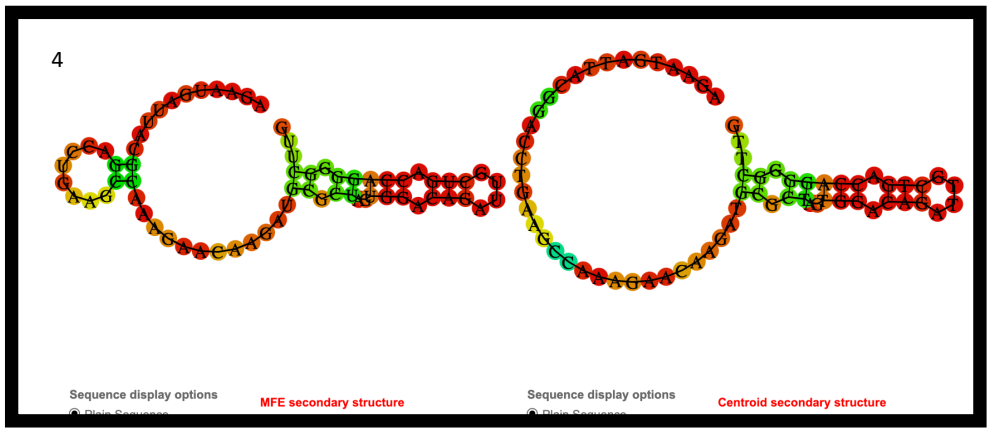
CTCF

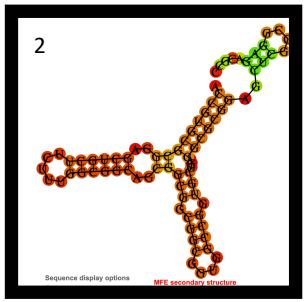
	chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr16	67562565	67562720	156	+	DDX3X	eCLIP	K562	ENCODE	23.994476	CTCF	hg38_genes_promoters
2	chr16	67562586	67562665	80	+	DDX3X	eCLIP	K562	ENCODE	7.839952	CTCF	hg38_genes_promoters
3	chr16	67562665	67562721	57	+	DDX3X	eCLIP	K562	ENCODE	5.006577	CTCF	hg38_genes_5UTRs
4	chr16	67571145	67571213	69	+	DDX3X	eCLIP	K562	ENCODE	4.668016	CTCF	hg38_genes_introns

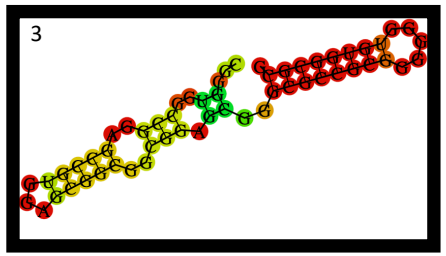
Binding sites 2 and 3 overlap with long higher-scoring binding site 1





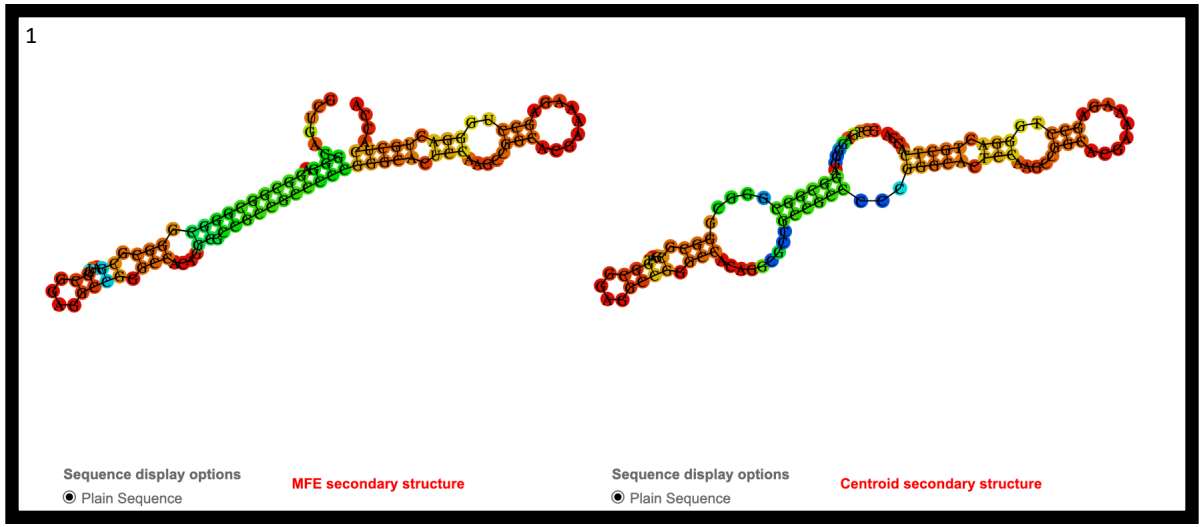






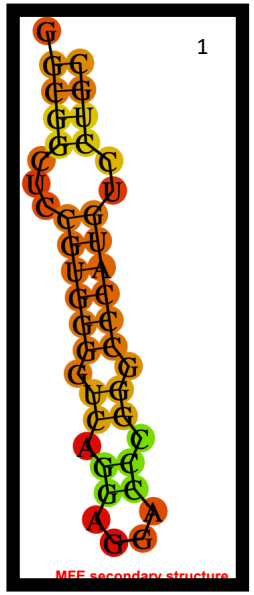
E2F4

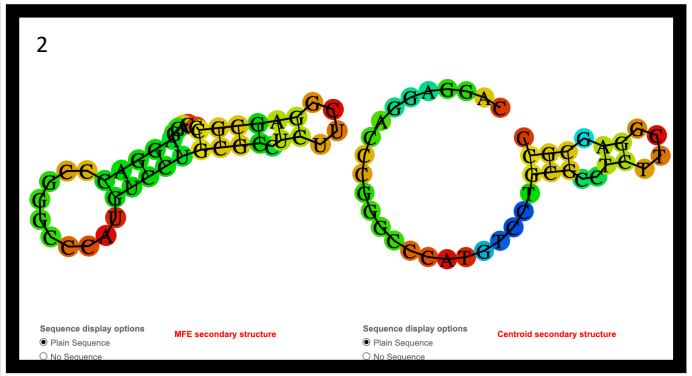
	chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr16	67192201	67192306	106	+	DDX3X	eCLIP	K562	ENCODE	31.31813	E2F4	hg38_genes_1to5kb

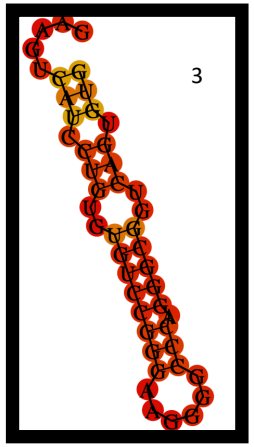


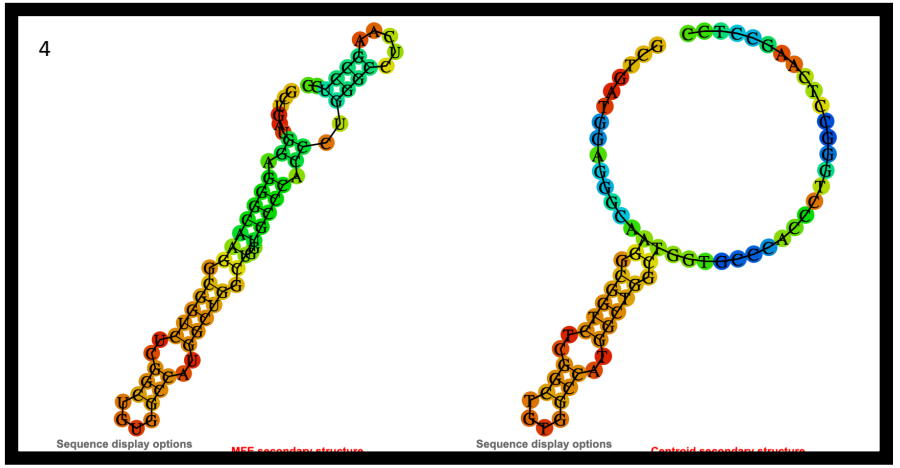
KLF1

	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



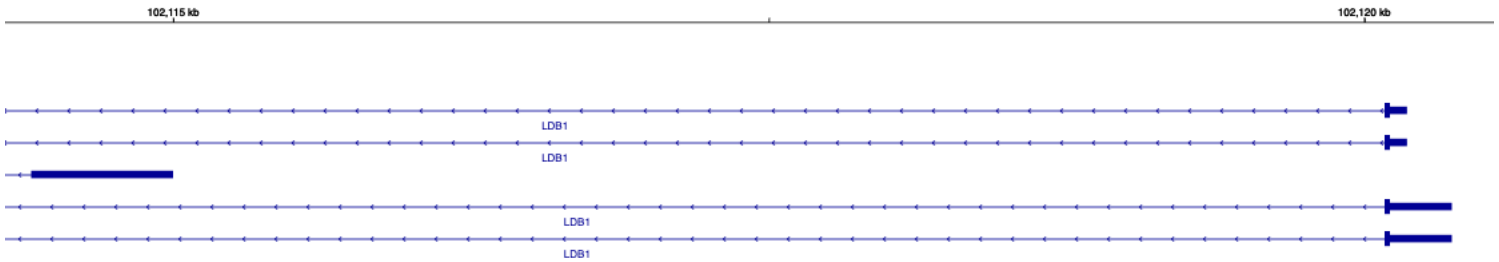


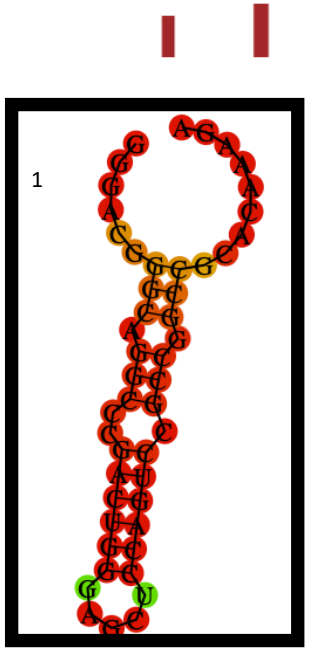


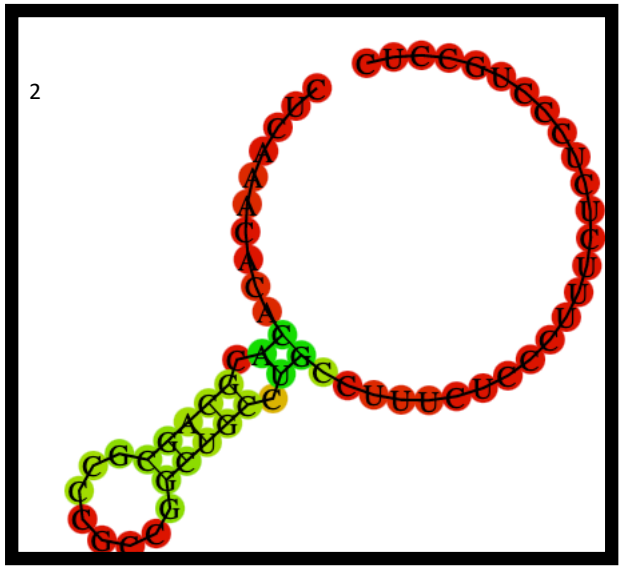


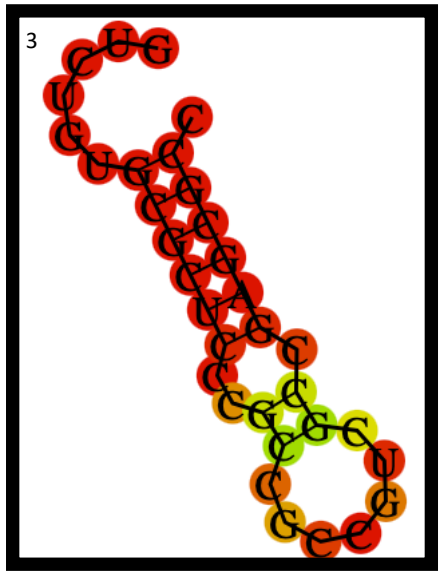
LDB1

	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



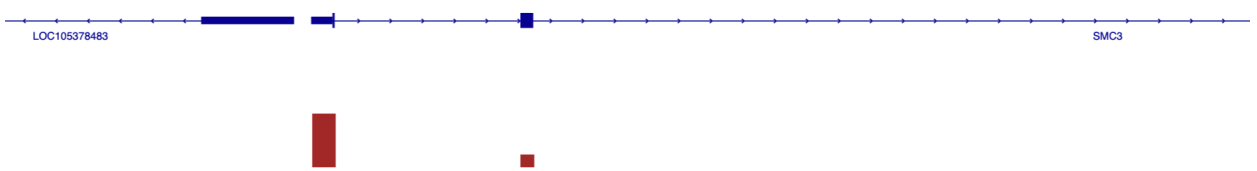


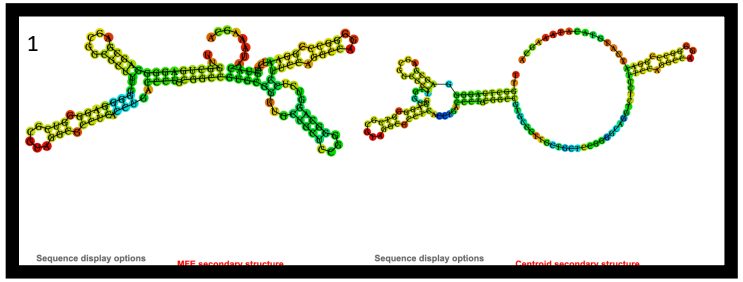


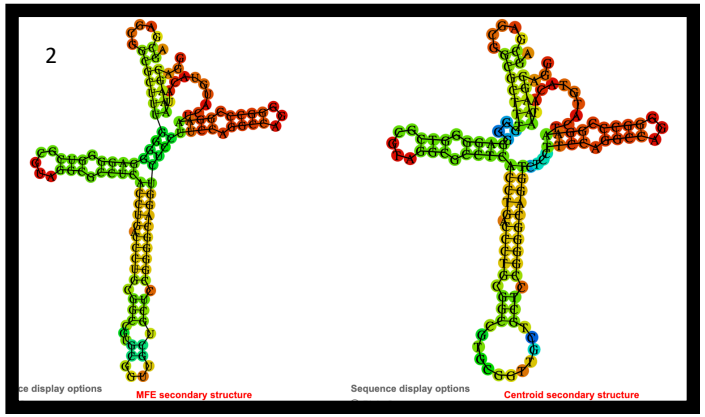


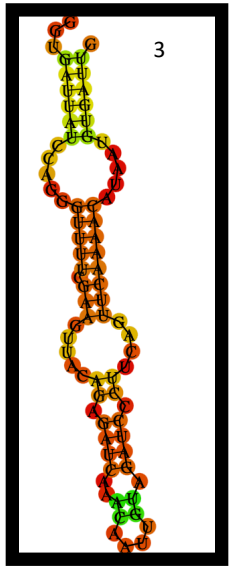
SMC3

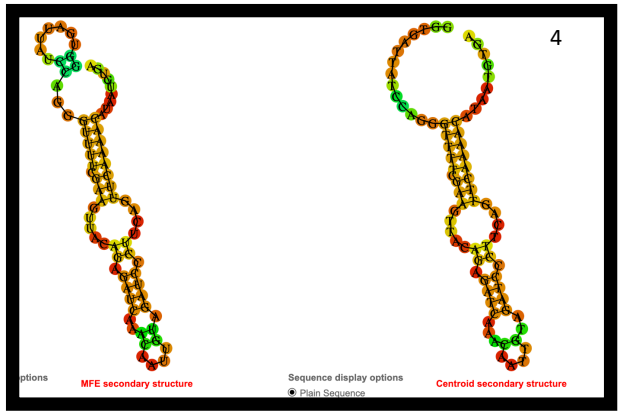
	chrom	start	endpos	width	strand	gene	method	celltype	accession	score	target	targetfeature
1	chr10	110567697	110567830	134	+	DDX3X	eCLIP	K562	ENCODE	44.68689	SMC3	hg38_genes_promoters
2	chr10	110567709	110567832	124	+	DDX3X	eCLIP	K562	ENCODE	29.39116	SMC3	hg38_genes_promoters
3	chr10	110568937	110569013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





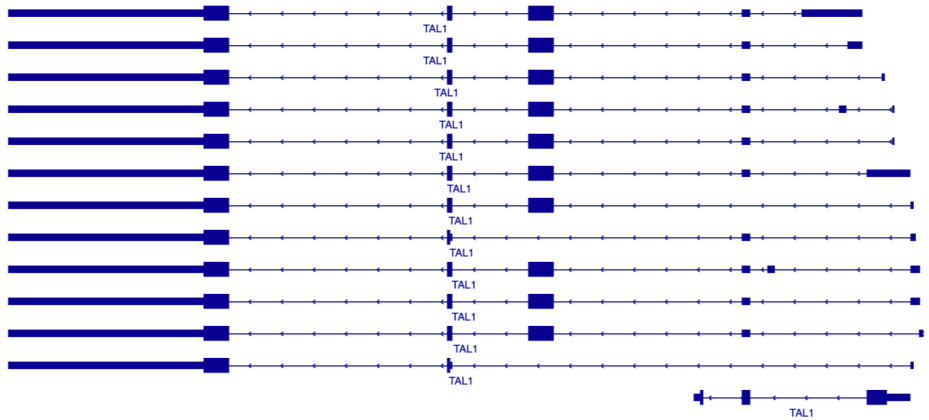


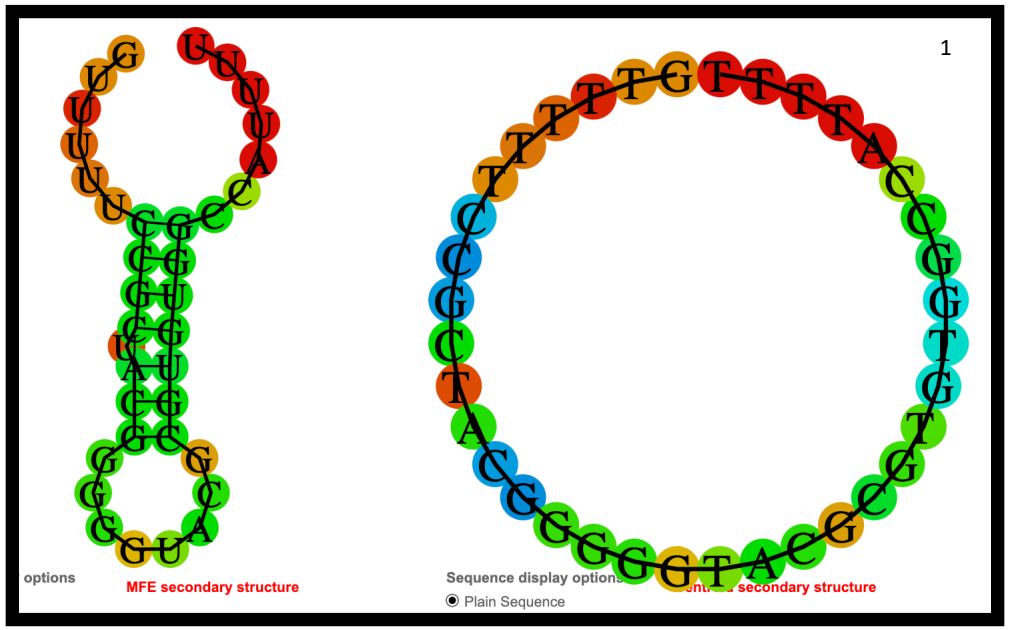




TAL1

	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



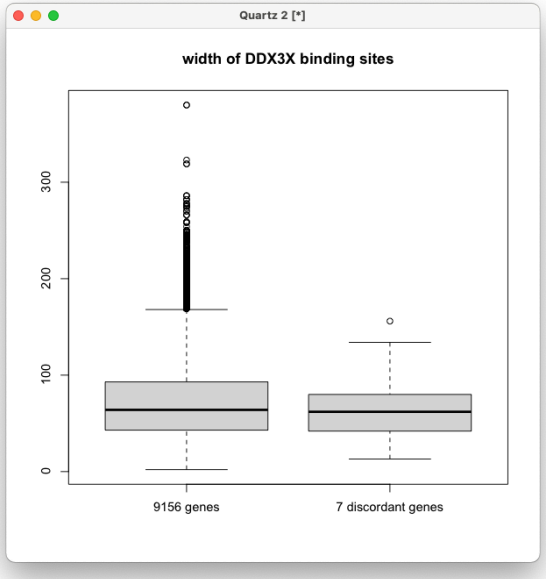


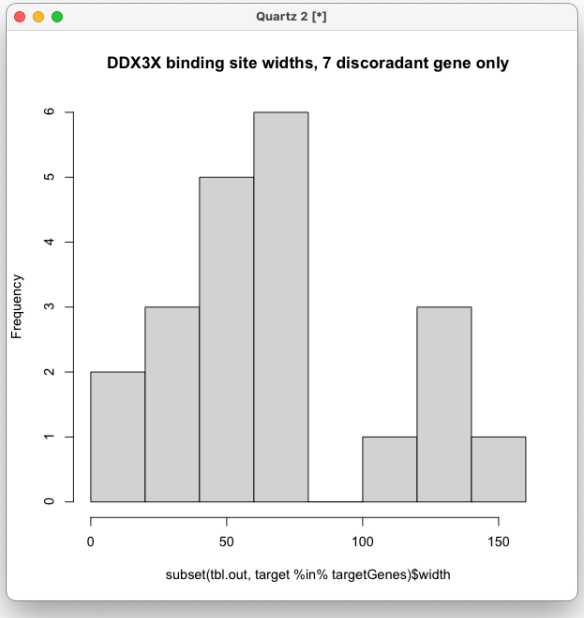
ZBTB7A

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 (padj = 2.83 × 10−5; Glass' Δ = 6.9) and the translational efficiency of 730 target genes (padj = 5.25 × 10−5; Glass' Δ= 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 (>=20bp) 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 ot endogenous 5' UTRs of similar length using a reporter assay (Supplementary Fig. 13d). Additionally, other transcripts with hig structured 5' UTRs showed reduced association with larger polysomes (Supplementary Fig. 14), while 5' UTRs that were unstructu 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 ribosome 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 x log2 (pi/qi), where p 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-value ≤ 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?

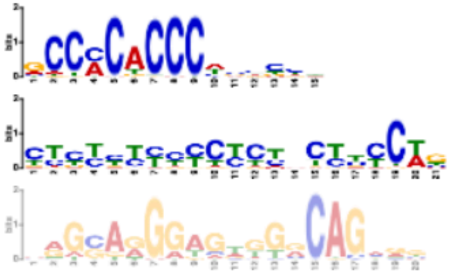
Regarding the eCLIP data processing presented in https://www.biorxiv.org/content/10.1101/2021.04.13.439465v2.full
do you think this processing step is worth trying?

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



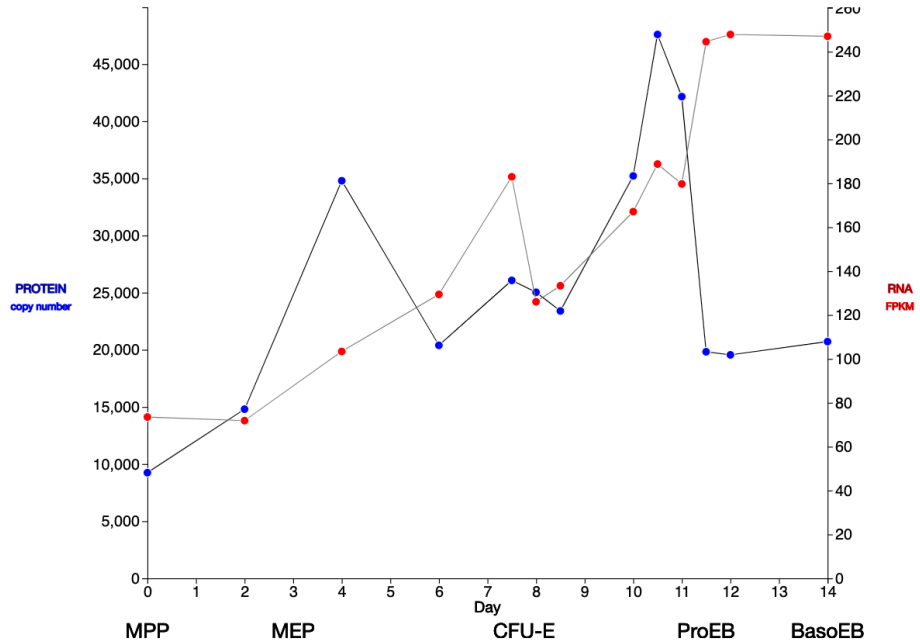
	Logo	E-value	Sites	Width
1.		5.4e-044	453	15
2.		4.0e-019	338	21
3.		1.6e+000	54	20

Stopped because requested number of motifs (3) found.

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

```
> head tests/common/Klf1.fa
>chr1:9738233-9738733
ACATCGCTCTCGGCTCTCGCTCGGCGTCCCAAGCGACCCTCAAAGGCACATAGCCAGACCTCCGCCGGCCCGACGAGGTCTAGTCCAGGCCCAGGGCGAATGGCTTTCCTTTTCCTACCAGTTCCTCCTCCTCCTAGGCGAAGGCGGAAGCGCCGGACGTTGTTTCTCTTCTCA
CTTCTCCACTGCCTTAGCCGTTGCCCCGAACGTAACGGCCACCACCCCACCCCGCACTCACACTCACTCACTCTCGCTCTCTCCCTCAGACACAGACATACACGCCCTCACTGAGACTGCGCAGGCGTAGCTTCTGCTCTGCCCTCTGGGAACCAGAGTCTTCCGGCTCCTCTC
TCGCGAAGGAGTGCTAGGCGCTTCCGTCCAGAAAGAACTTCAAATACCGTGATGCAAAGCGACTCGACGCACGCGGGGAGGAGGAAGAGGAAGGCACTCGGAAGGTGGGGGCAGAAAGGAAGTCGATGATTCGCGGAGCTTGTCGGTAGTTG
>chr1:9763391-9763891
TCACTATCTGCCTCCCACTTCGGCAGCTTAGGCAGTGGCCACTAAATAAACCAGTTCCCCTCTCTGAGAATACTGCTGTAGTCTGACCTACAAAGGTTGTGGTAAGAGTCTTGGTGGAATCTGAGAGAGGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNG
AAAGACAGACAGGCTTTTGAGAGACACTCCTACATGAAAAGAGCAATCTTGAGTTGTGGCATGTGACAAGTTCCTCTAGGACAGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCGTTGAAATGGTAACTCCGTTGCTTTCAGATCTGTGCCTTCACTTGACTTGAAAA
>chr1:17003945-17004445
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGAAGAGTCAACAACACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>chr1:22562530-22563030
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNAGTGTTCTTTACAATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGGCACAATCTAATACCTACAATTTCTGATTAAAACAAAATTATTTGACTAGGGTGAGATAATATA
TGACTTTATCCCTTTTCCATAATACTGTTTGCCATTGCTAGTTCTTCACTTGAGTAAAGTCAGCTTAGGTATATTGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>chr1:33631213-33631713
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NTCTGCTTTGCTGCTGGGCCACGCCCTGCGACCCACGGGTTGAAAGTCACTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

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

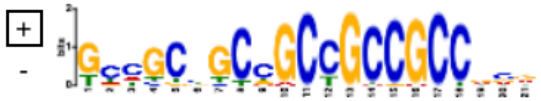
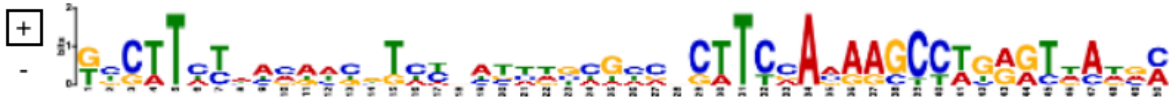
Motif discovery in DDX3X eCLIPs sequences

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

```
>KLF1-chr19:12887105-12887175-DDX3X-K562
GCTGATGGAGGGCAAGGCGGTCTCGGCTGTGGCCATGGCTGGCTGGTGCCCACCCTGGGCCTCAAGCCTCC
>KLF1-chr19:12891165-12891202-DDX3X-K562
CACTGTAGCCTCGGCAGTGAACCGGGAGGTACTACCAG
>KLF1-chr19:12891270-12891394-DDX3X-K562
GGTCGCCGTCGTTGCTCCGCTCGCTCTGAGAGAGCATGGCCCTGAGAGGCGTCTCCGTGCGGCTGCTGAGCCGCGGACCC
GGCCTGCACGTCCTTCGCACGTGGGTCTCGTCGGCGGCGCAGACC
>KLF1-chr19:12886081-12886122-DDX3X-K562
GGCGGCTCCGTGGGGTCAGGAGGACCCGGGCCCATGTCCTGC
>KLF3-chr4:38664217-38664307-DDX3X-K562
GTCATGTGACTGCCCGGAGTTGGTGCCAGGAGCCAGAGGGGAGCCAGGAGCGGAGCCGCGCGGAGCCGGGGCCCGAGCCG
GAGCGCAGCGA
>LDB1-chr10:102114520-102114567-DDX3X-K562
GGGACGGGCAGGCCCGACTGGGAGCTCCAGTCCGCCGGCCGCACAAAG
>LDB1-chr10:102120403-102120438-DDX3X-K562
GTGTCTGTGCGCTCCCGCCGCCGTCGCCGAGCGCCC
>LDB1-chr10:102114902-102114963-DDX3X-K562
ACTCAAACACACACGCAGCGCCCGCCGGCTGCCTGCCTTTCTCCCTTTCTCTCCCTGCCTCC
>MAFG-chr17:81927562-81927685-DDX3X-K562
GGGCGGGGCCGCGCGCGCTCCGAGGCCGCGCGGGGCAGGGACCGGCGCGAGGGGTCCGGGCCCGGGGCTCGGAGGACTCG
CCGCCTGCGCGGGCCGGGCCGAGCGCACTGGGAAGGCCGGGCCG
```

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

Discovered Motifs

	Logo	E-value	Sites	Width
1.		6.7e-035	37	21
2.		6.9e-018	9	50
3.		2.1e-007	6	50

Stopped because requested number of motifs (3) found.

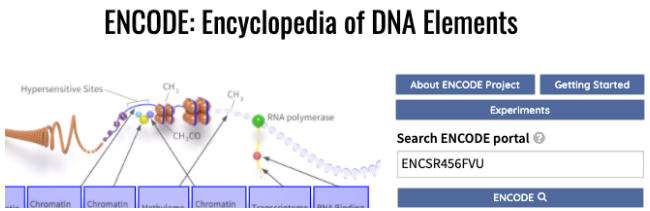
pval	leftFlank	sequence	rightFlank	motif	gene	score
1.91e-06	GTGTCTGTGC	GCTCCCGCCGCCGTCGCCGAG	CGCCC	motif.01	LDB1	5.718967

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:

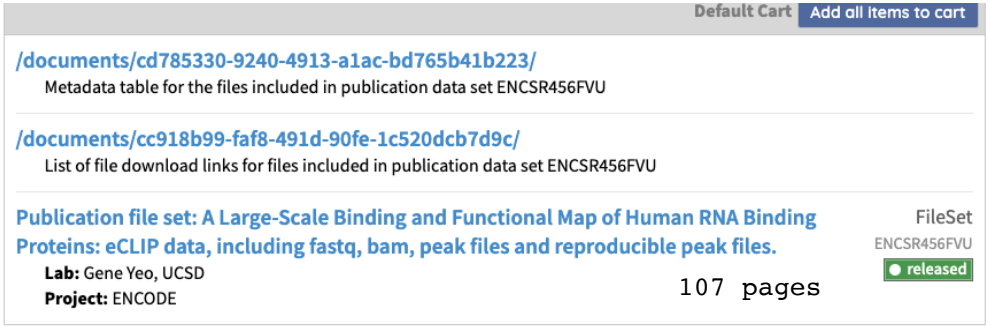
Raw and processed data sets are accessible using accession identifiers provided in Supplementary Data 2 or can be found using the following publication file set accession identifiers at the ENCODE Data Coordination Center (https://www.encodeproject.org): eCLIP (ENCSR456FVU), knockdown RNA-seq (HepG2: 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:




```
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
Dowloaded files.txt to
  ~/github/TrenaProjectErythropoiesis/inst/extdata/ddx3x/fromEncode/ENCSR456FVU/

 xargs -L 1 curl -O -J -L < files.txt
```



Reproducible peak files:
DDX3X K562 ENCFF667MXK
DDX3X HepG2 ENCFF885OWJ

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.

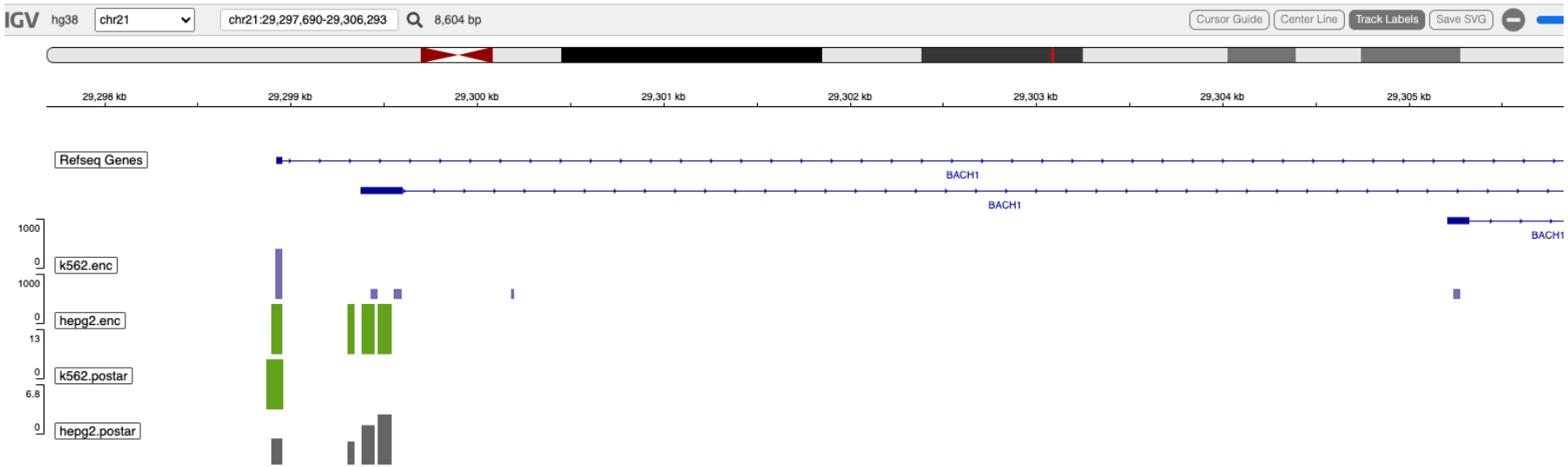
For about half of the RBPs assayed (37 of 78), we were able to identify highly enriched *k*mers of five nucleotides (nt) (k = 5) that could be clustered into a single motif. The remaining RBPs had more complex patterns of binding, best described by two motifs (32 of 78), or even three or more motifs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

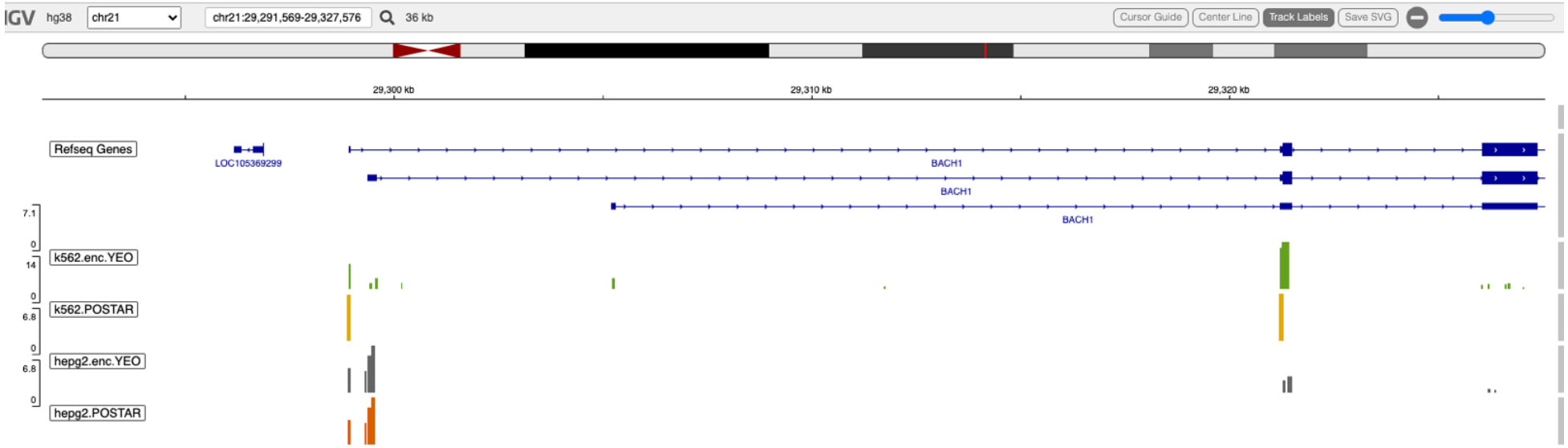
Two sources of data.

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)

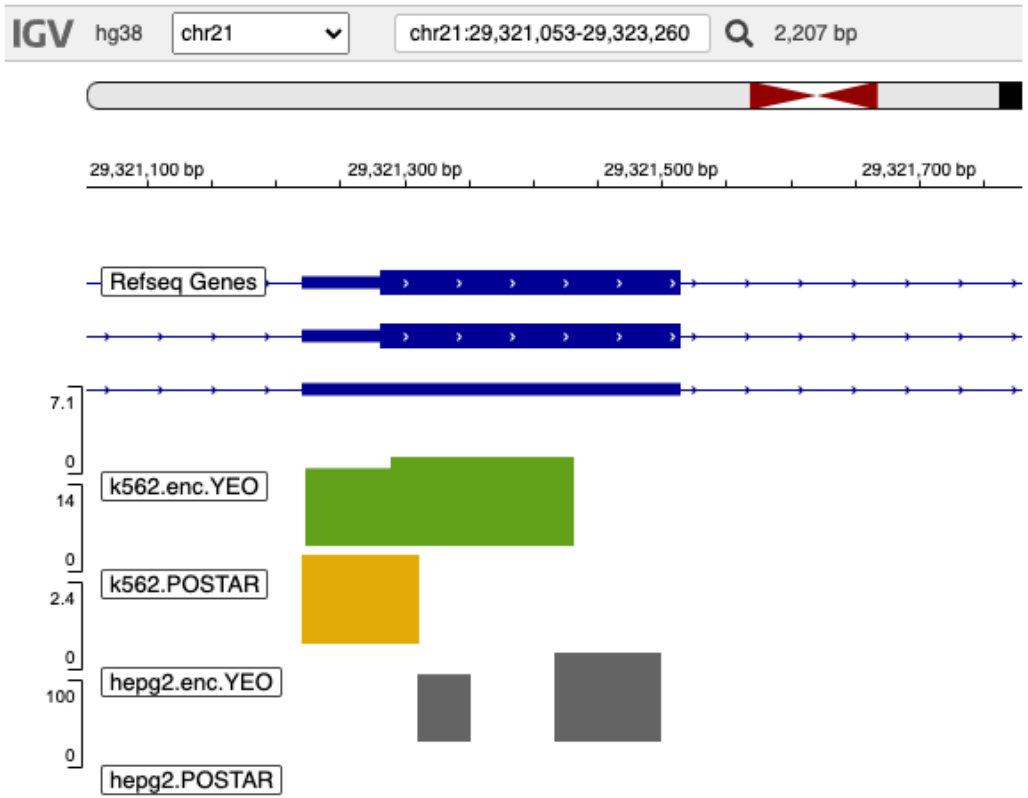
Second, from POSTAR2, by region (in the vicinity of BACH1's TSS) for the same two cell types.

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 log2(fold enrichment) and −log10(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
Immunofluoresence
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

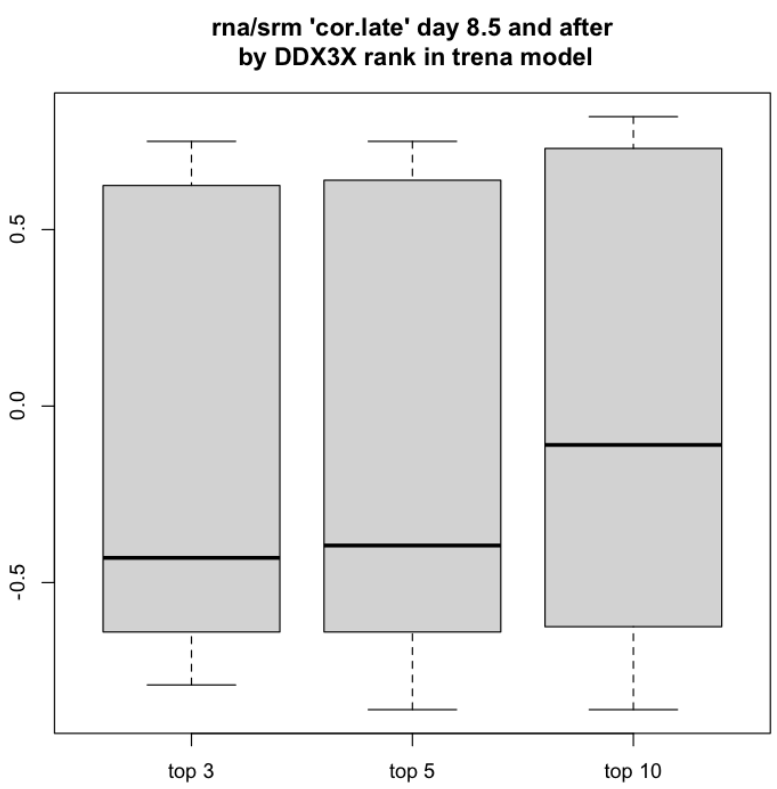
Some questions from Jeff (4 nov 2021)

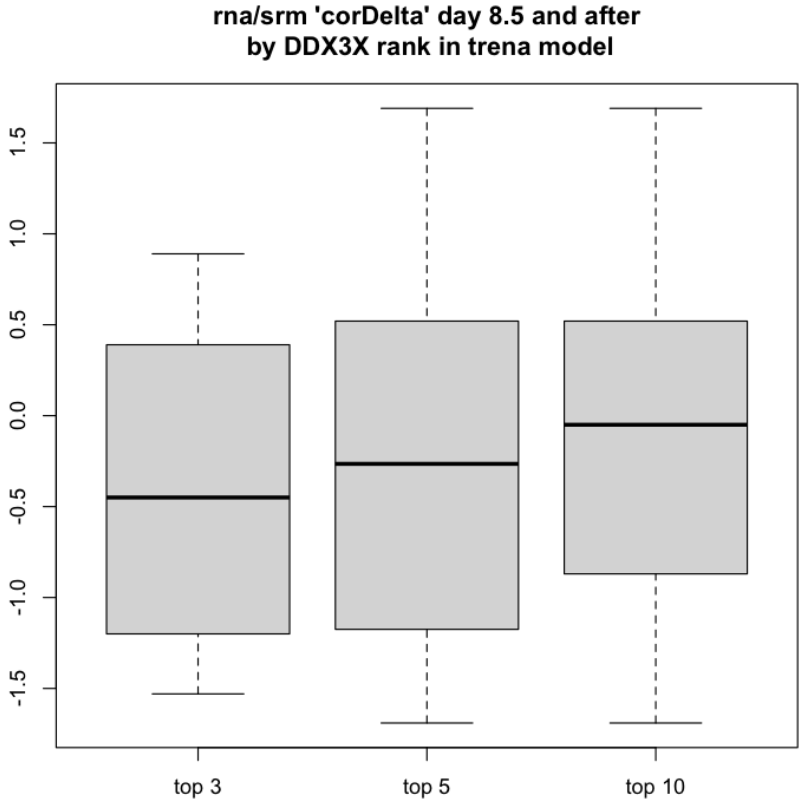
- I am thinking about a model where DDX3X represses translation in a way that doesn't necessarily involve binding to 5'UTRs. Trena RNA based models predict 36 DDX3X targets amongst our 100 genes. What are these genes? I wonder if they are enriched for discordant genes?

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 thetrena model



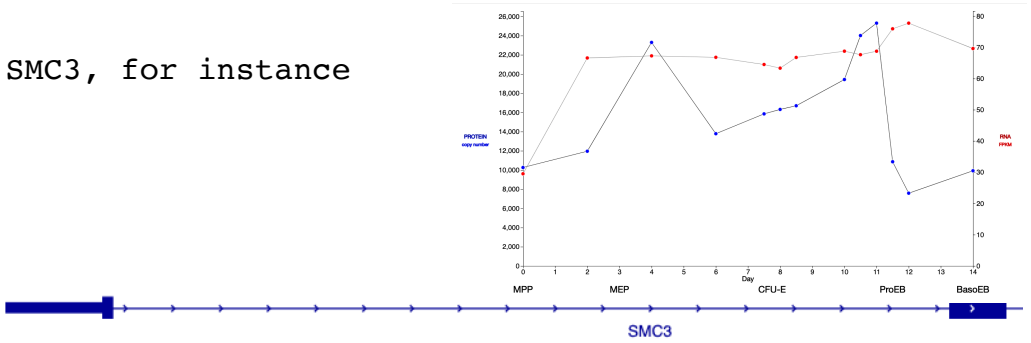


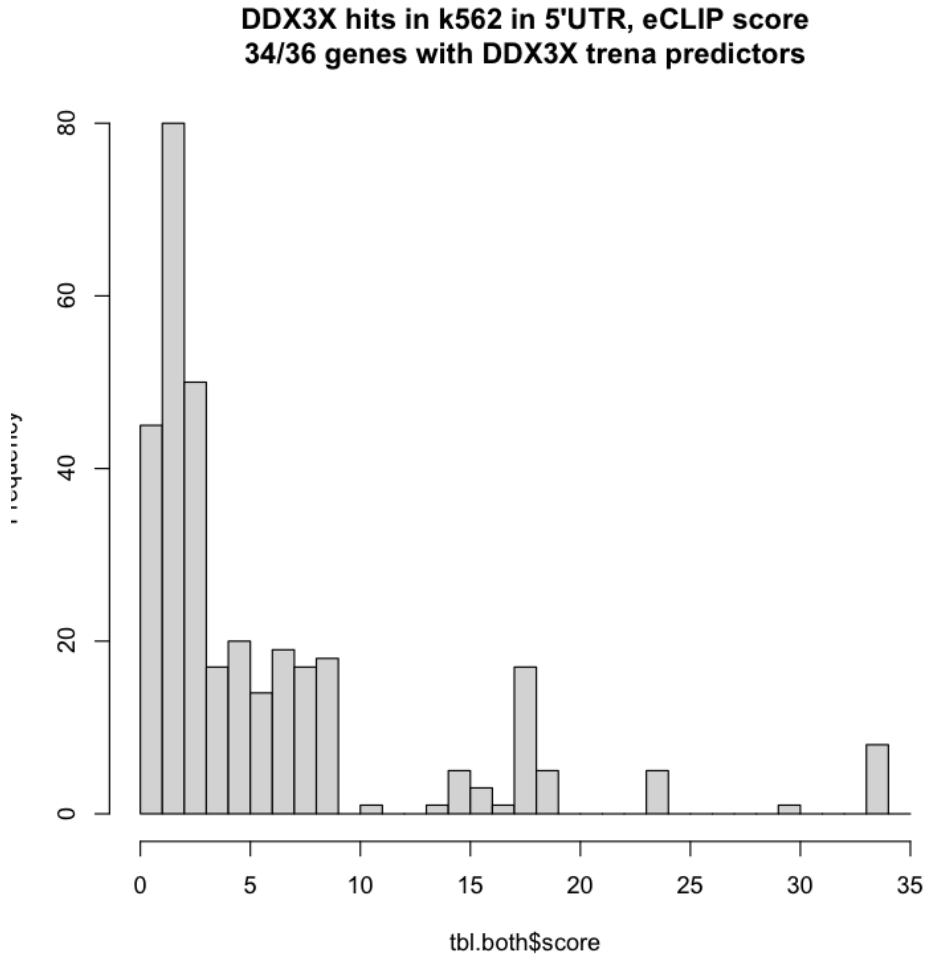
suggesting that higher trena rank for DDX3X (where mRNA predicts target protein level), leads to stronger median negative correlation.

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.

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

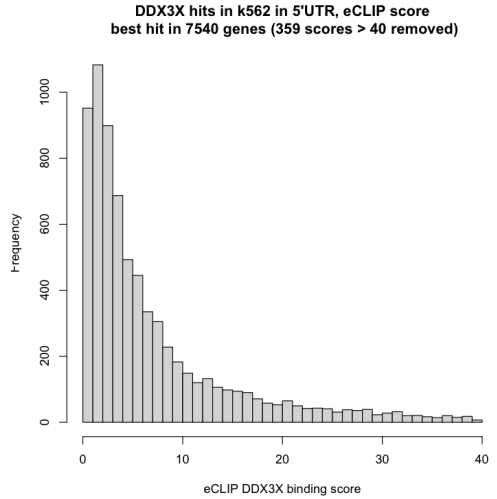
Amongst the remaining ~64 SRM genes (for which DDX3X is not a high ranking regulator) how many have 5'UTR DDX3X binding?
The score distribution would be interesting to see as well.
see my answer in the slide

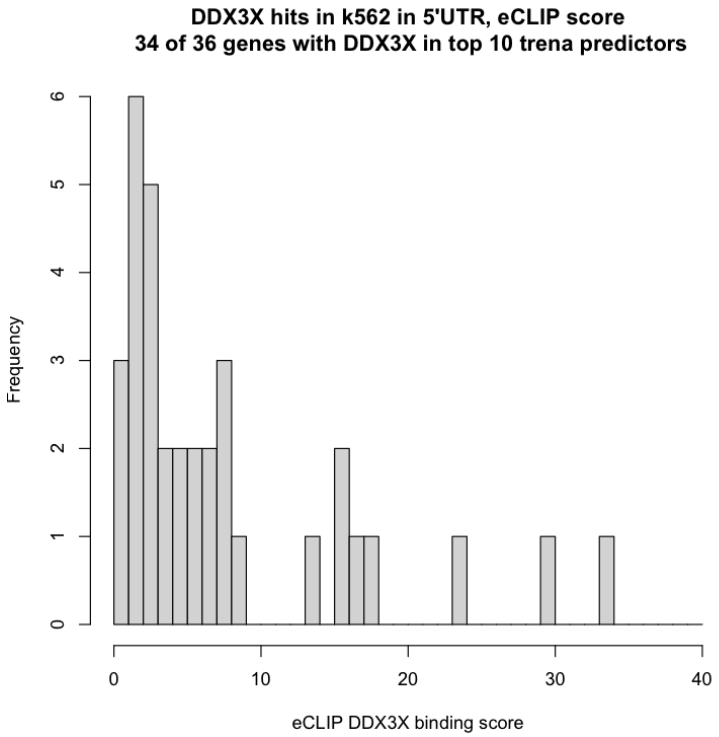
On seeing the results on the previous slide, Jeff asks (email, 513p, 4 nov 2021)

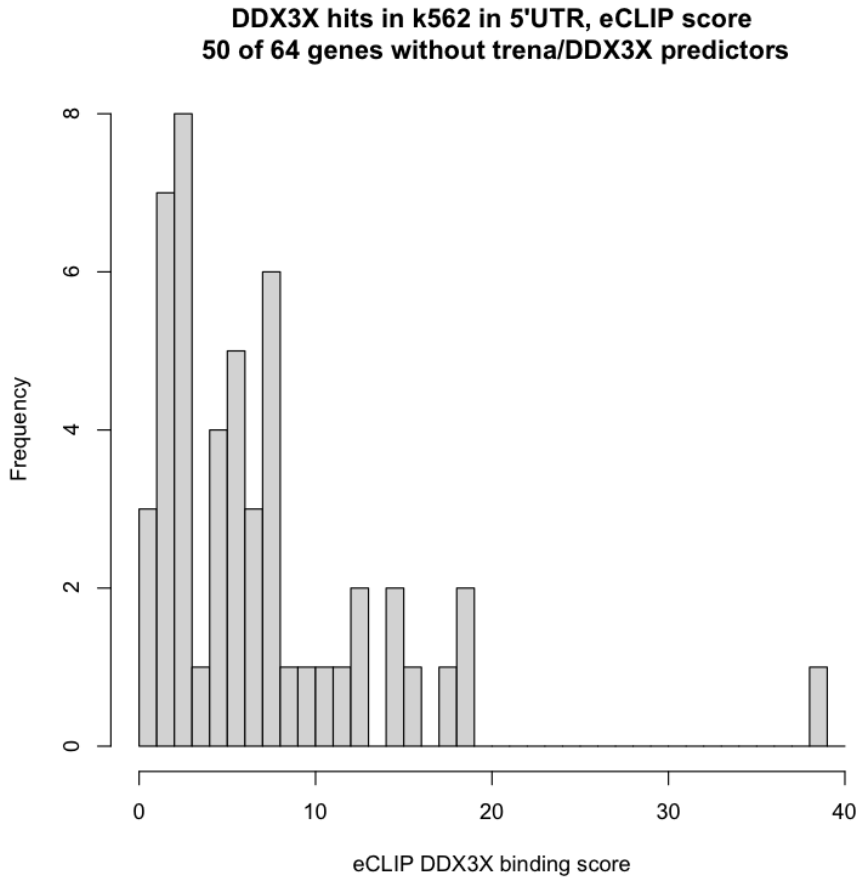
Amongst the remaining ~64 SRM genes (for which DDX3X is not a high ranking regulator) how many have 5'UTR DDX3X binding?

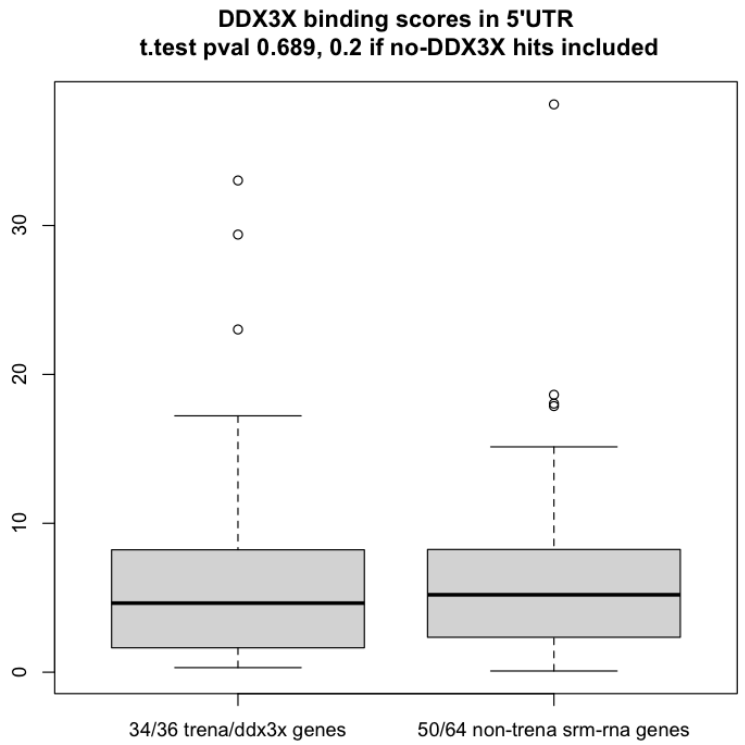
The score distribution would be interesting to see as well.

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

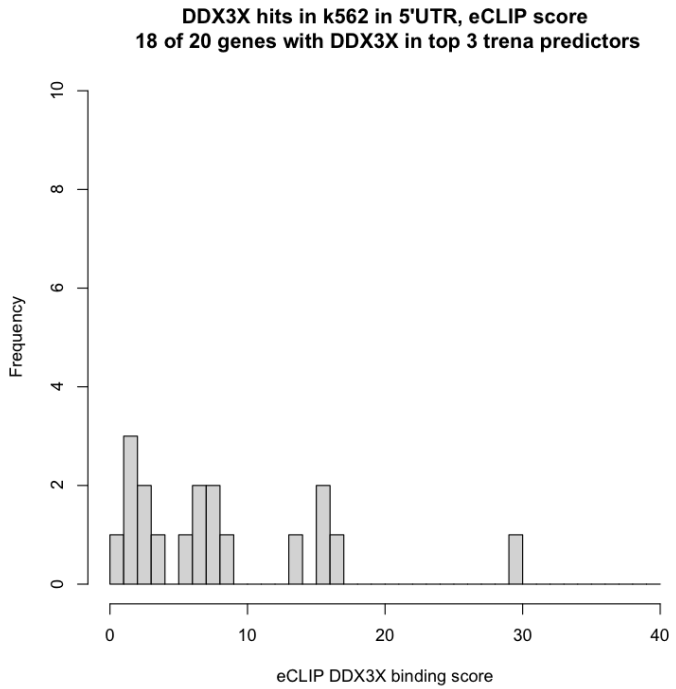


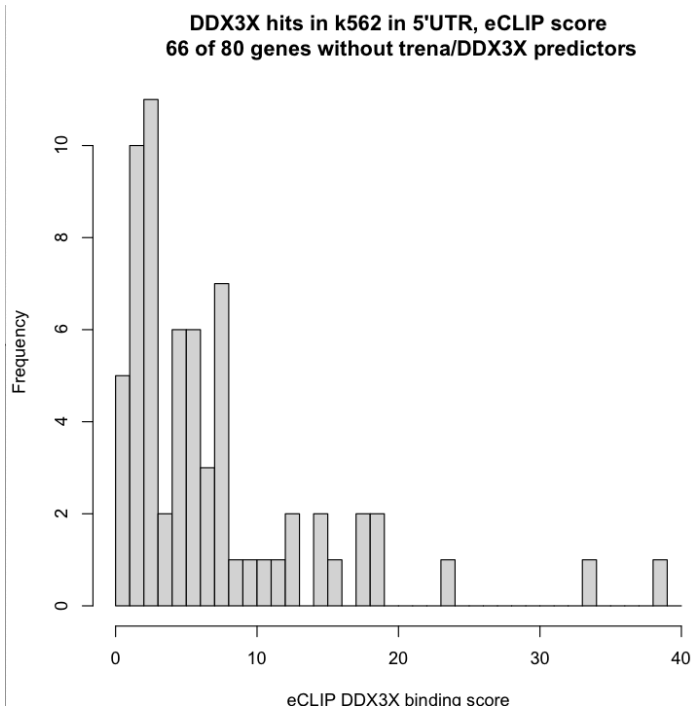


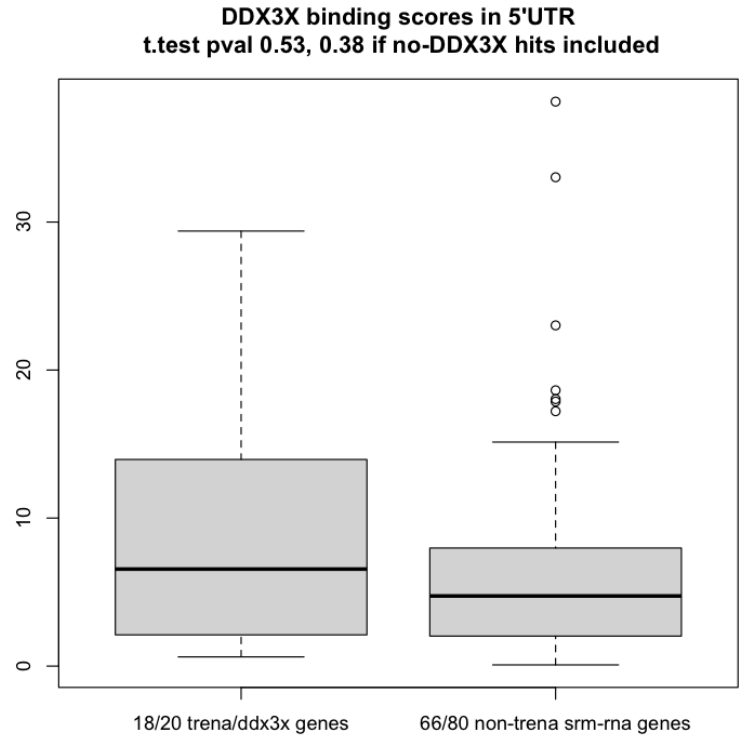




Compare DDX3X 5’UTR binding site scores in genes with and without DDX3X trena rank <= 3:
no signal

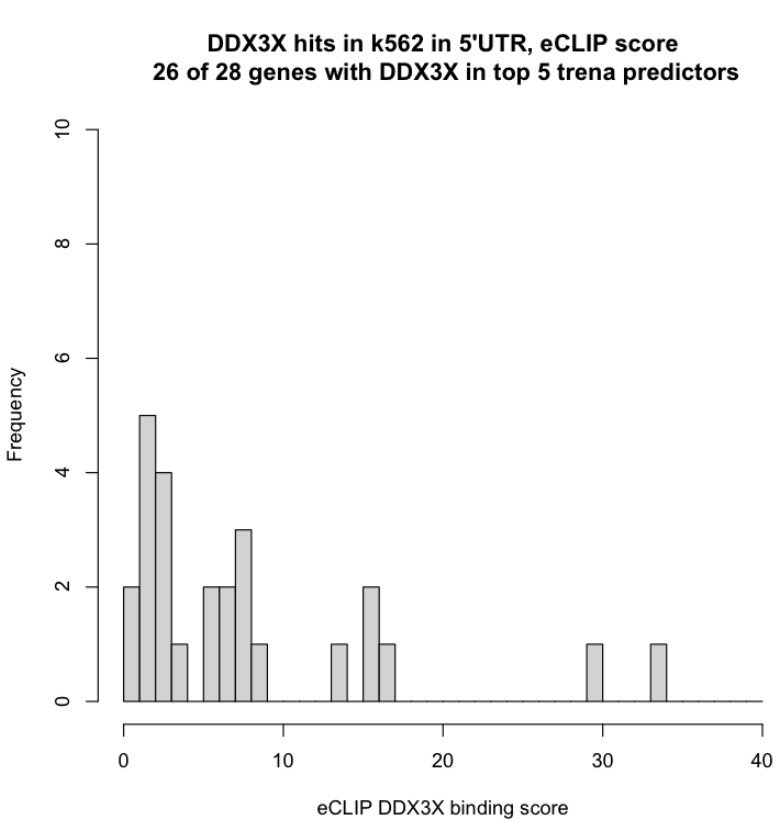


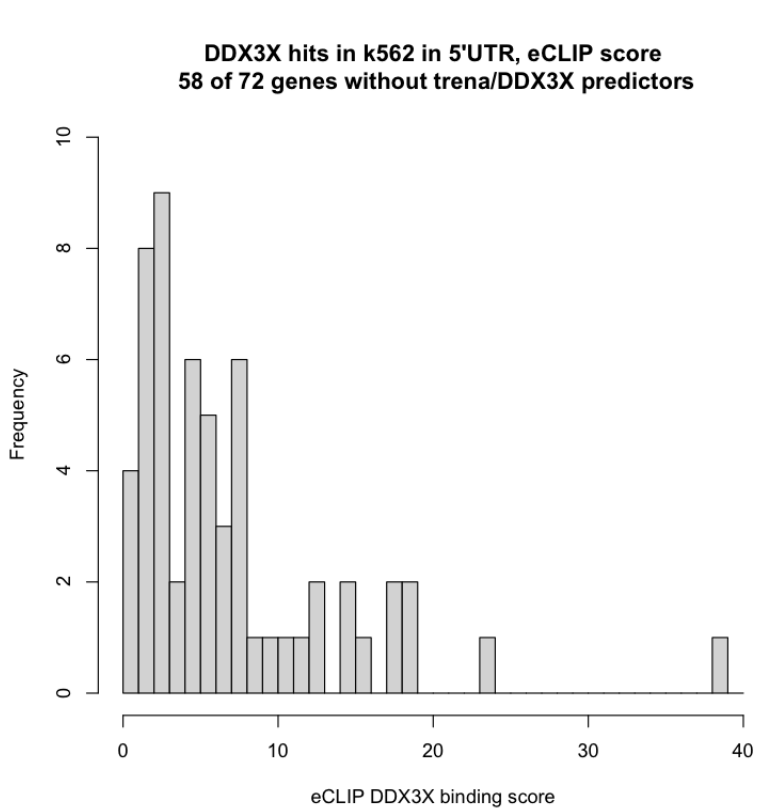


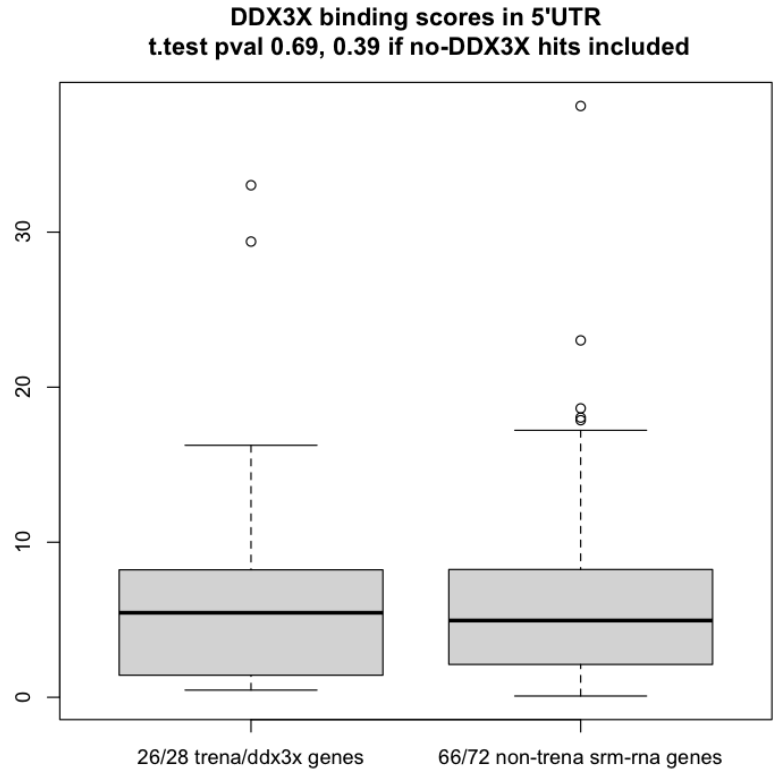


Compare DDX3X 5'UTR binding site scores in genes with and without DDX3X trena rank <= 5: again no signal

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.**





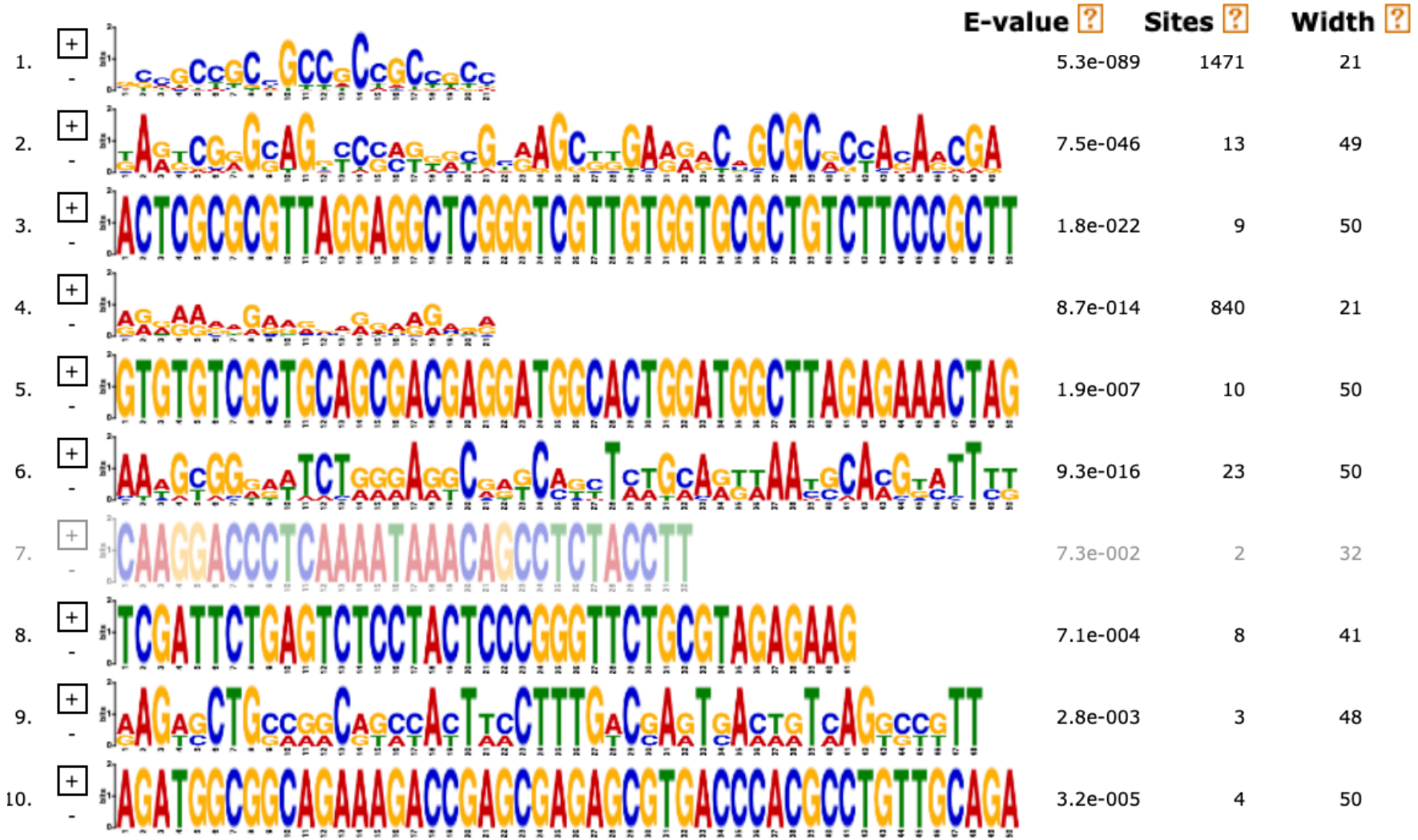


3/3 questions from Jeff (4 nov 2021)

- 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



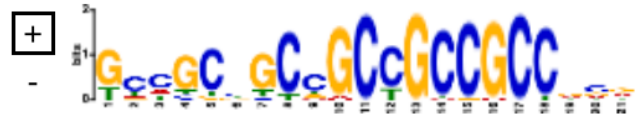
	E-value	Sites	Width
1.	5.3e-089	1471	21
2.	7.5e-046	13	49
3.	1.8e-022	9	50
4.	8.7e-014	840	21
5.	1.9e-007	10	50
6.	9.3e-016	23	50
7.	7.3e-002	2	32
8.	7.1e-004	8	41
9.	2.8e-003	3	48
10.	3.2e-005	4	50

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

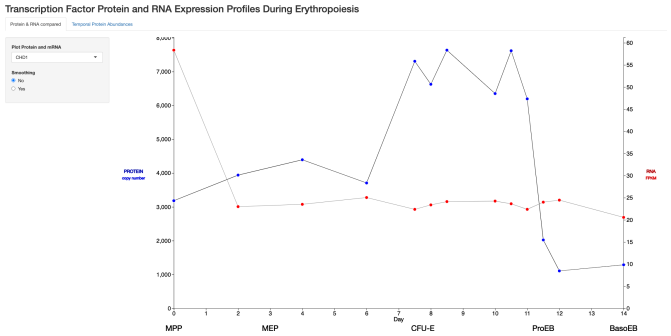
check DDX3X dominant motif & full eCLIP binding site with Dfam, maybe Arian

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.

	Logo	E-value	Sites	Width
1.		6.7e-035	37	21
1.		5.3e-089	1471	

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.



	Name	p-value	Motif Locations
4.	CEBPB-chr20:50190841-50190940-DDX3X-K562	1.20e-3	
7.	CHD1-chr5:98928817-98928954-DDX3X-K562	5.14e-4	
17.	CTCF-chr16:67562565-67562720-DDX3X-K562	7.47e-6	

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

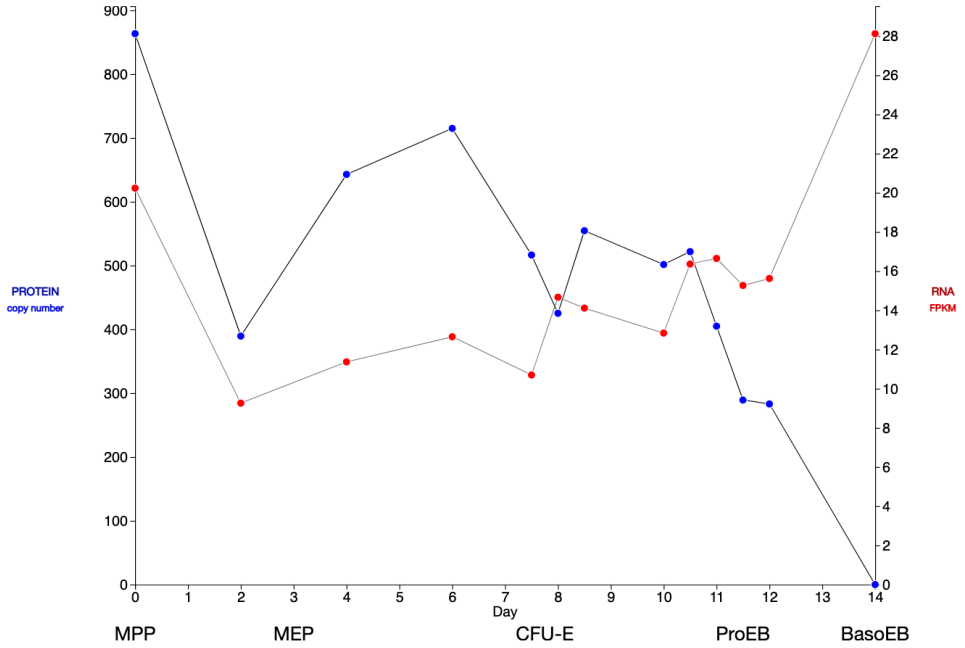
qtl_id	Lepik_2017.blood
molecular_trait_id.QTL	ENSG00000153922
chromosome.QTL	5
position.QTL	98928917
ref.QTL	A
alt.QTL	AGCC
variant.QTL	chr5_98928917_A_AGCC
ma_samples.QTL	48
maf.QTL	0.0509554
pvalue.QTL	0.0123618
beta.QTL	0.0495628
se.QTL	0.0197334
type.QTL	INDEL
ac.QTL	48
an.QTL	942
median_tpm.QTL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

Question from Jeff (11 nov 2021)

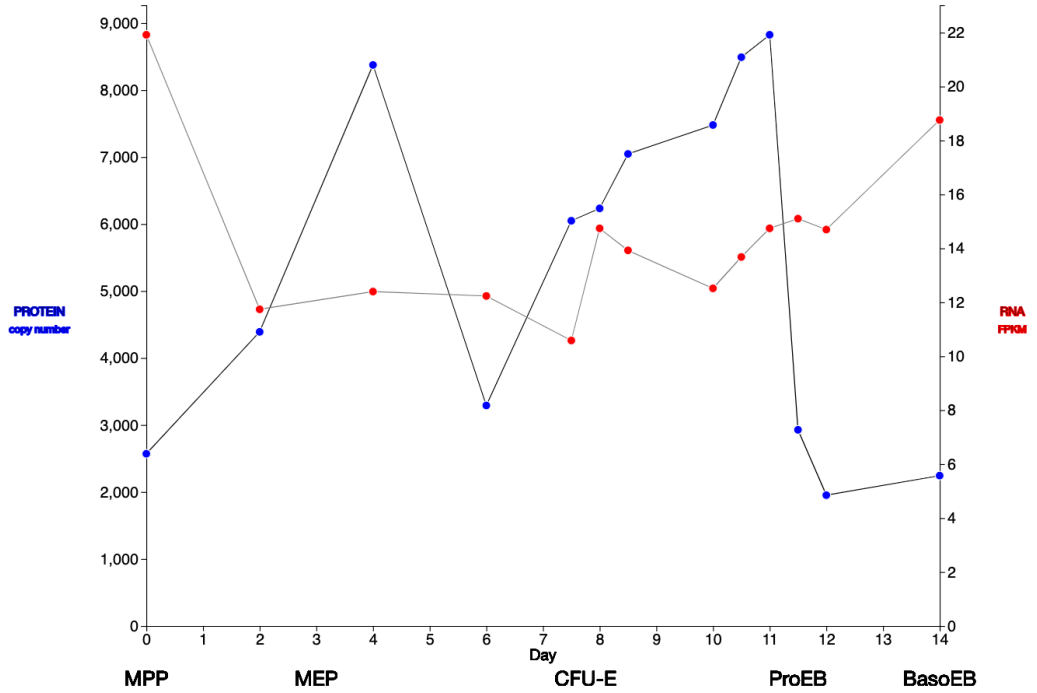
I'm curious about your analysis of DDX3X binding sites in 5'UTRs of the SRM genes. It seems like most of the genes have at least one binding site in their 5'UTR. Interestingly most of the proteins also decrease in abundance starting around day 10 or 11.

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?

BACH1, for instance:



CBP (CREBBP)



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?

	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 corrrelation 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		

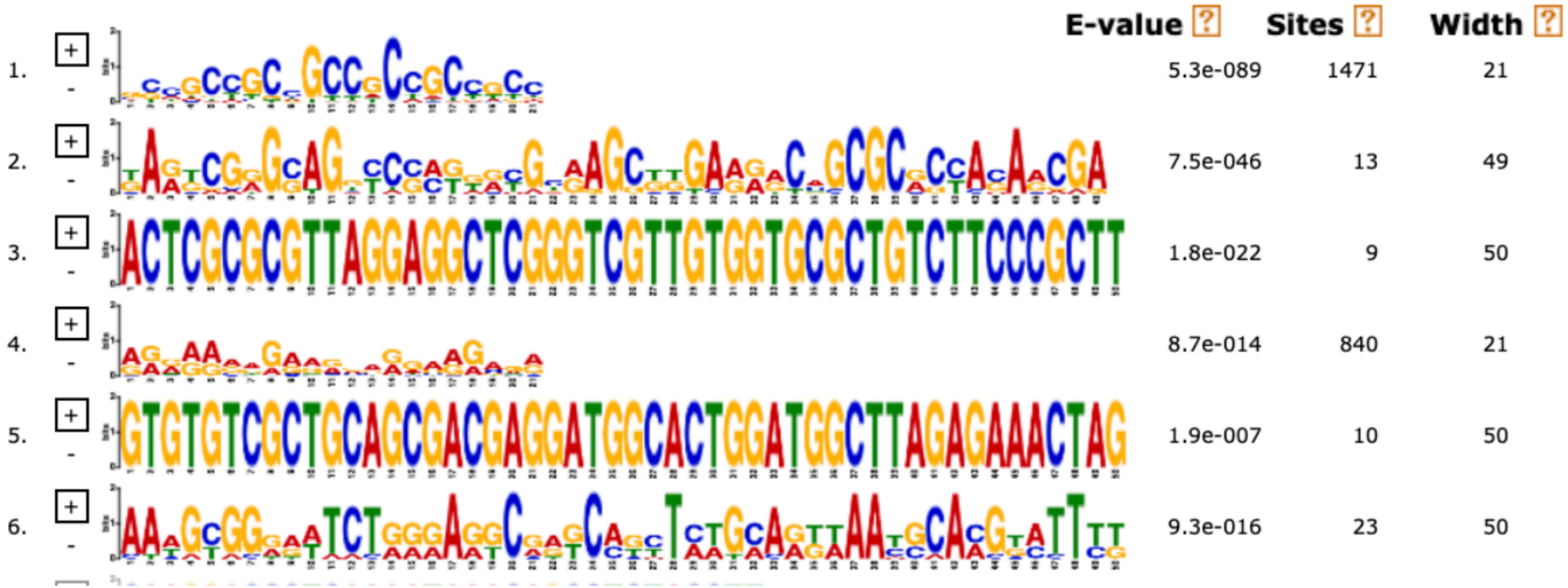
email to jeff (17 nov 2021)

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:

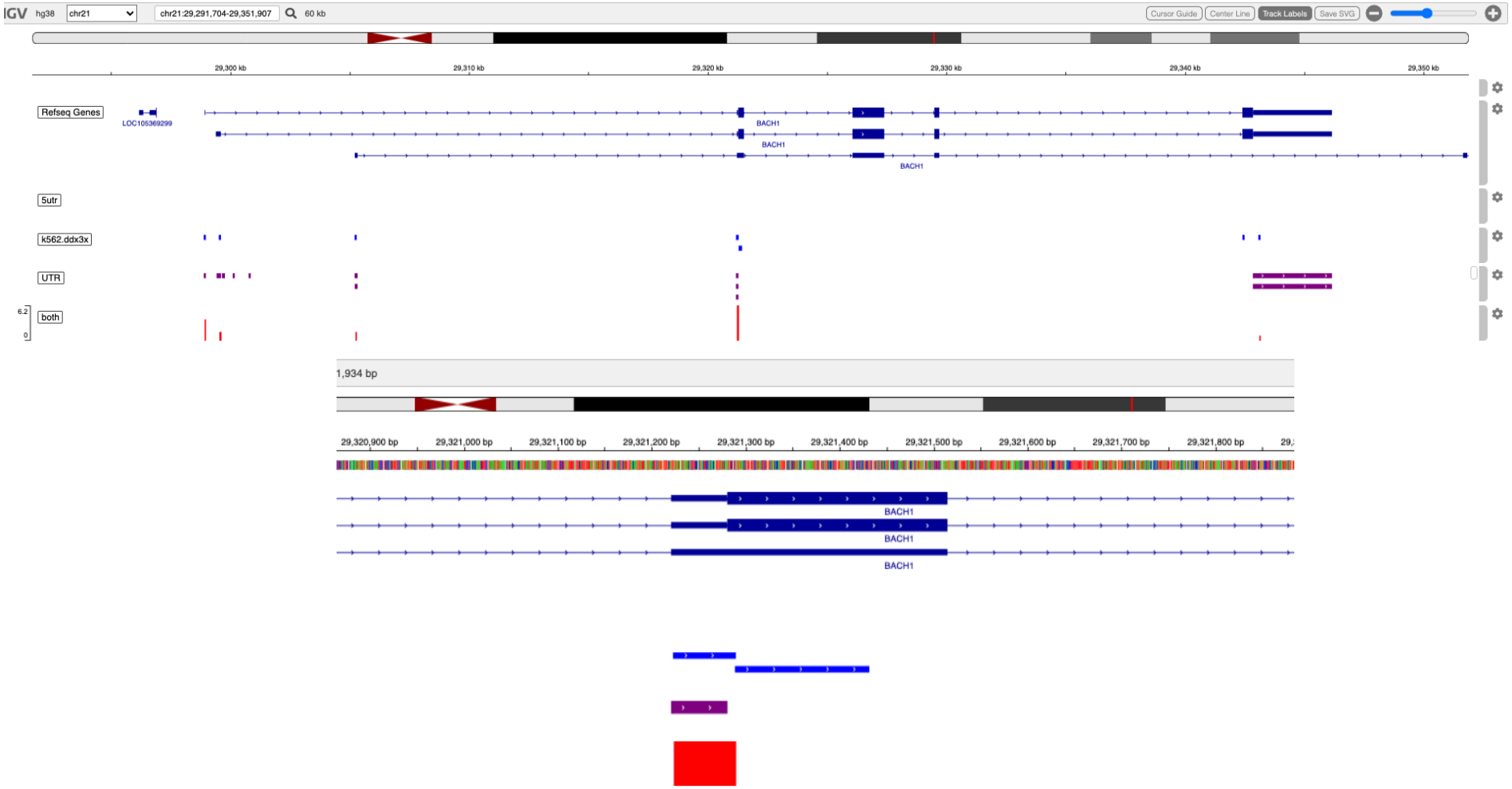
  ENCODE K562 reported 82K 5'UTR DDX3X hits across the genome, with multiple hits in many genes
  Each hit has a score, I kept only those hits in the top quartile.

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.

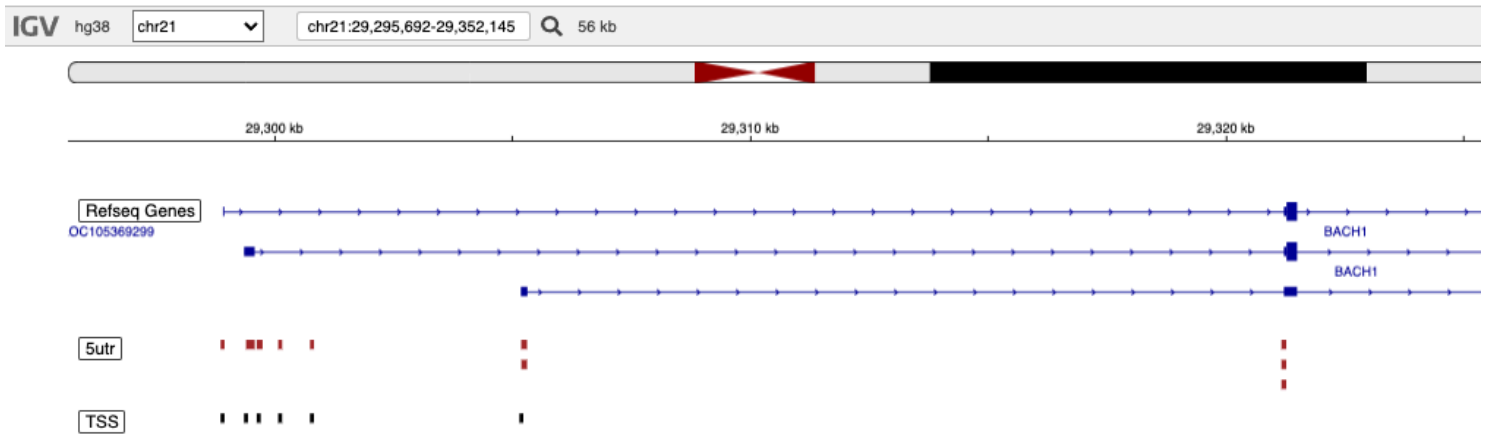
 - Paul



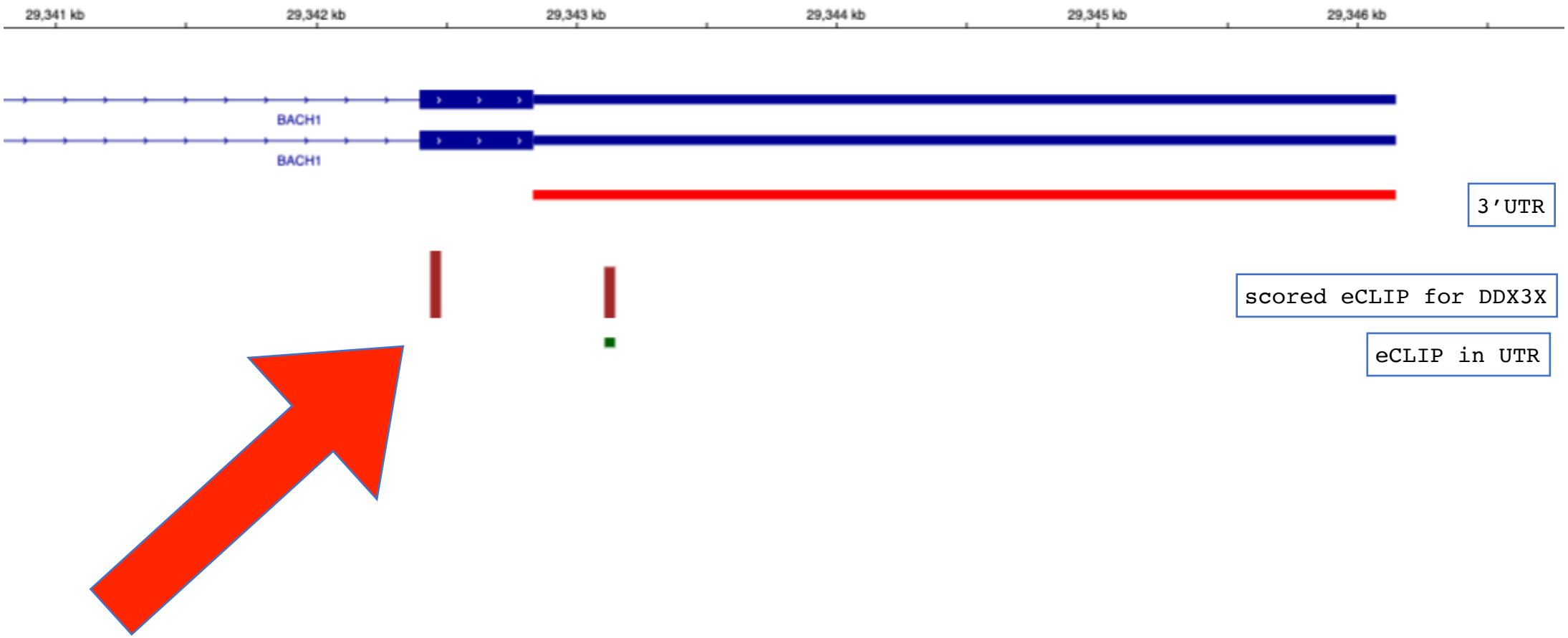
The method: find all binding sites for DDX3X (and soon, other RBPs also)
which overlap 5' or 3' UTRs. Interesting that the first BACH1 exon is annotated
as a 5'UTR. after discussing with Jeff, this resolution proposed: annotate exon intersections
(T/F) for each UTR/DDX3x intersection. should be infrequent.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



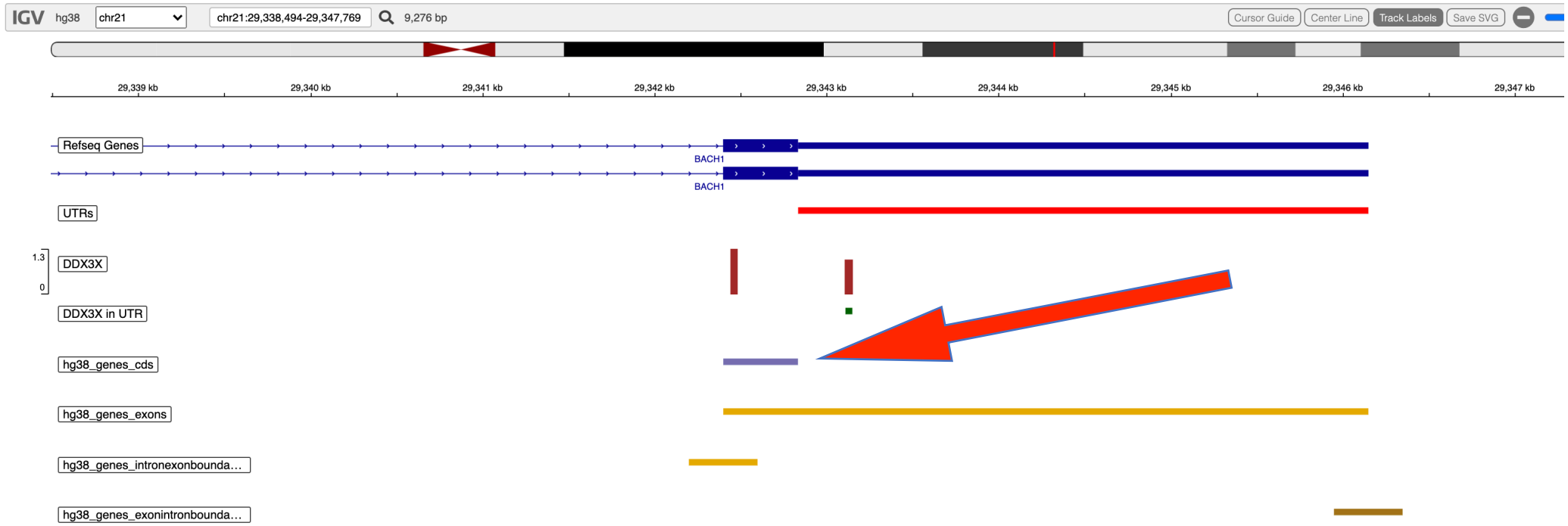
Jeff and me, email exchange (19 nov 330p); One more bit of information. I turned to Ensembl, queried them for 5'UTRs and for TSS for all BACH1 transcripts. Here's what I (Paul) got, pictured below.
We -could- drop 5'UTRs which have no TSS nearby. But I think you are saying, "keep them all, because after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



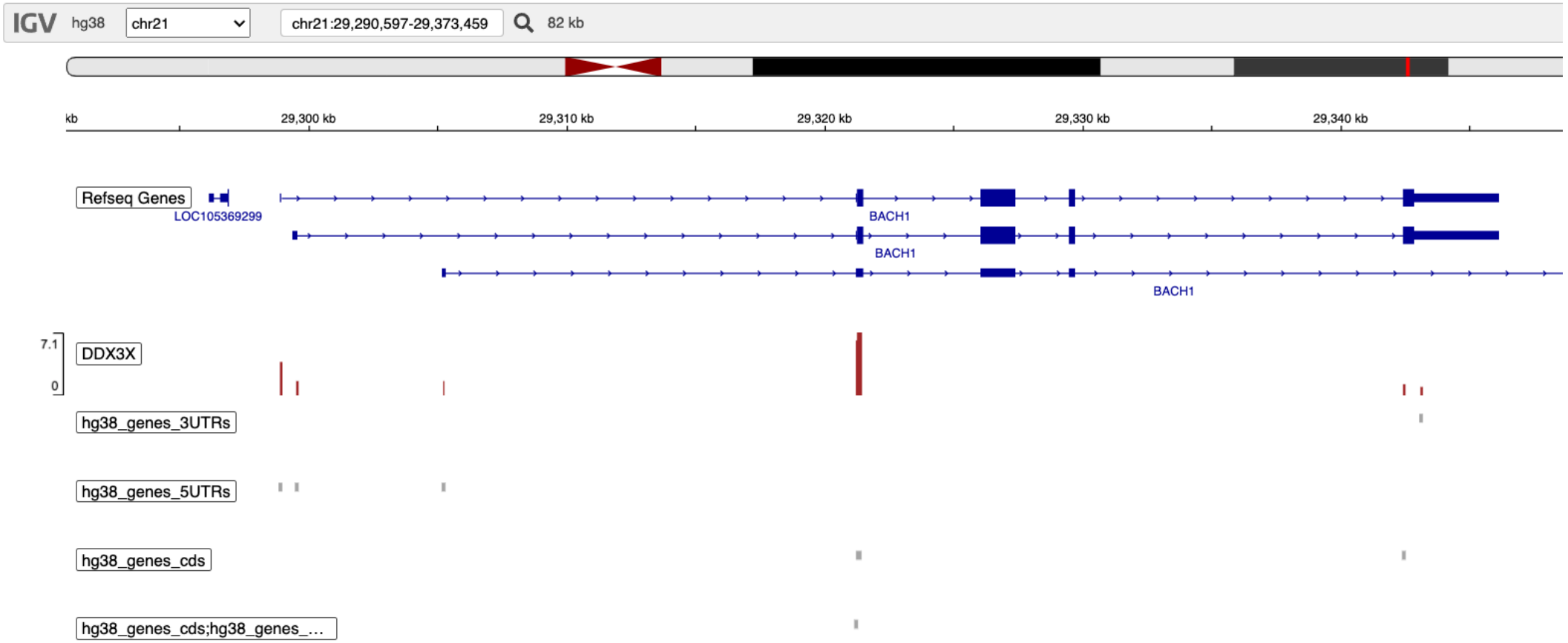
Question for Jeff: should this DDX3X binding site - pointed to by red arrow be included in our assessment of eCLIP in UTRs? not in a proper 3'UTR, but maybe interestingly functional nonetheless.



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.

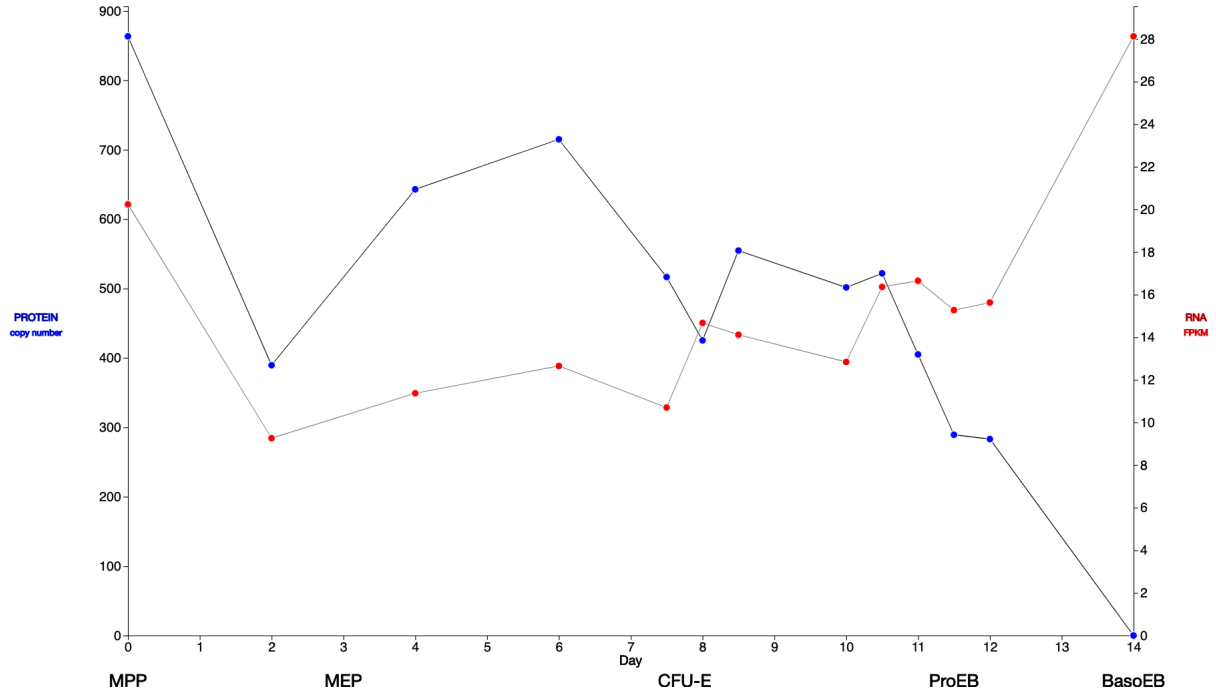
In our last meeting you reported that DDx3x could bind to 4/8 of the "no drop" genes. Could you tell which of these genes are DDX3x targets and which are not?

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

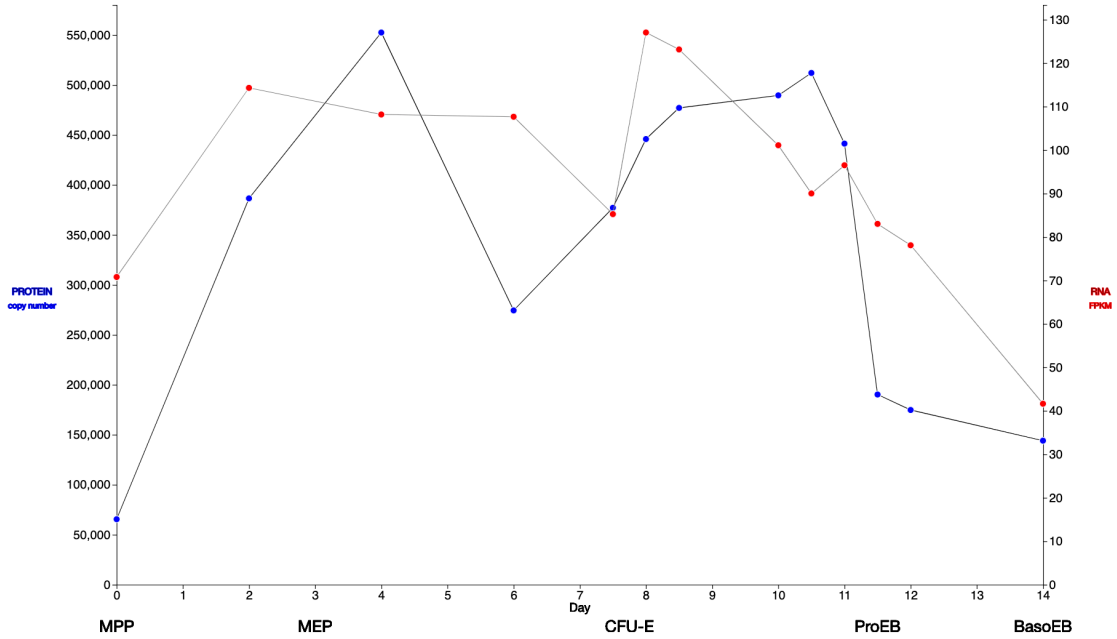
Acting on Jeff's email

clearProteinDrop	rnaProteinDiscord 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



BACH1



CHD4

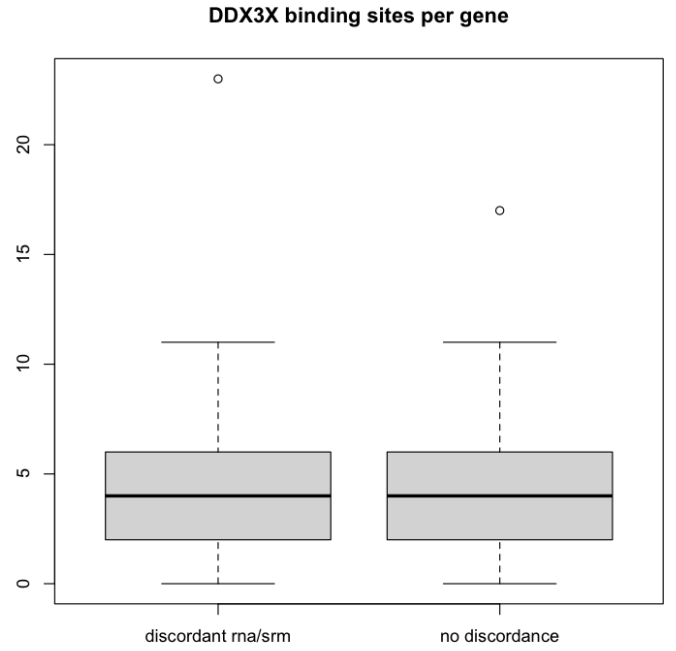
day 10-11 protein drop, with and without rna/srm discord
DDX3X binding sites do not predict discord.

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

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?

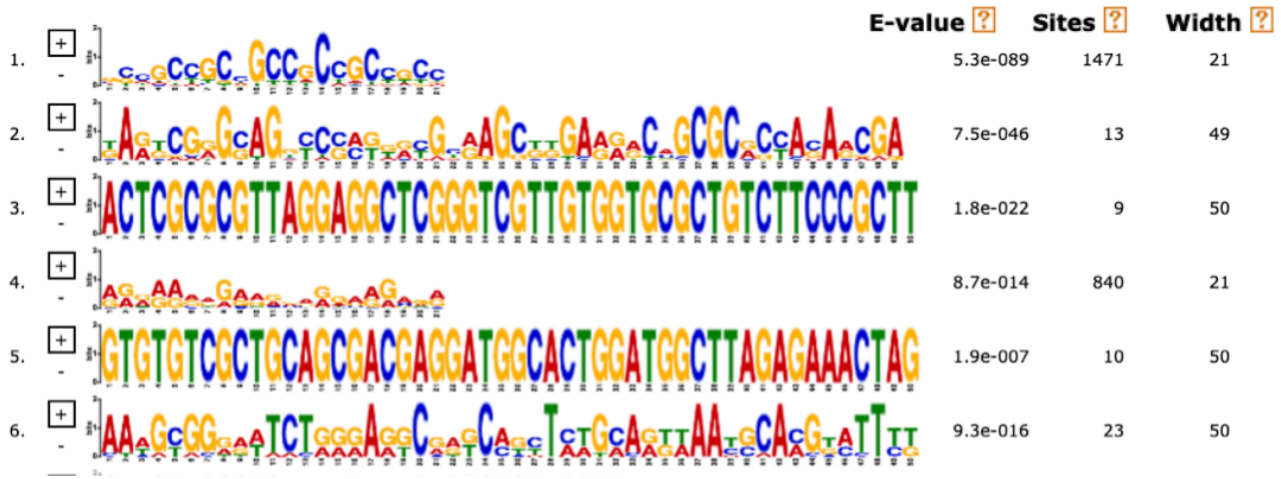
Jeff proposes these genes: ERG FLI1 GATA2 MNDA SOX6 SPI1 STAT1 STAT2 STAT3, of which ERG, FLI1, MNDA SOX6 have no binding sites.

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.

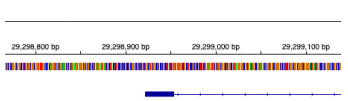
	chrom	start	end	width	score	regionType	targetGene	rbp
1	chr3	128479650	128479686	37	1.192377	hg38_genes_3UTRs	GATA2	DDX3X
2	chr3	128480413	128480438	26	1.360104	hg38_genes_3UTRs	GATA2	DDX3X
3	chr3	128485882	128485902	21	1.422952	hg38_genes_cds	GATA2	DDX3X
4	chr3	128487710	128487909	200	14.101637	hg38_genes_5UTRs	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

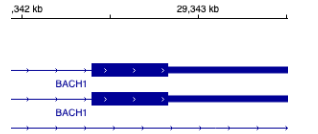
from slide 46: DDX3X binding motifs from 5'UTR in > 4k genes



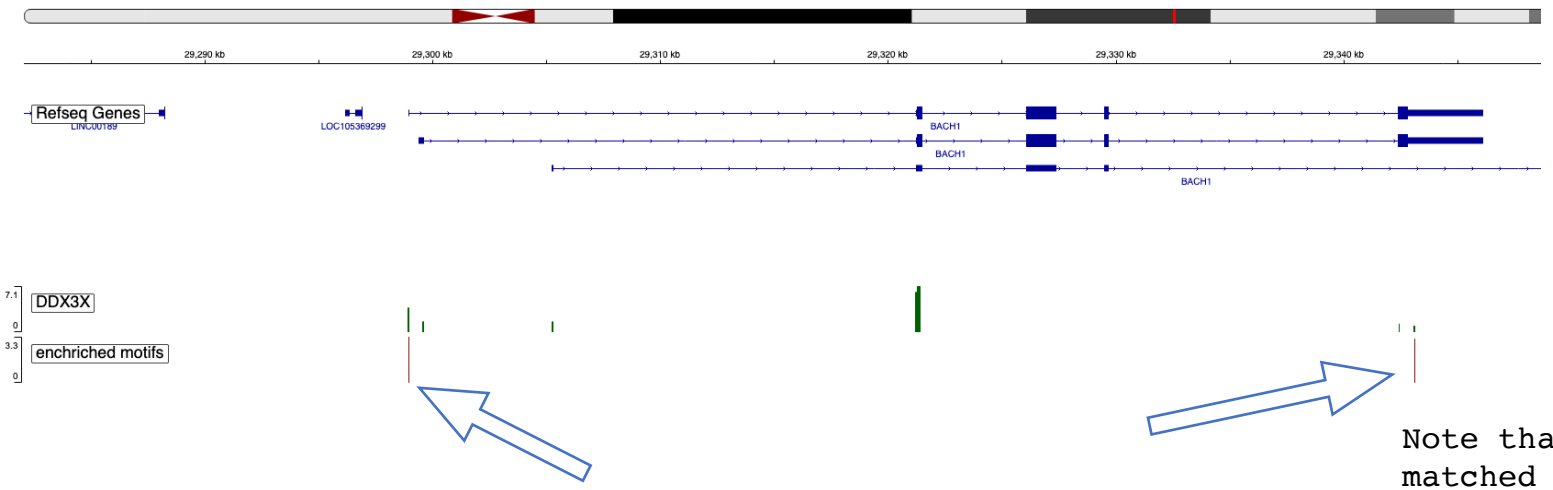
chrom	start	end	score	pval	strand	motif	sequence
chr21	29298934	29298954	3.341989	0.000455	+	ddx3x.01	CCGCGCCGCGCCTCAGCTCTG
chr21	29343125	29343145	3.206210	0.000622	+	ddx3x.04	AATAAGTAGGAGAATGAAAAC



motif 1 in 5'UTR



motif 4 in 3'UTR



chrom	start	end	width	score
chr21	**29298920**	**29298954**	**35**	**3.801192**
chr21	29299553	29299594	42	1.659932
chr21	29305242	29305275	34	1.659932
chr21	29321223	29321289	67	6.215896
chr21	29321289	29321431	143	7.116253
chr21	29342443	29342478	36	1.306399
chr21	**29343109**	**29343149**	**41**	**1.002156**

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.

email from jeff (9 dec 2021)

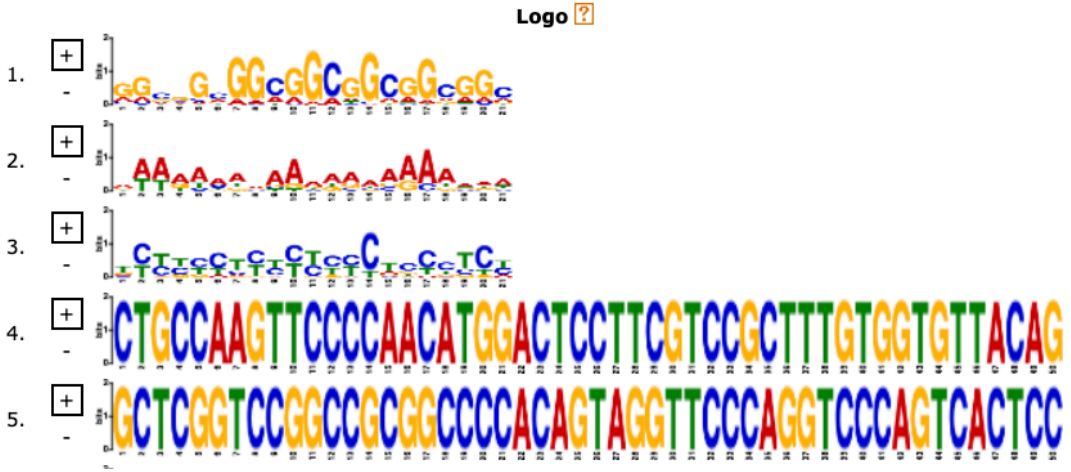
I agree that no clear correlation exists between DDX3X binding and the different group of SRM genes.

I think it’s a good idea to see if we can detect RBP motif enrichment in our different groups of genes - drop and no drop. with or without discord.

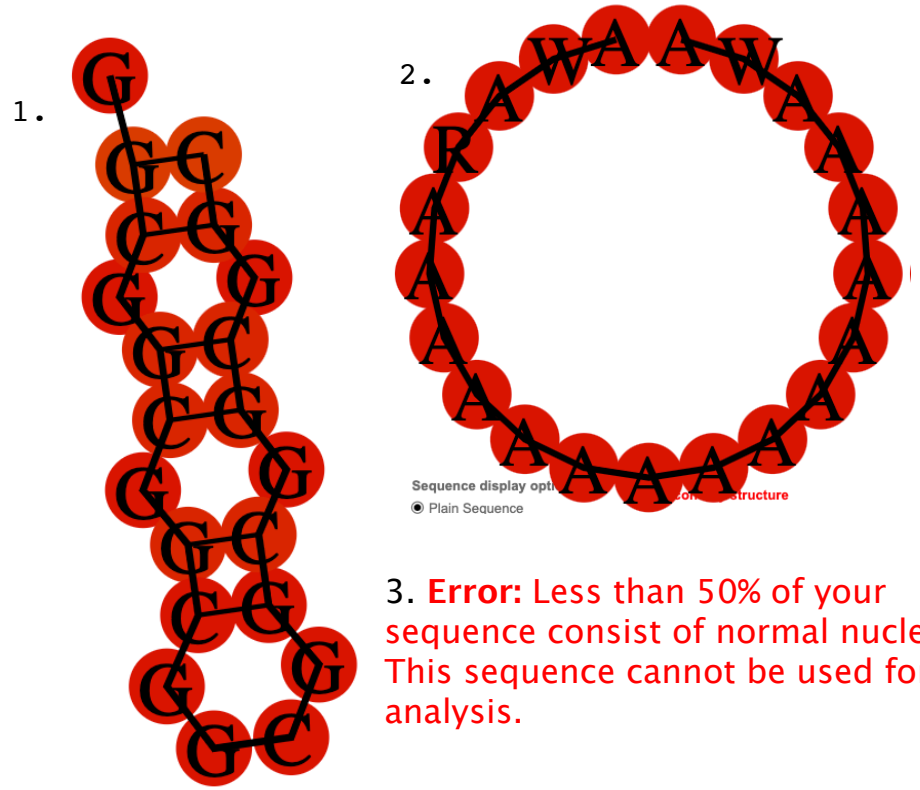
In the example that you show (slide 65)it appears that the DDX3X motif enrichment was performed on 5'UTRs. Would it be worthwhile to perform the enrichment analysis on 5' and 3' uTRs and CDSs and then use the region specific motifs to look for these sites in our gene set?

5′UTR and 3′UTR motifs

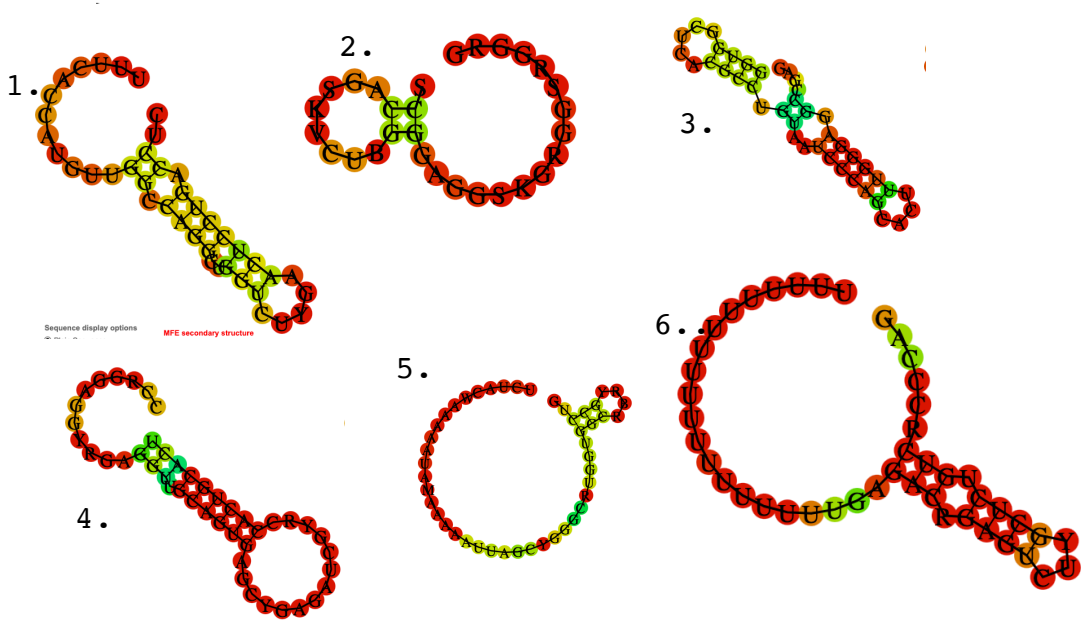
5′UTR: 32k binding sites



3′UTR: 20k binding sites



3. **Error:** Less than 50% of your sequence consist of normal nucleotides. This sequence cannot be used for analysis.



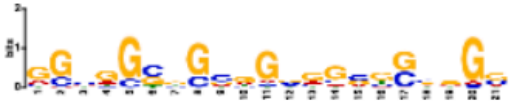
CDS - just one motif, no structure
39k binding sites, 5k of which are also annotated as UTRs

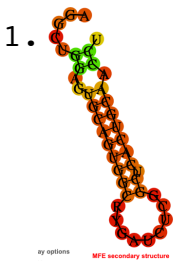
	Logo	E-value	Sites
1.		7.0e-026	22193
2.		1.9e+000	558
3.		2.0e+006	4
4.		7.5e+007	4
5.		2.1e+008	8

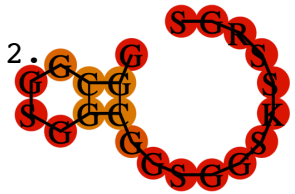
YCWBCWBCTBCWBCW

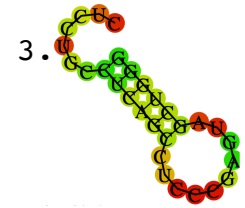
1. **Error:** Less than 50% of your sequence consist of normal nucleotides. This sequence cannot be used for analysis.

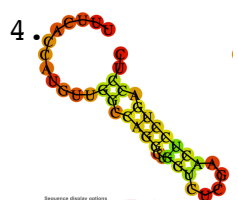
enriched motifs, 5′utr, 3′utr, cds combined: 86k binding sites

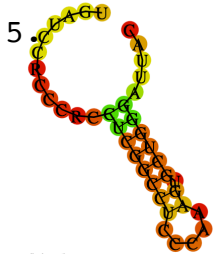
	Logo	E-value	Sites
1.		2.5e-056	411
2.		1.2e-048	43242
3.		8.4e-026	609
4.		9.3e-048	496
5.		6.0e-041	1136
6.		4.1e-032	249
7.		1.0e-020	311
8.		5.9e-007	14610

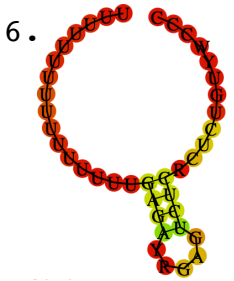
1.


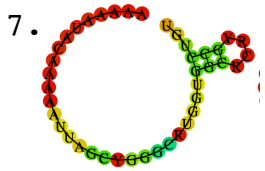
2.


3.


4.


5.


6.


7.


8 **Error:** Less than 50% of your sequence consist of normal nucleotides. This sequence cannot be used for analysis.

Neuron
Article

CRISPR-Cas9 Screens Identify the RNA Helicase DDX3X as a Repressor of *C9ORF72* (GGGGCC)n Repeat-Associated Non-AUG Translation

Cheng, Weiwei, Shaopeng Wang, Zhe Zhang, David W. Morgens, Lindsey R. Hayes, Soojin Lee, Bede Portz et al. "CRISPR-Cas9 screens identify the RNA helicase DDX3X as a repressor of C9ORF72 (GGGGCC) n repeat-associated non-AUG translation." *Neuron* 104, no. 5 (2019): 885-898.

We found that DDX3X, an RNA helicase, suppresses the repeat-associated non-AUG translation of GGGGCC repeats. DDX3X directly binds to (GGGGCC)n RNAs but not antisense (CCCCGG)n RNAs. Its helicase activity is essential for the translation repression. Reduction of DDX3X increases DPR (dipeptide repeat protein) levels in C9ORF72-ALS/FTD patient cells and enhances (GGGGCC)n-mediated toxicity in Drosophila. Elevating DDX3X expression is sufficient to decrease DPR levels

Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts

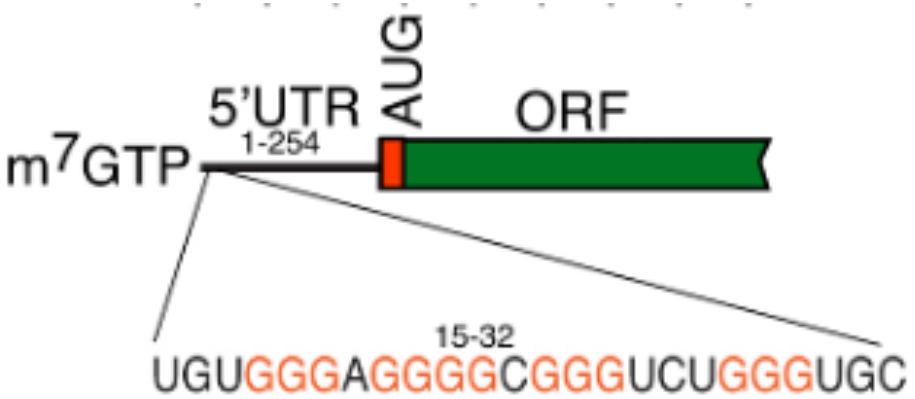
Barbara Herdy, Clemens Mayer, Dhaval Varshney, Giovanni Marsico, Pierre Murat, Chris Taylor, Clive D'Santos, David Tannahill, Shankar Balasubramanian ✉
Author Notes

Nucleic Acids Research, Volume 46, Issue 21, 30 November 2018, Pages 11592–11604,
https://doi.org/10.1093/nar/gky861

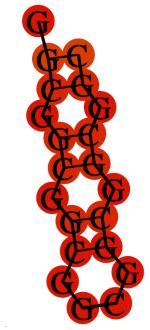
NRAS G-quadruplex (NRAS rG4)
UGUGGGAGGGGCGGGUCUGGGUGC

mutated NRAS G-quadruplex (NRAS mrG4)
UGUAGAAAGAGCAGAUCUAGAUGC

GUS mRNA stem-loop (SL)
ACAGGGCUCCGCGAUGGCGGAGCCCAA



Strongest 5'UTR motif. related to rG4, but has only paired Gs



New Direction: identify RBPs for Marjorie's KO experiments

Here are the tasks from the zoom call Jeff and I just had (22 dec 2021)

1) 4500 TMT protein measurements need interpolation with our rna-seq so that they may be compared.
   I will examine the TMT timepoints then propose (possibly) two strategies:
   a) interpolate TMT so that it has the same timepoints and replicates as rna-seq
   b) reduce the rna-seq matrix to match the TMT timepoints

2) The goals of task 1, which we can think of as matrix preparation are
   a) identify late (day 8.5-14) rna/protein discordants using simple spearman correlation
   b) build protein-protein trena models of 4500 genes

Since models are quick and easy to run, and since we are still in an exploratory mode, trying to get Marjorie strong candidates for RBP knockdown, I propose two kinds of models
   i) TFs and RBPs are included as candidate regulators only if there is some binding evidence (TF: motif match and open chromatin; RBP: eCLIPS)
  ii) remove binding evidence constraint on RBPs - use them all, thus looking only for correlation in expression

What did I miss? Or get wrong?

Revisions to this plan, from Marjorie & Jeff's feedback:

Jeff: However Marjorie asks a good question: Should we use the TMT data for the regulators, which is based on nuclear protein only, or should we use RNA for the regulators? I am now leaning towards the RNA data for the regulators until we have TMT data on both nuclear and cytoplasm since the regulators could be functioning in the cytoplasm and/or nucleus to control target protein expression.

Marjorie: Yes, I agree we should use the RNA data for the regulators until we have the TMT data on both nuclear and cytoplasm.

My conclusion: predict late time TMT levels from rna-seq

first set of models: rna-seq predicting tmt days 8-14.

First challenge: we need identical timepoints for rna & tmt

tmt colnames: d2 d7.5 d8 d10 d12 d14

rna-seq colnames:

day0.r1	day0.r2	day2.r1	day2.r2	day4.r1	day4.r2	day6.r1	day6.r2
day7.r1	day7.r2	day8.r1	day8.r2	day8.r1.2	day8.r2.2	day10.r1	day10.r2
day10.r1.2	day10.r2.2	day11.r1	day11.r2	day11.r1.2	day11.r2.2	day12.r1	day12.r2
day14.r1	day14.r2	day16.r1	day16.r2				

strategy:

from the rna-seq:
- average replicates from day2, day8, day10 (4 of them), day12, day14
- drop day0, day6, day11, day16
- treat day7 as day7.5?

first set of models: rna-seq predicting tmt days 8-14

First challenge: we need identical timepoints for rna & tmt. Marjorie and Jeff propose averaging rna-seq day2.r[12] to create a single day2, similarly for 8, 10, 12 and 14, and day[78].r[12] to for an approximation to day 7.5

tmt colnames: d2 d7.5 d8 d10 d12 d14

rna-seq colnames:

day0.r1	day0.r2	day2.r1	day2.r2	day4.r1	day4.r2	day6.r1	day6.r2
day7.r1	day7.r2	day8.r1	day8.r2	day8.r1.2	day8.r2.2	day10.r1	day10.r2
day10.r1.2	day10.r2.2	day11.r1	day11.r2	day11.r1.2	day11.r2.2	day12.r1	day12.r2
day14.r1	day14.r2	day16.r1	day16.r2				

Second challenge, deferred for now, since predictors (tfs and rbps) are rna. The problem, by example: HNRNPD codes for 4 RBPs, with very different tmt levels.

#	uniprot	ensg	d2	d7.5	d8	d10	d12	d14
# 404	Q14103	ENSG00000138668.18	7600.6	9826.0	10021.0	10197.6	6345.5	5576.1
# 649	H0YA96	ENSG00000138668.18	225.0	413.4	385.5	392.9	218.0	83.4
# 1232	D6RF44	ENSG00000138668.18	14.6	434.3	282.3	249.2	133.9	30.3
# 1550	D6RBQ9	ENSG00000138668.18	999.5	1203.8	1049.7	1199.5	1012.3	1065.6

Jeff & Marjorie's solution, by example:

```
 tbl.idMap[grep("HNRNPD$", tbl.idMap$SYMBOL),]
     UNIPROTID SYMBOL          name
293      D6RBQ9 HNRNPD HNRNPD.D6RBQ9
304      D6RF44 HNRNPD HNRNPD.D6RF44
557      H0YA96 HNRNPD HNRNPD.H0YA96
2402     Q14103 HNRNPD HNRNPD.Q14103
```

KLF1.Q13351 fpkm.tmt, mtx.fpkm.tmt.31760x6, d8, d10, d12, d14

rna filtered on mean > 100 & sd > 20

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	RPS3	-0.313	-0.107	-1.0	-0.937	521.936	0.201	rbp	1	KLF1.Q13351
2	PCBP2	-0.075	-0.083	-0.4	-0.854	487.219	0.000	rbp	2	KLF1.Q13351
3	JUND	0.000	0.045	0.8	0.739	608.855	0.555	tf	3	KLF1.Q13351
4	SERBP1	0.000	-0.067	-0.4	-0.591	181.868	0.000	rbp	4	KLF1.Q13351
5	METAP2	0.000	-0.110	-0.2	-0.589	610.770	0.000	rbp	5	KLF1.Q13351
6	MYC	0.000	-0.033	-0.4	-0.552	164.848	0.000	tf	6	KLF1.Q13351
7	TAL1	0.000	0.015	0.4	0.458	144.799	0.000	tf	7	KLF1.Q13351
8	U2AF2	0.000	-0.055	-0.2	-0.435	162.776	0.000	rbp	8	KLF1.Q13351
9	JUNB	0.000	0.014	0.4	0.326	171.512	0.000	tf	9	KLF1.Q13351
10	NFE2	0.000	0.005	0.4	0.258	109.990	0.000	tf	10	KLF1.Q13351
11	E2F4	0.000	0.009	0.4	0.236	206.079	0.000	tf	11	KLF1.Q13351
12	KLF1	0.000	0.002	0.4	0.214	249.073	0.243	tf	12	KLF1.Q13351

rna filtered on mean > 50 & sd > 10

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	RPS3	-0.502	-0.113	-1.0	-0.937	663.840	0.201	rbp	1	KLF1.Q13351
2	PCBP2	-0.210	-0.086	-0.4	-0.854	471.391	0.000	rbp	2	KLF1.Q13351
3	SERBP1	0.000	-0.072	-0.4	-0.591	193.782	0.000	rbp	3	KLF1.Q13351
4	METAP2	0.000	-0.113	-0.2	-0.589	586.050	0.555	rbp	4	KLF1.Q13351
5	TARDBP	0.000	-0.169	-0.4	-0.553	260.304	0.000	rbp	5	KLF1.Q13351
6	RTCB	0.000	-0.104	-0.2	-0.501	249.110	0.000	rbp	6	KLF1.Q13351
7	U2AF2	0.000	-0.061	-0.2	-0.435	182.008	0.000	rbp	7	KLF1.Q13351
8	NUDT21	0.000	-0.088	-0.2	-0.372	159.730	0.000	rbp	8	KLF1.Q13351
9	SF3A3	0.000	-0.027	-0.2	-0.280	160.265	0.000	rbp	9	KLF1.Q13351
10	GTF2F1	0.000	-0.079	-0.4	-0.256	241.722	0.000	rbp	10	KLF1.Q13351
11	KLF1	0.000	0.003	0.4	0.214	146.395	0.243	tf	11	KLF1.Q13351
12	YBX3	0.000	-0.026	-0.4	-0.186	164.845	0.000	rbp	12	KLF1.Q13351

rna filtered on mean > 20 & sd > 5

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	RPS3	-0.435	-0.102	-1.0	-0.937	333.629	0.201	rbp	1	KLF1.Q13351
2	PCBP2	-0.162	-0.078	-0.4	-0.854	526.545	0.000	rbp	2	KLF1.Q13351
3	SF3B1	0.000	-0.651	-0.8	-0.818	477.451	0.000	rbp	3	KLF1.Q13351
4	PPIG	0.000	-0.299	-0.8	-0.672	458.496	0.000	rbp	4	KLF1.Q13351
5	SLTM	0.000	-0.292	-0.4	-0.657	56.755	0.000	rbp	5	KLF1.Q13351
6	SERBP1	0.000	-0.057	-0.4	-0.591	125.721	0.000	rbp	6	KLF1.Q13351
7	METAP2	0.000	-0.102	-0.2	-0.589	528.639	0.555	rbp	7	KLF1.Q13351
8	TARDBP	0.000	-0.131	-0.4	-0.553	105.115	0.000	rbp	8	KLF1.Q13351
9	XRN2	0.000	-0.104	-0.4	-0.518	128.029	0.000	rbp	9	KLF1.Q13351
10	RTCB	0.000	-0.082	-0.2	-0.501	64.667	0.000	rbp	10	KLF1.Q13351
11	U2AF2	0.000	-0.045	-0.2	-0.435	60.929	0.000	rbp	11	KLF1.Q13351
12	EFTUD2	0.000	-0.079	-0.2	-0.417	87.225	0.000	rbp	12	KLF1.Q13351

Marjorie suggests (Jeff concurs):

It's not easy to choose between the 3 conditions. Maybe the second set of criteria (mean>50 sd>10) would be reasonable in terms of stringency.

I was surprised that DDX3X was not part of the list. I am wondering what the difference between these new TRENA models, and the very first ones where we identified DDX3X (on p.4)?

My reply: A few things explain the absence of DDX3X from the model.

1) I ordered the slide 75's trena model by pearson correlation, sometimes a bad idea since outliers can create a distorted picture. Spearman is better, but you can see that there is not much resolution in that column. Here is the table sorted on random forest score, with DDX3X at rank 7.
2) DDX3X comes in at rank 14 in the second table on slide 75, and I showed only the first 12 rows.
3) The KLF1p model in slide 5 predicts srm from rna-seq, using timepoints D8_5, D10, D10_5, D11, D11_5, D12, D14, with cor(KLF1p, DDX3X, method="spearman") -0.87, pearson at -0.93. If I include **D8**, these drop to -0.75 and -0.71 respectively, and DDX3X betaLasso falls to 0. Our TMT data is more sparse and starts at **D8**. For KLF1p and DDX3X, the strong anti-correlation is seen with SRM, and with 7 timepoints starting at day **8.5**.
4) Using 4 TMT timepoints (days 8, 10, 12, 14), see below, sorting by rfScore, DDX3X shows up at rank 7; has no betaLasso; and lags in betaRidge.
5) My experience, and my hunch, is that credible RBPs or TFs will have strong scores in random forest, spearman, and betaRidge. A non-zero betaLasso adds extra confidence, telling us that it can be a proxy for other regulators which have a similar expression pattern.
6) We -may- see a similar de-emphasis of DDX3X in the other TMT protein models.
7) 7 late time point SRM vs. 4 late time point TMT: which are you more confident in?

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	RPS3	-0.502	-0.113	-1.0	-0.937	663.840	0.201	rbp	1	KLF1.Q13351
2	METAP2	0.000	-0.113	-0.2	-0.589	586.050	0.555	rbp	2	KLF1.Q13351
3	PCBP2	-0.210	-0.086	-0.4	-0.854	471.391	0.000	rbp	3	KLF1.Q13351
4	SF3B4	0.000	0.004	-0.4	-0.133	261.558	0.000	rbp	4	KLF1.Q13351
5	TARDBP	0.000	-0.169	-0.4	-0.553	260.304	0.000	rbp	5	KLF1.Q13351
6	RTCB	0.000	-0.104	-0.2	-0.501	249.110	0.000	rbp	6	KLF1.Q13351
7	DDX3X	0.000	-0.099	0.4	-0.143	245.835	0.000	rbp	7	KLF1.Q13351
8	GTF2F1	0.000	-0.079	-0.4	-0.256	241.722	0.000	rbp	8	KLF1.Q13351
9	SERBP1	0.000	-0.072	-0.4	-0.591	193.782	0.000	rbp	9	KLF1.Q13351
10	U2AF2	0.000	-0.061	-0.2	-0.435	182.008	0.000	rbp	10	KLF1.Q13351
11	TFDP1	0.000	-0.001	0.4	0.093	171.968	0.000	tf	11	KLF1.Q13351
12	YBX3	0.000	-0.026	-0.4	-0.186	164.845	0.000	rbp	12	KLF1.Q13351
13	SF3A3	0.000	-0.027	-0.2	-0.280	160.265	0.000	rbp	13	KLF1.Q13351
14	NUDT21	0.000	-0.088	-0.2	-0.372	159.730	0.000	rbp	14	KLF1.Q13351
15	KLF1	0.000	0.003	0.4	0.214	146.395	0.243	tf	15	KLF1.Q13351
16	HNRNPC	0.000	0.000	-0.2	-0.167	134.515	0.000	rbp	16	KLF1.Q13351

Jeff and Marjorie reply

1) Would it help to have more data points in the TMT data by interpolating the values between these time points?. This would add 3 time points, 10.5, 11, 11.5.

Interpolation adds no new information to the TMT data, but it might help out indirectly. The interpolated TMT columns allow us to retain the corresponding (actually measured) timepoints from rna-seq.

2) If we are not satisfied with the current TMT data, and we still want to try to use the TMT data to identify candidate regulators of genes that drop in expression between days 10 and 12, would it be worthwhile to try an approach that doesn't rely on correlated expression as much as Trena? We could use the TMT and RNAseq data to identify the genes that display discordant expression between days 10 and 12. Then we can identify all RBPs that are increasing in protein expression between days 10 and 12. Next we can intersect eCLIPs data for these regulators and the discordant genes to identify candidate regulators. Maybe not as powerful as Trena, but also maybe a way to give us a list of candidate regulators with the low resolution TMT data. We can also check for in vitro defined RBP motifs amongst the eCLIPs sites to boost confidence.

Task 1: add interpolated 10.5, 11, 11.5 columns to mtx.tmt

Re-weave TMT1 data with rna-seq, a second try: this time interpolating missing TMT timepoints rather than excluding unmatched rna-seq timepoints

model with TMT timepoints interpolated: D8, D10, D10_5, D11, D11_5, D12, D14. order by rfScore

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	GFI1B	0.000	-0.047	-0.9	-0.917	222.531	0.000	tf	1	KLF1.Q13351
2	GATA1	0.000	-0.056	-0.9	-0.916	220.624	0.000	tf	2	KLF1.Q13351
3	TFDP1	0.000	-0.032	-0.9	-0.893	216.489	0.000	tf	3	KLF1.Q13351
4	XBP1	0.104	0.084	0.8	0.909	208.828	0.000	tf	4	KLF1.Q13351
5	KLF1	-0.014	-0.020	-1.0	-0.937	199.952	0.000	tf	5	KLF1.Q13351
6	MYB	0.017	0.084	0.9	0.940	195.396	0.000	tf	6	KLF1.Q13351
7	METAP2	0.000	-0.065	-0.9	-0.932	187.363	0.000	rbp	7	KLF1.Q13351
8	ELF1	0.000	0.097	0.9	0.907	183.398	0.000	tf	8	KLF1.Q13351
9	SF3B1	0.000	-0.073	-0.8	-0.784	164.413	0.000	rbp	9	KLF1.Q13351
10	GATA2	0.385	0.039	1.0	0.962	160.667	0.000	tf	10	KLF1.Q13351
11	NFE2	0.000	-0.035	-0.9	-0.918	156.803	0.000	tf	11	KLF1.Q13351
12	HNRNPC	0.000	0.032	0.8	0.894	143.994	0.000	rbp	12	KLF1.Q13351
13	DDX3X	0.000	-0.060	-0.6	-0.803	132.992	0.000	rbp	13	KLF1.Q13351
14	ATF4	0.000	-0.046	-0.6	-0.133	131.423	0.615	tf	14	KLF1.Q13351
15	MYBL2	0.000	-0.064	-0.9	-0.882	129.229	0.000	tf	15	KLF1.Q13351
16	SF3A3	0.000	0.117	0.8	0.856	126.309	0.000	rbp	16	KLF1.Q13351
17	TCF3	0.000	-0.111	-0.7	-0.854	123.019	0.000	tf	17	KLF1.Q13351
18	TAL1	0.000	-0.050	-0.7	-0.879	111.308	0.000	tf	18	KLF1.Q13351
19	RPS3	0.000	-0.029	-0.7	-0.474	106.853	0.000	rbp	19	KLF1.Q13351
20	JUNB	0.000	-0.049	-0.9	-0.886	96.491	0.000	tf	20	KLF1.Q13351

model with TMT timepoints interpolated: D8, D10, D10_5, D11, D11_5, D12. ordered by abs(betaRidge)

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	LYL1	0.000	0.157	0.543	0.460	74.520	0.000	tf	1	KLF1.Q13351
2	SF3A3	0.000	0.121	0.771	0.852	141.832	0.000	rbp	2	KLF1.Q13351
3	SF3B1	0.000	-0.117	-0.771	-0.778	128.641	0.000	rbp	3	KLF1.Q13351
4	SERBP1	0.000	0.091	0.086	-0.031	56.330	0.000	rbp	4	KLF1.Q13351
5	ELF1	0.000	0.079	0.771	0.856	145.462	0.000	tf	5	KLF1.Q13351
6	TCF3	0.000	-0.078	-0.600	-0.790	83.438	0.000	tf	6	KLF1.Q13351
7	DDX3X	0.000	-0.073	-0.600	-0.804	95.911	0.000	rbp	7	KLF1.Q13351
8	XBP1	0.000	0.064	0.657	0.860	179.600	0.000	tf	8	KLF1.Q13351
9	SREBF1	0.000	-0.063	-0.771	-0.654	125.450	0.000	tf	9	KLF1.Q13351
10	MYBL2	0.000	-0.052	-0.771	-0.837	80.453	0.000	tf	10	KLF1.Q13351
11	TAL1	0.000	-0.046	-0.543	-0.855	95.853	0.000	tf	11	KLF1.Q13351
12	GATA2	0.345	0.039	1.000	0.944	193.664	0.427	tf	12	KLF1.Q13351
13	JUNB	0.000	-0.039	-0.771	-0.842	185.663	0.000	tf	13	KLF1.Q13351
14	HNRNPC	0.014	0.039	0.714	0.897	127.623	0.000	rbp	14	KLF1.Q13351
15	METAP2	0.000	-0.037	-0.771	-0.815	182.460	0.000	rbp	15	KLF1.Q13351
16	E2F4	0.000	-0.035	-0.600	-0.799	134.528	0.405	tf	16	KLF1.Q13351
17	MAZ	0.000	-0.034	-0.486	-0.491	38.499	0.000	tf	17	KLF1.Q13351
18	GATA1	0.000	-0.031	-0.771	-0.798	171.445	0.000	tf	18	KLF1.Q13351
19	RPS3	0.000	-0.030	-0.829	-0.485	107.771	0.000	rbp	19	KLF1.Q13351
20	MYB	0.000	0.022	0.771	0.697	208.348	0.000	tf	20	KLF1.Q13351

EIF3D

slide 4 shows EIF3D in 3/91 late-time, rna-seq/srm models (DDX3X has 36). Jeff asks:

Since EIF3D is another RBP of interest, and we don't know ground truth for any RBP, I think it might be helpful to compare how the models based on interpolation of TMT data compare to the SRM-based models for an EIF3D target (slide 4) [that is, with EIF3D as a regulator - pshannon]

gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
EIF3D	364.947	28.983	0.893	0.922	116133750.76	0	rbp	3	GATAD2Bp
EIF3D	0.000	1.444	0.714	0.840	14842.83	0	rbp	7	GTF2F2p
EIF3D	31.892	2.983	0.857	0.946	542278.74	0	rbp	5	PSIP1p

first result - not promising

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	PAX5	0	-762.171	-0.900	-0.928	815.860	0	tf	1	GATAD2B.Q8WXI9
2	ELF4	0	0.327	0.900	0.911	764.718	0	tf	2	GATAD2B.Q8WXI9
3	WTAP	0	-0.206	-0.800	-0.796	729.220	0	rbp	3	GATAD2B.Q8WXI9
4	NONO	0	0.033	1.000	0.957	715.682	0	rbp	4	GATAD2B.Q8WXI9
5	MAX	0	0.099	1.000	0.948	697.031	0	tf	5	GATAD2B.Q8WXI9
6	MAF	0	30.477	0.800	0.688	638.970	0	tf	6	GATAD2B.Q8WXI9
7	YTHDC1	0	-32.510	-0.600	-0.854	634.157	0	rbp	7	GATAD2B.Q8WXI9
8	FOXL1	0	-31.924	-0.975	-0.948	623.628	0	tf	8	GATAD2B.Q8WXI9
9	KLF16	0	-0.015	0.600	0.340	596.853	0	tf	9	GATAD2B.Q8WXI9
10	HNRNPC	0	0.010	0.800	0.894	594.447	0	rbp	10	GATAD2B.Q8WXI9

```
> grep("EIF3D", tbl.trena.p$gene) # [1] 166
```

	gene	betaLasso	betaRidge	spearmanCoeff	pearsonCoeff	rfScore	xgboost	class	rank	target
1	GATA2	0.184	0.029	0.950	0.973	6750.703	0.067	tf	1	GATAD2B.Q8WXI9
2	MEIS2	0.000	-0.928	-0.883	-0.568	5994.546	0.000	tf	2	GATAD2B.Q8WXI9
3	DICER1	0.000	0.477	0.950	0.919	5018.258	0.000	rbp	3	GATAD2B.Q8WXI9
4	MXI1	0.000	-0.043	-0.933	-0.788	4996.448	0.000	tf	4	GATAD2B.Q8WXI9
5	XBP1	0.000	0.055	0.817	0.956	4389.053	0.000	tf	5	GATAD2B.Q8WXI9
6	HNF1A	0.000	13.787	0.850	0.879	4341.541	0.000	tf	6	GATAD2B.Q8WXI9
7	NFIA	0.000	-0.202	-0.850	-0.942	4203.177	0.000	tf	7	GATAD2B.Q8WXI9
8	ZBED1	0.000	225.246	0.467	0.655	4168.523	0.000	tf	8	GATAD2B.Q8WXI9
9	SF3B4	0.000	0.075	0.767	0.803	3950.818	0.000	rbp	9	GATAD2B.Q8WXI9
10	QKI	0.000	-0.251	-0.783	-0.805	3824.396	0.000	rbp	10	GATAD2B.Q8WXI9

```
> grep("EIF3D", tbl.trena.p$gene) # [1] 260
```

Next task for Paul: find binding sites, motifs, maybe functional effects for (initially) 4 RBPs in 103 genes, eCLIPs and in vitro (5 jan 2022)

Van Nostrand, p712: We generated high-quality eCLIP profiles for 120 RBPs in K562 cells and 103 in HepG2 cells (**73** in both cell types, a total of 150 profiled RBPs.

To obtain insight into the functions of eCLIP peaks, we used short hairpin RNA (shRNA) or CRISPR to deplete individual RBPs followed by RNA sequencing (RNA-seq). We depleted 235 RBPs in K562 cells and 237 RBPs in HepG2 cells (**209** in both cell types, a total of 263 RBPs;

Marjorie (email 28 dec 21) suggests: **EIF3D** and **RPS3.** Here are the top 10 trena predictors from slide 4, where SRM protein targets were predicted from mRNA:

	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

Jeff replies:

1) Lets start with the 4 RBPs that Marjorie pointed out: **DDX3x**, **FXR2**, **EIF3D** and **IGF2BP2**. These proteins are all increasing at day 10-12 by TMT1 and they were part of the list on slide 4. Should we include **RPS3**?

2) extract eCLIP information for SRM 103 or (my preference) focus on SRM proteins that drop in expression between day 10 and 12, or those that drop in protein but increase in RNA. I think we can use the eCLIP data from both cell lines.

3) I think the RBP depletion information could be useful. It is possible for example that an RBP that causes target protein to drop might sequester and stabilize target RNA. So we see both RBP and Target RNA increasing together under normal conditions. KnockDown might cause the target RNA to drop. KD could also change the splicing pattern of the target which could affect its stability ior translatability. On the other hand, KD of the RBP might not affect RNA levels or splicing at all, but could still impact the translatability of the RNA. So I would use the depletion data as a secondary source of information after eCLIPs and expression profiles.

4) in vitro RBP binding motifs could be used to complement the eCLIPs data. However von Nostrand reports in vitro motifs for 78 RBPs. I think IGF2BP2 is included but not the other RBPs

5) One we are satisfied with the framework, I hope that we can then expand to all RBPs that are increasing by RNAseq and all proteins that are dropping by SRM or TMT1 (or that show discordant expression). My hope is that by looking at many RBPs that have similar expression patterns at late time points, and many targets that have similar expression patterns at late time points, we might be able to find a common set of RBPs that could explain their expression pattern.

No eCLIP for EIF3D or IGF2BP2 in ENCODE, in either cell line. Here's what I will download. Suggestions welcome.

BUD13 DDX3X FXR2 RPS3

All available RBPs can be seen here in the box titled **Target of assay.**

oRNAment

rna motifs enrichment in complete transcriptomes:
http://rnabiology.ircm.qc.ca/oRNAment/dashboard/
top 4 rbps, all 103 srm target genes, survey for RBP binding sites

oRNAment is a database of putative RBP binding site instances in both coding and non-coding RNA in various species.

Protein-RNA interactions play important roles in most aspects of RNA metabolism. Several in vitro selection methods, in particular RNAcompete and RNA Bind-n-Seq (RBNS), have defined the consensus target motifs of hundreds of RBPs. However, information about the distribution features of these motifs across the transcriptome is lacking.

Here, we developed a computational tool to enable the global mapping of putative RBP target motifs across full transcriptomes, including messenger RNAs (mRNAs) and non-coding RNAs, in multiple species. The oRNAment database provides a user-friendly interface to parse the motif scan data. See tutorial for further detail.

This database catalogues the motif instances of 223 RBPs in 525, 718 complete coding (UTR and CDS) and non-coding transcripts at various thresholds of motif similarity across the transcriptomes (excluding introns) of human and 4 model organisms. You can search the database for a specific species, gene or group of genes, a particular RBP, or distinct transcript attributes.

rbp binding by eCLIP: DDX3X, FXR2, EIF3D, IGF2BP2, RPS3, across 103 erythro genes

```
ENCFF333JSY-eclip-bud13-k562.bigBed       5157513
ENCFF357WBG-eclip-bud13-hepg2.bigBed      4293724

ENCFF562CTF-eclip-ddx3x-hepg2.bigBed      2743632
ENCFF565FNW-eclip-ddx3x-k562.bigBed       2434733

ENCFF733BQQ-eclip-eif3d-hepg2.bigBed      6588596

ENCFF483HFI-eclip-fxr2-hepg2.bigBed       2168751
ENCFF323QTM-eclip-fxr2-k562.bigBed        2128433

ENCFF416JIV-eclip-igf2bp2-k562.bigBed     2955924

ENCFF199OFQ-eclip-rps3-k562.bigBed        1471282
ENCFF848LEJ-eclip-rps3-hepg2.bigBed       1760199
```

encode had no k562 eCLIP for eif3d, nor hepg2 for igf2bp2

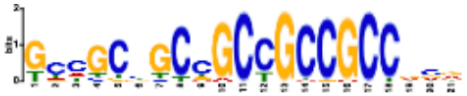
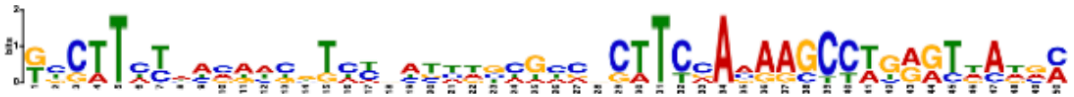
jeff asks for DDX3x, FXR2, EIF3D and IGF2BP2, marjorie RPS3,

cory suggests: Paul, once you've extracted the motifs for the 103 genes, I'd suggest looking at your eQTL database to see if any eQTLs exist in any of these binding sites. They could be added circumstantial evidence for being functional, especially if we see changes upon knockdown of the RBPs.

5 RBP, 103 gene, K562 & HepG2, mostly UTRs
fails to find deep enrichment of the GCCGCC motif

```
 --- confusion: slide 28 shows 37 sites for the GCCGCC motif, from 220 sites,
     97 discordant genes, eval 6.7e-35,  way more significant that i see above
    can i reproduce that slide? yes...
cd ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants/
~/meme/bin/meme targetGenes-97-ddx3x-k562.fa -dna -oc meme.out.targetGenes-97-ddx3x-k562\
     -nostatus -time 14400 -mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun  classic -revcomp -markov_order 0
open ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants/meme.out.targetGenes-97-ddx3x-k562/meme.html
cd ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants
source("findMotifs.R")
quick.demo()
 as.data.frame(table(tbl.out$targetfeature))
                 Var1 Freq
1    hg38_genes_1to5kb   63
2     hg38_genes_5UTRs    9
3     hg38_genes_exons   72
4   hg38_genes_introns   73
5 hg38_genes_promoters   42
     table(tbl.out$method)  #       eCLIP       259
```

Discovered Motifs

	Logo	E-value	Sites	Width	More	Submit/Download
1.		6.7e-035	37	21	↧	⇢
2.		6.9e-018	9	50	↧	⇢
3.		2.1e-007	6	50	↧	⇢

Stopped because requested number of motifs (3) found.

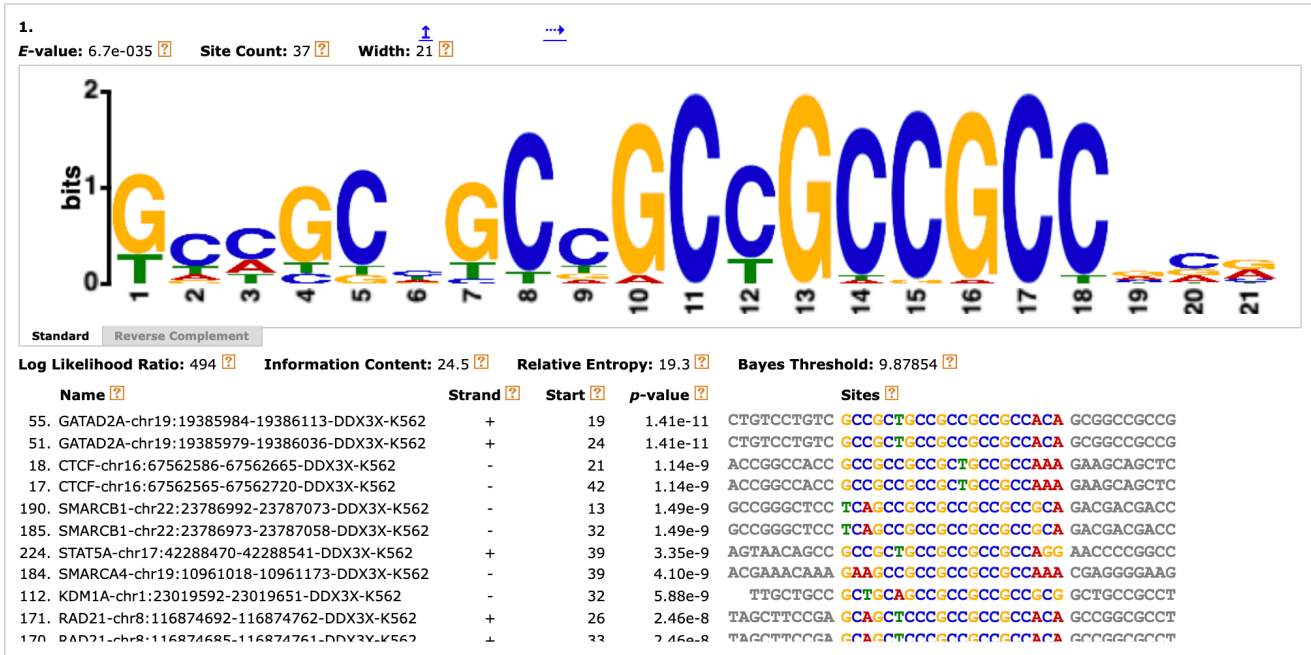
5 RBP, 103 gene, K562 & HepG2, mostly UTRs
fails to find deep enrichment of the GCCGCC motif

--- confusion: slide 28 shows 37 sites for the GCCGCC motif, from 220 sites,
97 discordant genes, eval 6.7e-35, way more significant that i see above
can i reproduce that slide? yes...

```
cd ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants/
~/meme/bin/meme targetGenes-97-ddx3x-k562.fa -dna -oc meme.out.targetGenes-97-ddx3x-k562\
    -nostatus -time 14400 -mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun  classic -revcomp
-markov_order 0
open ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants/meme.out.targetGenes-97-ddx3x-
k562/meme.html
cd ~/github/TrenaProjectErythropoiesis/explore/rbp/discordants
source("findMotifs.R")
quick.demo()
 as.data.frame(table(tbl.out$targetfeature
                 Var1 Freq
1    hg38_genes_1to5kb   63
2     hg38_genes_5UTRs    9
3     hg38_genes_exons   72
4   hg38_genes_introns   73
5 hg38_genes_promoters   42
    table(tbl.out$method)
#       eCLIP       259
```

possibly related GGGGCC repeat motif described on slide 50

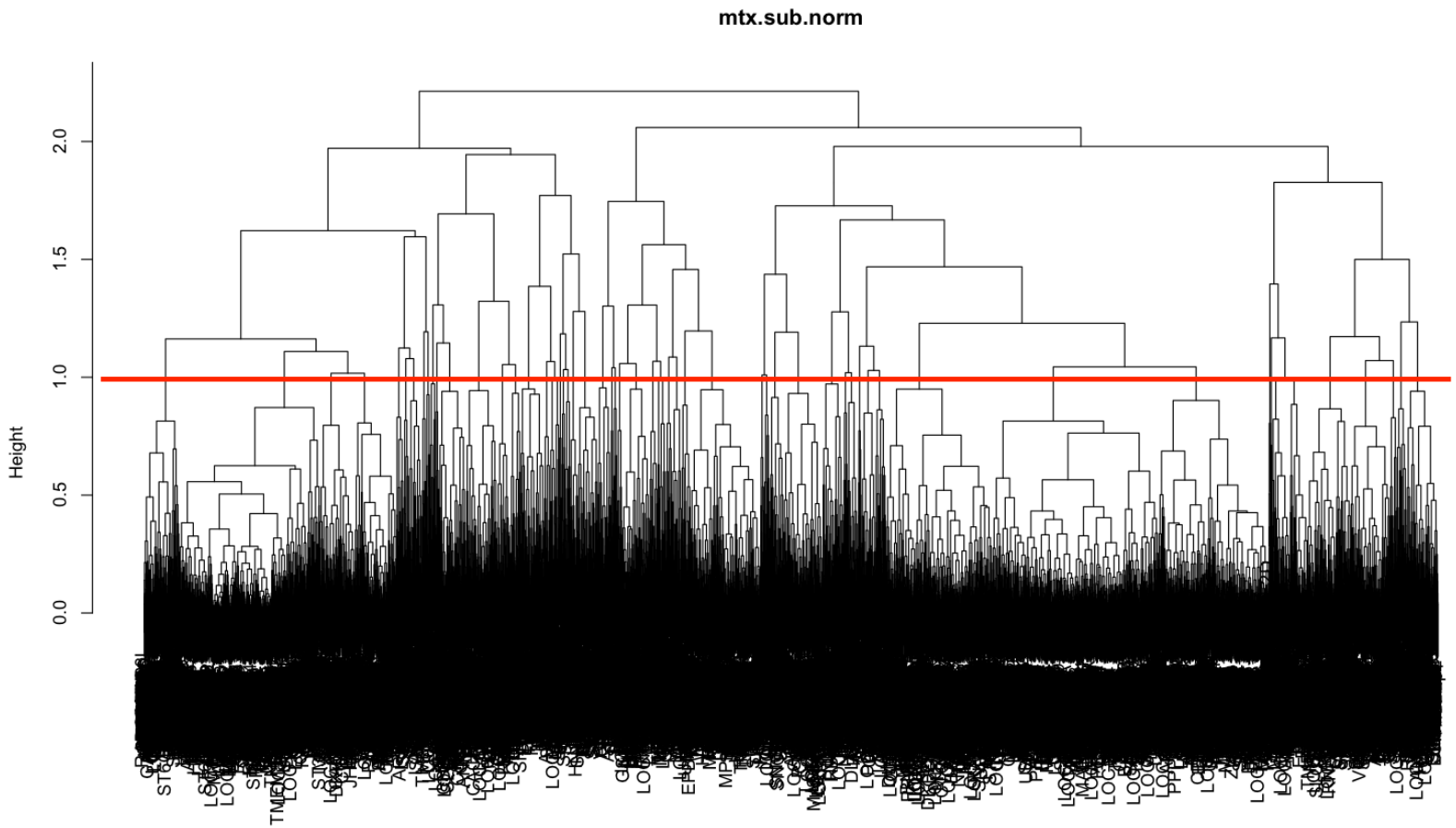
suggesting that DDX3X binds mostly in non-UTR regions



todo list from 10 jan 2022 zoom meeting

1. research meme's approach to multiple hypothesis testing: is it handled properly?
2. run meme on full transcript of each 103 & 47 genes, 5 RBPs, 2 cell-lines
3. provide ranked-list of
4. count eCLIP binding events on each full transcript, by RBP & cell-line
5. WGCNA on all full genome, Marjorie's rna-seq, day 8 - 14?
6. 82k ddx3x binding sites from eCLIP: which genes? is there WGCNA pattern in day 8-14?
7. survey 47 discordant vs 56 concordant gene sets for rbp binding sites across full transcripts
8. identify target genes for any rbp binding event, run on all proteins, all genes, find clusters, then examine clusters for rbp binding sites.
    1. do TMT-1 wgcna on this set, across what time interval? -> **10-12**
    2. do rna-seq wgcna on this set, 10-12?
    3. any correlation between # of binding sites, # or rbps, and wgcna clusters?
9. erythro wgcna: https://www.biorxiv.org/content/10.1101/234062v1.abstract, mouse. use their data. R package accompanies. find any rbps in those clusters look for binding events in human eCLIP: are they enriched in the mouse cluster to which the rbp belongs? iterativeWGCNA: iterative refinement to improve module detection from WGCNA co-expression networks: https://github.com/cstoeckert/iterativeWGCNA
10. *or* run WGCNA on marjorie's rna-seq, find clusters, look for differential representation of our 5 RBPs (present/absent or counts) in the clusters - especially if an RBP is in the cluster
    1. use WGCMA if our data meets WGCNA's requirements
    2. of not, look for good-enough alternative.
    3. increase RBPs - the signal may be elsewhere. wgcna ids those which are dynamic and significant. get ENCODE eCLIPs for these rbps.
    4. incorporate protein data: how do they relate to the forthcoming clusters?

cluster expression, mtx-rna-srm-27270x13, D10-12, normalized

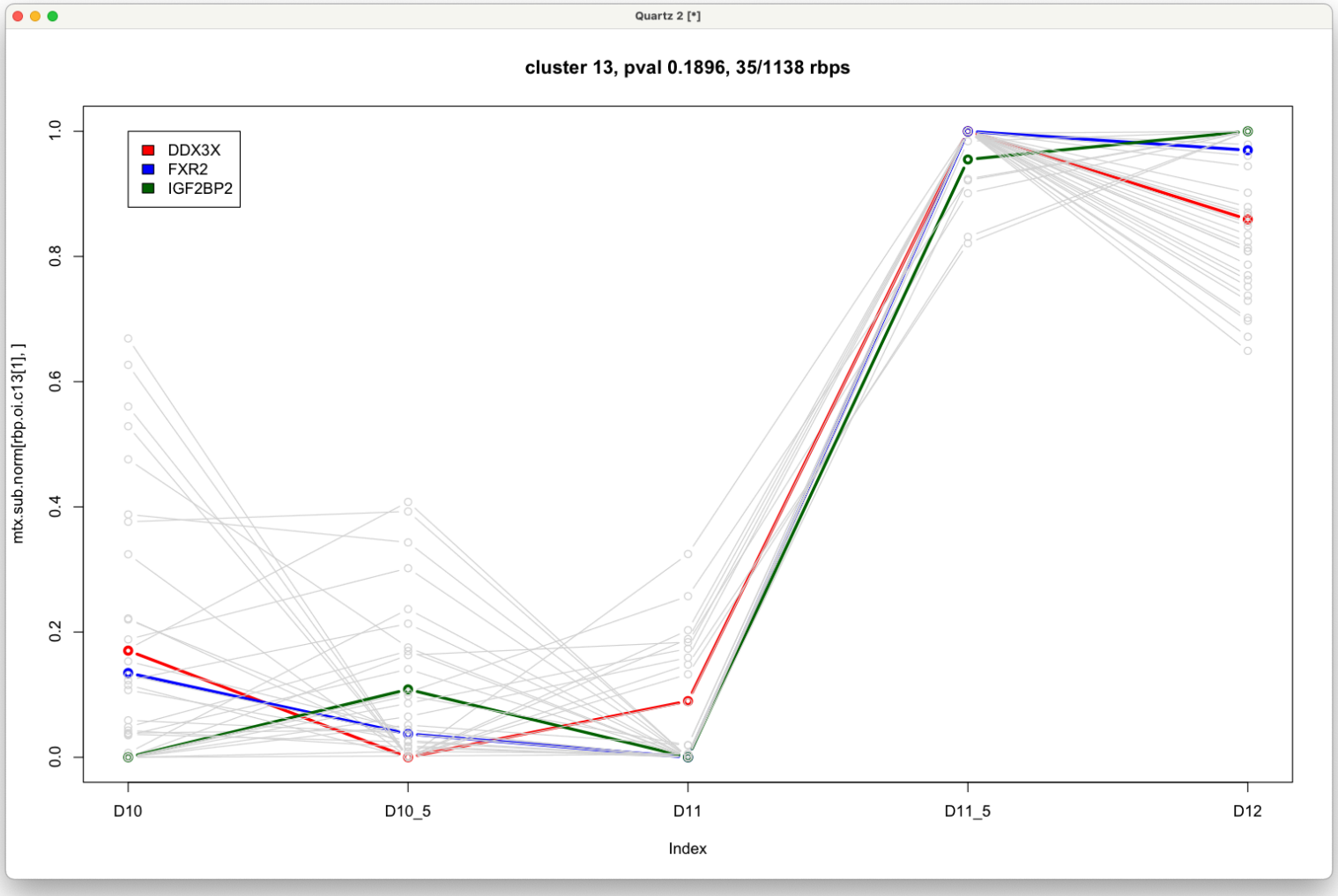


best of 55 clusters at height=1, mtx-rna-srm-27270x13, D10-12, normalized

	cluster	size	rbp	percent	pval
1	17	320	19	5.937500	0.0008
2	53	40	4	10.000000	0.0206
3	15	157	9	5.732484	0.0233
4	31	287	13	4.529617	0.0410
5	16	53	4	7.547170	0.0508
6	2	102	6	5.882353	0.0528
7	7	564	21	3.723404	0.0701
8	51	30	2	6.666667	0.1872
9	13	1138	35	3.075571	0.1896

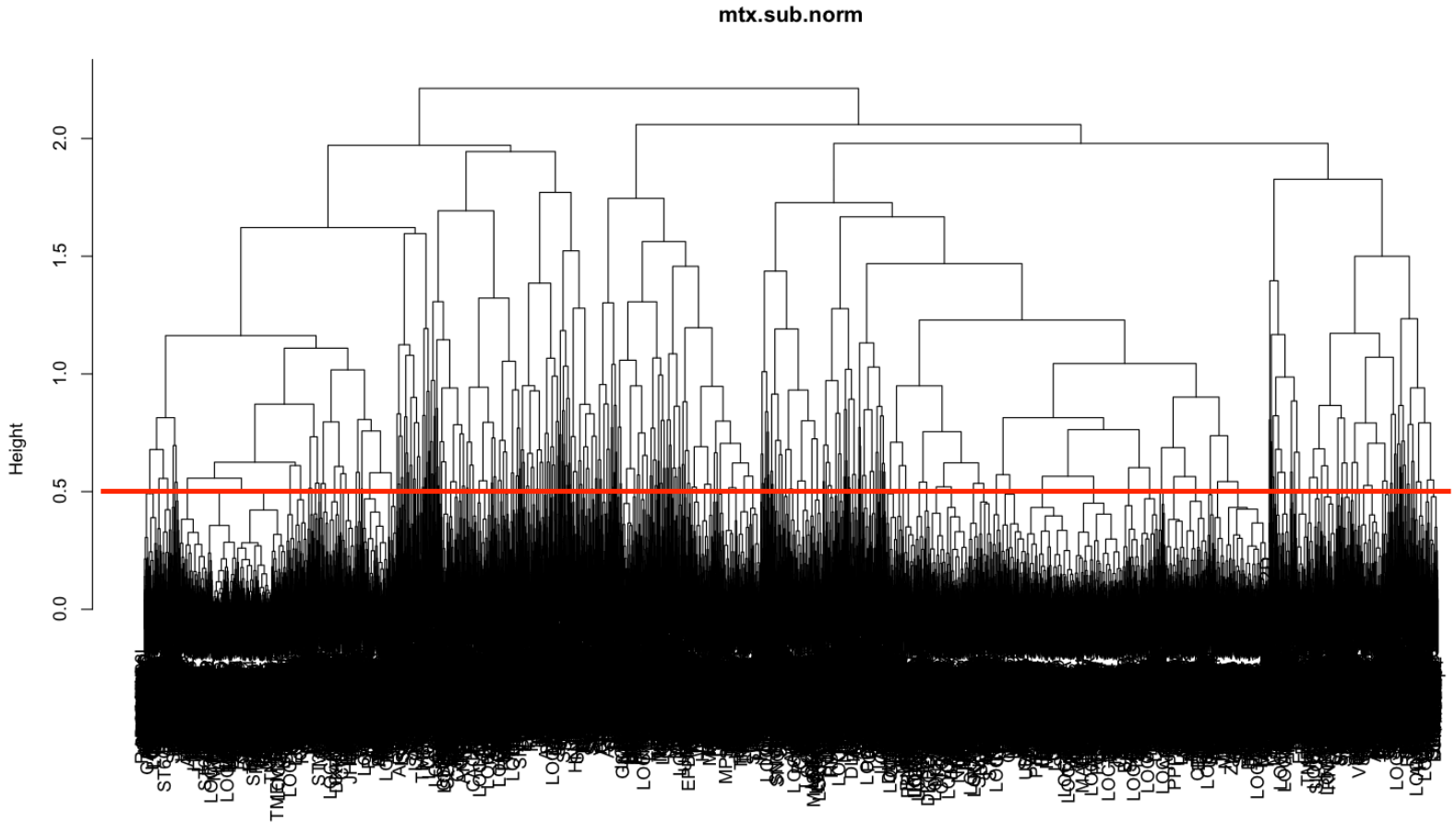
RBP	cluster
DDX3X	13
FXR2	13
EIF3D	6
IGF2BP2	13
RPS3	36

cluster 13 expression, mtx-rna-srm-27270x13, D10-12, normalized



CDC40	CELF1	CNBP
CNOT4	CPEB4	CSDE1
DDX3X	EIF4G2	EWSR1
FMR1	**FXR2**	HNRNPAB
HNRNPH1	HNRNPK	**IGF2BP2**
PABPC1	PABPC3	PCBP1
PCBP2	PCBP4	PUM1
QKI	RBM38	RBM39
RBM6	RBMS1	RBPMS
SF3B1	TARBP2	TIAL1
TNRC6A	TNRC6C	VIM
YBX2	YTHDF3	

cluster expression, mtx-rna-srm-27270x13, D10-12, normalized



best of 326 clusters at height=0.5, mtx-rna-srm-27270x13, D10-12, normalized

	cluster	size	rbp	percent	pval
1	53	20	4	20.000000	0.0016
2	18	22	4	18.181818	0.0024
3	284	18	3	16.666667	0.0111
4	133	55	5	9.090909	0.0147
5	164	40	4	10.000000	0.0206
6	112	24	3	12.500000	0.0245
7	137	10	2	20.000000	0.0272
8	229	11	2	18.181818	0.0326
9	181	27	3	11.111111	0.0334
10	81	31	3	9.677419	0.0475
11	21	53	4	7.547170	0.0508
12	86	105	6	5.714286	0.0592
13	62	86	5	5.813953	0.0768
14	202	18	2	11.111111	0.0805
15	135	20	2	10.000000	0.0966
16	254	4	1	25.000000	0.1015
17	317	4	1	25.000000	0.1015
18	221	21	2	9.523810	0.1050
19	79	22	2	9.090909	0.1136
20	74	72	4	5.555556	0.1222
21	54	5	1	20.000000	0.1252
22	146	50	3	6.000000	0.1450
23	258	6	1	16.666667	0.1483
24	291	6	1	16.666667	0.1483
25	29	235	9	3.829787	0.1695
26	300	7	1	14.285714	0.1708
27	59	116	5	4.310345	0.1922
28	306	8	1	12.500000	0.1927

cluster membership for "our" RBPs

RBP	cluster
DDX3X	78
FXR2	17
EIF3D	162
IGF2BP2	17
RPS3	110

todo from 20 jan 2022 zoom meeting

1. 11-11.5 spike: do we find that in the parent clusters of 13? (see slide 88)
2. do GO enrichment of cluster 13 rbps, also look for complex membership, via protein-protein interactions.
3. cluster 17: same spike? [since close to 13, we may] [do plot of rbps like slide 88]
4. anti-correlation of srm against that spike [web search to see if hclust handles that]
5. rbps in c17 which intersect with: "CDC40" "CELF1" "FXR2" "IGF2BP2" "PCBP1" "PCBP4" "RBPMS" "TNRC6C"
6. find binding sites of slide 88 rbps in cluster 3 and 4 proteins:

```
> grep("p$", names(clusters[clusters==3]), value=TRUE)
 [1] "CREBBPp"  "CHD1p"    "RCOR1p"   "CTCFp"    "TCF3p"    "E2F4p"    "EGR1p"    "ELF1p"
 [9] "ETV6p"    "GATAD2Ap" "NR3C1p"   "HDAC2p"   "HMGB3p"   "JUNDp"    "KLF1p"    "KLF3p"
[17] "LDB1p"    "NELFEp"   "NFE2p"    "NRF1p"    "POU2F1p"  "RAD21p"   "RFX5p"    "POLR2Ap"
[25] "SMC3p"    "GTF2F1p"  "TAL1p"    "GTF2F2p"  "WDR5p"    "ZBTB7Ap"
> grep("p$", names(clusters[clusters==2]), value=TRUE)
character(0)
> grep("p$", names(clusters[clusters==4]), value=TRUE)
[1] "BACH1p"   "DOT1Lp"   "HCFC1p"   "SMARCA4p" "TTF2p"    "UBTFp"
```

7. Jeff asks: Could you tell me in which clusters the following proteins reside? They all drop starting around day 11 like KLF1

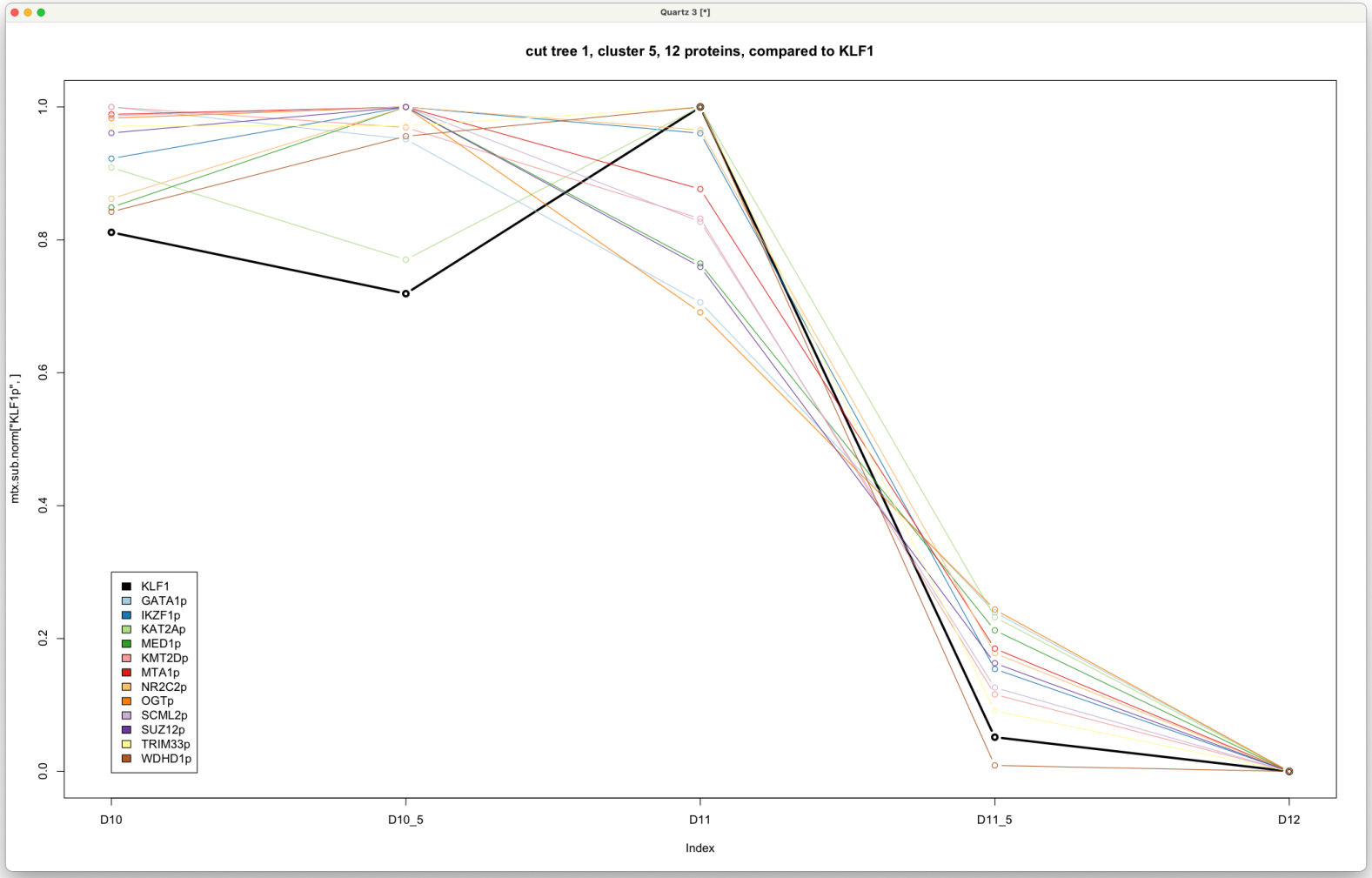
BCL11A_XL_L, GATA1, GFI1B, HDAC3, HLTF, IKZF1, KAT2A, MED1, MLL3, MLL4 (KMT2D), MTA1, NFE2, NR2C2, OGT, PSIP1, SCML2, SSRP1, SUZ12, TRIM33, USF1, UTX, WDHD1

Response to Jeff (21 jan 2022)

23 proteins of interest (sharp day 11 drop):
BCL11Ap GATA1p GFI1Bp HDAC3p HLTFp IKZF1p KAT2Ap MED1p KMT2Cp KMT2Dp MTA1p NFE2p
NR2C2p OGTp PSIP1p SCML2p SSRP1p SUZ12p TRIM33p USF1p KDM6Ap WDHD1p KLF1p
clustering, cut dendrogram at 1.0, protein distribution across 55 clusters, average size 190:
cluster: **3 5 6 20 22 24 26 36 54**
count: 2 12 2 2 1 1 1 1 1
at 1.7, in hopes of getting all 23 into cluster, 10 clusters, average size 1050
cluster: **2 3 4 6**
count: 17 3 2 1
KLF1 ends up in cluster 3 with NFE2p, and cluster 5 has the lion's share.

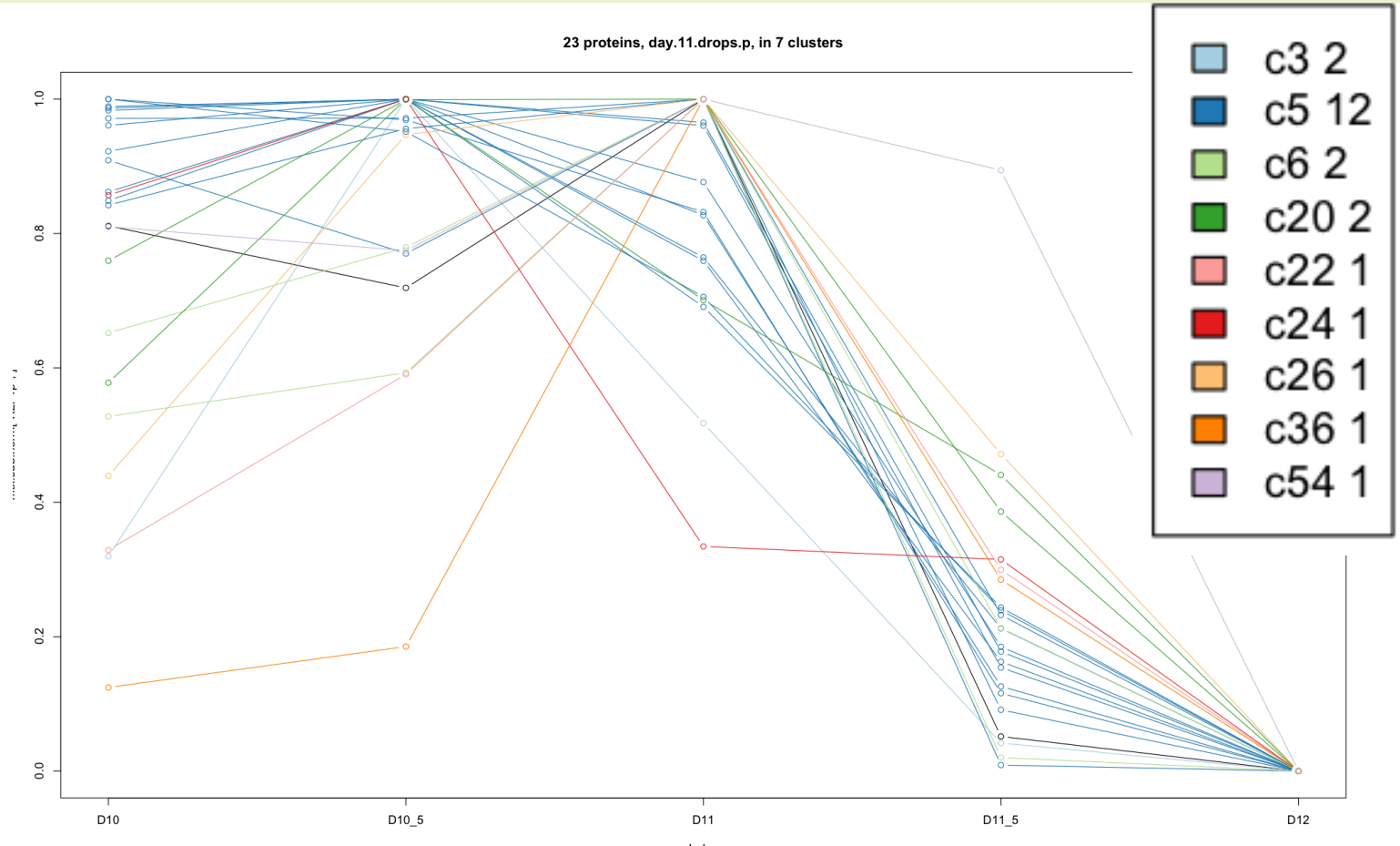
The plot below seems to explain why. We might suppose that KAT2Ap (light green) would cluster with KLF1p rather than end up in cluster 5. KLF1 and KAT2PAp have similar shape from day 10 to day 11. But the algorithm puts in in cluster 5, apparently because KAT2PAp otherwise mostly agrees with the average shape which defines the cluster.

We could transform the data to highlight the day11-11.5 drop, by averaging day 10, 10.5, 11, giving all three timepoints that mean value for each protein. Then even small clusters put most or all of your 23 proteins into the same cluster.



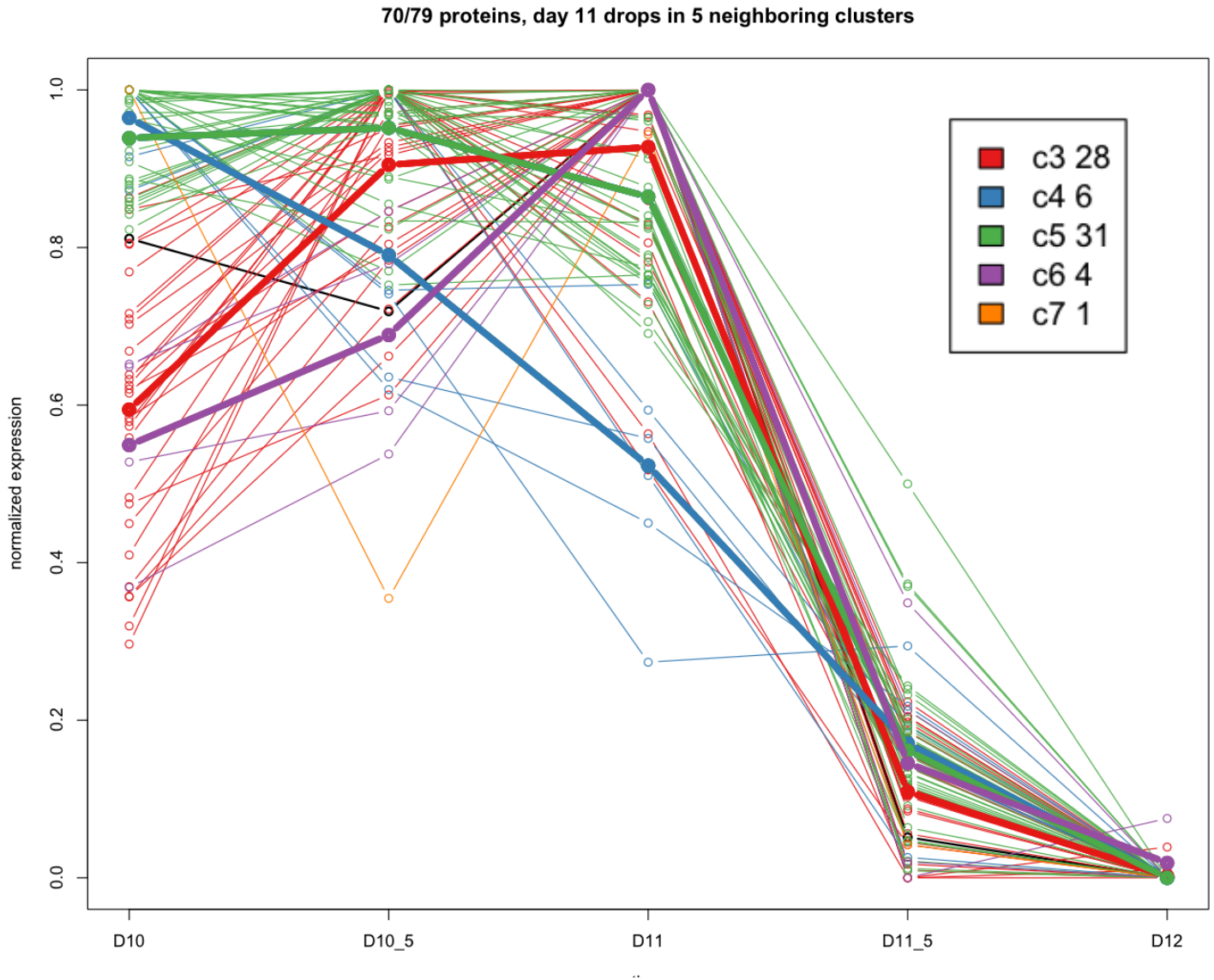
23 proteins, 7 clusters in 7 colors, cluster size in the legend.
blue & green clusters are stronger. klf1 in black

```
 c3: NFE2p KLF1p
 c5: GATA1p  IKZF1p  KAT2Ap  MED1p   KMT2Dp  MTA1p   NR2C2p  OGTp    SCML2p
     SUZ12p  TRIM33p WDHD1p
 c6: USF1p  KDM6Ap
c20: HDAC3p SSRP1p
c22: PSIP1p
c24: HLTFp
c26: BCL11Ap
c36: GFI1Bp
c54: KMT2Cp
```



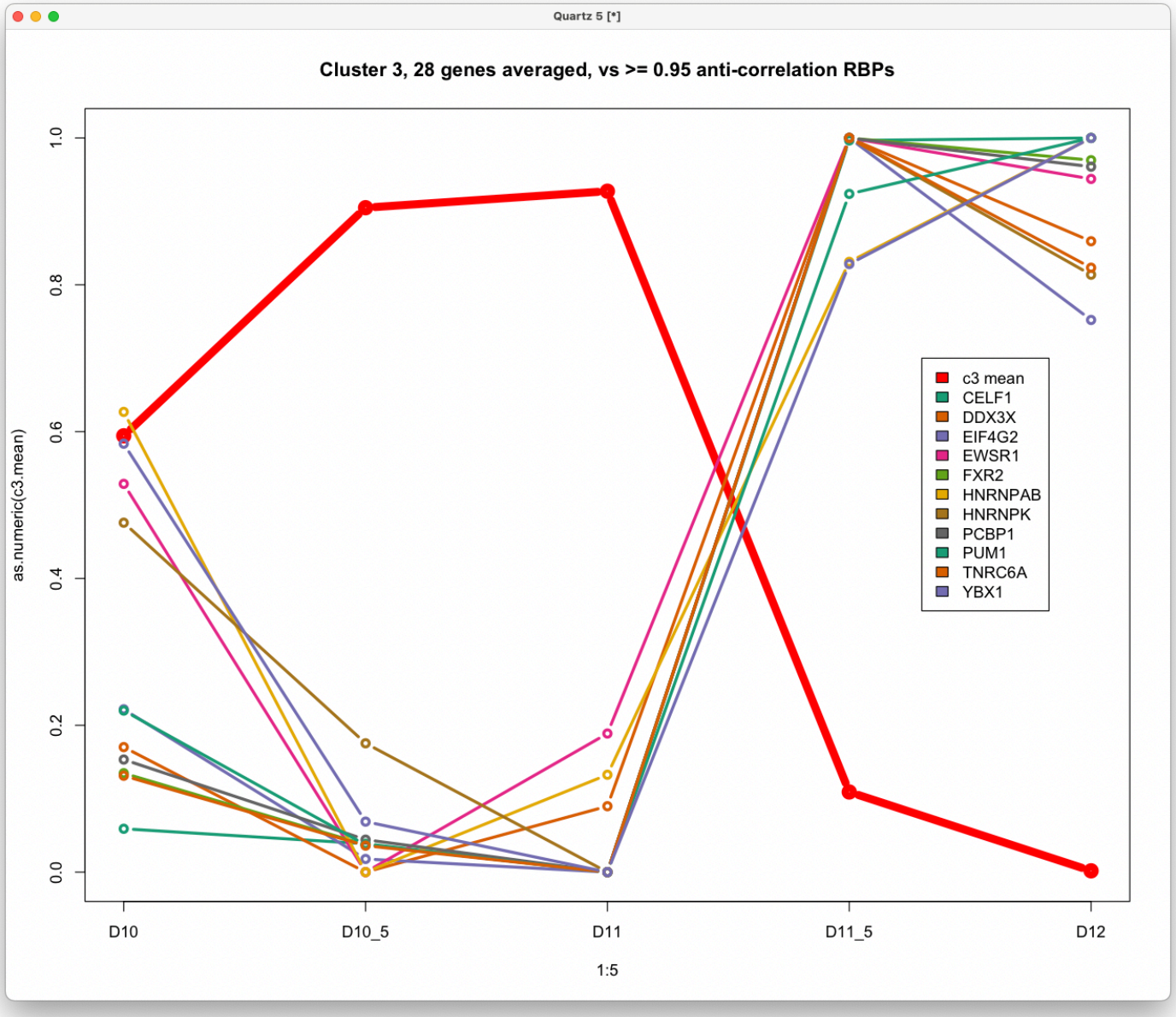
Redo clustering with Jeff's full set of 80
(actually 79: we apparently have no srm for GTF2Bp)

```
table(unlist(lapply(jeff.80.proteins, function(p) clusters[[p]])))
 3  4  5  6  7 20 22 24 26 36 54
28  6 31  4  1  3  1  2  1  1  1
```



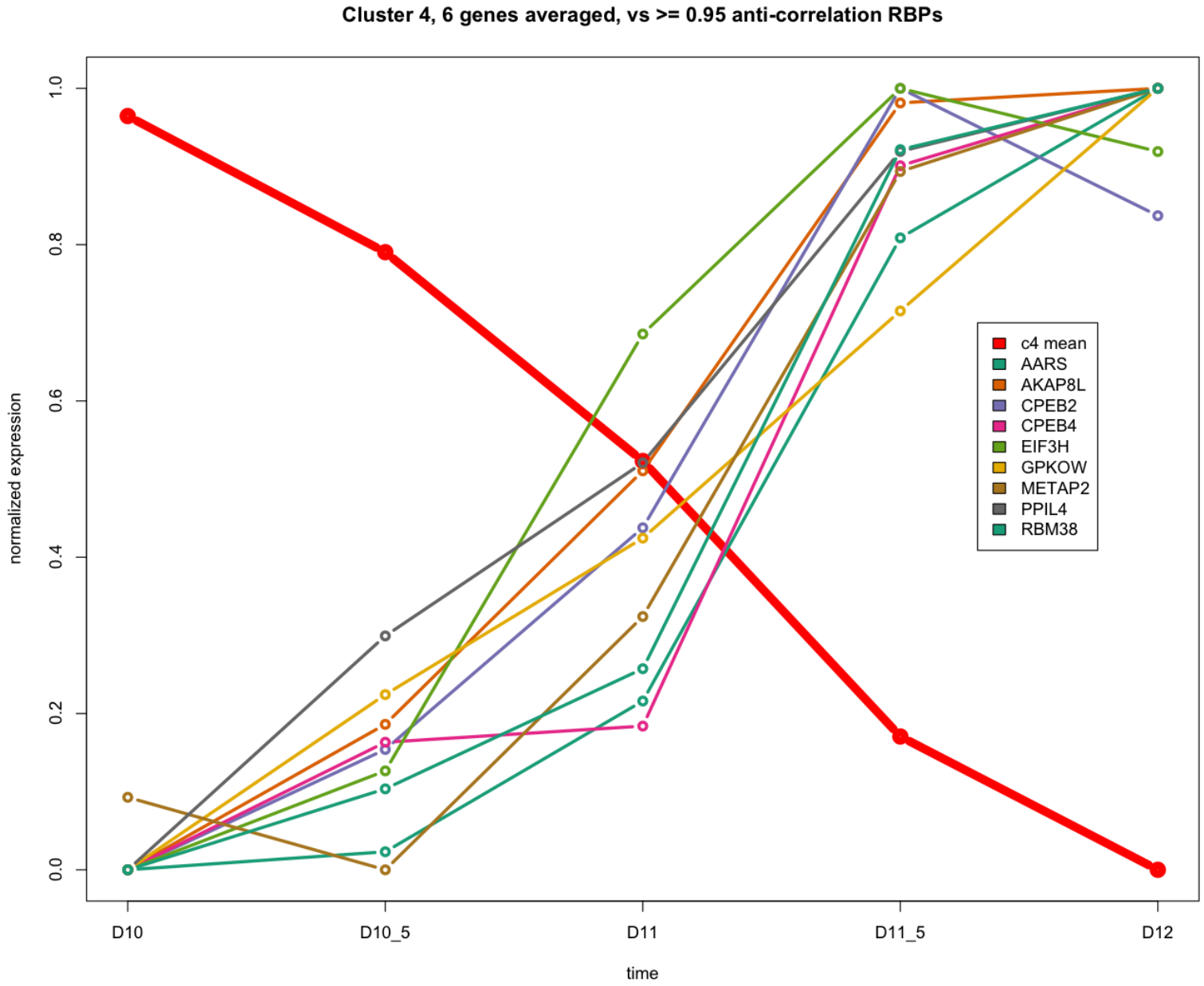
cluster 3 mean protein expression(n=28) vs anti-correlated RBP mRNA

CHD1p	CREBBPp	CTCFp	E2F4p	EGR1p	ELF1p	GATAD2Ap	GTF2F1p
HDAC2p	HMGB3p	JUNDp	KLF1p	KLF3p	LDB1p	NELFEp	NFE2p
NR3C1p	NRF1p	POLR2Ap	POU2F1p	RAD21p	RCOR1p	RFX5p	SMC3p
TAL1p	TCF3p	WDR5p	ZBTB7Ap				



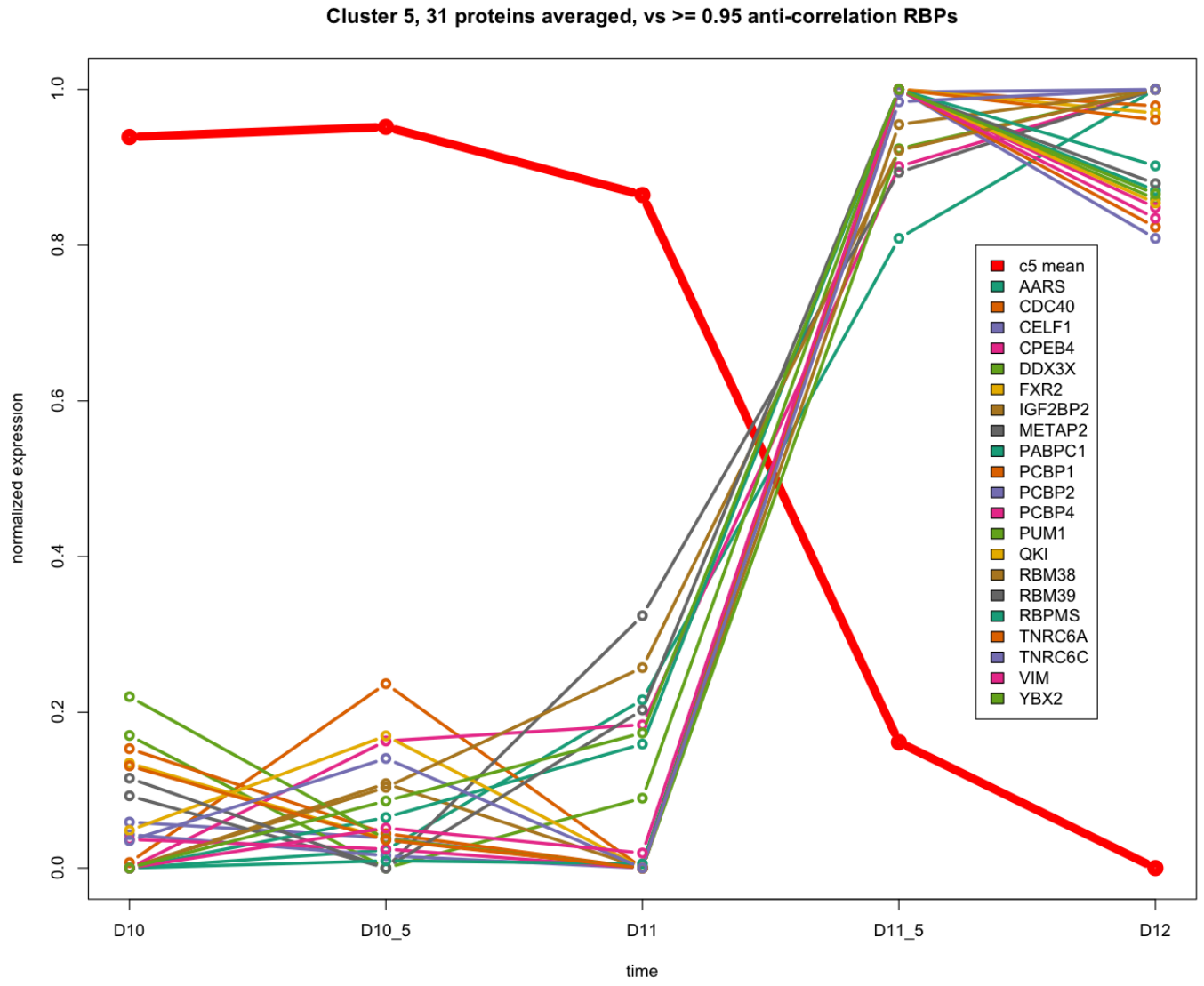
cluster 4 mean protein expression(n=6) vs anti-correlated RBP mRNA

BACH1p DOT1Lp HCFC1p SMARCA4p TTF2p UBTFp



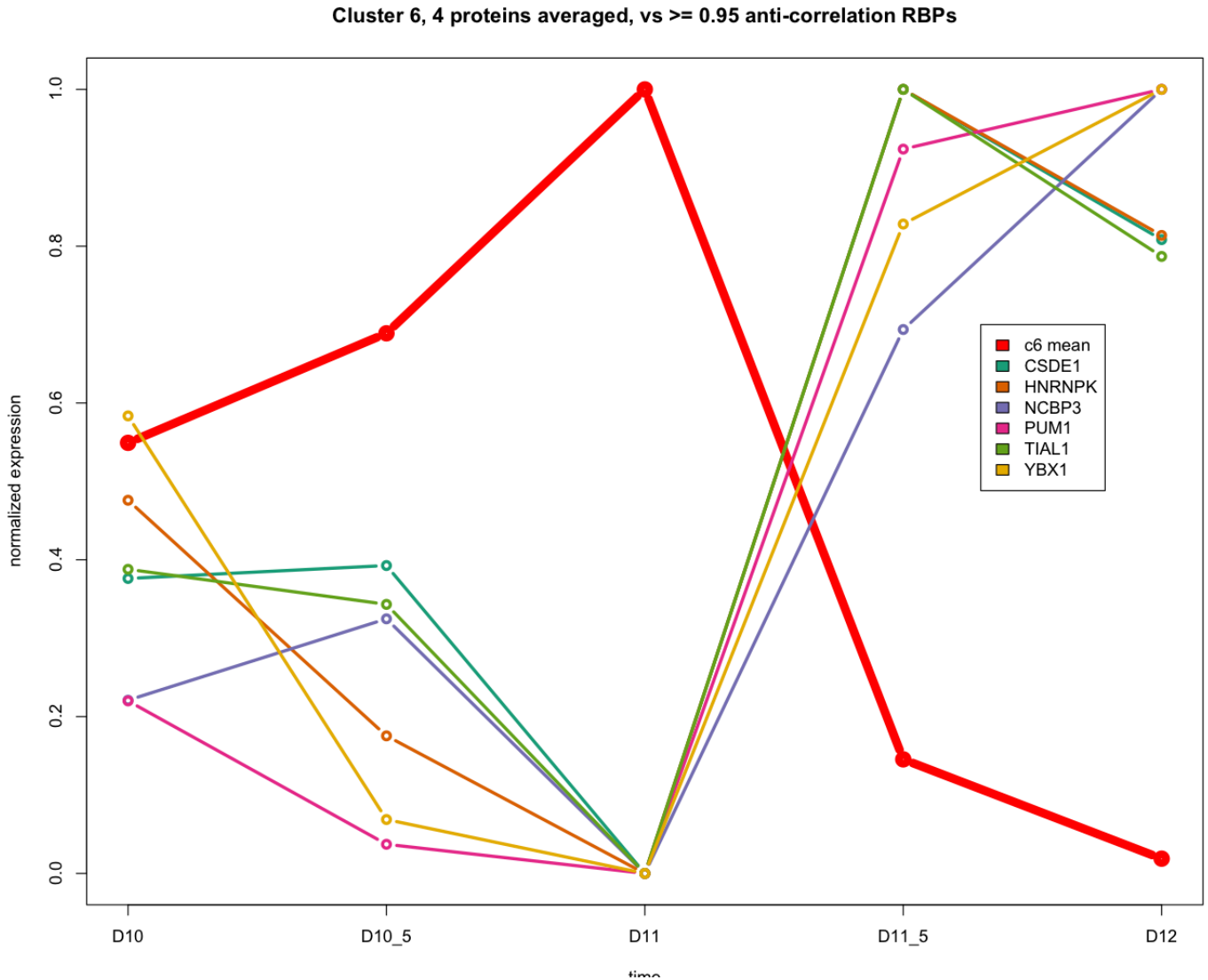
cluster 5 mean protein expression(n=31) vs anti-correlated RBP mRNA

CHD4p	CTBP2p	DNMT1p	E2F8p	FOXO3p	GATA1p	GTF3C2p	HOXB4p
IKZF1p	KAT2Ap	KMT2Dp	MED1p	MTA1p	NR2C2p	OGTp	PARP1p
RUNX1p	RXRBp	SAP130p	SCML2p	SETD1Bp	SETDB1p	SIN3Ap	SIRT6p
SMARCB1p	SMARCC1p	STAT5Ap	SUZ12p	TRIM33p	WDHD1p	ZC3H11Ap	



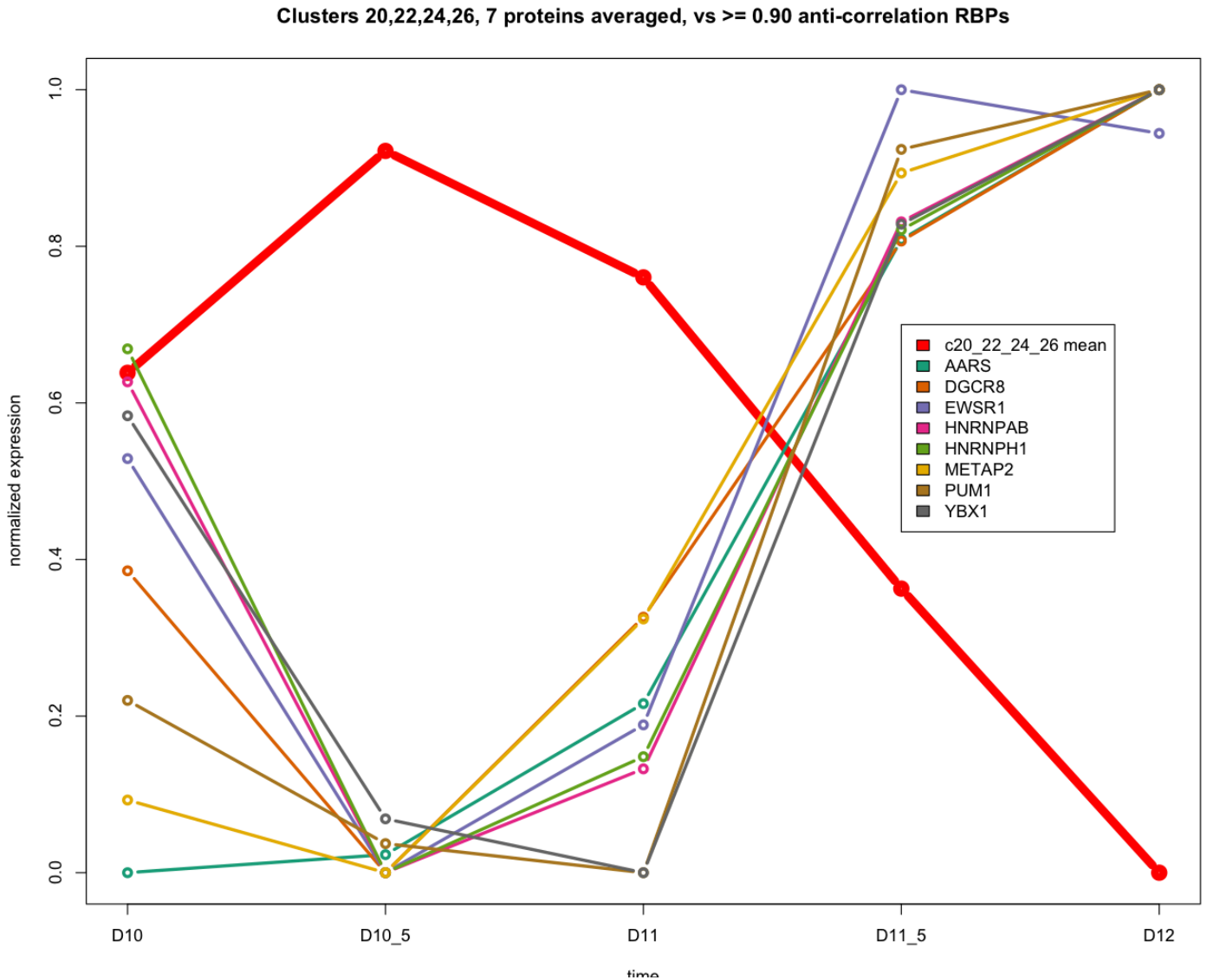
cluster 6 mean protein expression(n=4) vs anti-correlated RBP mRNA

KDM1Ap KDM6Ap KLF13p USF1p



clusters 20,22,24,26 mean protein expression(n=4) vs anti-correlated RBP mRNA

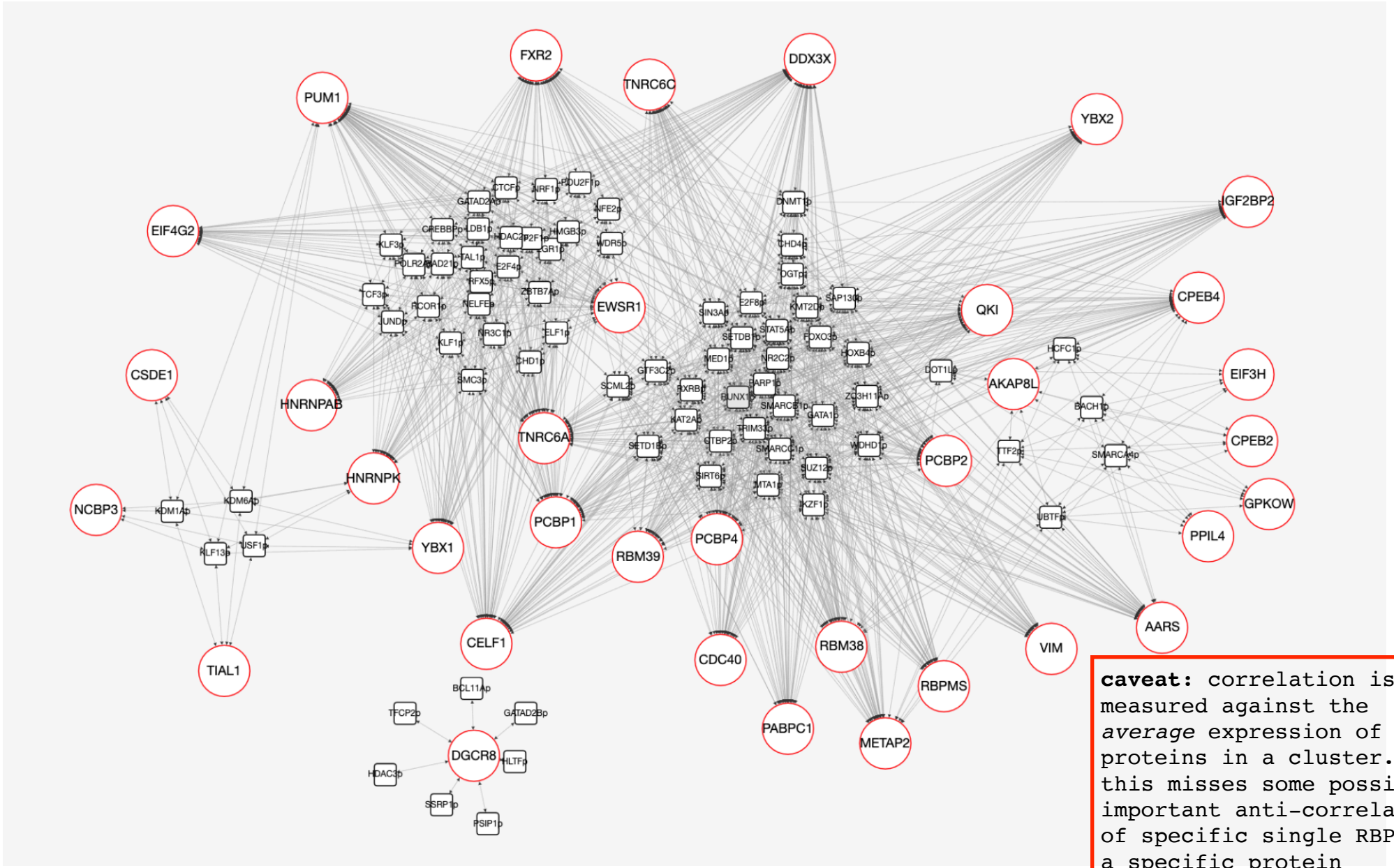
BCL11Ap GATAD2Bp HDAC3p HLTFp PSIP1p SSRP1p TFCP2p



clusters 20, 22, 24, 26 are near neighbors, and have few members, so are collected together here.

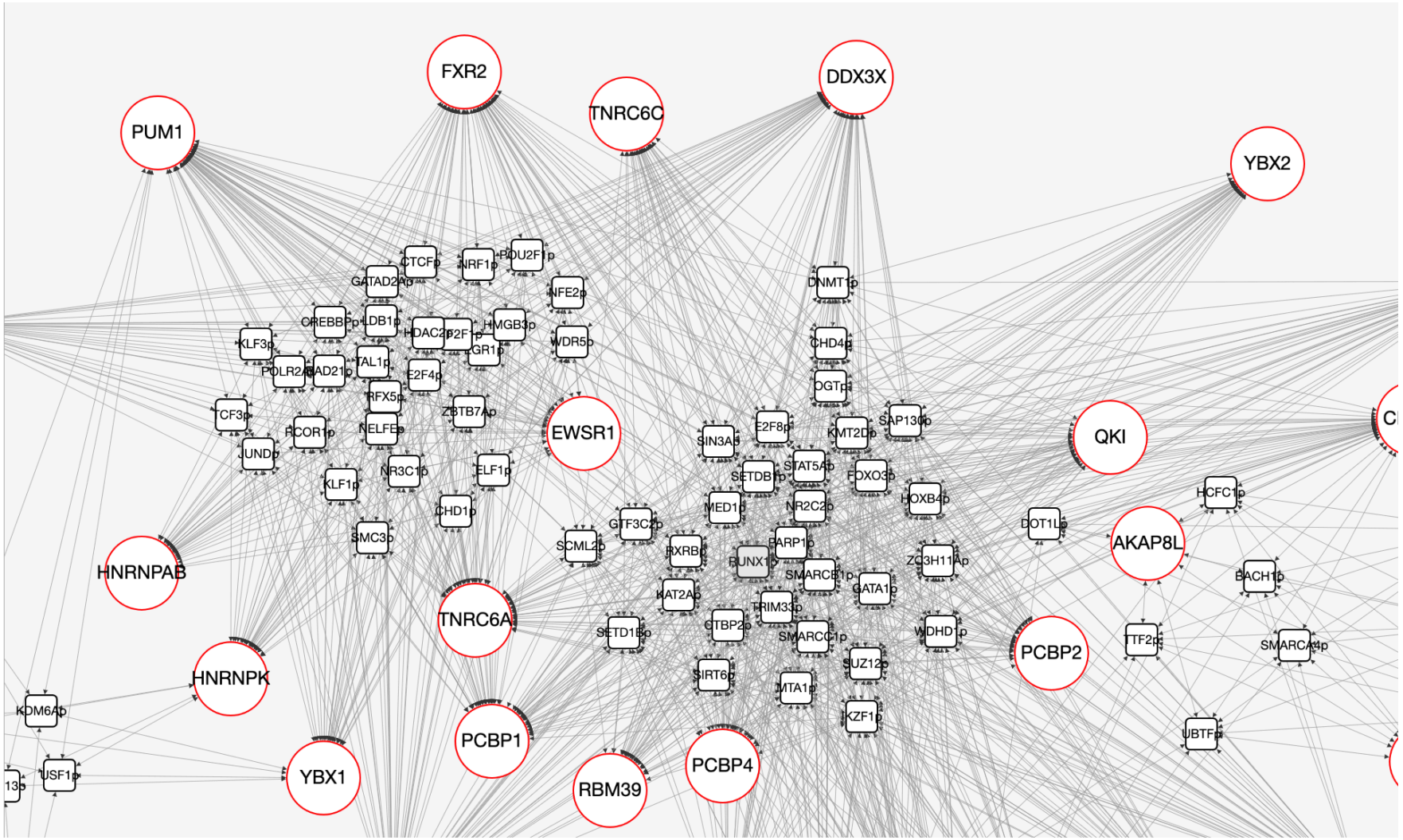
only DGCR8 meets the customary < -0.95 correlation, so I dropped the threshold to -0.90.

overview all rbp/protein anti-correlations, <= 0.95
todo: provide navigability to Marjorie and Jeff

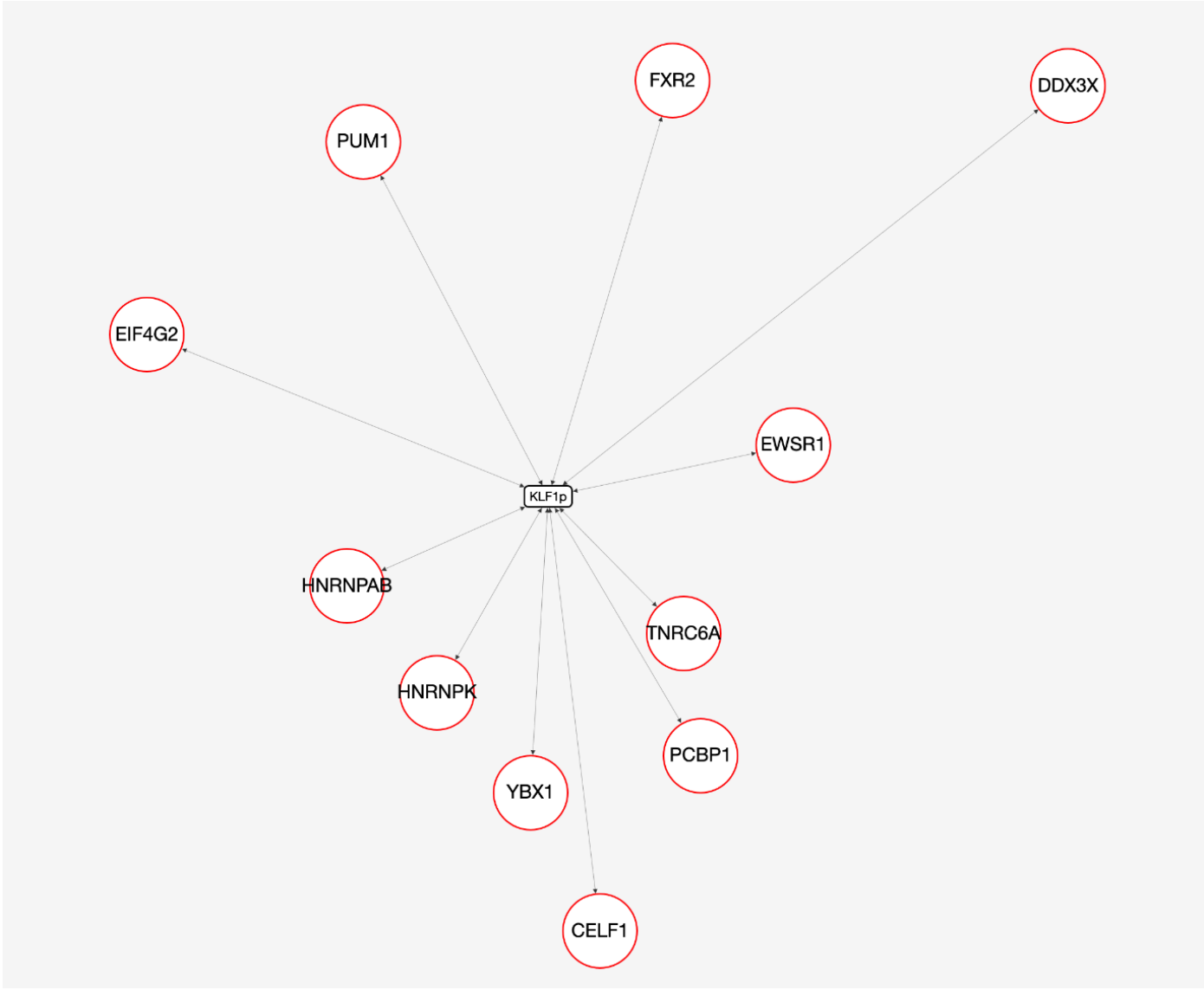


caveat: correlation is measured against the *average* expression of the proteins in a cluster. this misses some possibly important anti-correlations of specific single RBPs to a specific protein

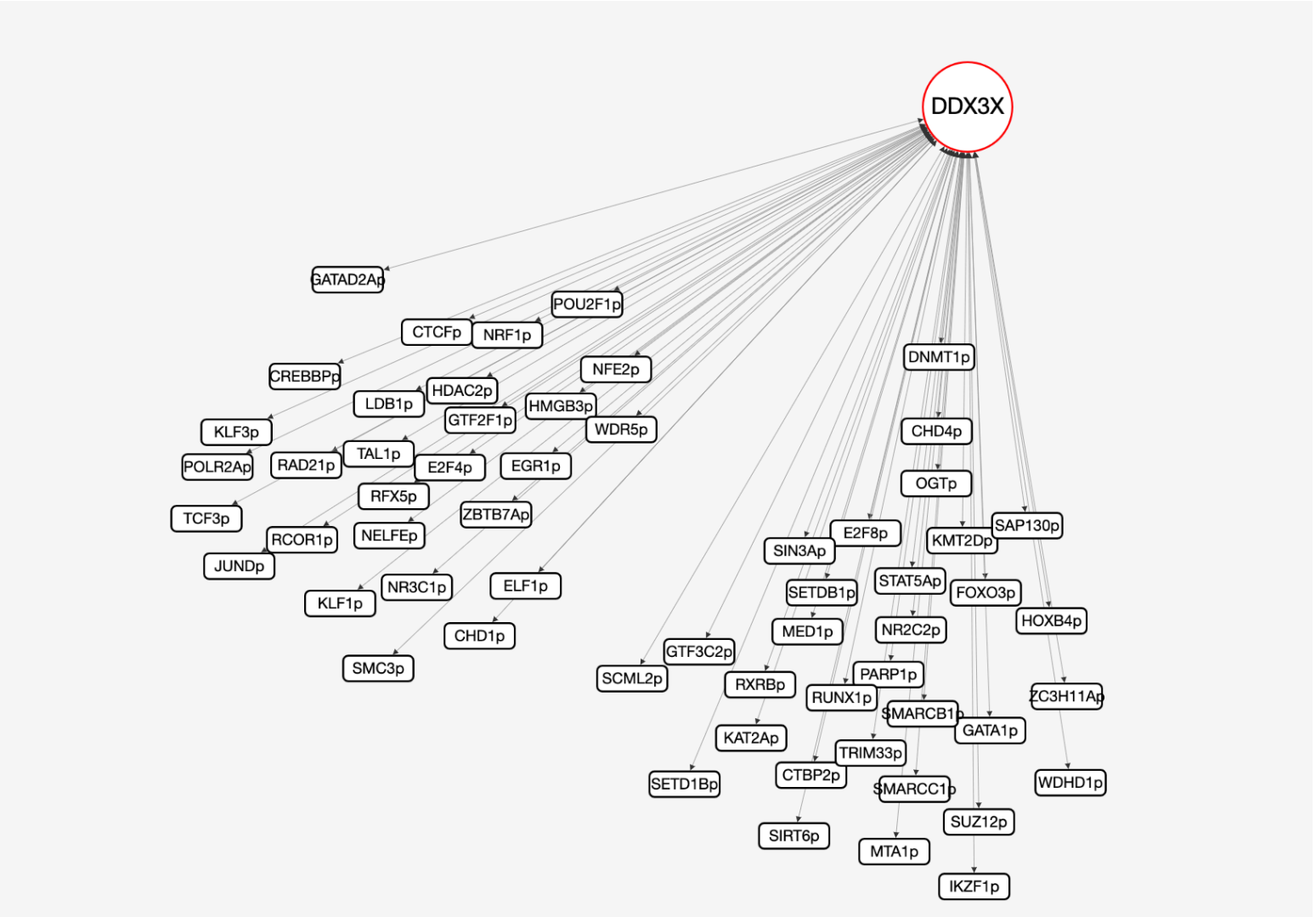
zoomed in a little



all RBPs< -0.95 correlation with KLF1p



all proteins whose cluster-mean is < -0.95 correlation with DDX3X



Next? email from paul 28 jan

Here's a thought on what to do next - based upon the two additional data sets Jeff mentions:

1) mRNA levels of the target proteins. if the target mRNA is follows the RBP rna, then the RBP-target[gene|protein] relationship seems stronger
2) eCLIP binding data

We can put more information into the rbp/protein anti-correlations network:
- new edges for eCLIP binding (perhaps different edges by cell-type and binding region: 5', 3').
- bring along eCLIP score
- add an edge for positively correlated RBP/target rna (see KLF1 example below)

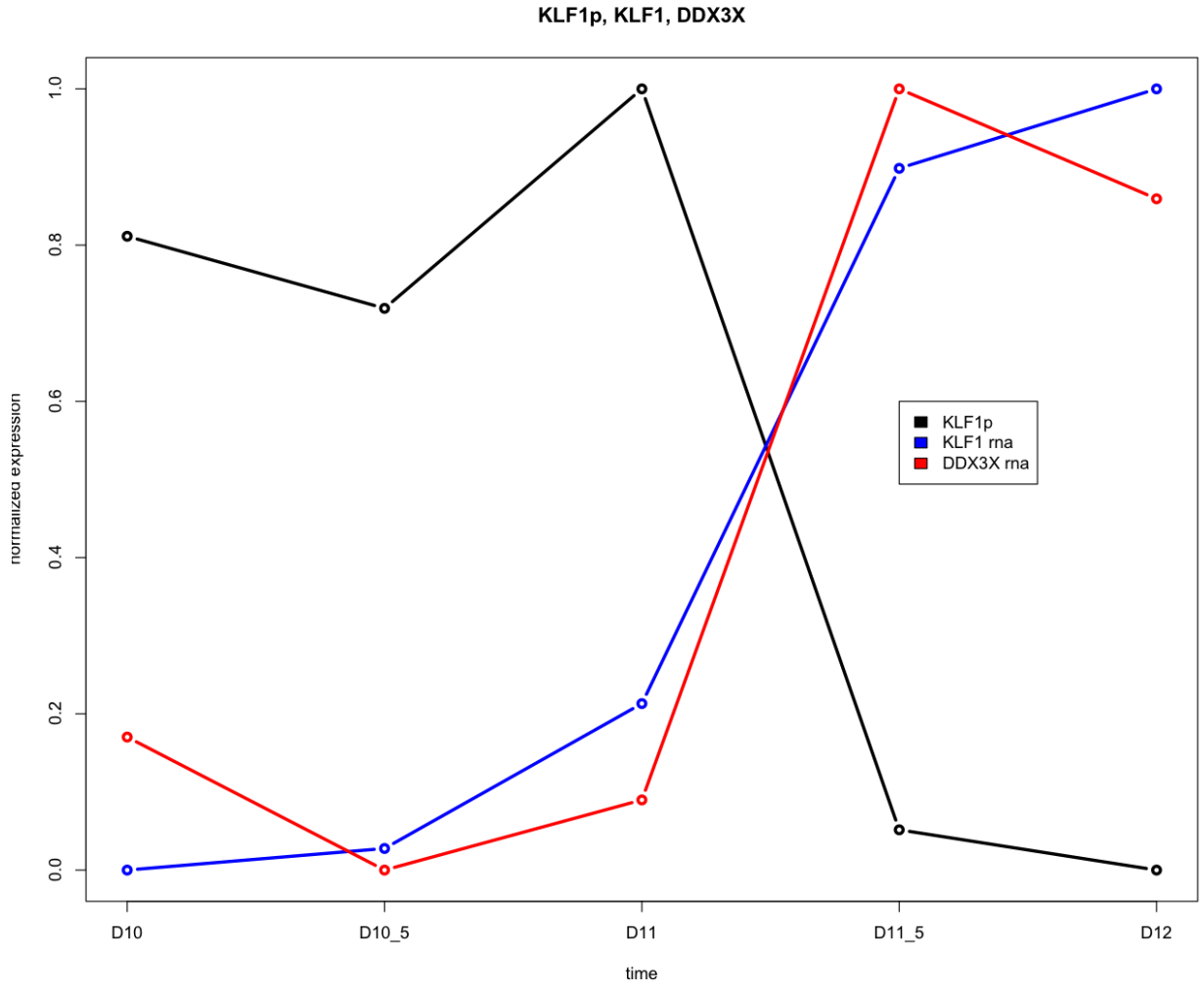
If an approach like this seems warranted, then a cytoscape shiny app would let you select and filter on the different kinds of data.

For instance: filter then display only the RBP-targetProtein edges which have
1) high anti-correlation
2) in cluster 3
3) 5' binding score > 10
4) target's rna is > someThreshold correlated with RBP rna.

Here is one instance of the scenario Jeff mentioned, filtering criterion #4, just above
- RBP rna is anti-correlated to protein expression
- RBP rna is positively correlated to expression of that protein's rna (see preceeding slide)

Jeff suggests integrating target rna, and eCLIP data (email 27jun2022)

If the RBP functions by interacting with target RNA it is possible that target RNA has a similar profile as the RBP RNA, or that target RNAs of a cluster have similar profiles (perhaps a sub-cluster). Furthermore, we know that about half of the targets have discordant RNA- protein profiles. It is possible that for some RBPs, RBP and target RNA expression profiles correlate whereas target protein is anti-correlated.



support exploration with cyjShiny?

email to jeff, marjorie, cory (28 jan 2022)

Here's a thought on what to do next - based upon the two additional data sets Jeff mentions:

1) mRNA levels of the target proteins. if the target mRNA is follows the RBP rna, then the RBP-target[gene|protein] relationship seems stronger
2) eCLIP binding data

We can put more information into the rbp/protein anti-correlations network:
- new edges for eCLIP binding (perhaps different edges by cell-type and binding region: 5', 3').
- bring along eCLIP score
- add an edge for positively correlated RBP/target rna (see KLF1 example below)

If an approach like this seems warranted, then a cytoscape shiny app would let you select and filter on the different kinds of data.

For instance: filter then display only the RBP-targetProtein edges which have
1) high anti-correlation
2) in cluster 3
3) 5' binding score > 10
4) target's rna is > someThreshold correlated with RBP rna.

To which Jeff replies:
The approach that you describe sounds great to me. For the eCLIPS data, could you include binding in the ORF as well as 5' and 3'?

DDX3X binding sites in 28 cluster 3 genes and in multiple sets of size ~28 randomly selected genes for comparison

	total	binding sites		gene counts	
group	genes	tx	5' only	tx	5' only
cluster 3	28	88	**41**	23/28	19/28
random #1	31	92	24	31/31	19/31
random #2	26	68	20	25/26	20/26
random #3	17	42	11	17/17	7/17
random #4	28	72	25	25/28	14/28
random #5	26	56	22	26/26	14/26
random #6	25	79	30	25/25	18/25
random #7	26	71	25	26/26	16/26
random #8	29	69	16	29/29	16/29
random #9	25	58	16	25/25	14/25
random #10	34	84	23	34/34	14/25
random #11	32	89	31	32/32	20/32
random #12	27	73	27	27/27	15/27
random #13	26	72	20	26/26	13/26

tx: DDX3X binding sites in any of 5', 3', CDS

DDX3X binding sites in 28 cluster 3 genes and in multiple sets of size ~28 randomly selected genes for comparison

study	candidates	recognized	sites	genes.all	5utr.sites	3utr.sites	cds.sites	5utr.genes	3utr.genes	cds.genes
cluster3	28	23	88	23	41	19	27	19	9	15
random.01	50	24	66	24	24	11	23	12	8	12
random.02	50	25	65	25	22	7	21	15	6	14
random.03	50	30	72	30	29	10	24	15	7	15
random.04	50	23	56	23	18	9	22	14	6	13
random.05	50	30	69	30	20	14	21	13	10	15
random.06	50	27	70	27	25	15	25	14	8	16
random.07	50	19	42	19	16	5	16	10	4	10
random.08	50	26	67	26	18	9	29	11	4	15
random.09	50	32	78	32	28	17	29	20	8	19
random.10	50	24	51	24	15	8	19	12	5	12

need to improve the randomly selected gene sets
some failures are due to, e.g.
check, in addition to random sets, the other clusters
  - must be expressed in some standard k562 (and other celltype) rna-seq dataset
  - form large pool of non-descending TMT proteins. draw 28 at a time, look for rbp binding sites by motif

goal is to **score** each rbp/motif in the genes in each cluster (or each random set of ~28 genes)
   rbp: all those whose rna is anti-correlated to the cluster's mean protein profile
   motif: find significant motifs in the binding sites in the genes in the cluster.

for all rna-binding-proteins, score each of the 55 expression clusters for rbp/motif enrichment.

jeff's summary, after the meeting:
We start with the clusters of RNAs and SRM proteins based on expression similarities during late times. There are ~50 clusters.

For each cluster, search the eCLIPs data for all RBPs to identify binding sites in the genes in the cluster. Once the sites are identified, search for binding motifs and then calculate an enrichment score for each motif for the genes in the cluster.

By surveying all RBPs from von Nostrand for each cluster, we have a way to assess RBP-cluster relationships independent of the RBP-cluster expression correlation.

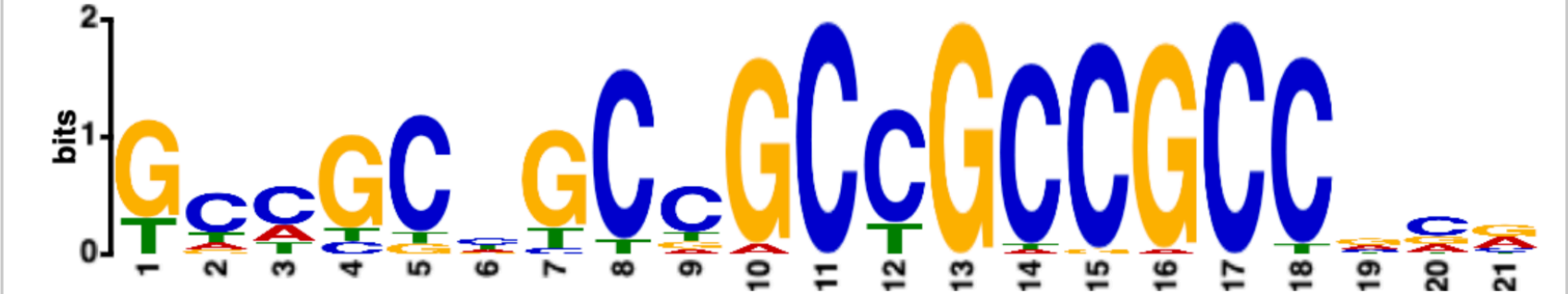
My plan (Paul):

postar2 database, 259 binding sites, 74/97 submitted genes
GCC * 4 motif, 6.7e-035 - but not with ENCODE K562 direct

Discovered Motifs

1.

***E*-value:** 6.7e-035 **Site Count:** 37 **Width:** 21



Standard | Reverse Complement

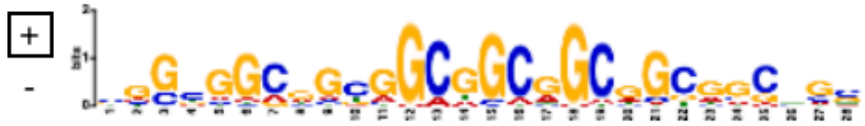
Log Likelihood Ratio: 494 **Information Content:** 24.5 **Relative Entropy:** 19.3 **Bayes Threshold:** 9.87854

	Name	Strand	Start	*p*-value	Sites
55.	GATAD2A-chr19:19385984-19386113-DDX3X-K562	+	19	1.41e-11	CTGTCCTGTC GCCGCTGCCGCCGCCGCCACA GCGGCCGCCG
51.	GATAD2A-chr19:19385979-19386036-DDX3X-K562	+	24	1.41e-11	CTGTCCTGTC GCCGCTGCCGCCGCCGCCACA GCGGCCGCCG
18.	CTCF-chr16:67562586-67562665-DDX3X-K562	-	21	1.14e-9	ACCGGCCACC GCCGCCGCCGCTGCCGCCAAA GAAGCAGCTC
17.	CTCF-chr16:67562565-67562720-DDX3X-K562	-	42	1.14e-9	ACCGGCCACC GCCGCCGCCGCTGCCGCCAAA GAAGCAGCTC
190.	SMARCB1-chr22:23786992-23787073-DDX3X-K562	-	13	1.49e-9	GCCGGGCTCC TCAGCCGCCGCCGCCGCCGCA GACGACGACC
185.	SMARCB1-chr22:23786973-23787058-DDX3X-K562	-	32	1.49e-9	GCCGGGCTCC TCAGCCGCCGCCGCCGCCGCA GACGACGACC
224.	STAT5A-chr17:42288470-42288541-DDX3X-K562	+	39	3.35e-9	AGTAACAGCC GCCGCTGCCGCCGCCGCCAGG AACCCCGGCC
184.	SMARCA4-chr19:10961018-10961173-DDX3X-K562	-	39	4.10e-9	ACGAAACAAA GAAGCCGCCGCCGCCGCCAAA CGAGGGGAAG
112.	KDM1A-chr1:23019592-23019651-DDX3X-K562	-	32	5.88e-9	TTGCTGCC GCTGCAGCCGCCGCCGCCGCG GCTGCCGCCT
171.	RAD21-chr8:116874692-116874762-DDX3X-K562	+	26	2.46e-8	TAGCTTCCGA GCAGCTCCCGCCGCCGCCACA GCCGGCGCCT

	Logo	E-value	Sites	Width	More	Submit/Download
2.	+ -	6.9e-018	9	50		
3.	+ -	2.1e-007	6	50		

ENCODE Van Nostrand, DDX3X 7307, K562 sequences, score >= 6.01
round(fivenum(tbl.sites$score), digits=2)
1.00 1.42 2.53 6.02 400.00
GCC * 4 motif, 1.0e-063
GGA * 6 motif, 6.1e-018

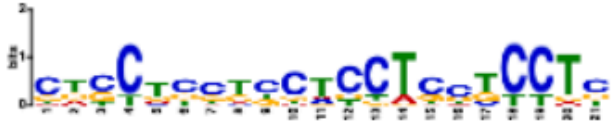
DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.		1.0e-063	400	28		
2.		6.1e-018	423	20		
3.		1.4e+002	2	28		
4.		2.6e+003	3	29		
5.		2.5e+005	100	11		

Stopped because requested number of motifs (5) found.

ENCODE Van Nostrand, DDX3X 7307, **HEPG2** sequences, score >= 8.9
top quartile of eCLIPs score
all.sites.HEPG2.count.12728.score.ge008.9.round
(fivenum(tbl.sites$score), digits=2)
[1] 1.00 1.40 2.78 8.91 400.00
GCC * 5 motif, 1.0e-055

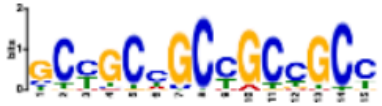
Discovered Motifs

	Logo	E-value	Sites	Width	More	Submit/Download
1.		8.8e-055	1499	15		
2.		2.6e-003	425	21		
3.		4.5e+005	5	28		
4.		2.4e+005	4	15		
5.		3.9e+006	2	21		

Stopped because requested number of motifs (5) found.

ENCODE Van Nostrand, DDX3X, 2176 sequences, score >= 20
GCC * 4 motif, 1.0e-063
GGA * 6 motif, 6.1e-018

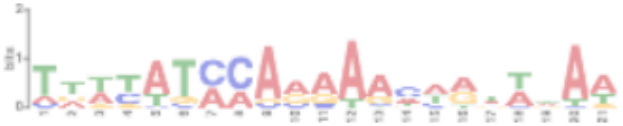
Discovered Motifs

	Logo	E-value	Sites	Width	More	Submit/Download
1.		1.3e-029	13	29		
2.		5.6e-011	11	28		
3.		1.7e-013	5	50		
4.		3.1e-006	191	15		
5.		1.1e+005	2	21		

Stopped because requested number of motifs (5) found.

ENCODE Van Nostrand, 14735 sequences, score >= 2.5 genes
even top motifs have poor significance

DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.		1.4e+006	533	15		
2.		2.0e+004	186	11		
3.		3.4e+002	688	15		
4.		5.3e+006	33	21		
5.		4.0e+007	2	14		

Stopped because requested number of motifs (5) found.

Van Nostrand bigBed files added to the RnaBindingProtein R package

The list is here: https://igv-data.systemsbiology.net/static/slides/erythropoiesis/encodeCatalog.txt

Unless you have other suggestions, I will next

- find enriched motifs in the top quartile by score of these full-genome binding sites, for each of the 150 RBPs, in each cell type
- do the same for only binding sites for genes in cluster 3
- see what these results suggest, look to you all for next steps

77 of the 150 RBPs were assessed in just one cell type, with this breakdown:

HepG2	K562
30	47

We have had some focus on these RBPs, all of which I'm glad to say are in the van Nostrand encode data:

	id	cellType	rbp
ENCFF562CTF		HepG2	DDX3X
ENCFF565FNW		K562	DDX3X
ENCFF733BQQ		HepG2	EIF3D
ENCFF323QTM		K562	FXR2
ENCFF483HFI		HepG2	FXR2
ENCFF416JIV		K562	IGF2BP2
ENCFF199OFQ		K562	RPS3
ENCFF848LEJ		HepG2	RPS3

Van Nostrand bigBed files added to the RnaBindingProtein R package

Jeff asks: van nostrand also reported 50 addtional eCLIPs datasets that they said "were not included in further analyses as they did not meet these stringent standards but contained a reproducible signal (Gene Expression Omnibus (GSE107768); Extended Data Fig. 1c, Supplementary Data 9)."

I'm wondering if we should include these datasets in our analysis? The attached table lists the RBPs corresponding to these addtional datasets. If some of these RBPs are of interest it might be worthwhile including the relevant data.

see next slide for more on known motifs

From Van Nostrand, Nature 2020, https://www.nature.com/articles/s41586-020-2077-3%C2%A0

We performed 488 eCLIP experiments, each including biological duplicate IPs and a paired size-matched input (Extended Data Fig. 1, Supplementary Data 4, 8–10, Supplementary Figs. 9,10). Manual quality assessment was based on IP validation, library yield, presence of reproducible peak or repeat family signal, motif enrichment (for **RBPs with known binding motifs**) and consistency with established biological functions, and yielded the 223 high-quality eCLIP data sets described here and released at the ENCODE Data Coordination Center (https://www.encodeproject.org). An additional 50 data sets were not included in further analyses as they did not meet these stringent standards but contained a reproducible signal (Gene Expression Omnibus (GSE107768); Extended Data Fig. 1c, Supplementary Data 9). Automated metrics also accurately classified quality for 83% of eCLIP data sets (Extended Data Fig. 1d, e, Supplementary Text, Supplementary Fig. 11). Data sets that passed manual but not automated quality assessment were released with specific exceptions noted (Supplementary Data 8). Although we have observed that stringent IP wash conditions generally limit the recovery of indirect interactions, we note that the eCLIP experiments described here did not include visualization of protein-associated RNA and thus independent validation of eCLIP profiles through comparison with in vitro motifs and knockdown-responsive changes provides essential validation of authentic binding.

Known RBP motifs

Published: 10 July 2013

A compendium of RNA-binding motifs for decoding gene regulation

Debashish Ray, Hilal Kazan, ... Timothy R. Hughes + Show authors

Nature **499**, 172–177 (2013) | Cite this article

Abstract

RNA-binding proteins are key regulators of gene expression, yet only a small fraction have been functionally characterized. Here we report a systematic analysis of the RNA motifs recognized by RNA-binding proteins, encompassing 205 distinct genes from 24 diverse eukaryotes. The sequence specificities of RNA-binding proteins display deep evolutionary conservation, and the recognition preferences for a large fraction of metazoan RNA-binding proteins can thus be inferred from their RNA-binding domain sequence. The motifs that we identify *in vitro* correlate well with *in vivo* RNA-binding data. Moreover, we can associate them with distinct functional roles in diverse types of post-transcriptional regulation, enabling new insights into the functions of RNA-binding proteins both in normal physiology and in human disease. These data provide an unprecedented overview of RNA-binding proteins and their targets, and constitute an invaluable resource for determining post-transcriptional regulatory mechanisms in eukaryotes.

As we look for enriched binding of specific RBPs in the erythropoiesis time series, initially ignorant of eCLIPs score thresholds, and of what constitutes significant enrichment, we will be well-served by studying known motifs: their size, their variability.

https://www.nature.com/articles/nature12311

known motif for PUM2
(see next slide also)

cell 2010: Transcriptome-wide identification of RNA-bind
https://www.sciencedirect.com/science/article/pii/S00928

PAR-CLIP is a transcriptome-wide crosslinking method for RNA-binding proteins (RBP) ▶ It is based on incorporation of photoactivatable nucleoside analogs into nascent RNA ▶ Characteristic sequence transitions in the prepared cDNA reveal the precise binding site ▶ We deduced binding motifs and preferences for 5 different RBP families.

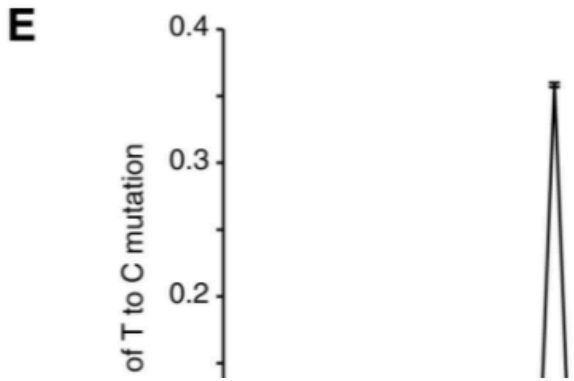
RNA transcripts are subject to posttranscriptional gene regulation involving hundreds of RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) expressed in a cell-type dependent fashion. We developed a cell-based crosslinking approach to determine at high resolution and transcriptome-wide the binding sites of cellular RBPs and miRNPs. The crosslinked sites are revealed by thymidine to cytidine transitions in the cDNAs prepared from immunopurified RNPs of 4-thiouridine-treated cells. We determined the binding sites and regulatory consequences for several intensely studied RBPs and miRNPs, including **PUM2**, QKI, IGF2BP1-3, AGO/EIF2C1-4 and TNRC6A-C. Our study revealed that these factors bind thousands of sites containing defined sequence motifs and have distinct preferences for exonic versus intronic or coding versus untranslated transcript regions. The precise mapping of binding sites across the transcriptome will be critical to the interpretation of the rapidly emerging data on genetic variation between individuals and how these variations contribute to complex genetic diseases.

3'UTR of ELF1

```
AAATGTTTTTAGATTACTTTTTCAACTGTAAATAATGTACATTTAATGTCACAAGAAAA # reads  error
-----------ATTACTTTTTCAACTGTAAACAATGTACATTT--------------- 581      1
-----------ATTACTTTTTCAACTGTAAATAATGTACACTT--------------- 239      1
-----------ATTACTTTTTCAACTGTAAATAATGTACATTT--------------- 113      0
--------------ACTTTTTCAACTGTAAACAATGTACATTTAAT------------ 82       1
-----------ATTACTTTTTCAACTGTAAATAATGTACATCT--------------- 67       1
```

3'UTR of HES1

```
GTGACTGACCATGCACTATATTTGTATATATTTTATATGTTCATATTGGATTGCGCCTT # reads  error
-------------CACTATATTTGTATACATTTTATATG------------------- 527      1
-------------CACTATATTTGTATACATTTTATATGT------------------ 130      1
-------------CACTATATTTGTATACATTTTATA-------------------- 48       1
-------------ACTATATTTGTATACATTTTATATG------------------- 40       1
-------------CACTATATTTGTATATATTTTATATGTTCACA------------ 22       1
```



https://zhanglab.c2b2.columbia.edu/mCrossBase/rbp.php?id=K562.PUM2

Data and motif summary of RBP **PUM2**

RBP summary

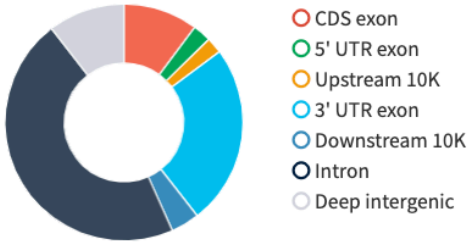
RBP Name:	PUM2
RBP description:	Pumilio RNA Binding Family Member 2
UniProt ID:	Q8TB72
Entrez Gene:	23369
Species:	Homo Sapiens
Protein domains:	PUMx1
Subcellular localization (1 (nucleus) – 5 (cytoplasm):	4(HepG2) 3(HeLa)

CLIP data summary

Data source: ENCODE **Cell-type/Tissue:** K562

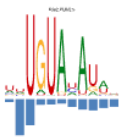
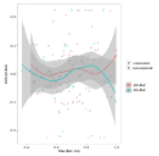
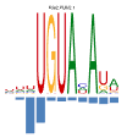
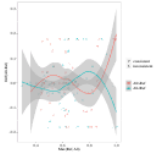
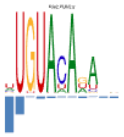
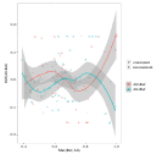
Unique tags: 4314871 **Peaks:** 18721

Genomic distribution:



Crosslink sites: 1221 (CIMS, deletion) ; 1154 (CIMS, insertion); 8447 (CIMS, substitution); 6431 (CITS)

Motif summary

Cluster	Motif Name	Consensus	Motif logo	Score	No. Sites	AI scatter plot	AI consistency	AI p-value
1	K562.PUM2.0	• TGTACAG • TGTACAT • TGTATAT • TGTACAA		374.8431186	2526		0.466666667 (7/15)	0.696380615
1	K562.PUM2.1	• TTGTACA • CTGTACA • GTGTACA • ATGTACA		368.0471196	1836		0.6 (9/15)	0.303619385
2	K562.PUM2.2	• GTACAGA • GTACAAA		257.3104898	736		0.625 (10/16)	0.227249146

email from me, then jeff (9 feb 2022)

paul: Meanwhile, keeping on with the "RBP motifs in clustered binding sites, enriched or not?" question, I reckoned I should start from, or at least consider, known RBP binding sites. Two resources came up. Please see slides 116 - 119.

https://igv-data.systemsbiology.net/static/slides/erythropoiesis/ddx3x.pdf

One lesson in all of this prior work is that RBP binding sites tend to be about 6 bases long. I did not realize that.

jeff: I think this is a good place to start. Van Nostrand suggests combining eCLIPs-derived motif information with binding motifs derived from other sources such as in vitro data.

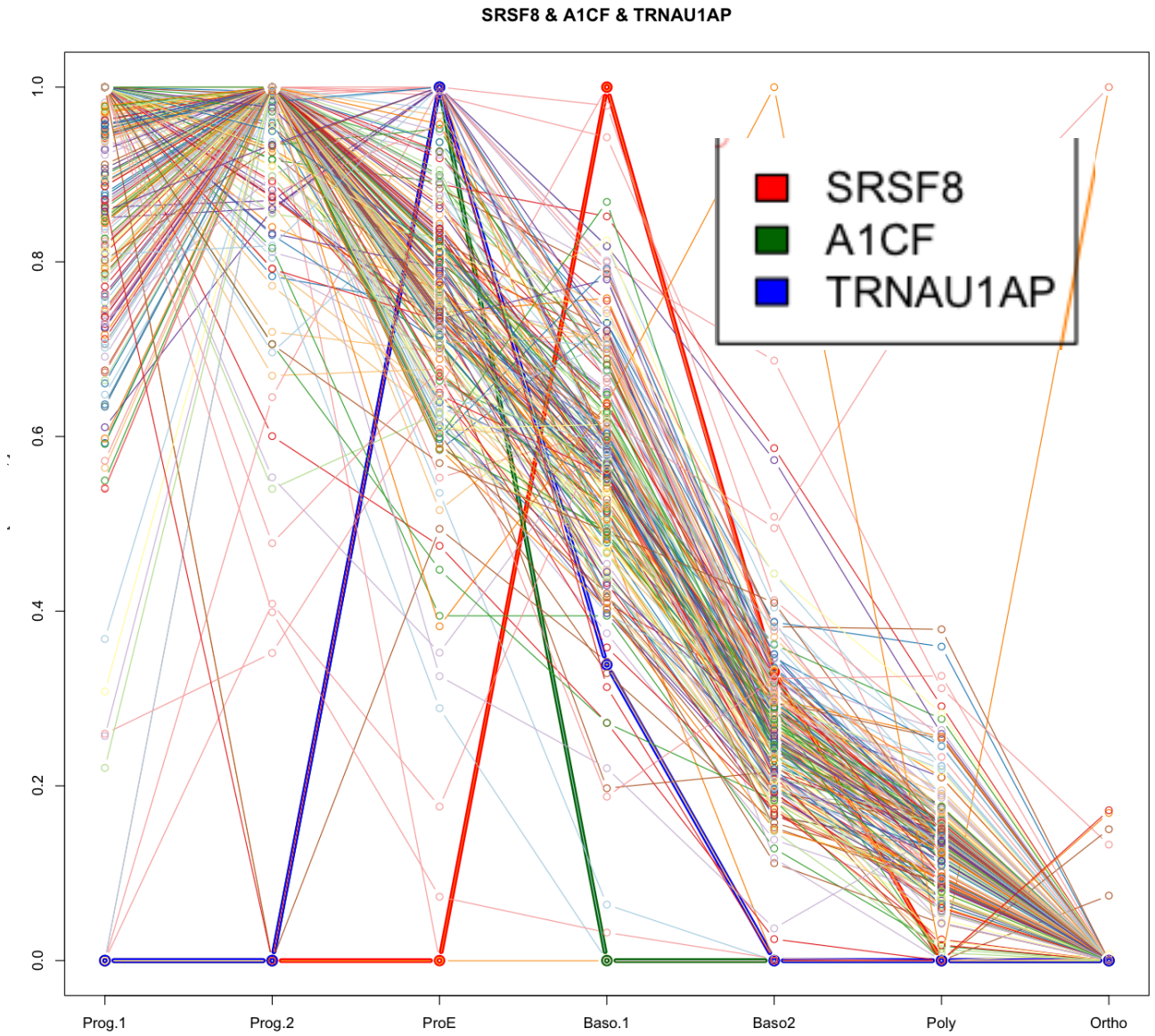
I attach a file from van Nostrand that contains in vitro derived motifs from 78 RBPs that could be used in conjunction with the resources that you found. The motifs found in this study also appear to be short; 5 bases.

The Mcross database appears to contain the eCLIPs data from van Nostrand. Looks like a good resource.

todo from zoom meeting (15 feb 2022)

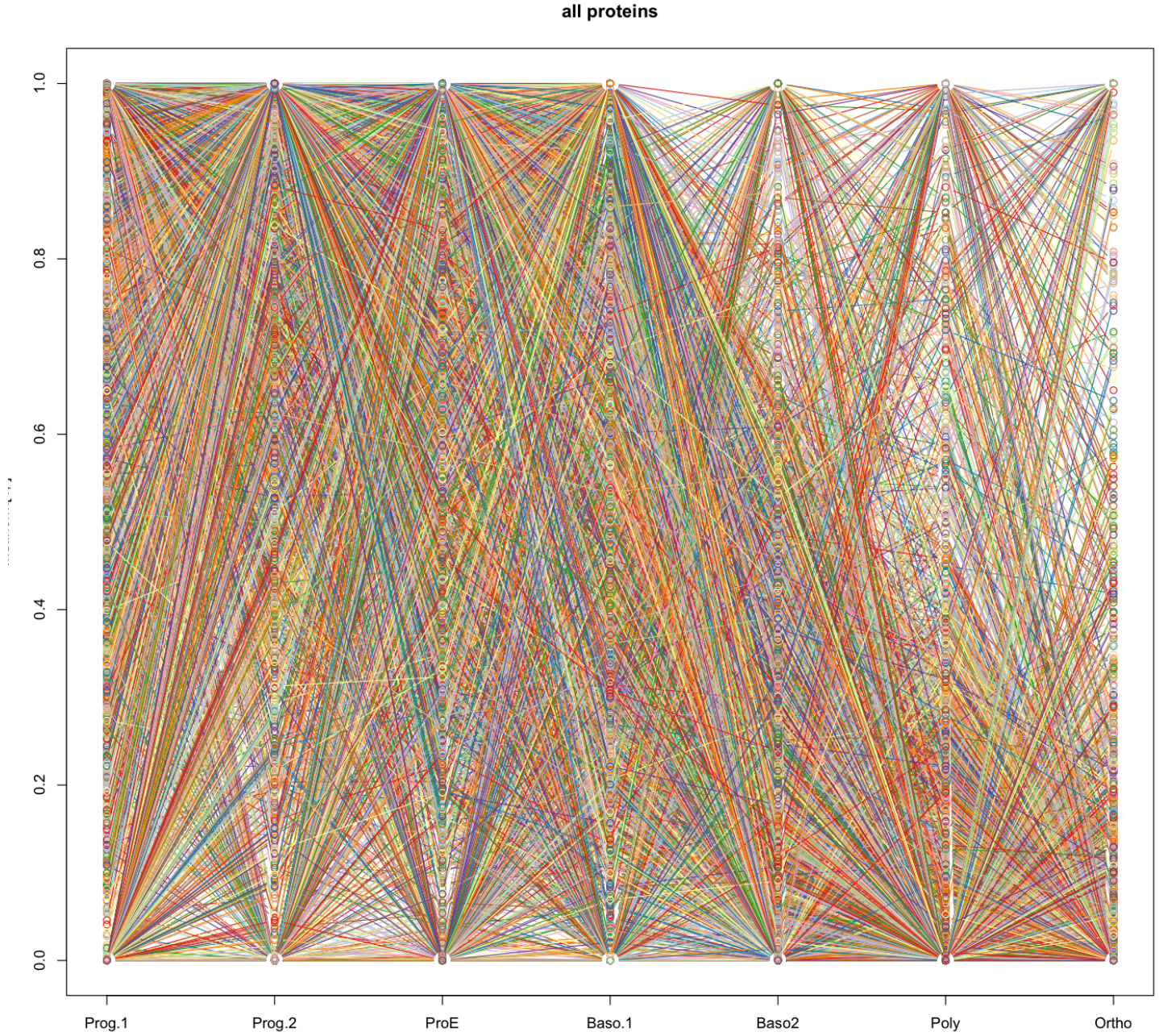
1. obtain new external protein data set: gautier
2. cluster on all groups
3. do the same on our own rna data, see if there may be (smaller) clusters which share regulatory sites, motifs. or something. especially interesting if gautier or brand small tight clusters have members which have our usual protein/rna discordance at proEB - EB.
4. do this clustering FAST! report back in.
5. try clustering our proteins on their own.
6. "gautier could lead us to rbp priortization, since g provides our first protein rbp data"
7. also: gautier provides our first opportunity to do protein-only clustering.
8. for instance a gautier rbp anti-correlated to a TF protein at one time period would suggest new functional insight.

gautier 6130 proteins
columns: Prog.1, Prog.2, ProE, Baso.1, Baso2, Poly, Ortho

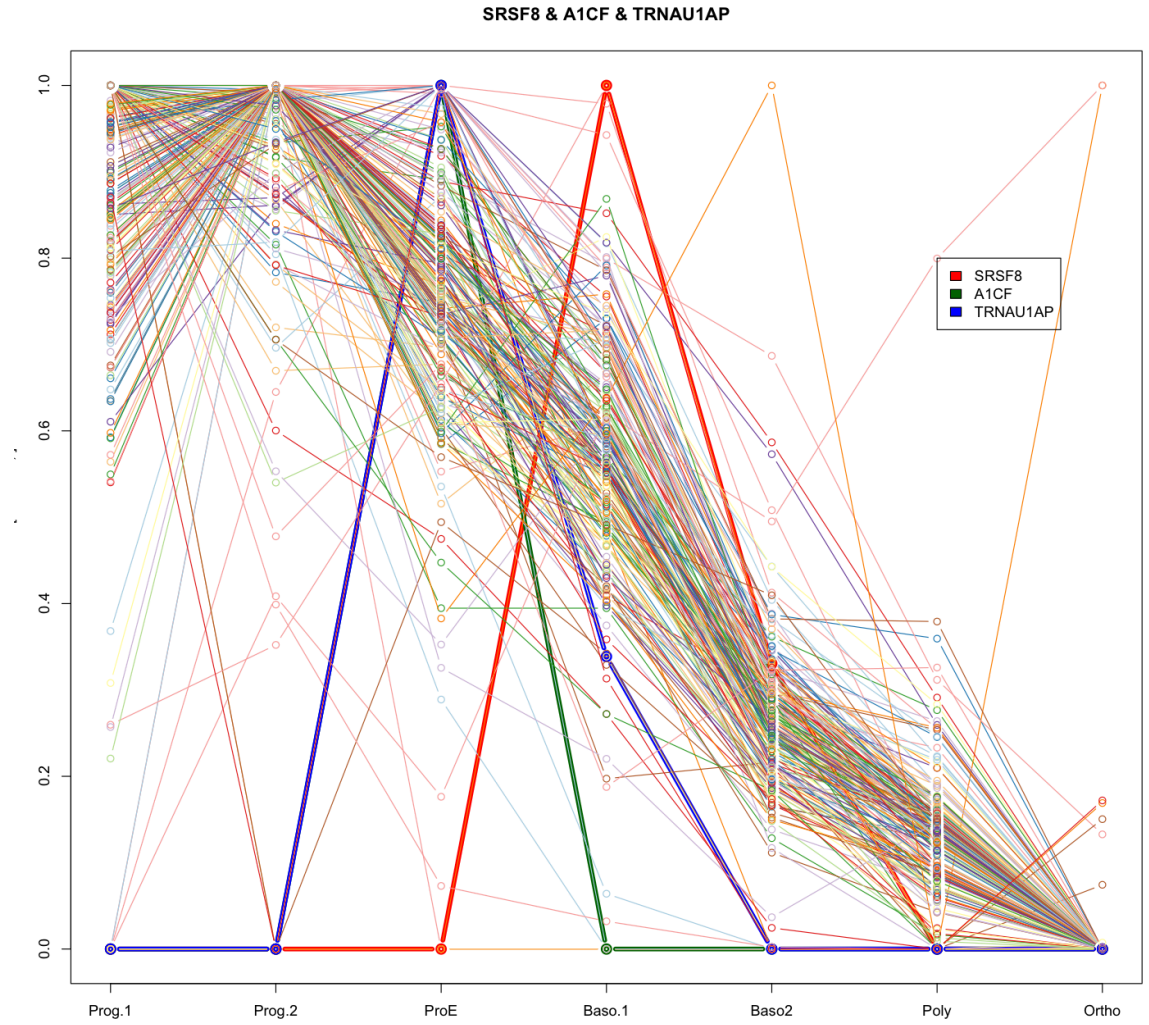


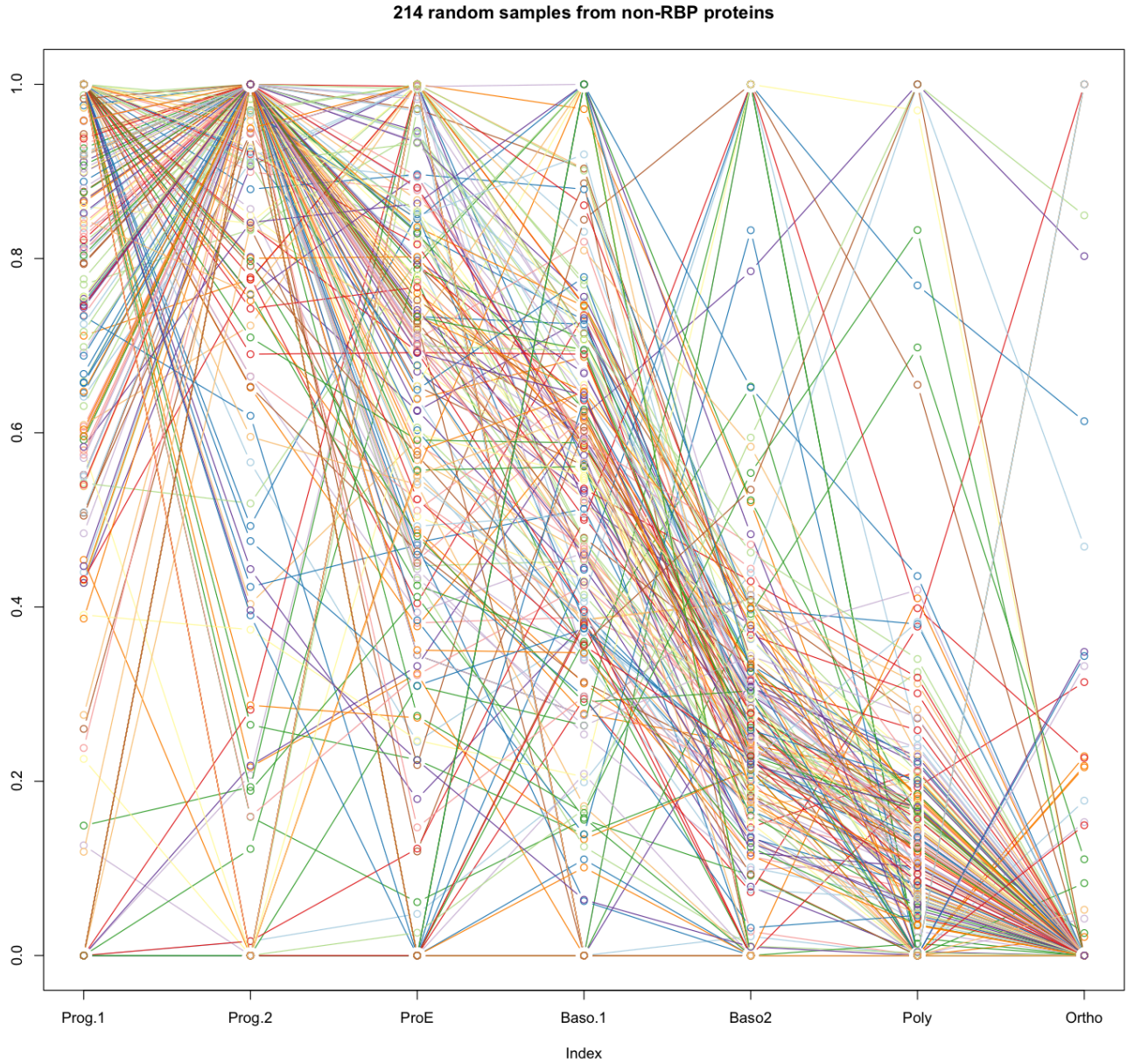
1st exploration of the gautier data

- normalized all data to 0-1 range
- calculate ProE-Baso.1 delta
- only these three had abs(delta) > 0.8
- overall similarity of the trajectory of most rbps

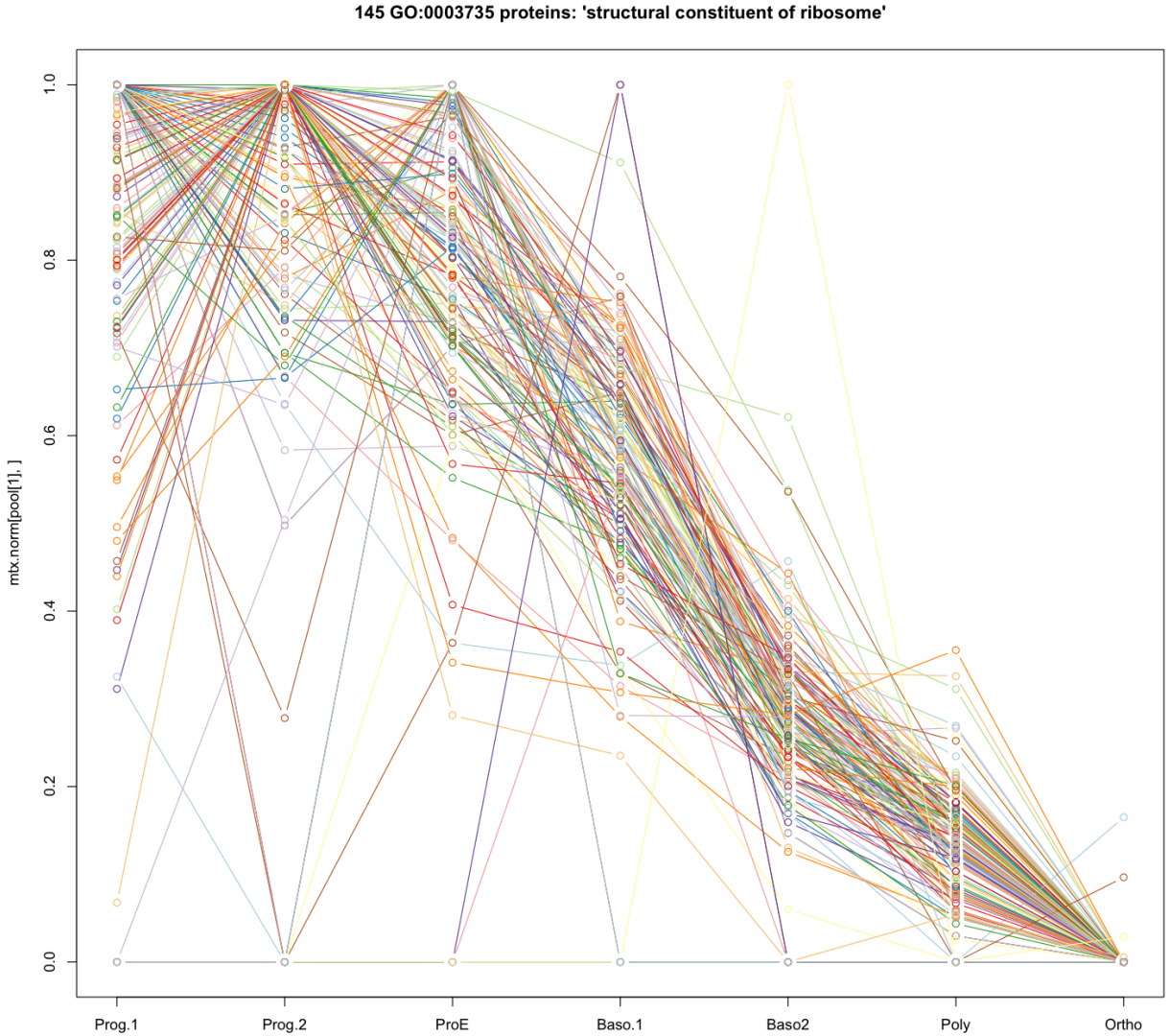

all proteins
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

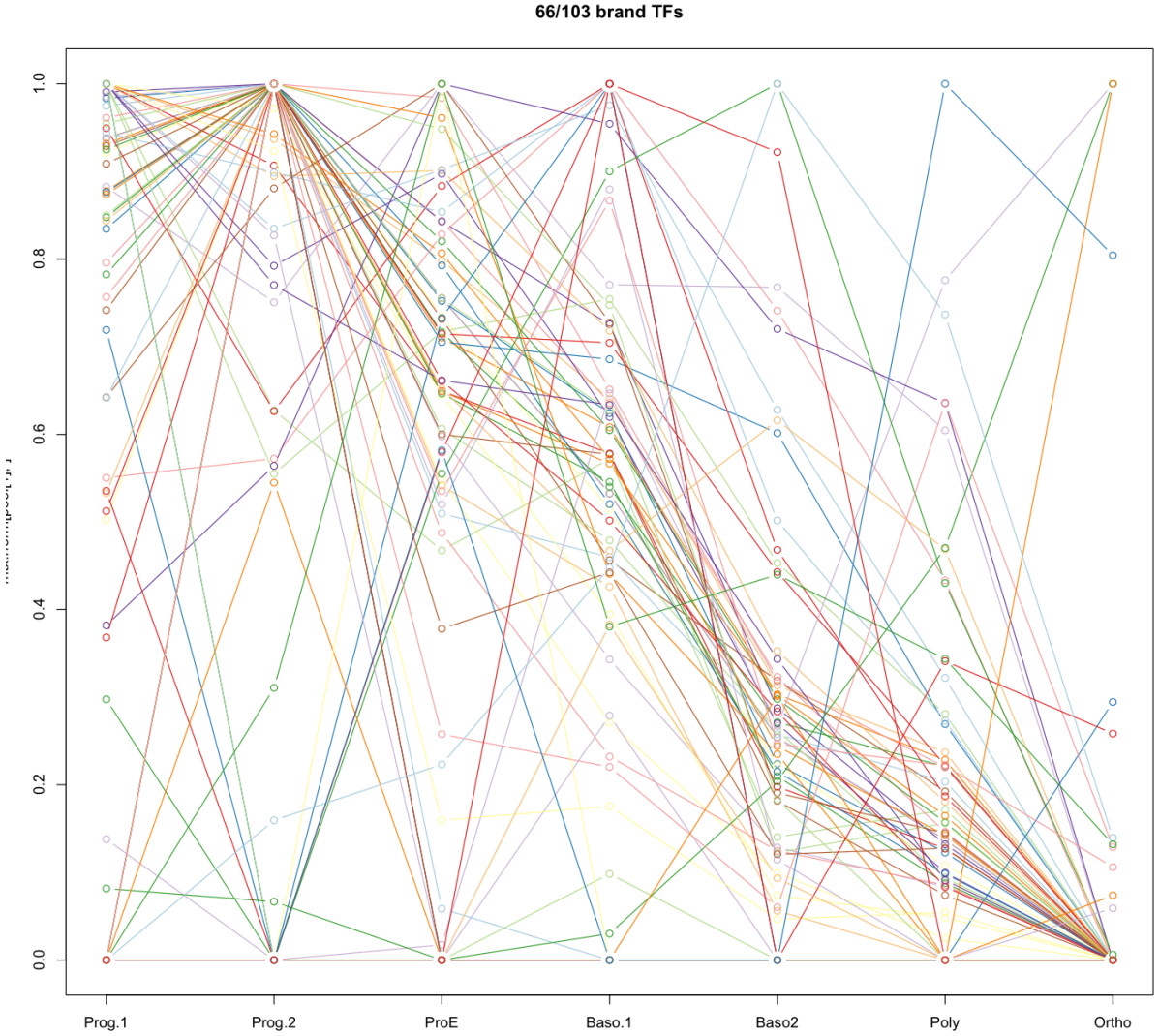
214 RBPs vs a representative random sample of 214 non-RBPs





Ribosomal Proteins in Gautier, 66/103 from Ranish srm





Jeff proposes ribosomal protein units

email 2/24, 3pm: So the profiling of ribosomal subunits, RBPs and our SRM measured TFs is quite interesting. I will summarize here some ideas for possible next steps.

1) Classify ribosomal subunits based on protein expression profiles. For example some subunits appear to be decreasing faster than others, and other subunits are actually increasing. This could mean that the composition of ribosomes is changing during differentiation. This analysis could suggest different ribosomal complex compositions which could allow us to focus on specific subunits for IP analysis. We can also do RIP-seq to see if the ribosome with different compositions are specifically associated with those proteins.

2) Classify RBPs based on protein expression profiles. similar rationale as described above.

3) Compare Gautier expression profiles to our SRM expression profiles. Mainly a sanity check

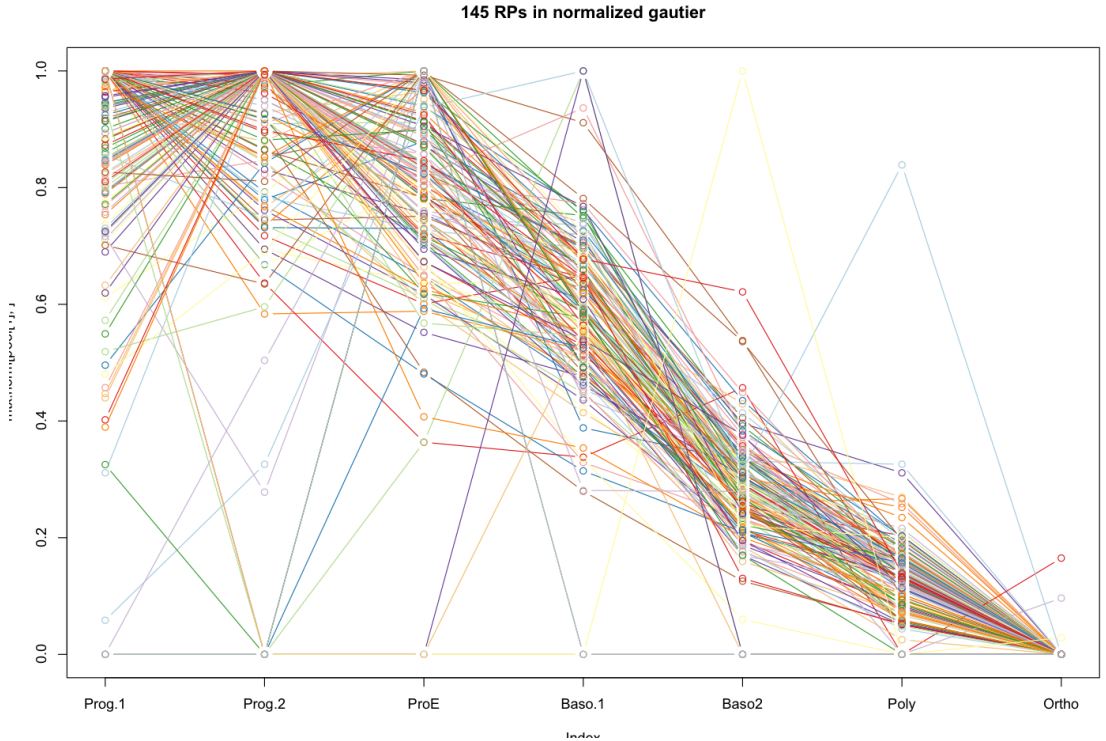
4) Identify the proteins that are increasing during the proEB transition. This can give us a target list to try to identify RNA regulatory elements in these genes that could explain why these proteins are not decreasing.

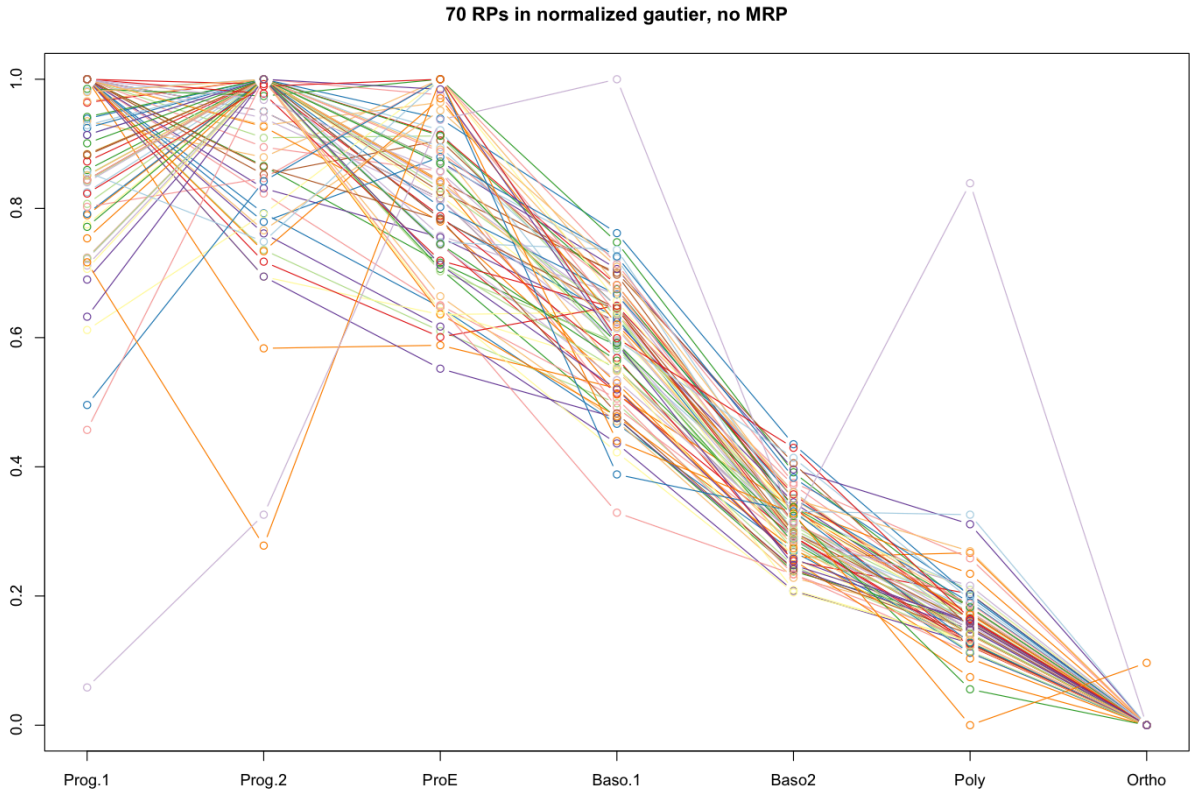
email 2/24, 5pm: Once we know which ribosomal SUs are differentially expressed we can check the structure of the ribosome to see where they are located. Also mutations in a few ribosomal SUs (RPL11/uL5, RPS7/eS7, and RPS26/eS26) are associated with defects in erythropoiesis. Also, Maria Barna's group has shown that the ribosomes with different compositions exist in mouse ES cells and these ribosomes translate specific types of transcripts.

The polysome machine could be very useful in the near future. Jacob is designing mass spec assays to quantify many ribosomal subunits. It would be great to isolate polysomes for these measurements.

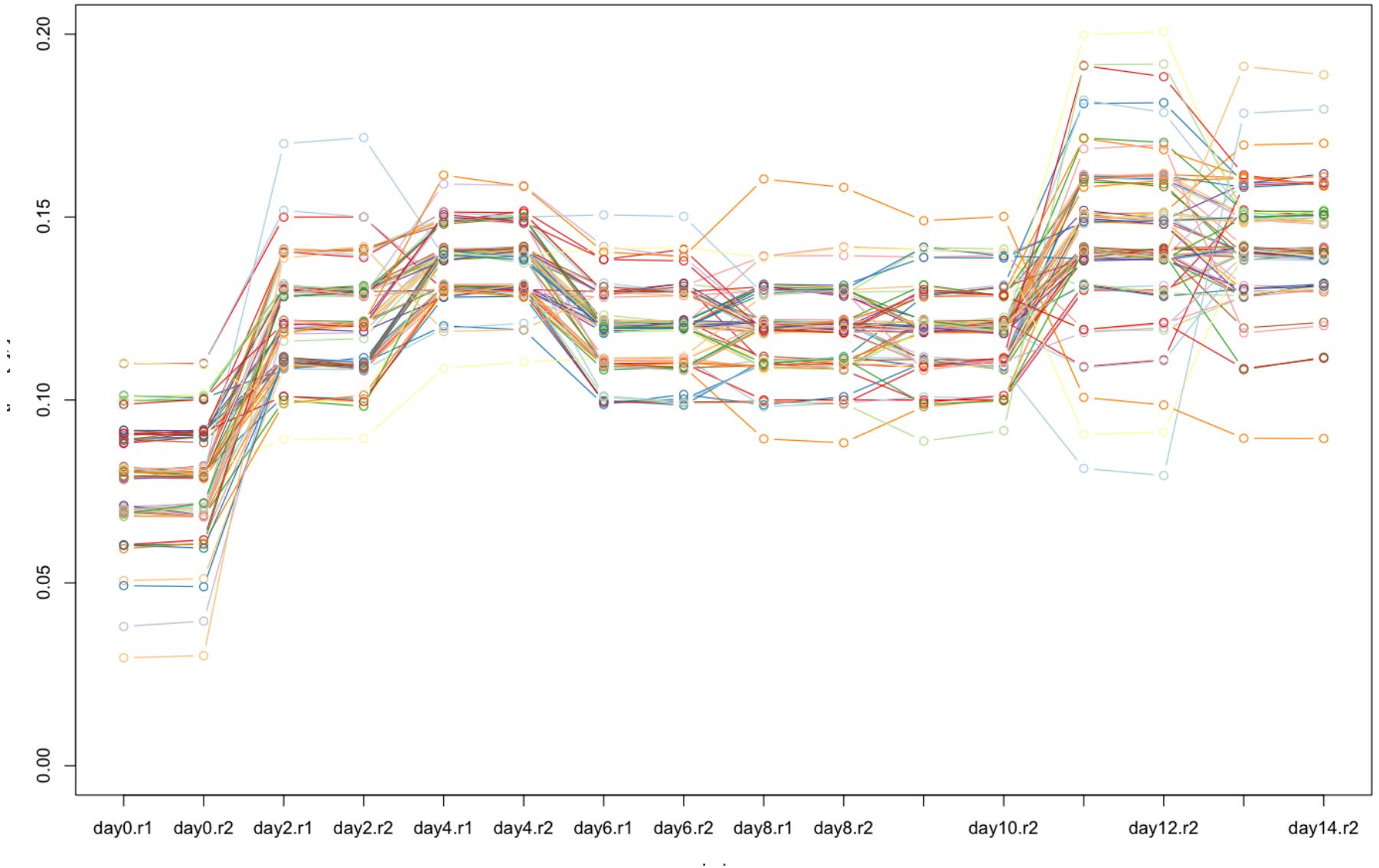
genuth and barna 2018: The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. https://pubmed.ncbi.nlm.nih.gov/30075139/

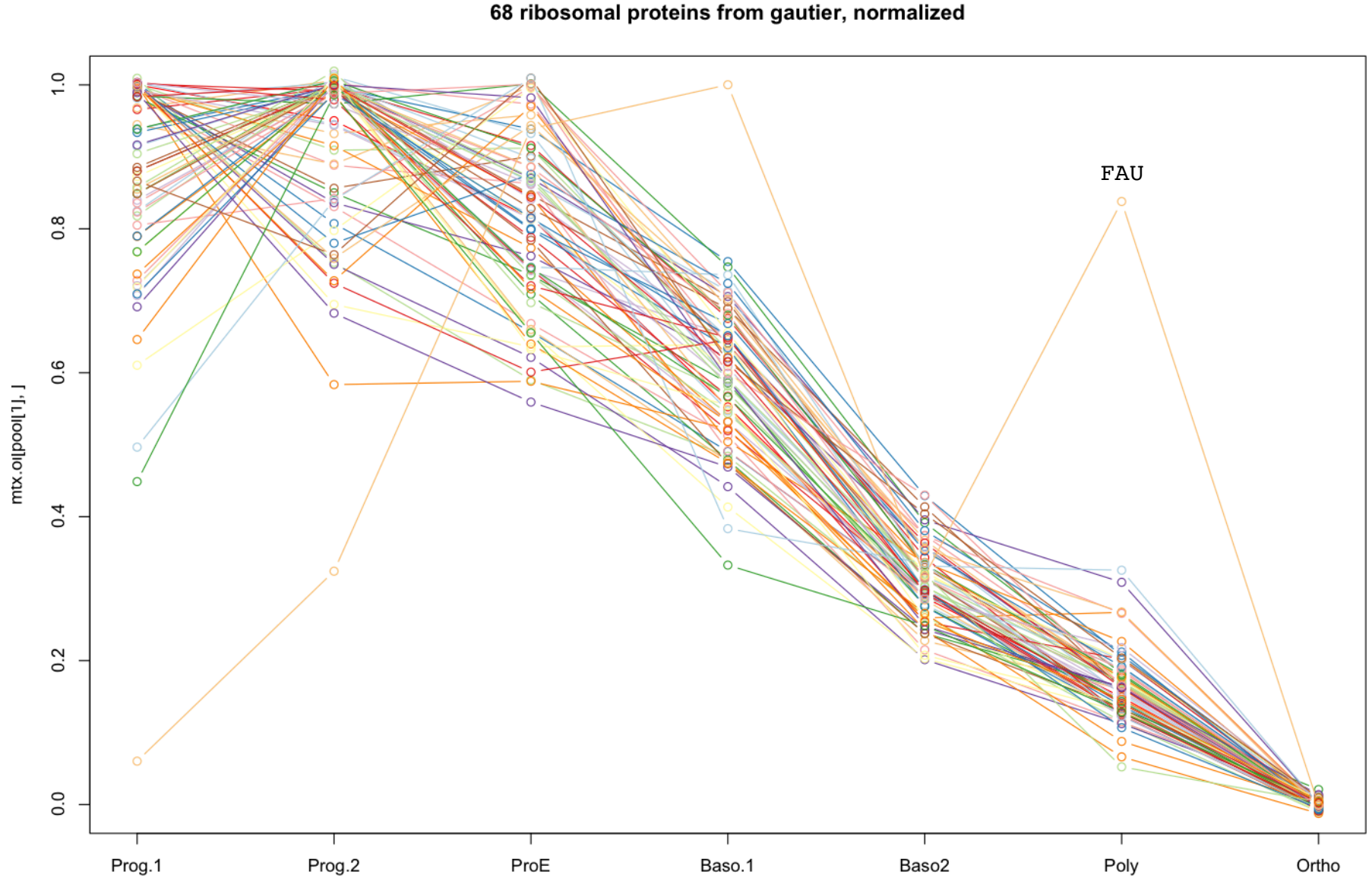
they mention 80 "core RPs"
https://www.genenames.org/cgi-bin/genegroup/download?id=1054&type=branch lists 168, 54 of which are mitochondrial, 54 (SU=L) + 35 (SU=S) -> 89.
gautier: 6161 gene symbols





only itraq exhibits late departures in RP pattern

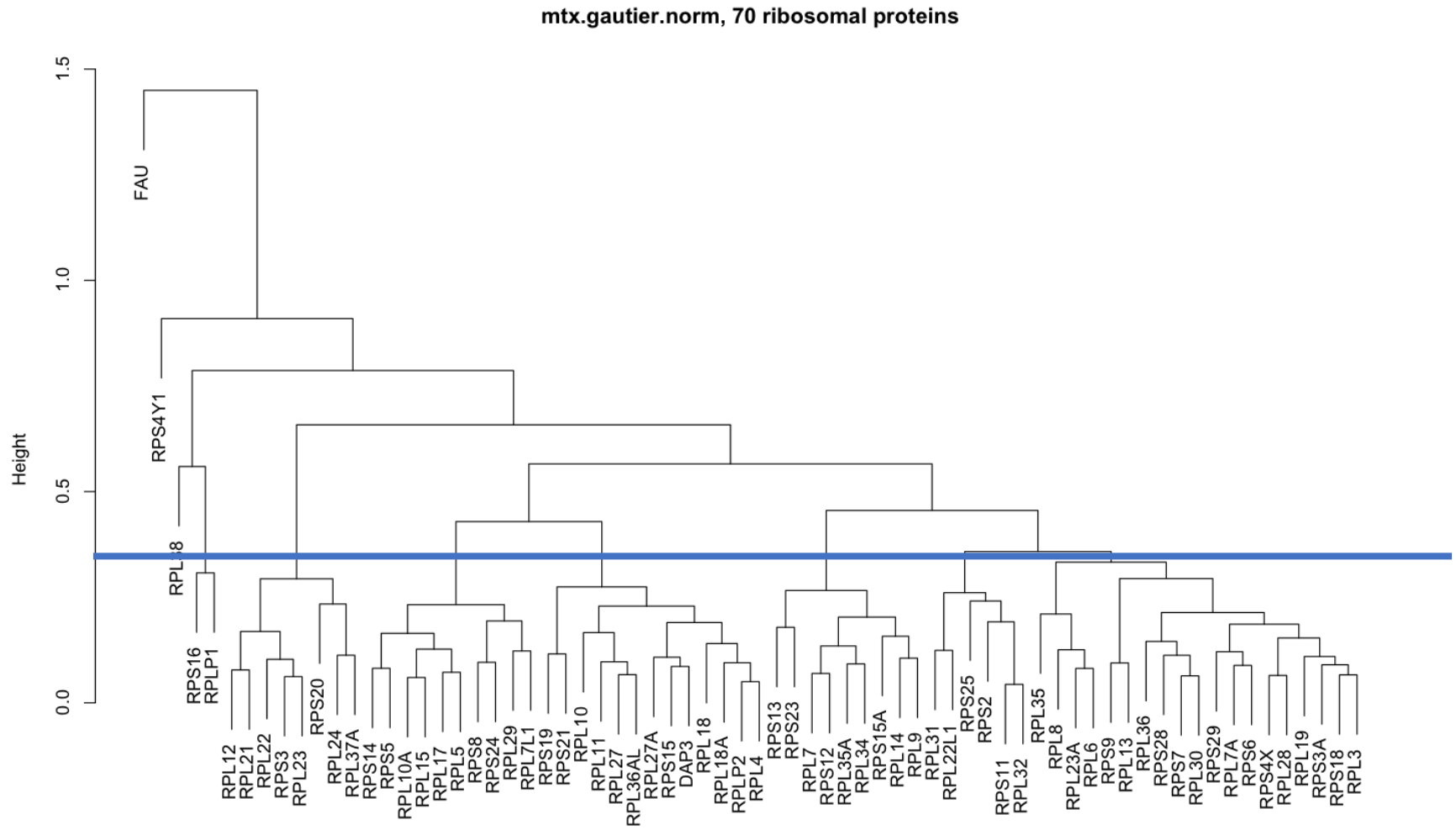


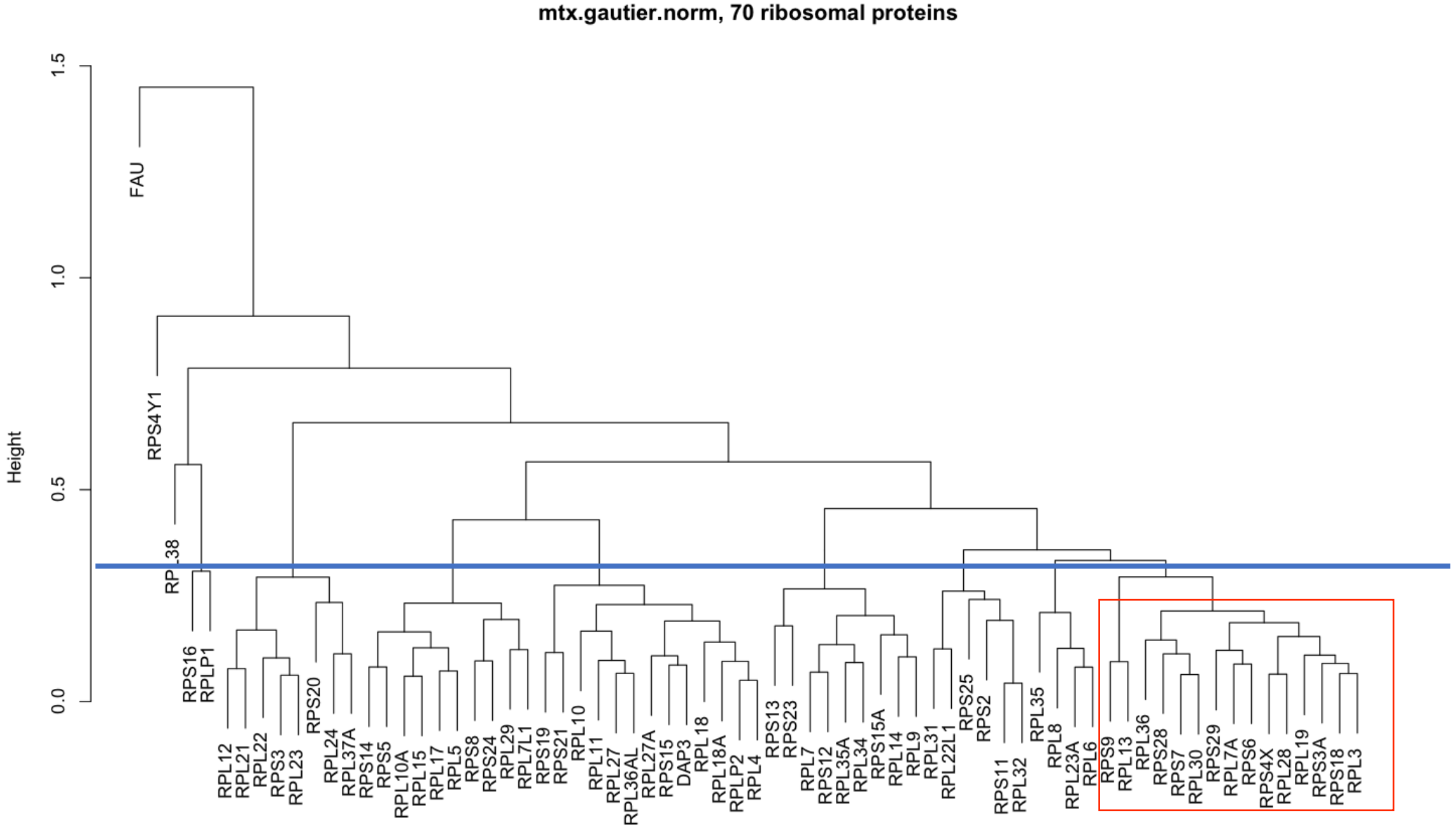

68 ribosomal proteins from gautier, normalized
mtx.ol[pool[1],]
0.0
0.2
0.4
0.6
0.8
1.0
FAU
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

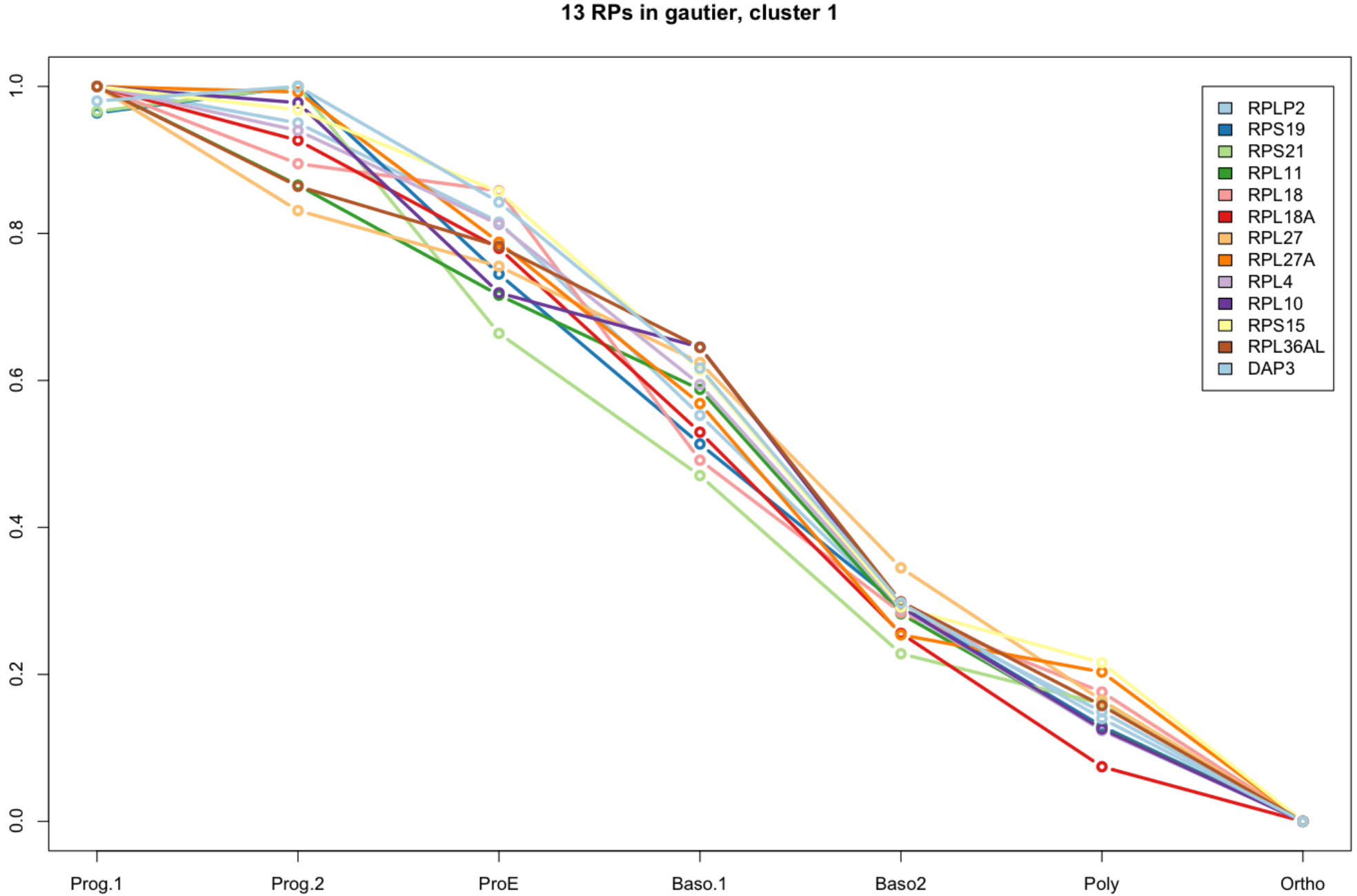
todo (zoom meeting 2/28)

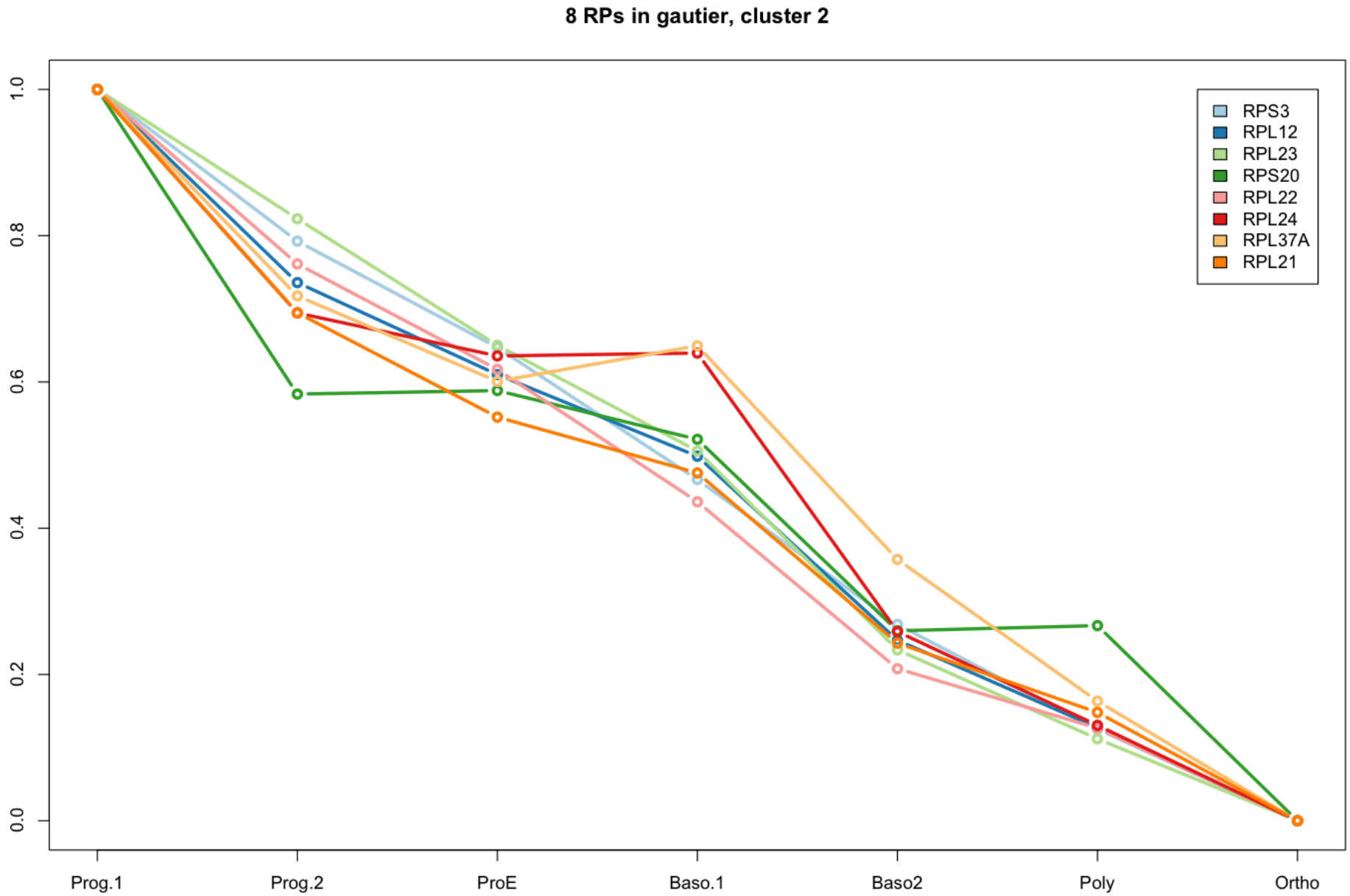
extract tight, aberrant, outlier clusters in both gautier ribosomal & rna-binding proteins

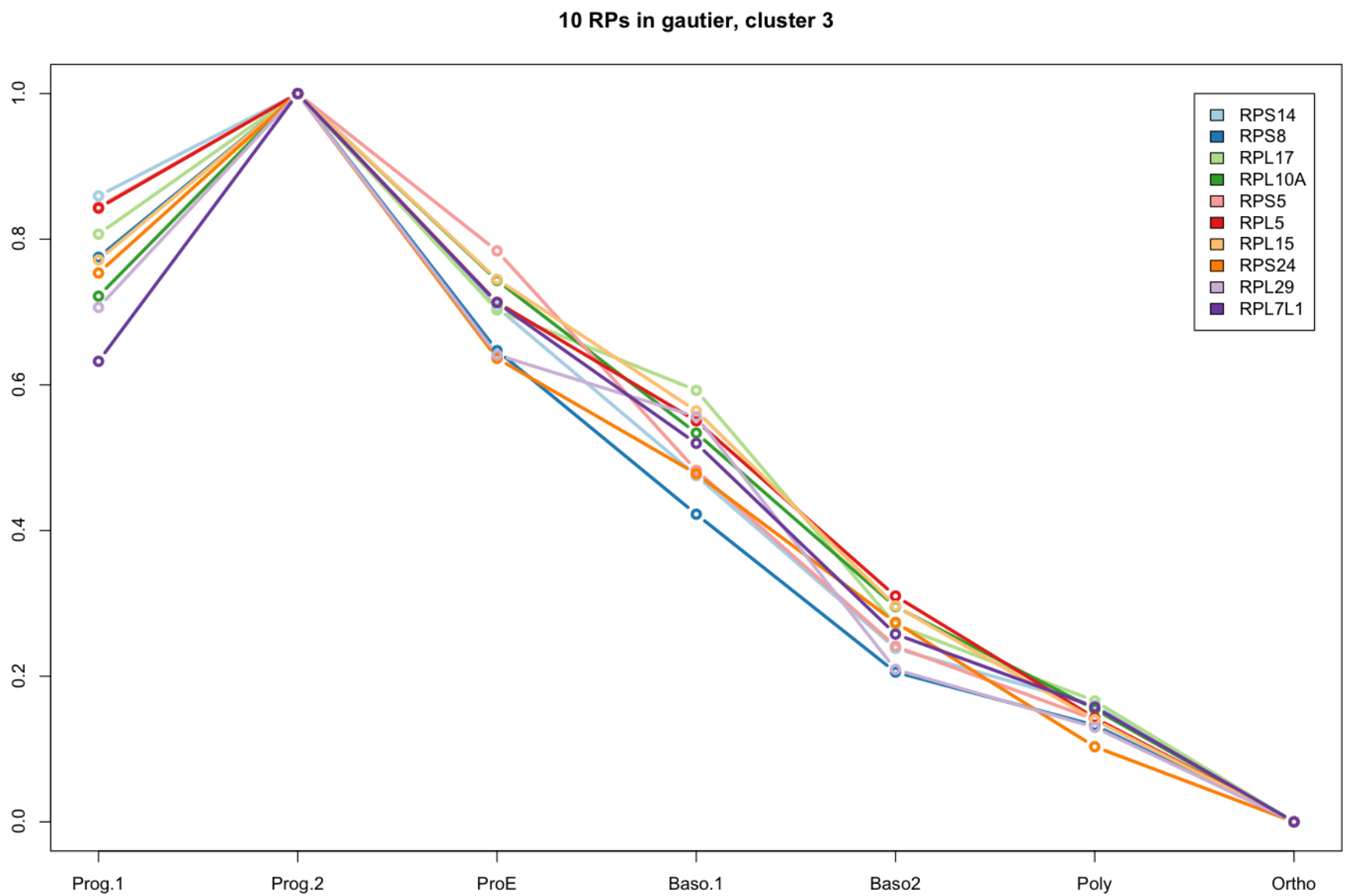
we should find RP FAU almost alone in its cluster. For RPs, do two versions: normalized and raw. RBPs just normalized.

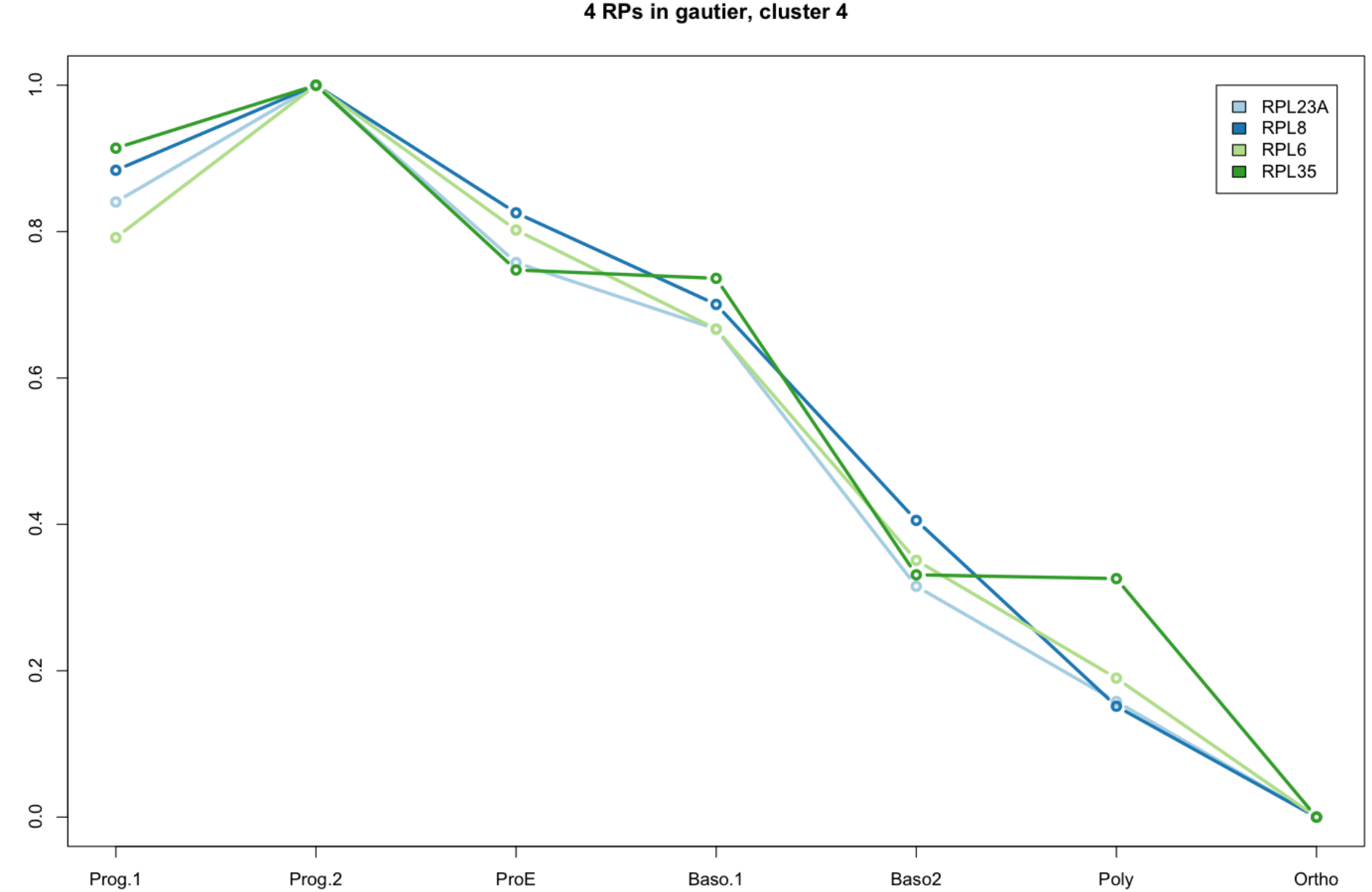


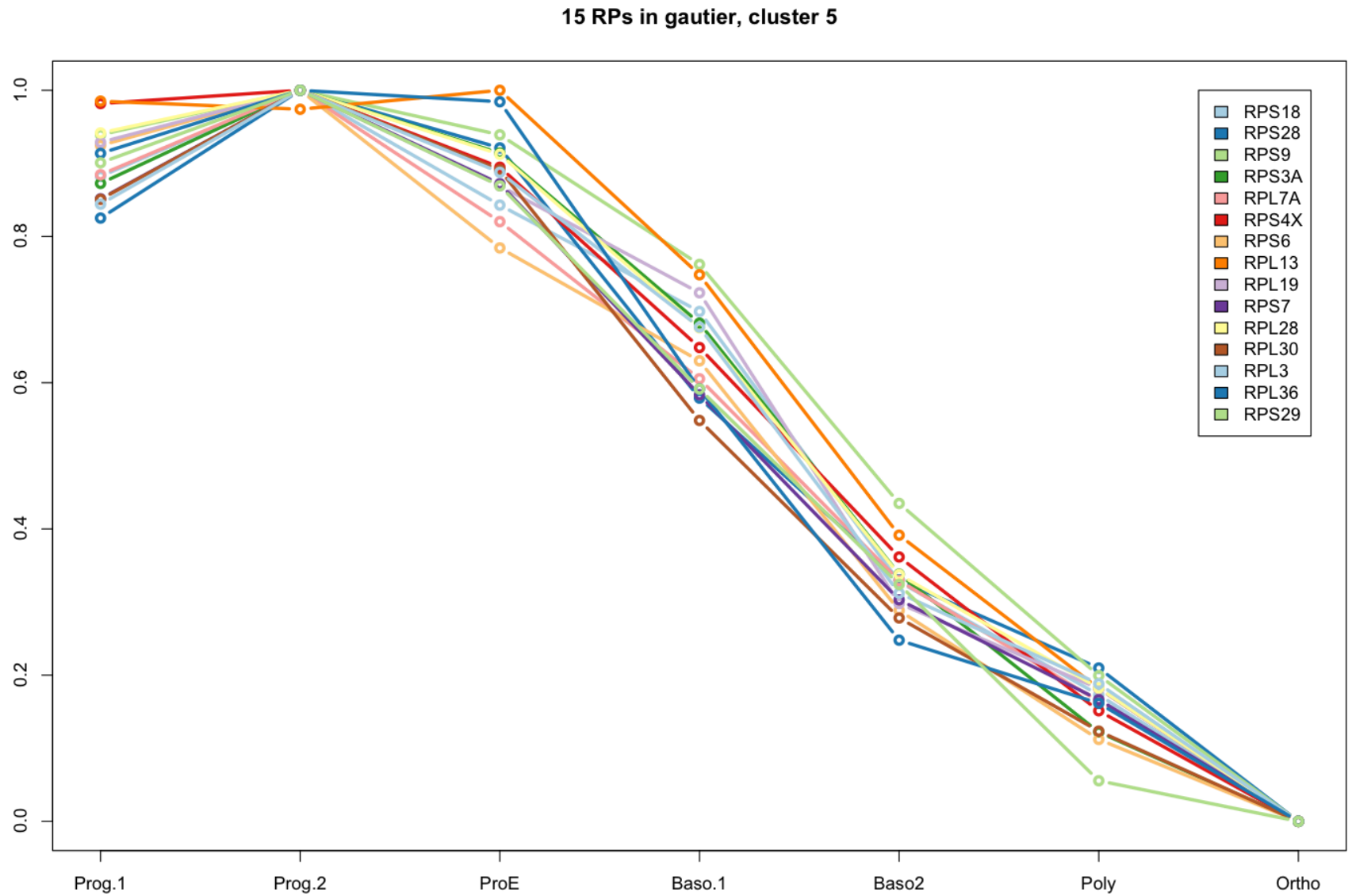

mtx.gautier.norm, 70 ribosomal proteins
Height
0.0
0.5
1.0
1.5
FAU
RPS4Y1
RP_38
RPS16
RPLP1
RPL12
RPL21
RPL22
RPS3
RPL23
RPS20
RPL24
RPL37A
RPS14
RPS5
RPL10A
RPL15
RPL17
RPL5
RPS8
RPS24
RPL29
RPL7L1
RPS19
RPS21
RPL10
RPL11
RPL27
RPL36AL
RPL27A
RPS15
DAP3
RPL18
RPL18A
RPLP2
RPL4
RPS13
RPS23
RPL7
RPS12
RPL35A
RPL34
RPS15A
RPL14
RPL9
RPL31
RPL22L1
RPS25
RPS2
RPS11
RPL32
RPL35
RPL8
RPL23A
RPL6
RPS9
RPL13
RPL36
RPS28
RPS7
RPL30
RPS29
RPL7A
RPS6
RPS4X
RPL28
RPL19
RPS3A
RPS18
RPL3

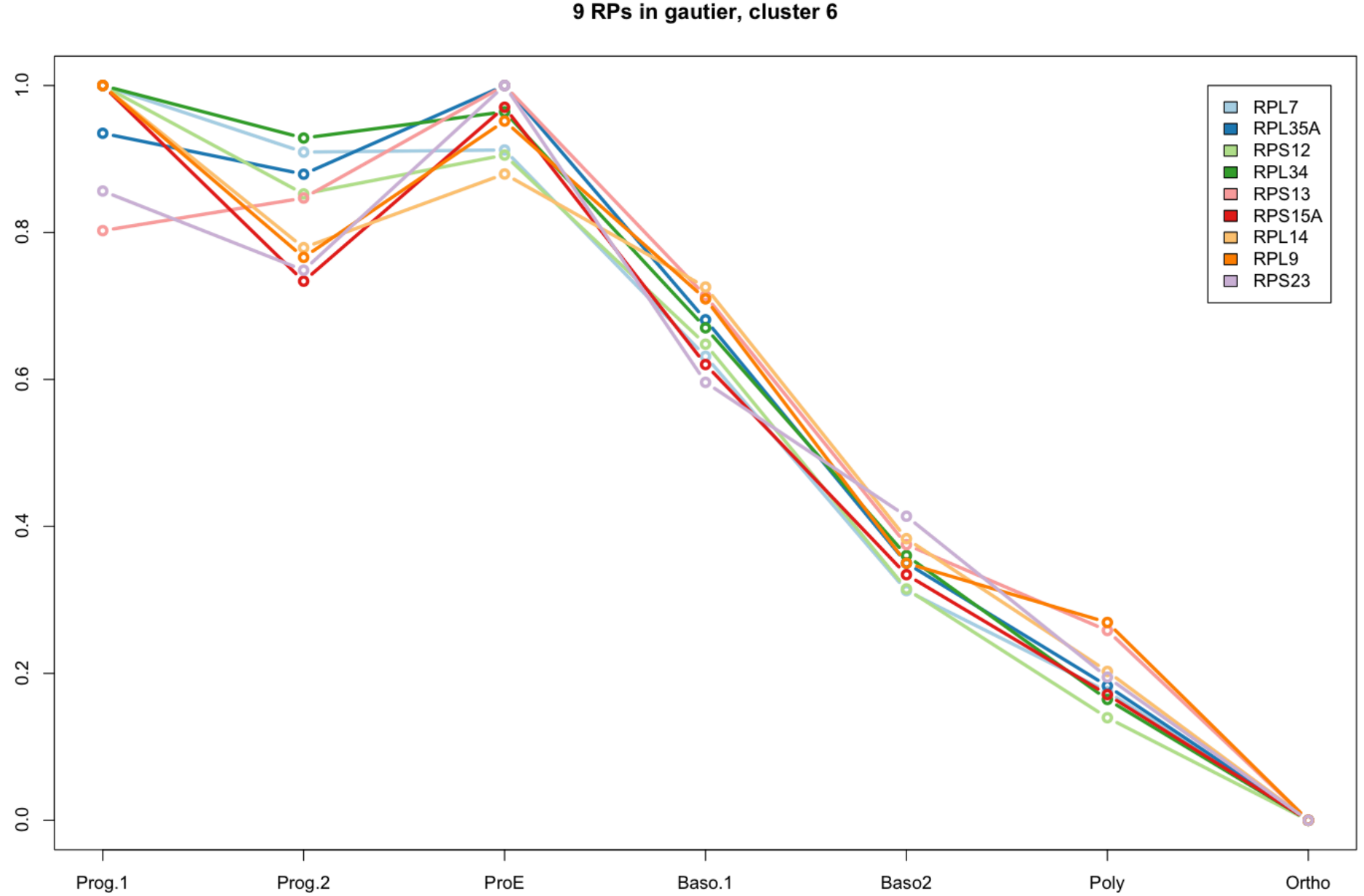

13 RPs in gautier, cluster 1
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
RPLP2
RPS19
RPS21
RPL11
RPL18
RPL18A
RPL27
RPL27A
RPL4
RPL10
RPS15
RPL36AL
DAP3

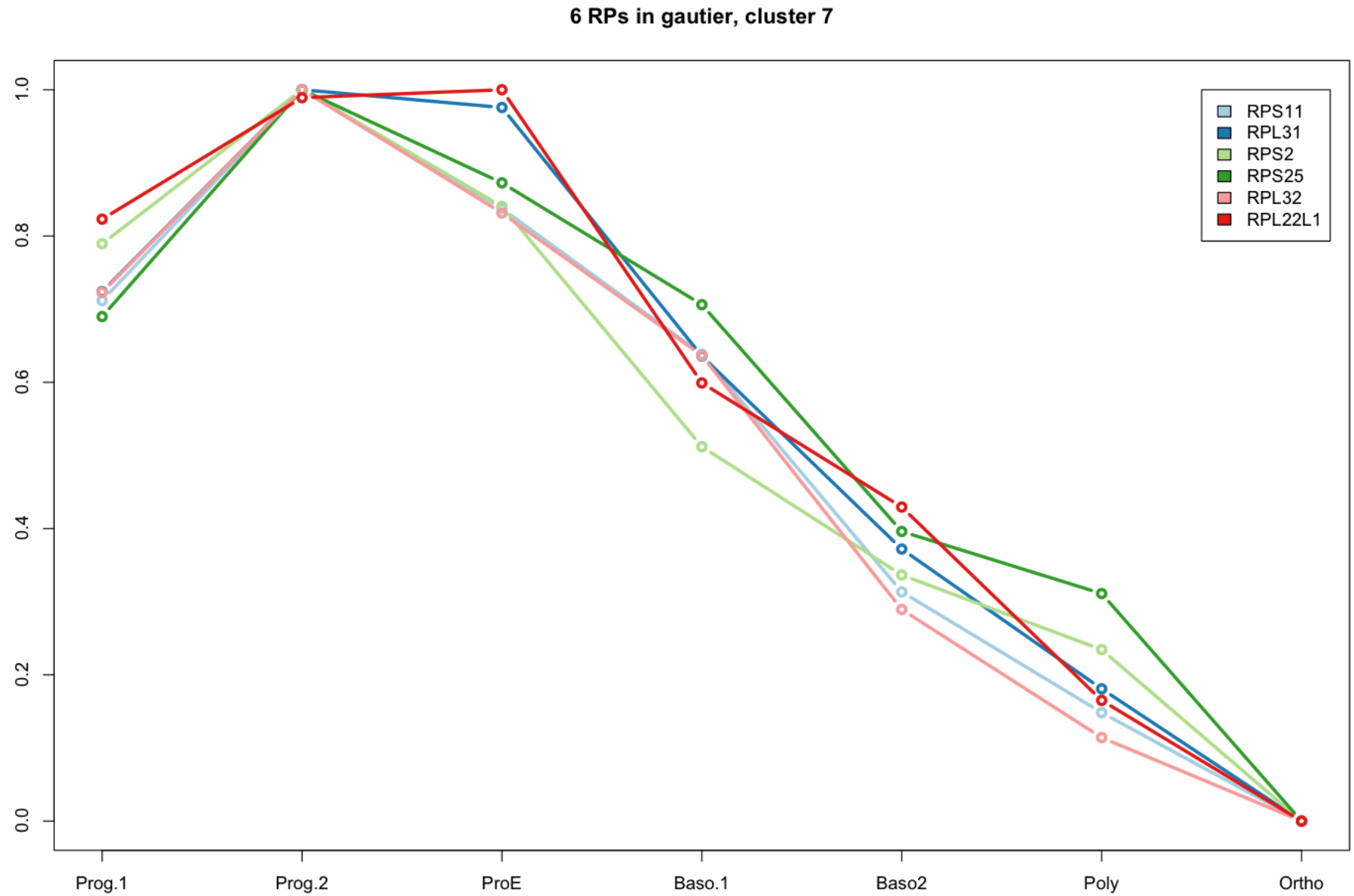

8 RPs in gautier, cluster 2
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
RPS3
RPL12
RPL23
RPS20
RPL22
RPL24
RPL37A
RPL21

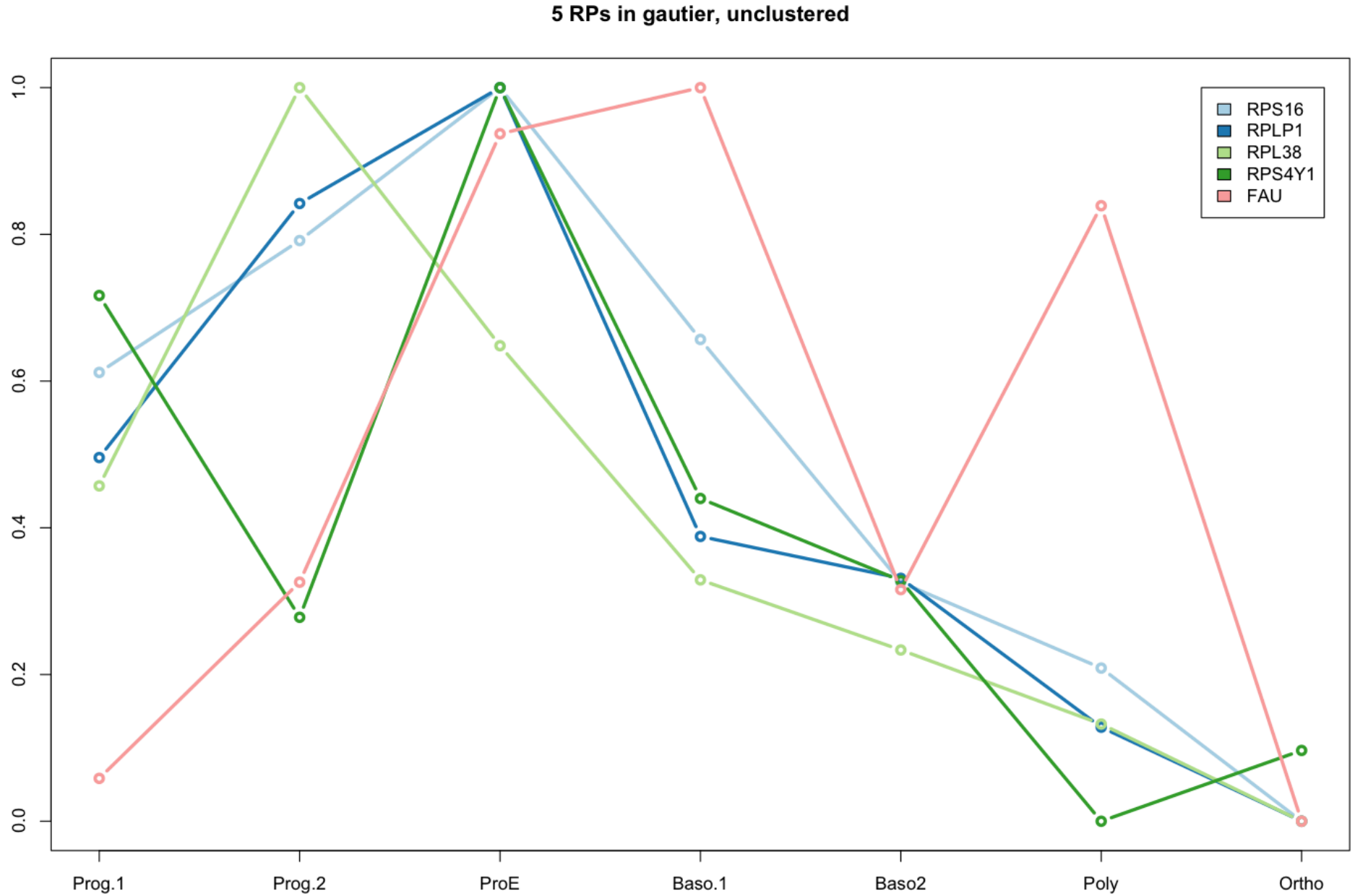

10 RPs in gautier, cluster 3
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
RPS14
RPS8
RPL17
RPL10A
RPS5
RPL5
RPL15
RPS24
RPL29
RPL7L1

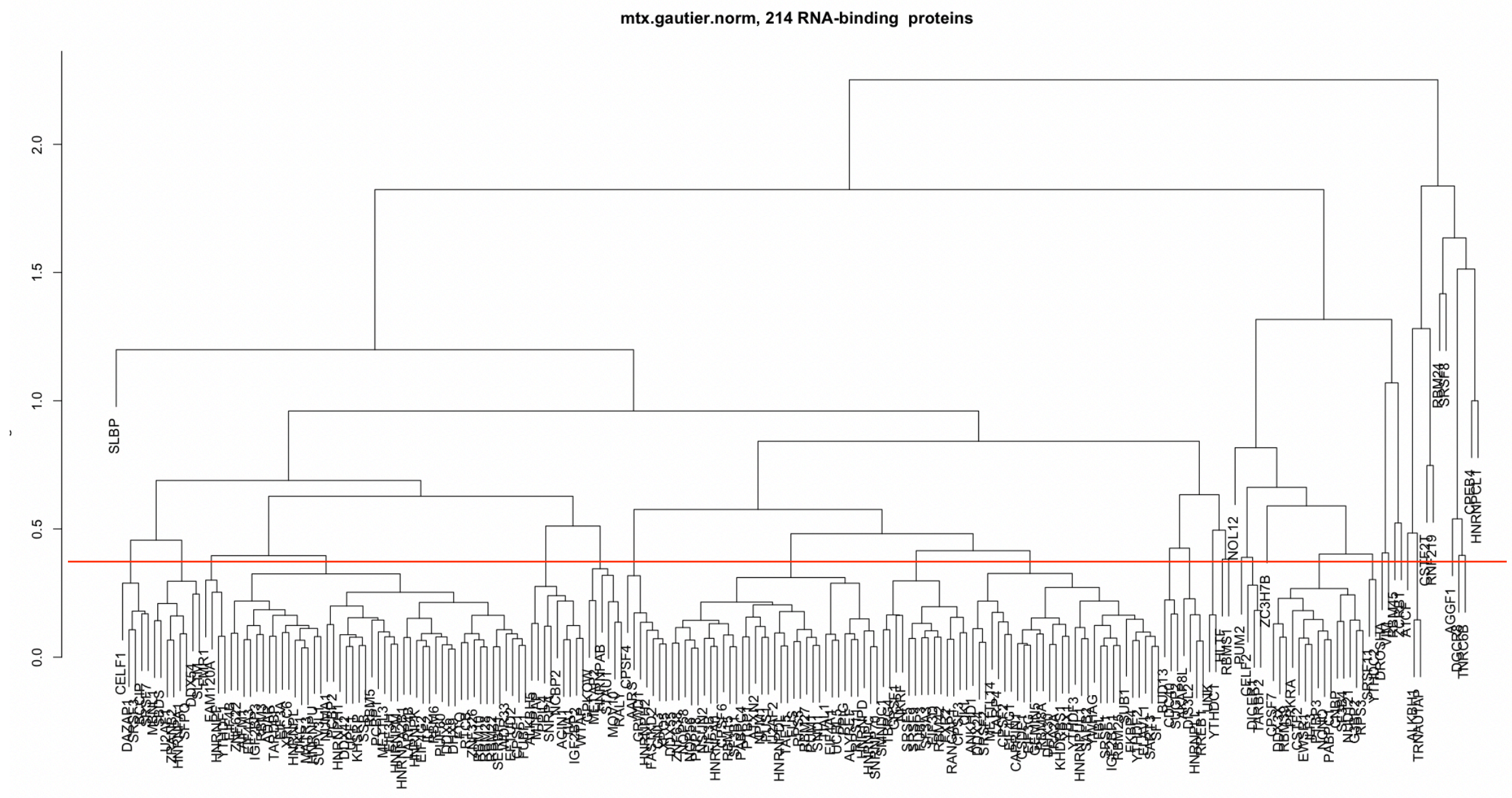

4 RPs in gautier, cluster 4
RPL23A
RPL8
RPL6
RPL35
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

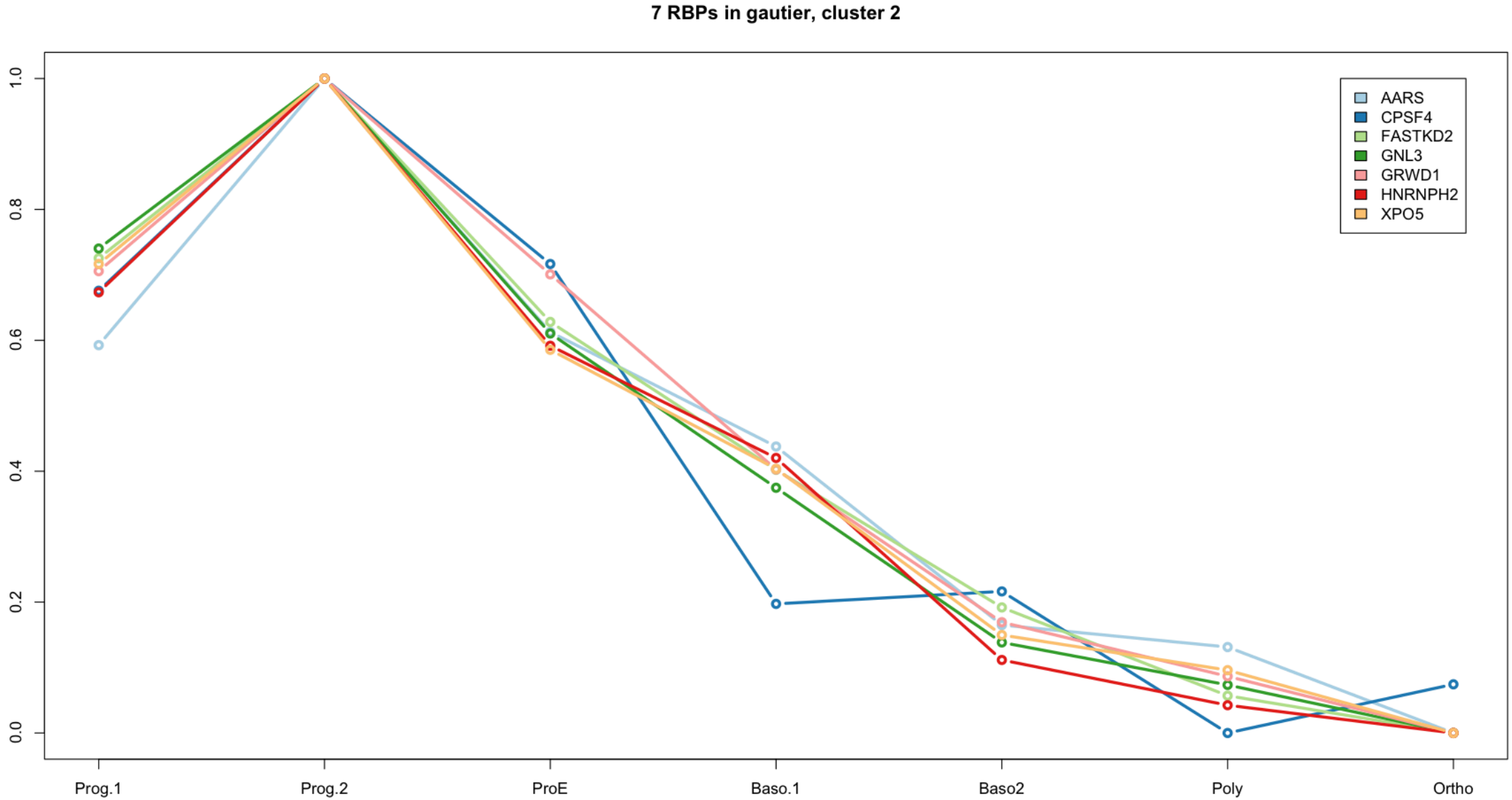

15 RPs in gautier, cluster 5
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
RPS18
RPS28
RPS9
RPS3A
RPL7A
RPS4X
RPS6
RPL13
RPL19
RPS7
RPL28
RPL30
RPL3
RPL36
RPS29

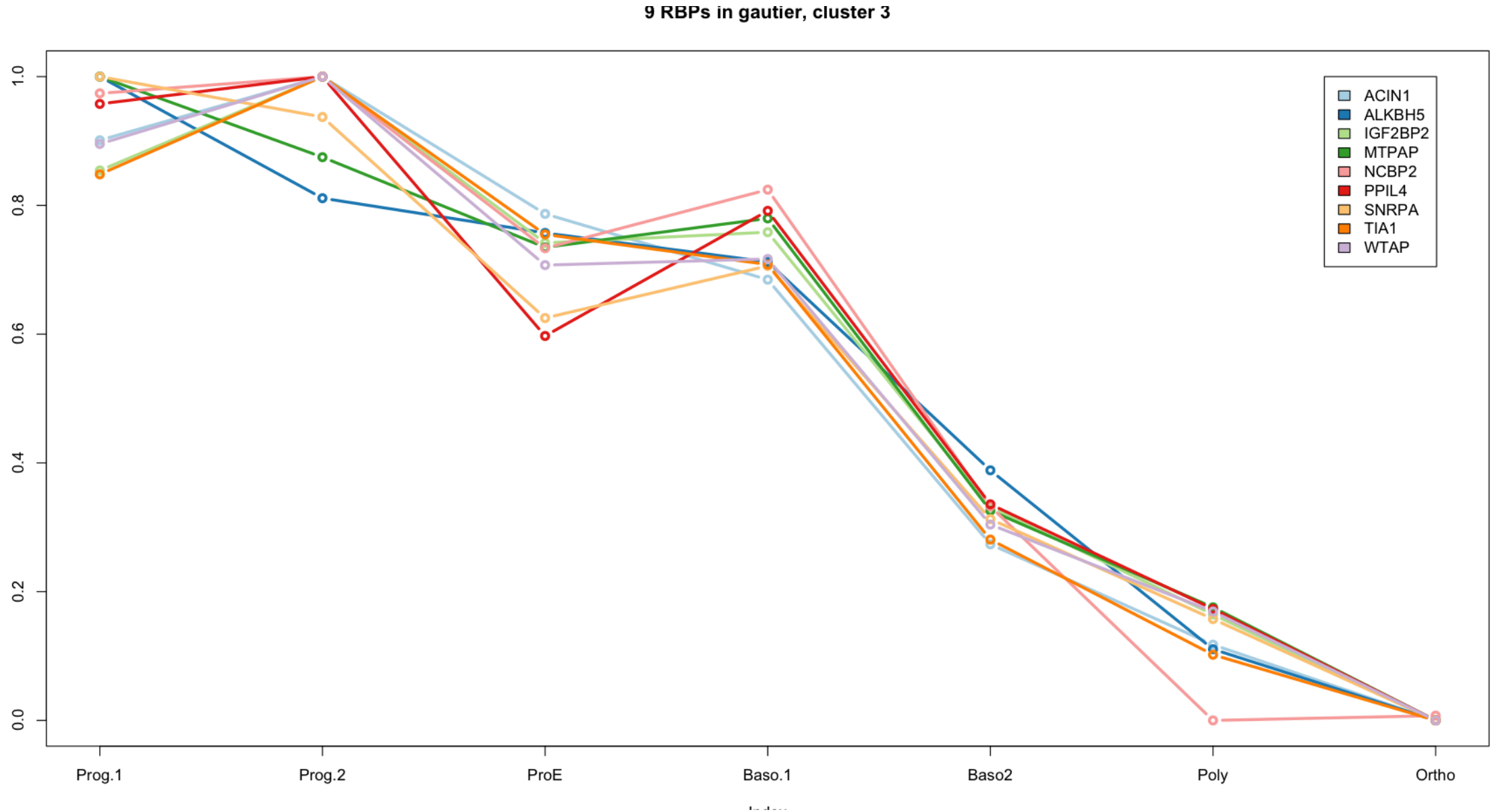

9 RPs in gautier, cluster 6
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
RPL7
RPL35A
RPS12
RPL34
RPS13
RPS15A
RPL14
RPL9
RPS23

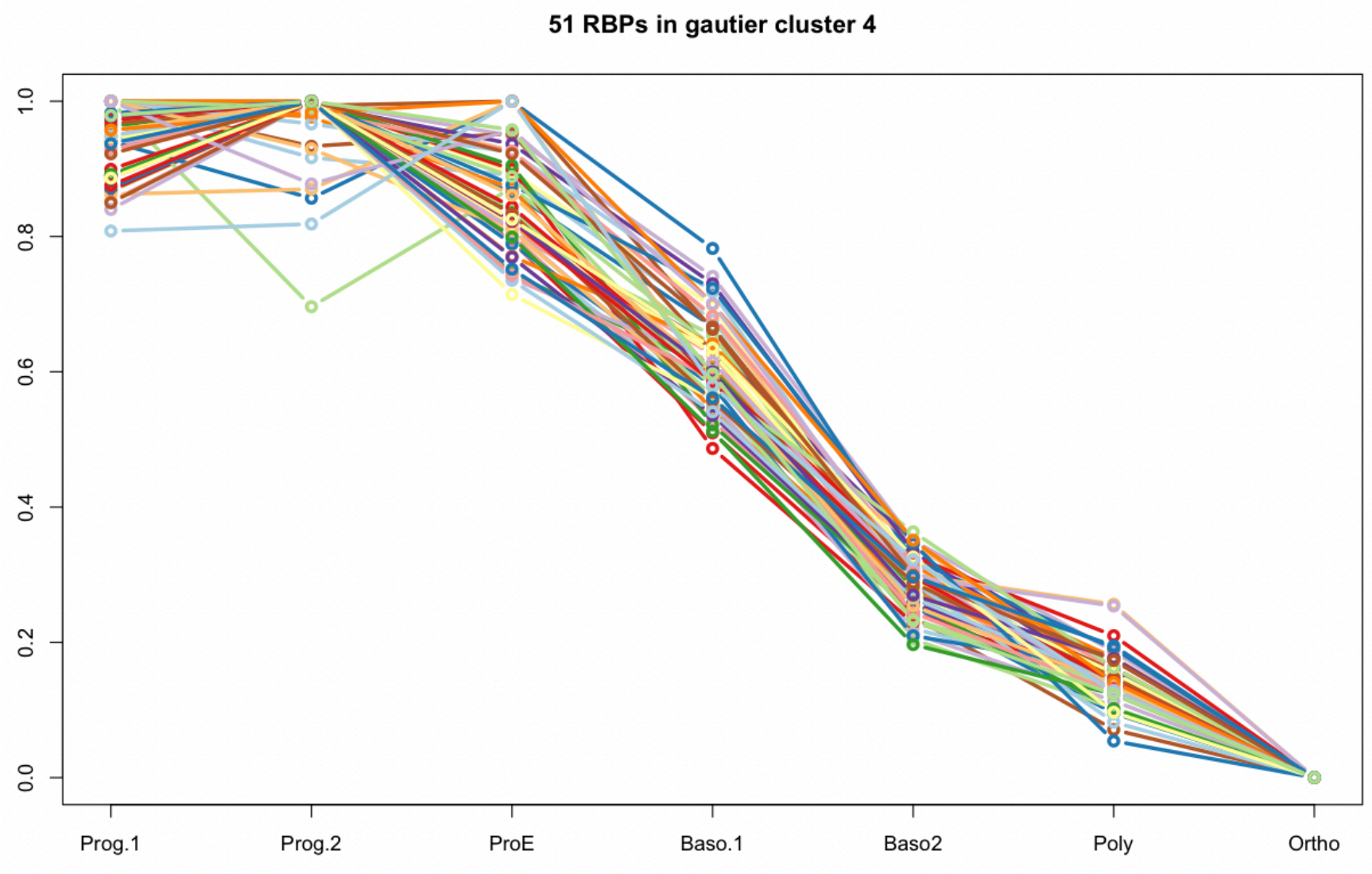

6 RPs in gautier, cluster 7
RPS11
RPL31
RPS2
RPS25
RPL32
RPL22L1
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho


5 RPs in gautier, unclustered
RPS16
RPLP1
RPL38
RPS4Y1
FAU
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

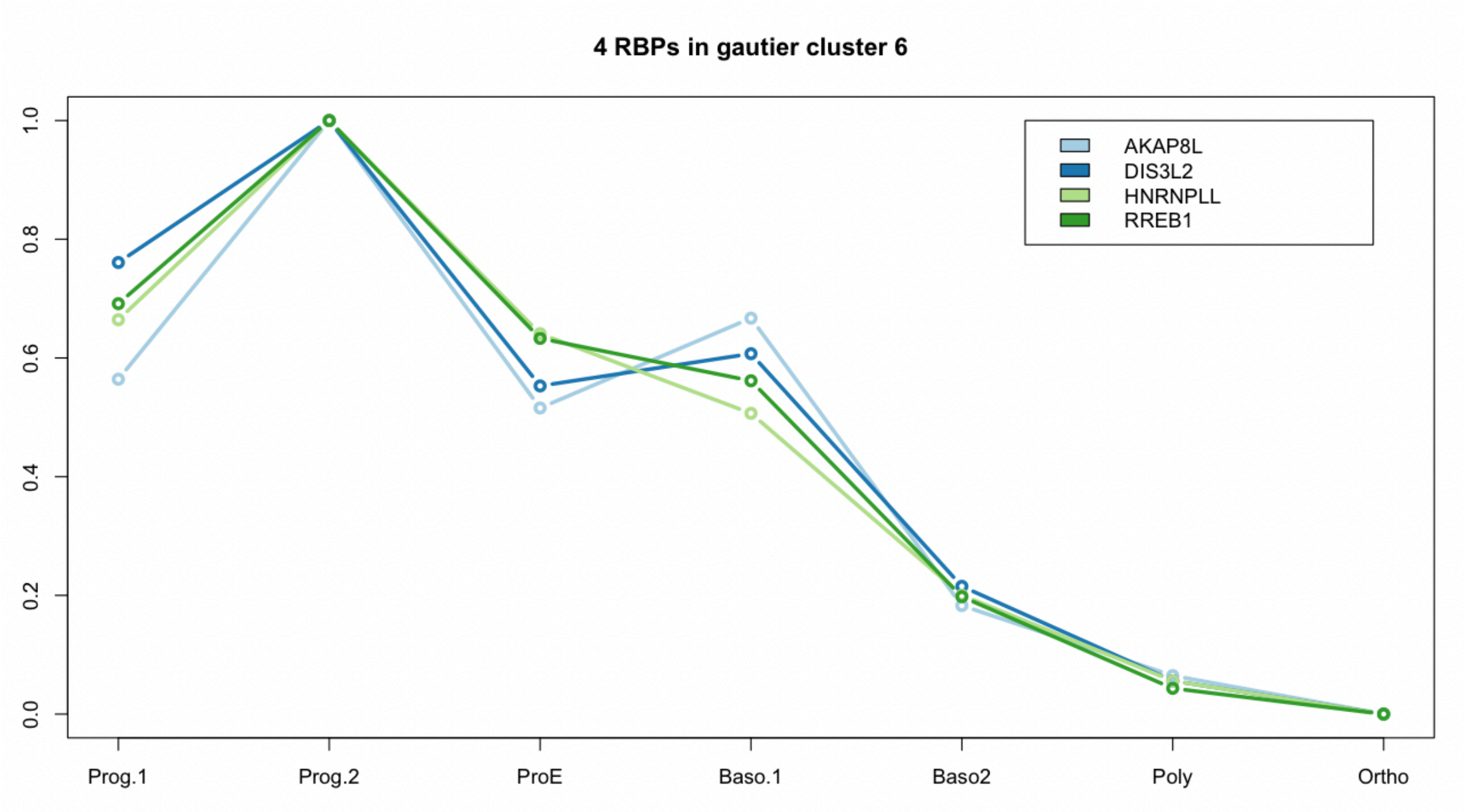

mtx.gautier.norm, 214 RNA-binding proteins
0.0
0.5
1.0
1.5
2.0
SLBP
CELF1
NOL12
ZC3H7B
RBM24
SRSF4
HNRNPCL1
AGGF1

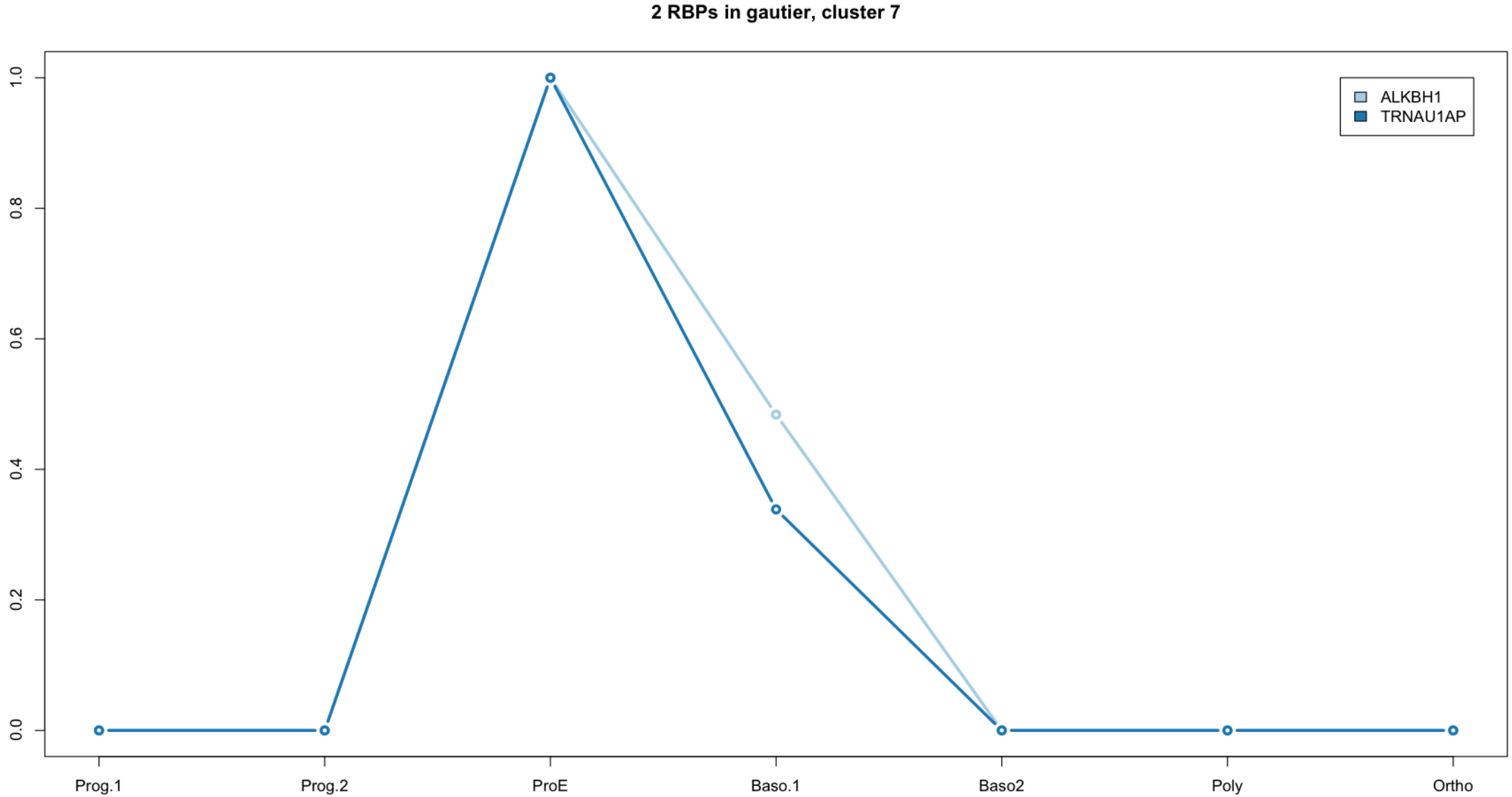

7 RBPs in gautier, cluster 2
AARS
CPSF4
FASTKD2
GNL3
GRWD1
HNRNPH2
XPO5
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

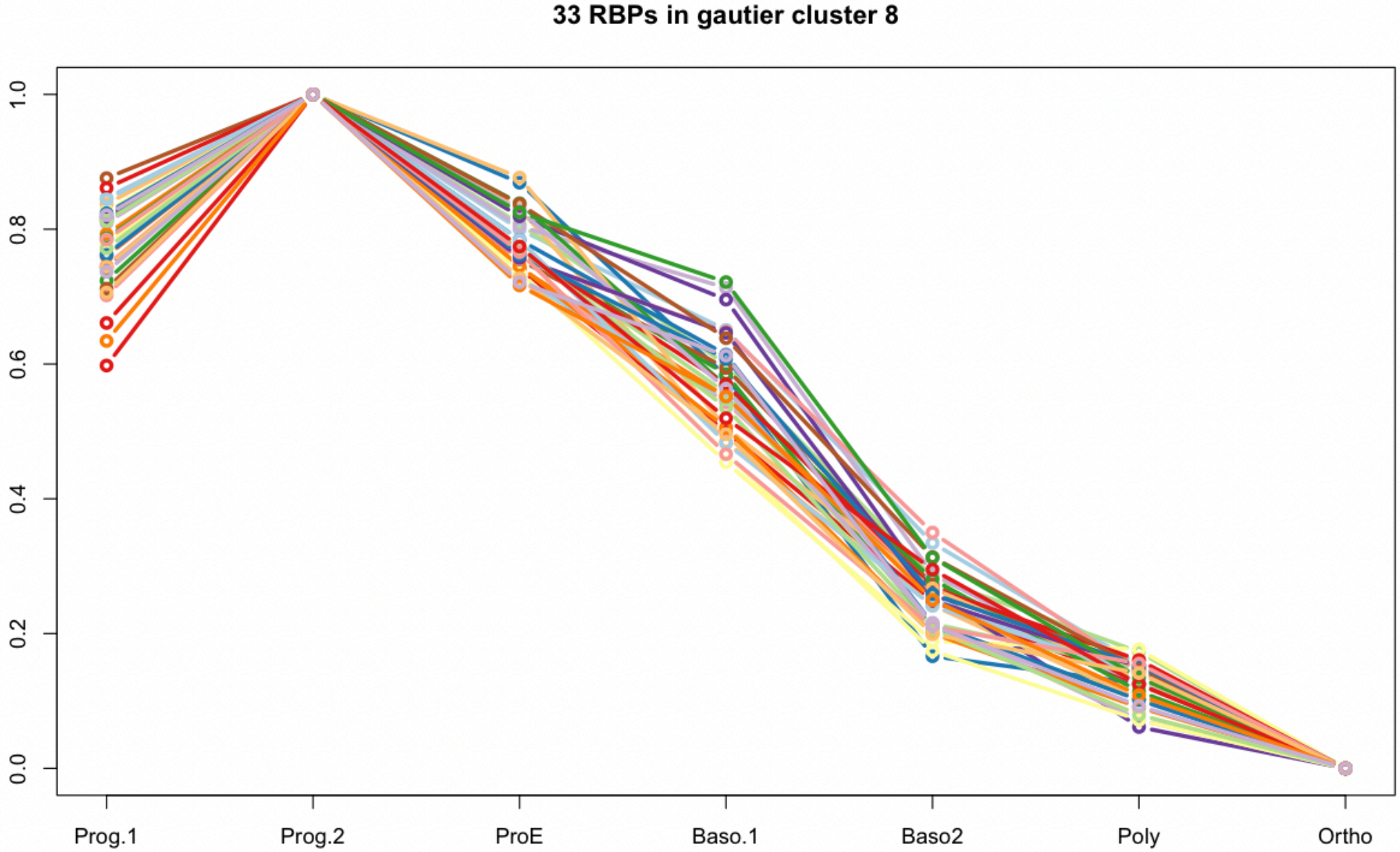

9 RBPs in gautier, cluster 3
ACIN1
ALKBH5
IGF2BP2
MTPAP
NCBP2
PPIL4
SNRPA
TIA1
WTAP
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho



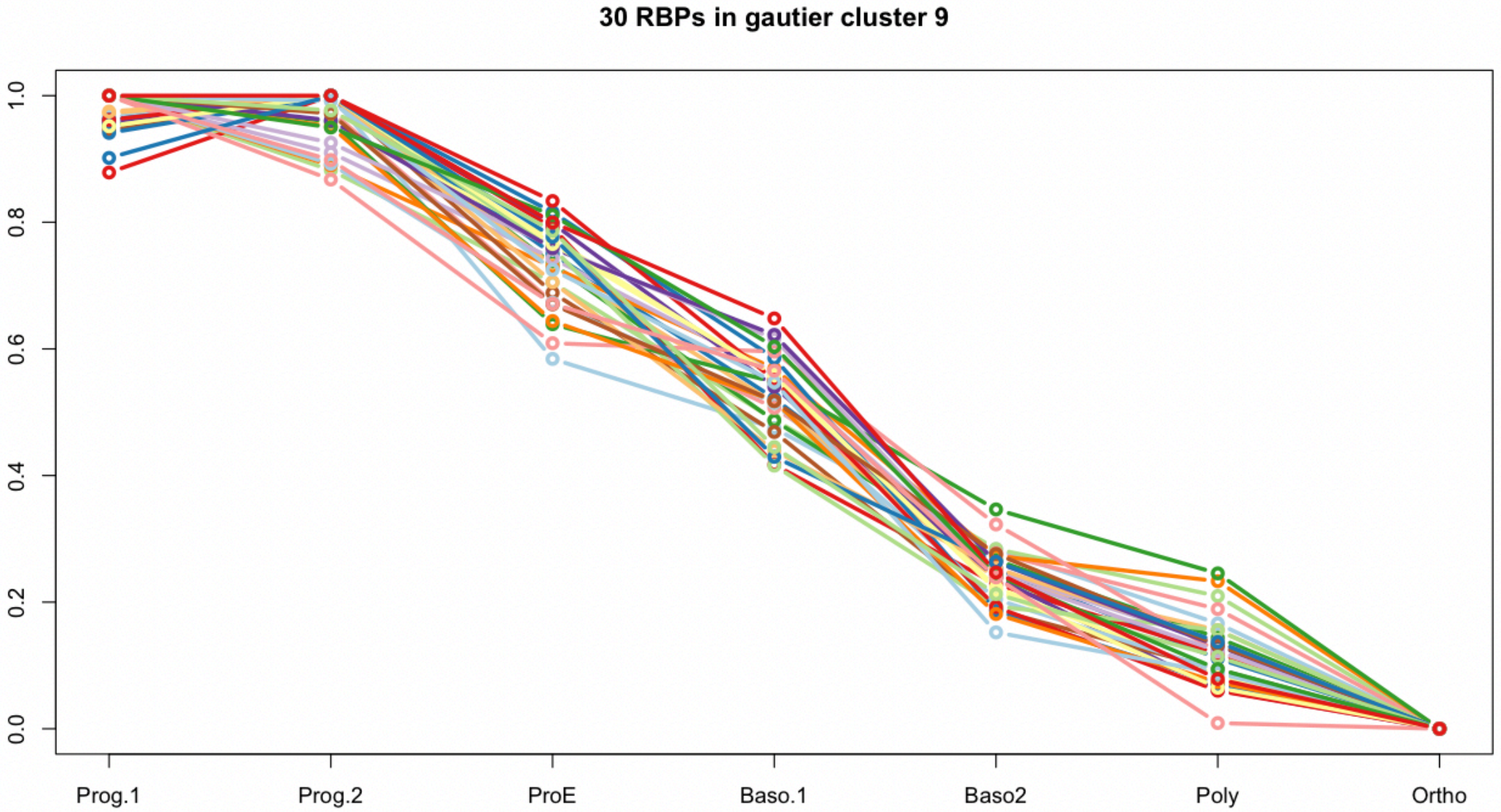
ADAR	AIFM1	CPSF1	DDX42	DDX6	DHX9	DKC1	EFTUD2	EIF3A
EIF3H	EIF4A3	EIF4B	EIF4G2	FAM120A	FMR1	FTO	FUBP1	HNRNPA2B1
HNRNPF	HNRNPH1	HNRNPK	HNRNPL	HNRNPM	HNRNPU	IGF2BP3	ILF2	KHSRP
LARP4B	MATR3	METTL3	NUMA1	PCBP1	PCBP2	PUF60	RBM10	RBM22
RBM28	RBM3	RBM5	RBM6	RTCB	SERBP1	SSB	SUPV3L1	TARDBP
UPF1	WDR33	XRCC6	YTHDF1	ZNF326	ZNF622			


4 RBPs in gautier cluster 6
AKAP8L
DIS3L2
HNRNPLL
RREB1
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

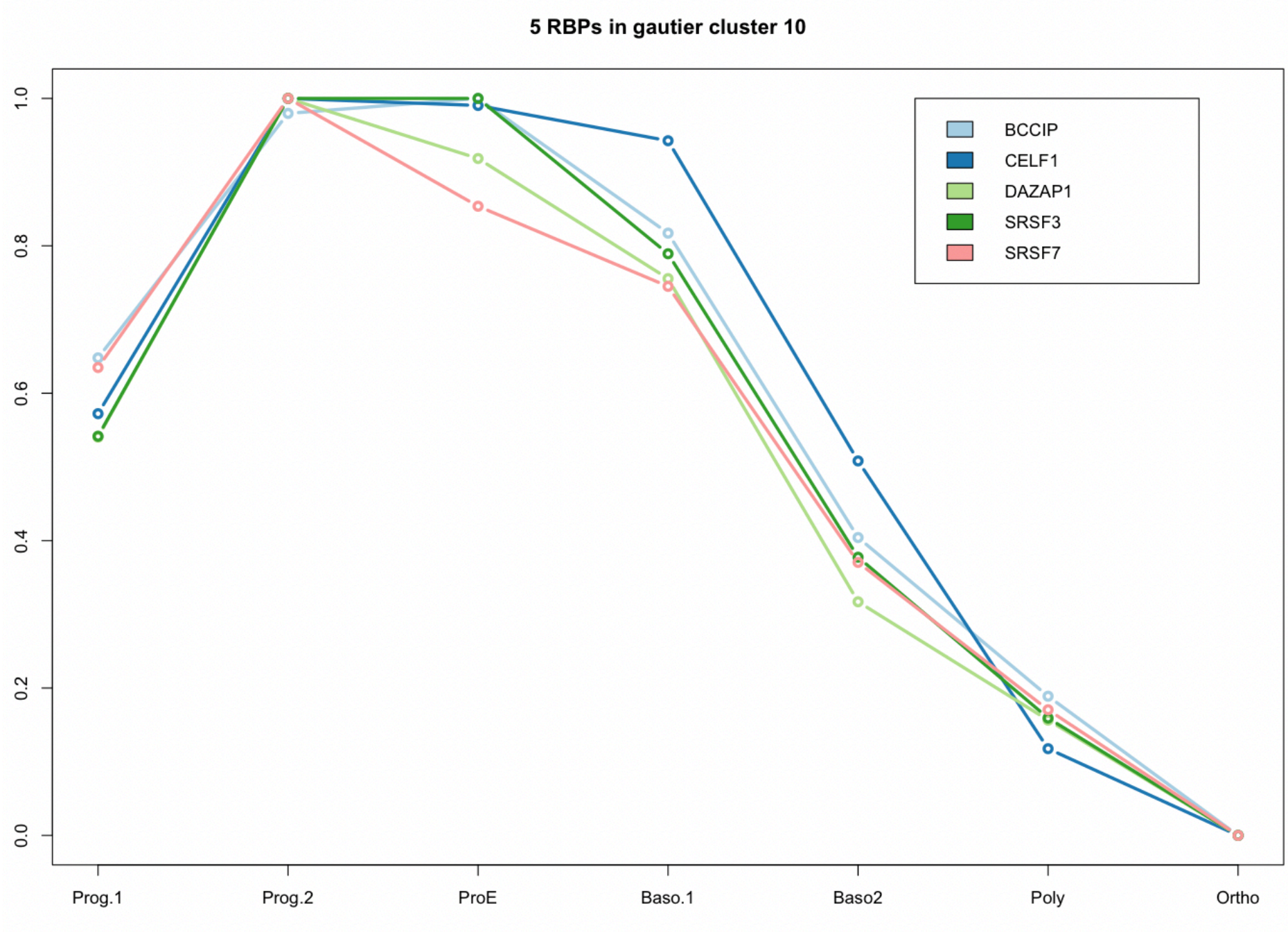

2 RBPs in gautier, cluster 7
ALKBH1
TRNAU1AP
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

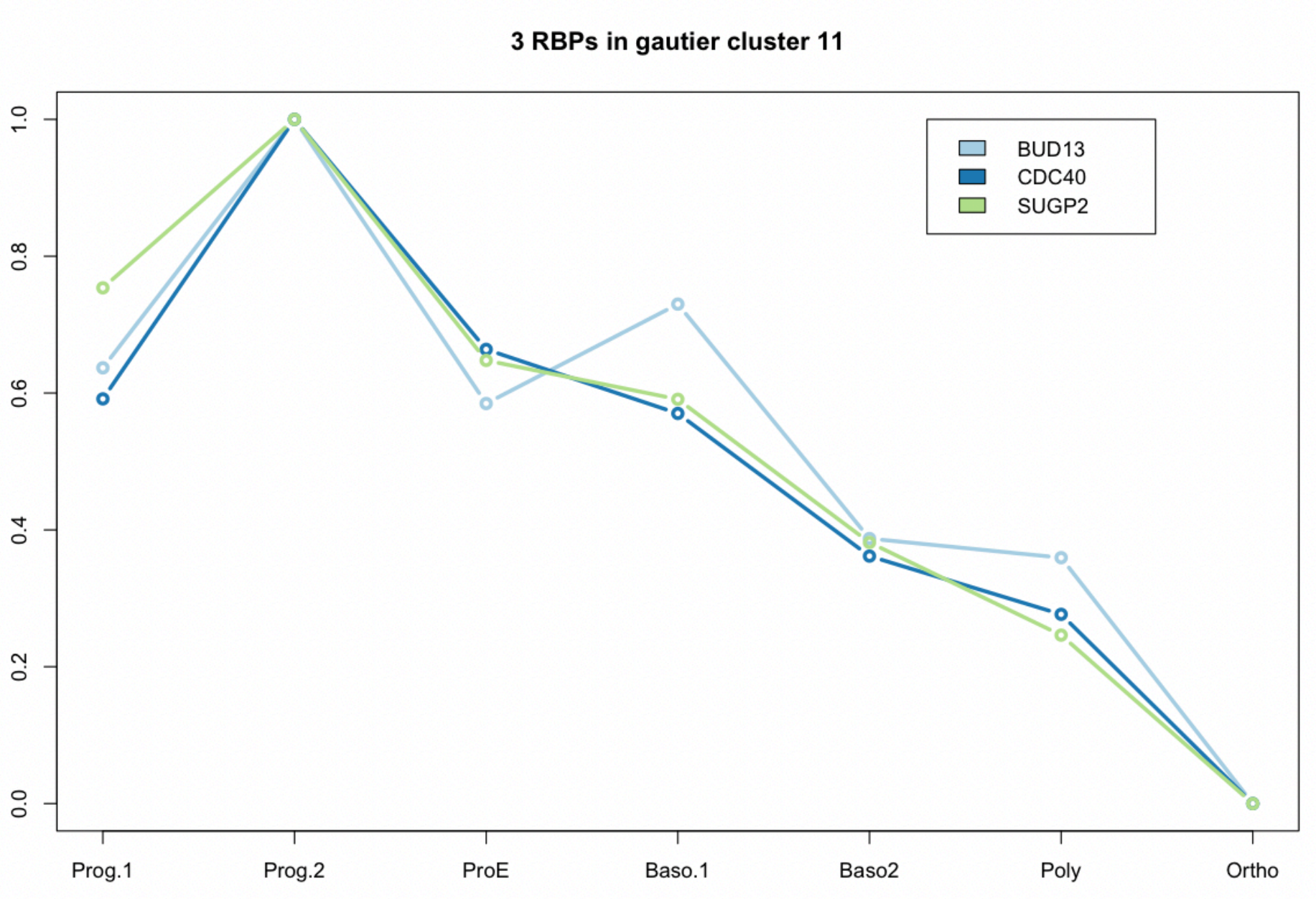


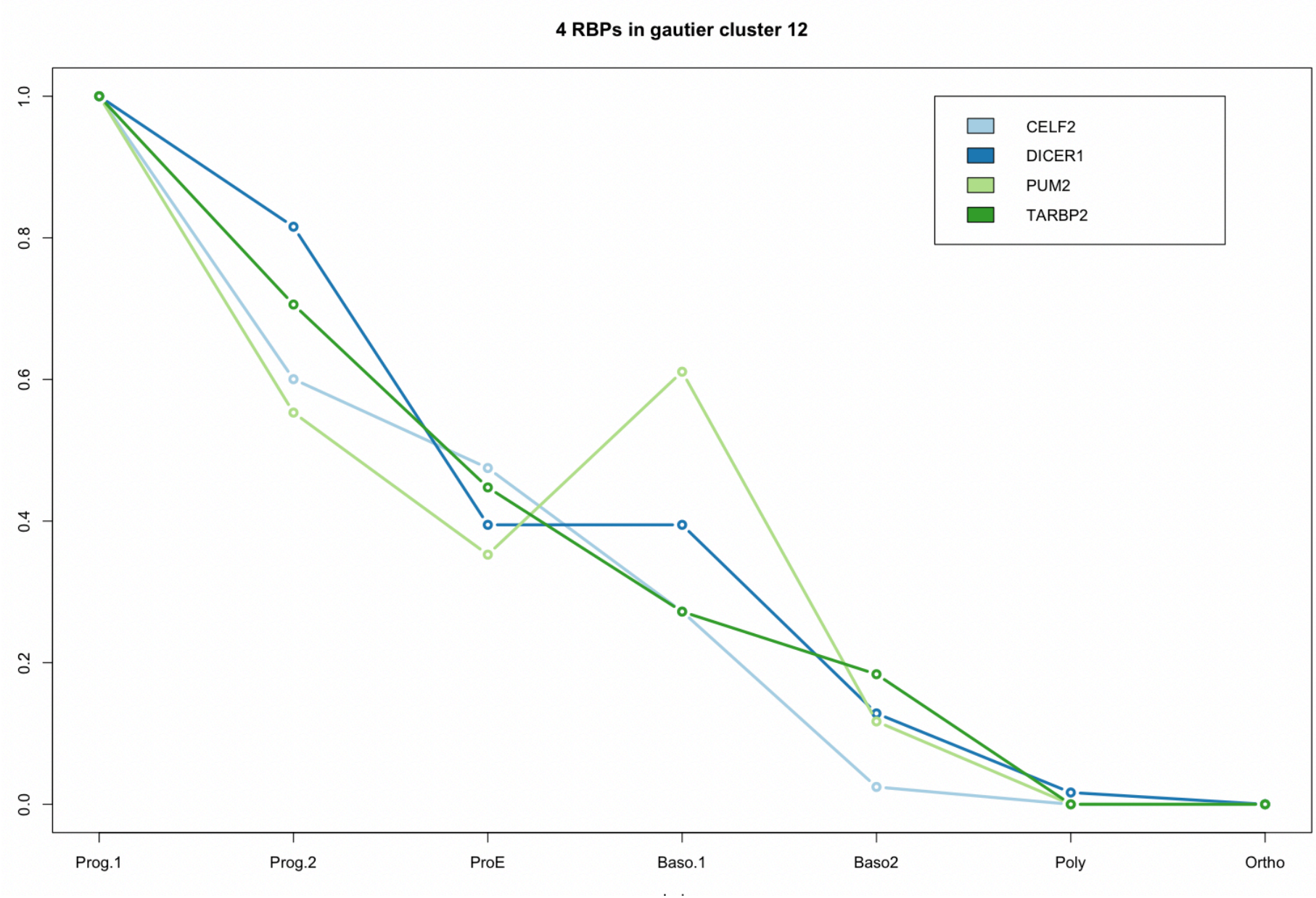
ALYREF	ATXN2	CPSF6	DDX55	EIF3B	EIF4A1	FUS	HNRNPA0	HNRNPC
HNRNPD	HNRNPDL	NOP56	NOP58	NPM1	NSUN2	PABPC4	PPIG	PRPF8
PTBP1	PUM1	PUS1	RBM15	RBM27	RPS11	RPS5	SF3B1	SND1
SNRNP70	TAF15	TIAL1	U2AF2	UCHL5	ZNF638			

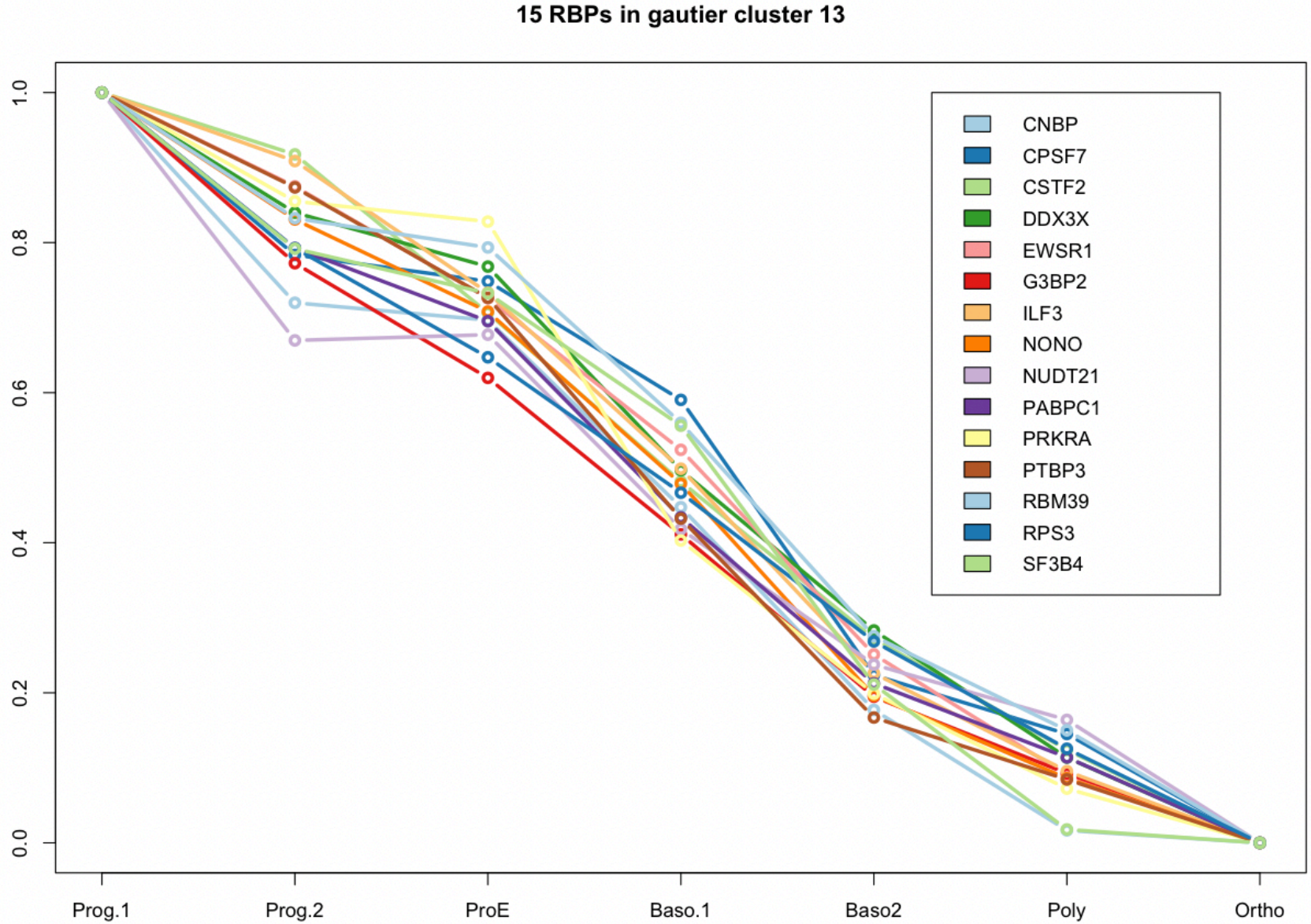


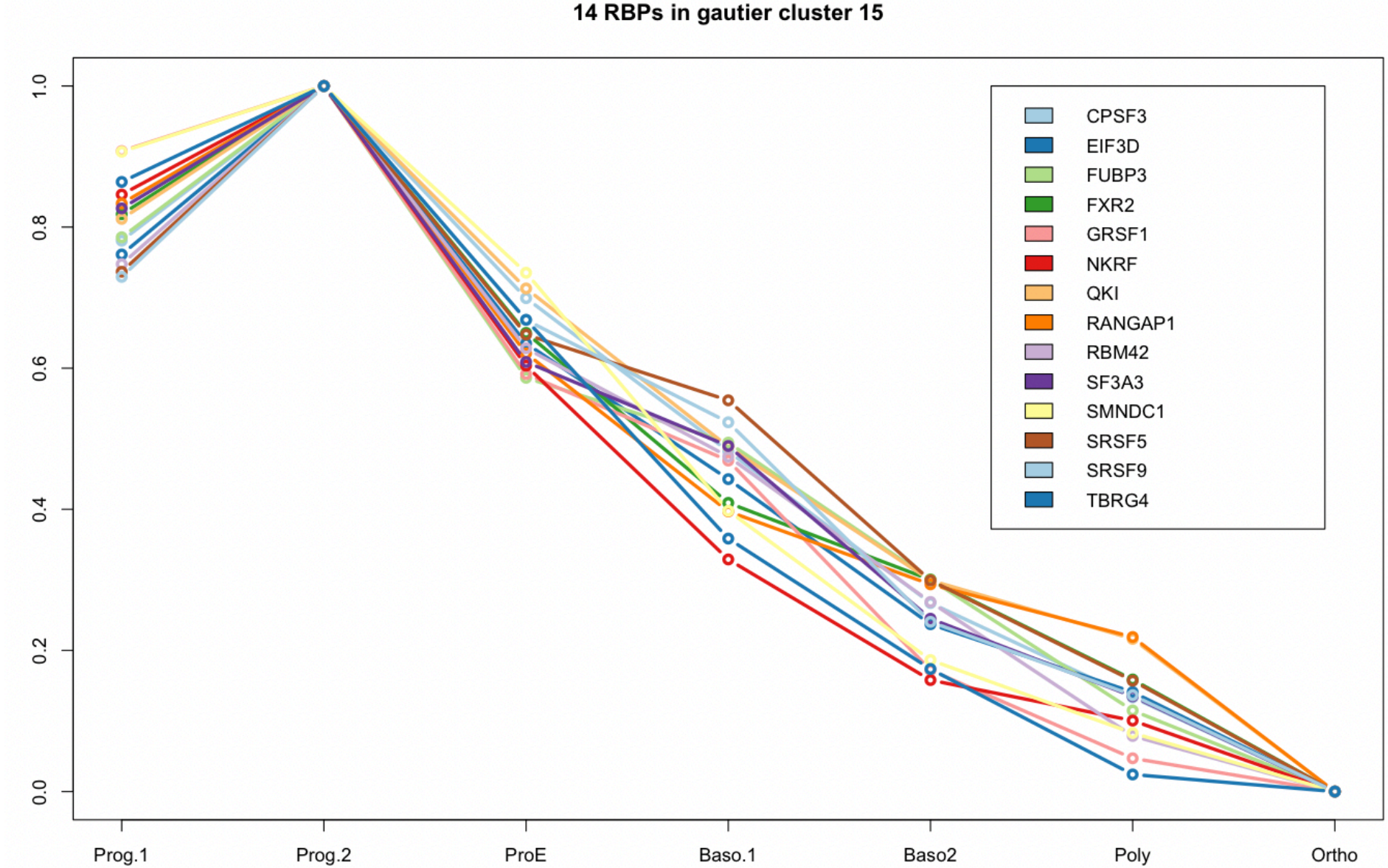
ANKHD1	CAPRIN1	CPSF2	CSDE1	DDX24	DHX30	EIF3G	EIF4G1
ELAVL1	FBL	FKBP4	GEMIN5	HNRNPUL1	IGF2BP1	KHDRBS1	LARP4
METTL14	RBM25	RBM8A	SAFB2	SART3	SF1	SRSF1	SRSF2
SUB1	TRA2A	YBX1	YTHDF2	YTHDF3	YWHAG		

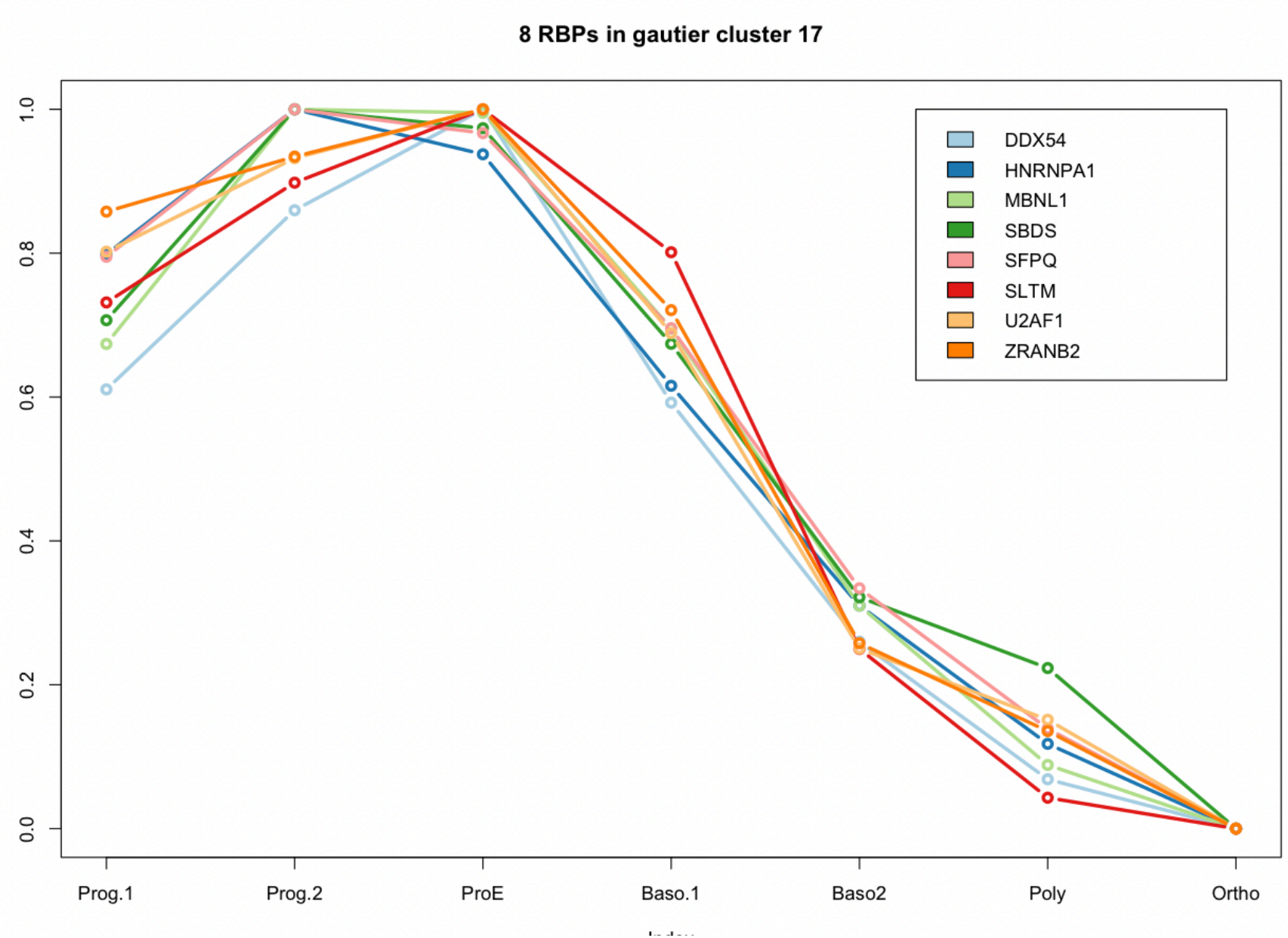

5 RBPs in gautier cluster 10
1.0
0.8
0.6
0.4
0.2
0.0
BCCIP
CELF1
DAZAP1
SRSF3
SRSF7
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

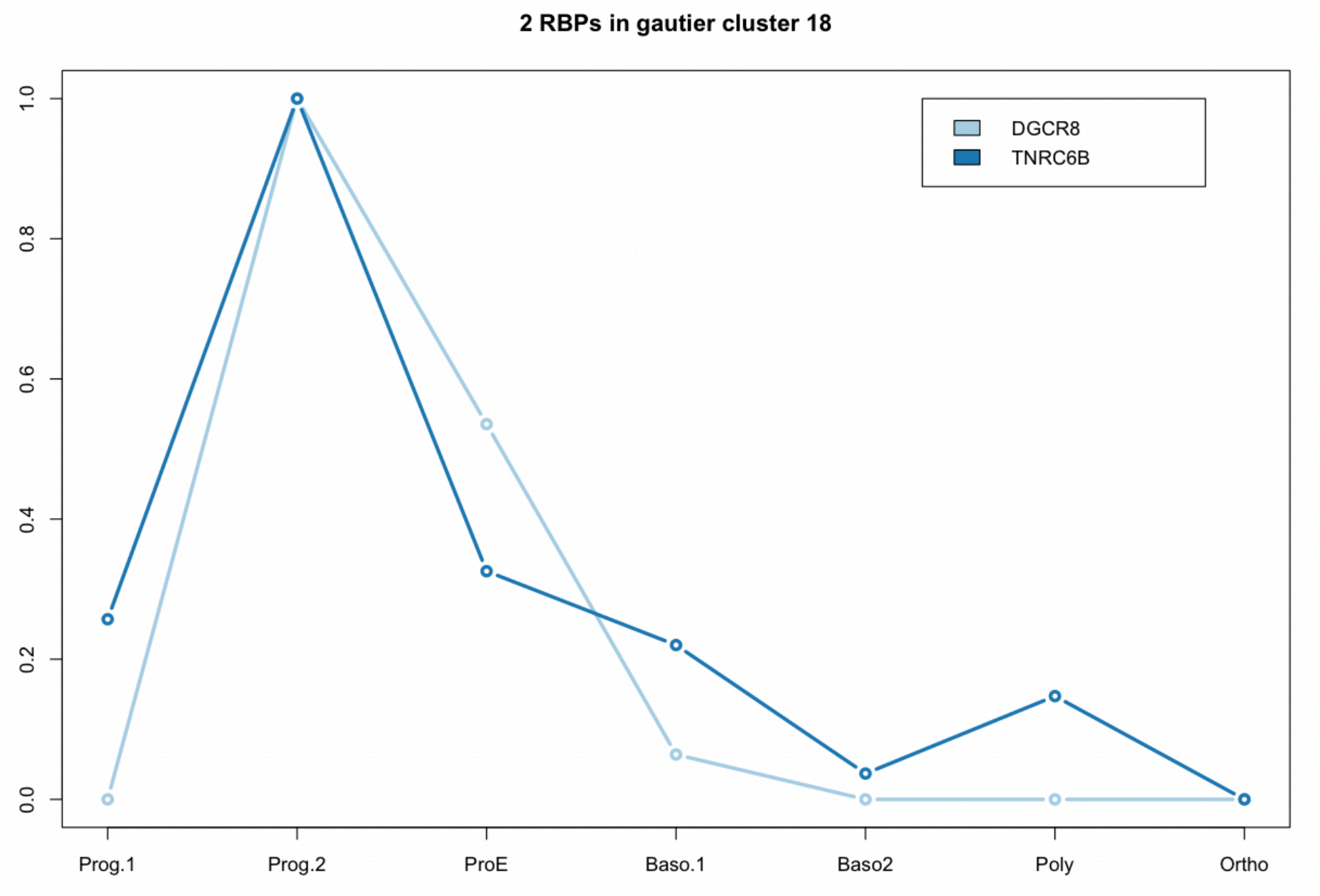

3 RBPs in gautier cluster 11
BUD13
CDC40
SUGP2
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

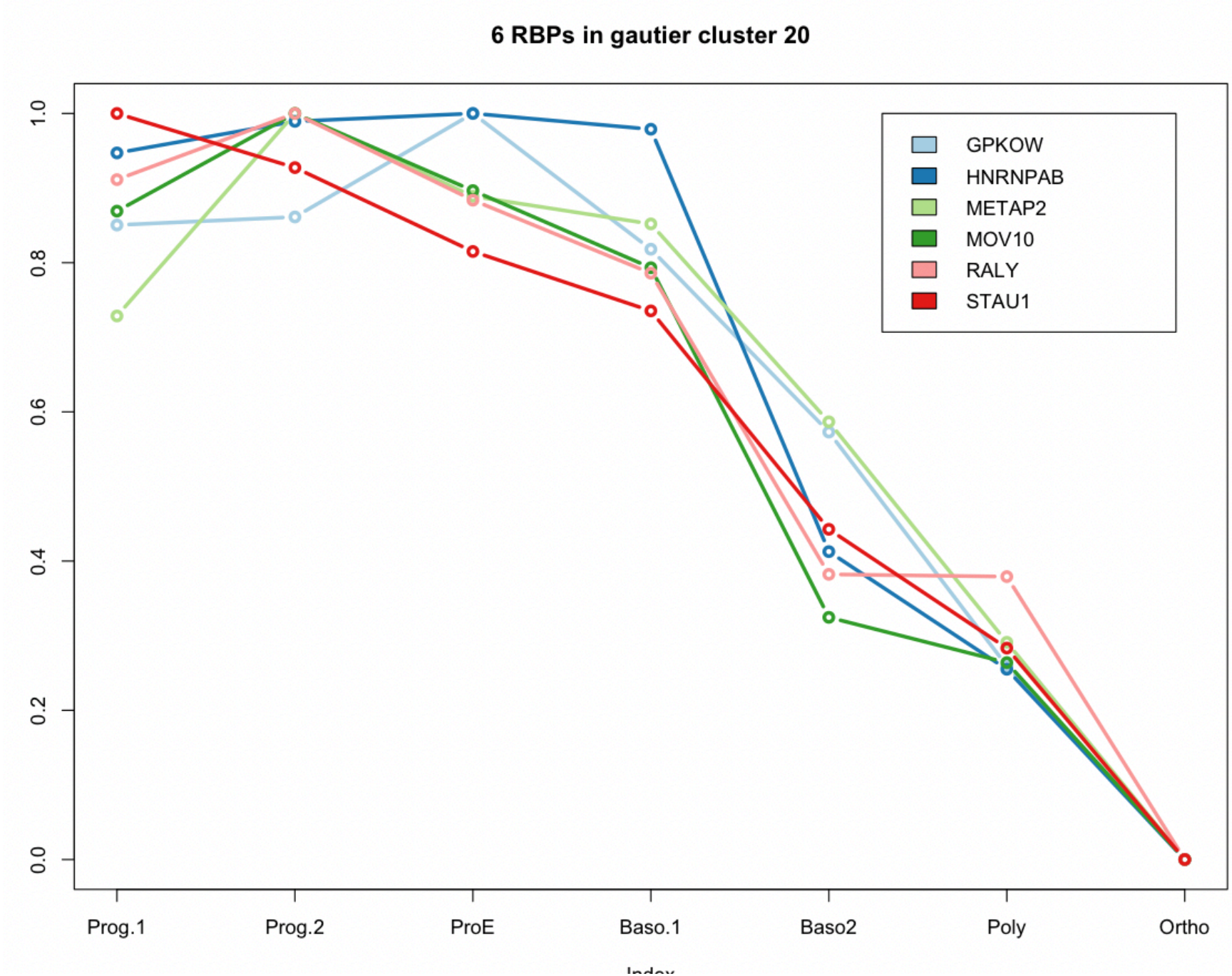

4 RBPs in gautier cluster 12
CELF2
DICER1
PUM2
TARBP2
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

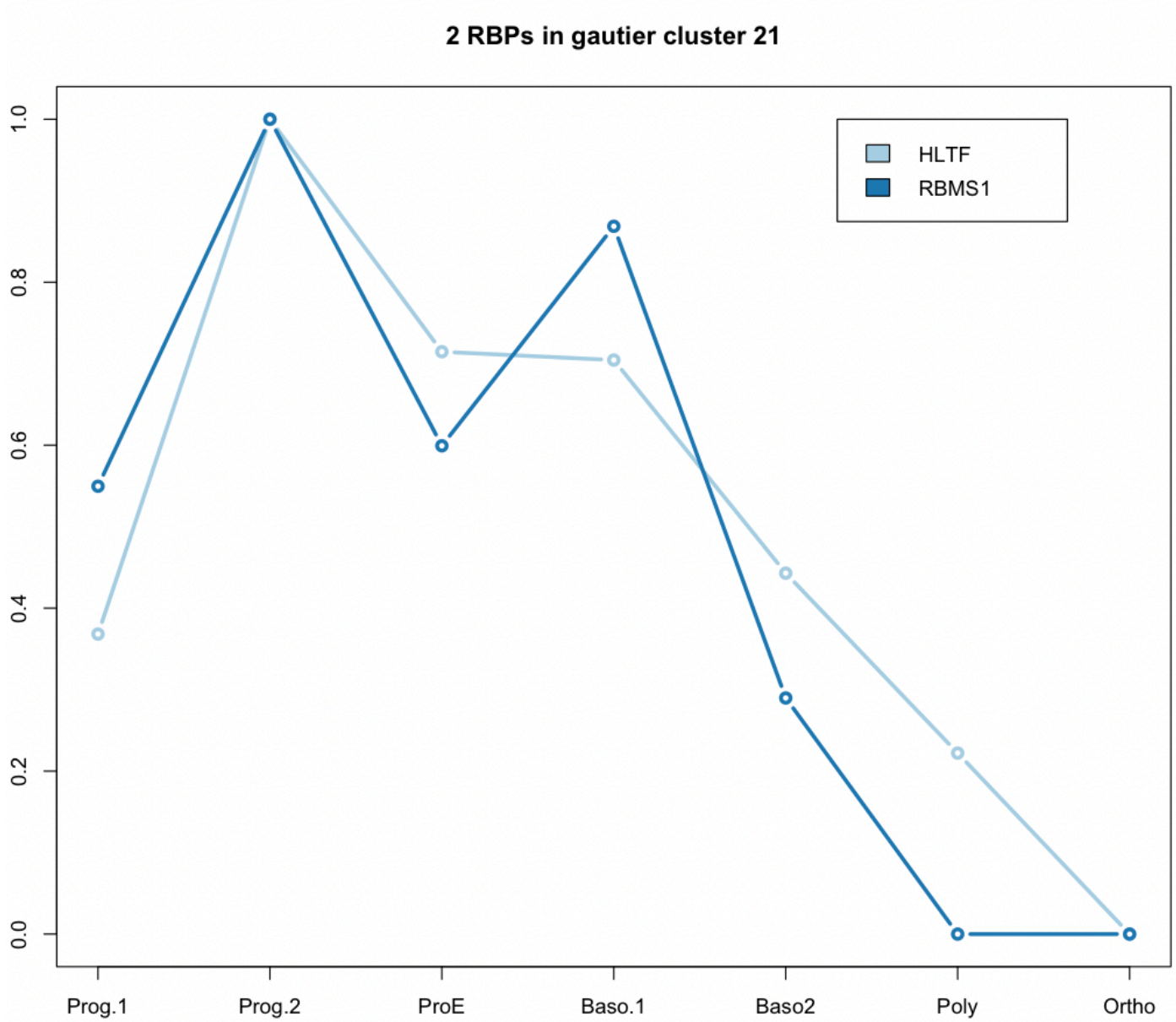

15 RBPs in gautier cluster 13
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
CNBP
CPSF7
CSTF2
DDX3X
EWSR1
G3BP2
ILF3
NONO
NUDT21
PABPC1
PRKRA
PTBP3
RBM39
RPS3
SF3B4

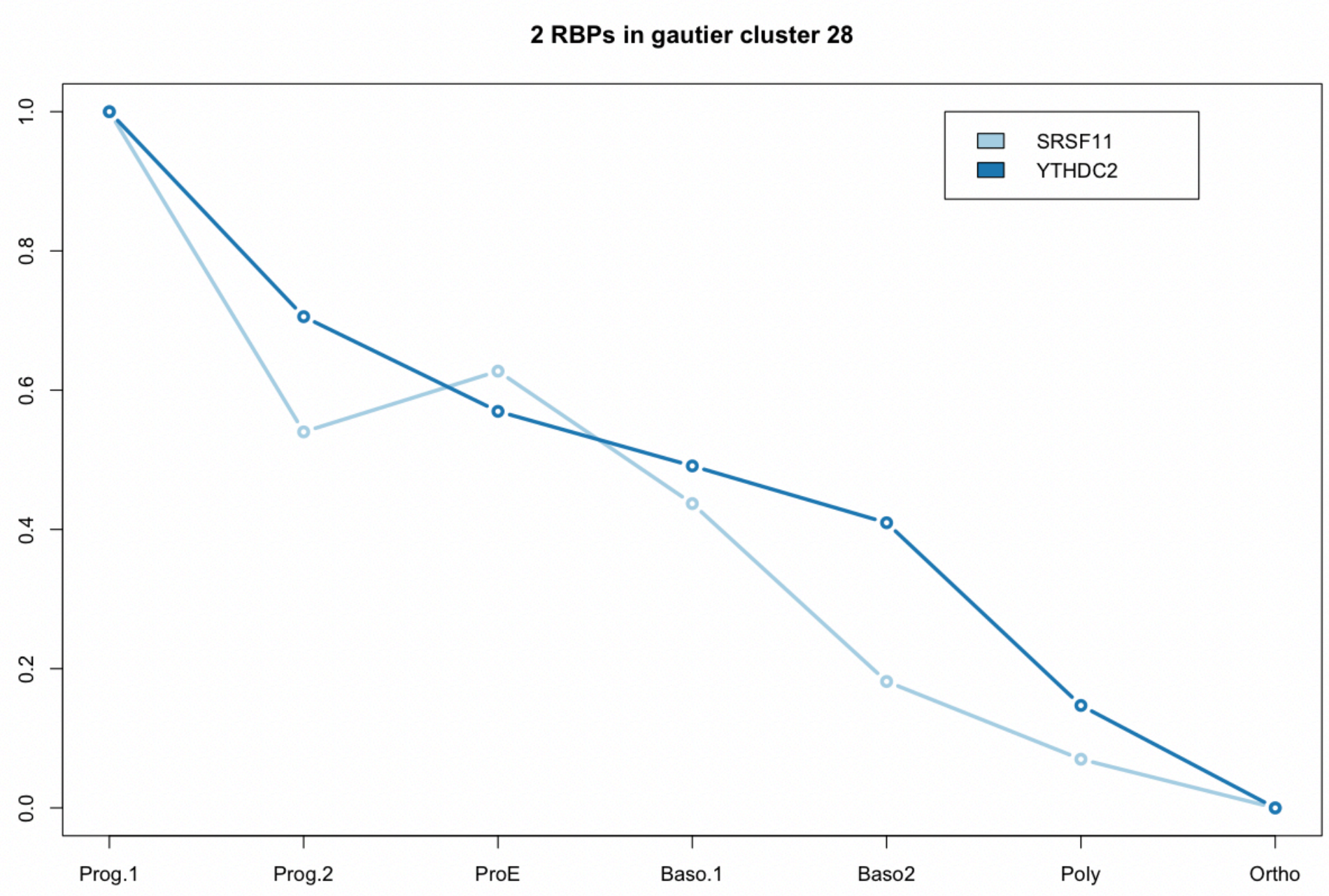

14 RBPs in gautier cluster 15
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
CPSF3
EIF3D
FUBP3
FXR2
GRSF1
NKRF
QKI
RANGAP1
RBM42
SF3A3
SMNDC1
SRSF5
SRSF9
TBRG4

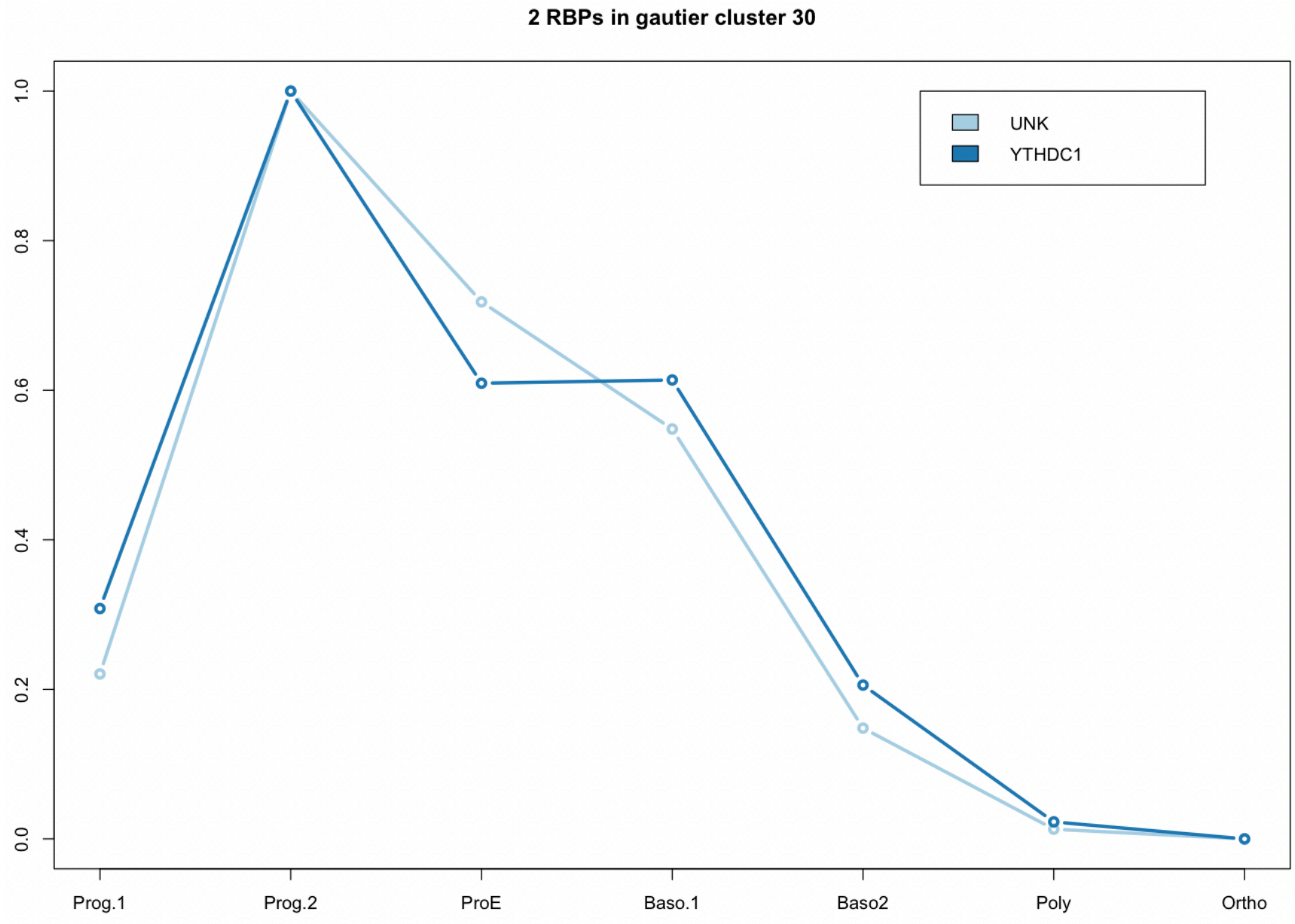

8 RBPs in gautier cluster 17
1.0
0.8
0.6
0.4
0.2
0.0
DDX54
HNRNPA1
MBNL1
SBDS
SFPQ
SLTM
U2AF1
ZRANB2
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

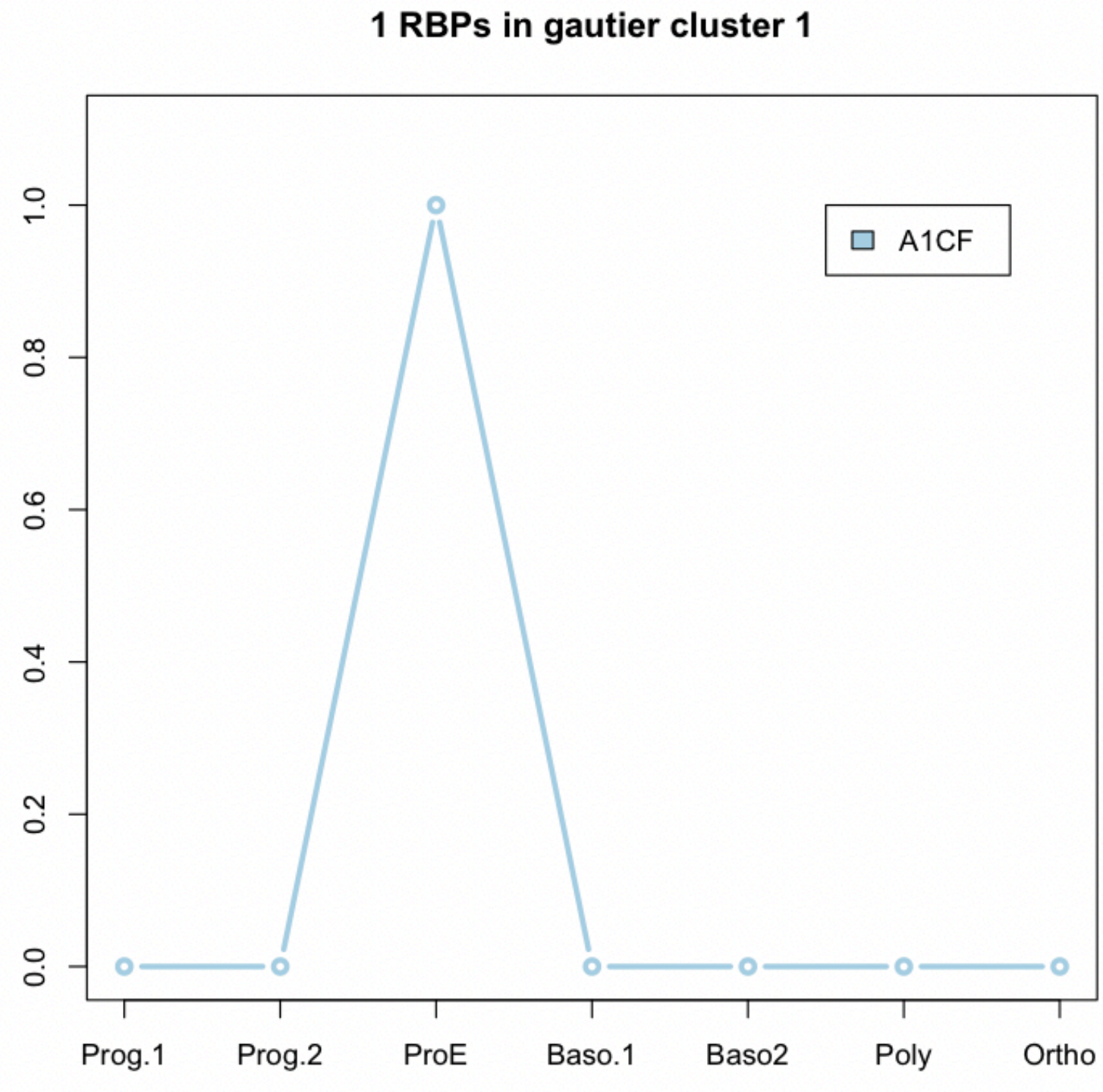

2 RBPs in gautier cluster 18
DGCR8
TNRC6B
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

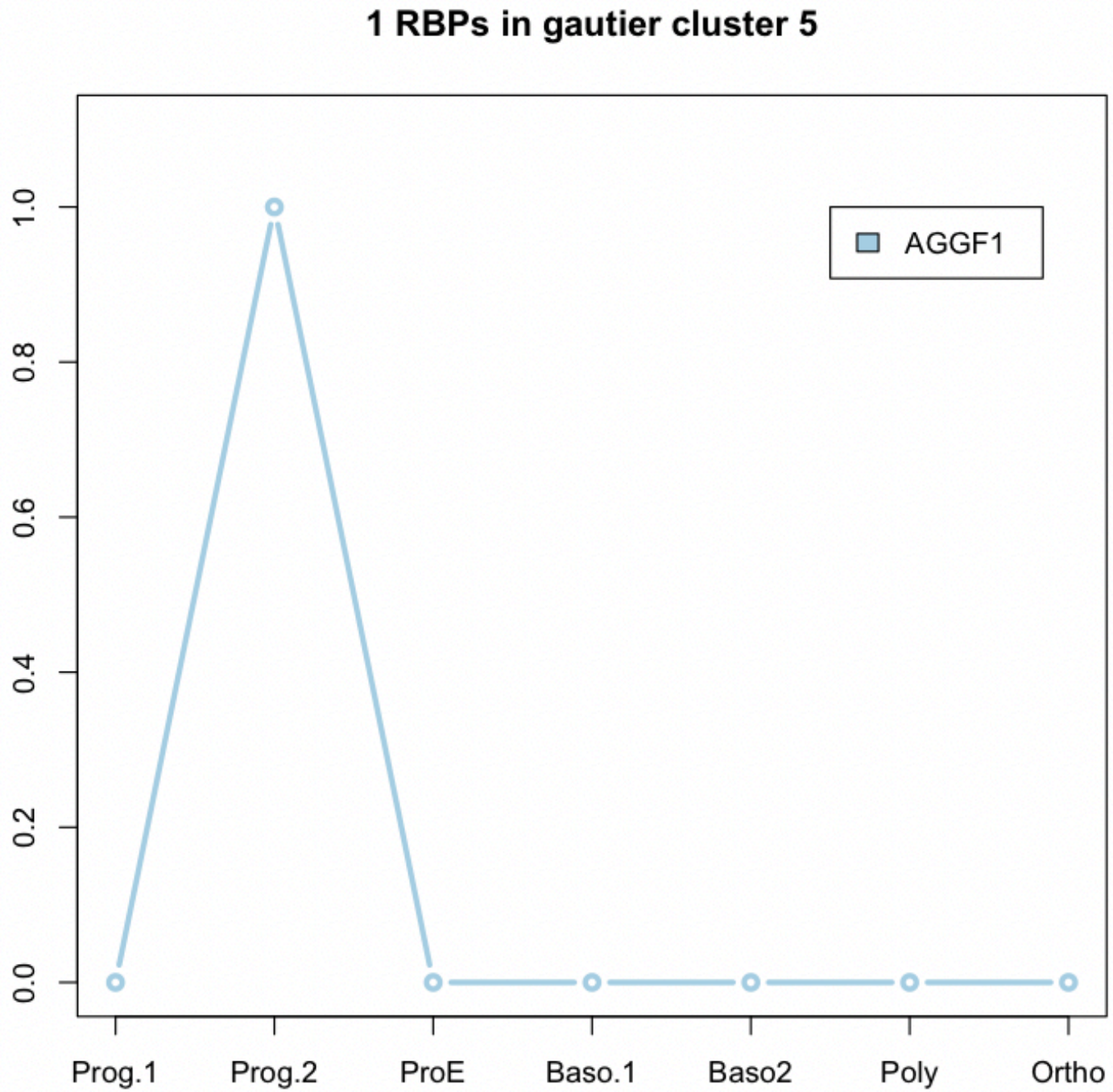

6 RBPs in gautier cluster 20
1.0
0.8
0.6
0.4
0.2
0.0
GPKOW
HNRNPAB
METAP2
MOV10
RALY
STAU1
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho
Index

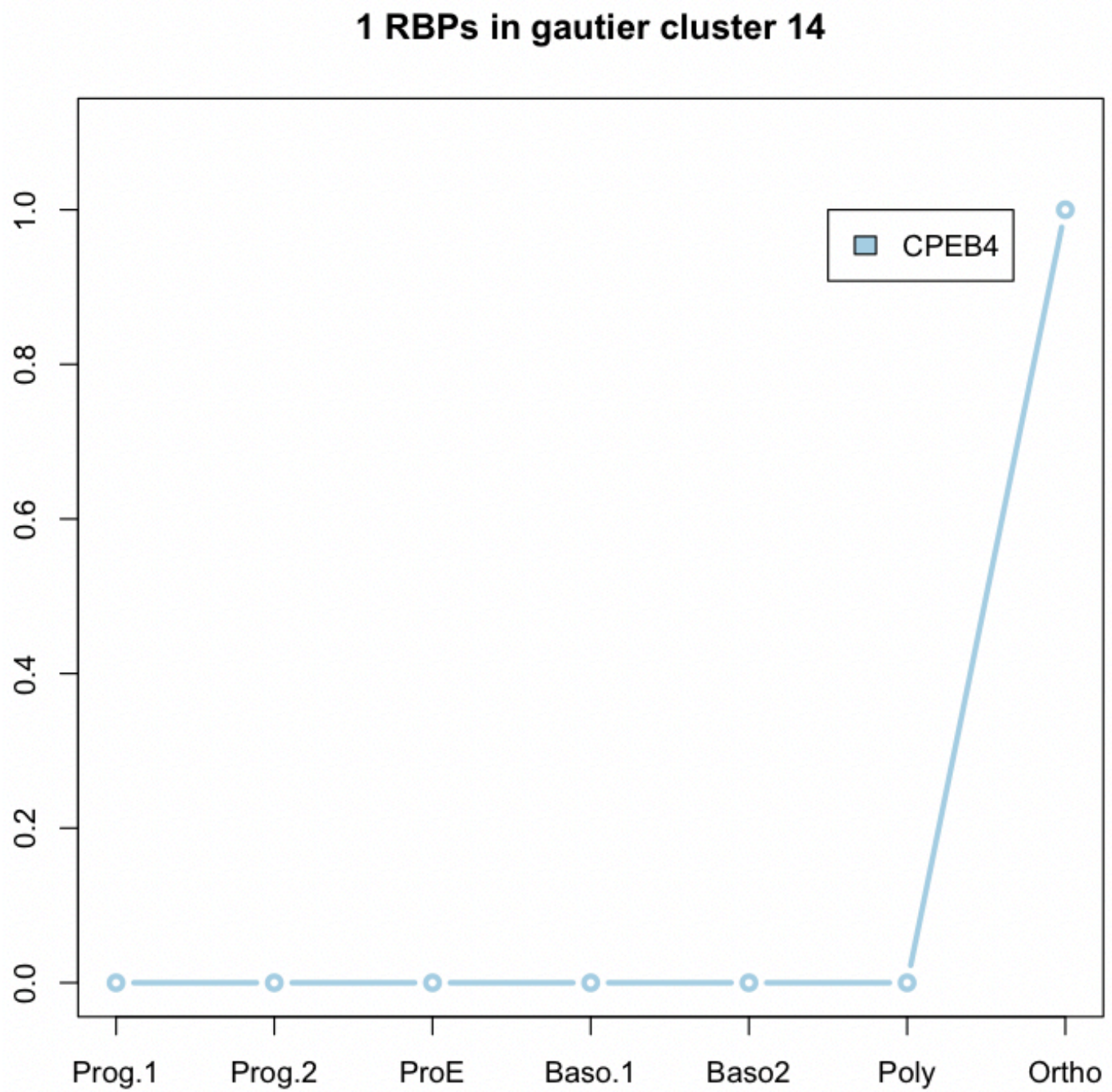

2 RBPs in gautier cluster 21
HLTF
RBMS1
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

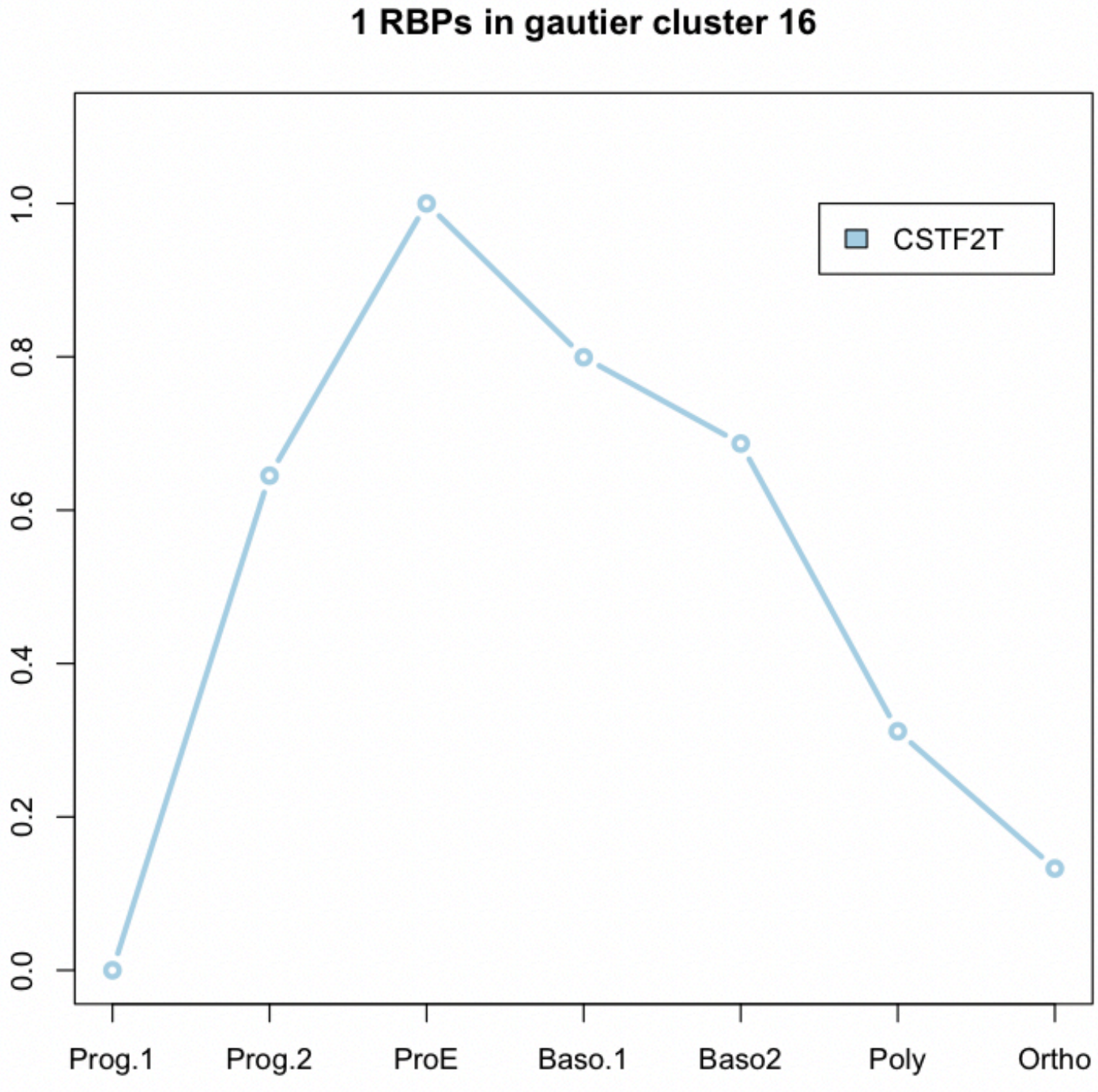

2 RBPs in gautier cluster 28
SRSF11
YTHDC2
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

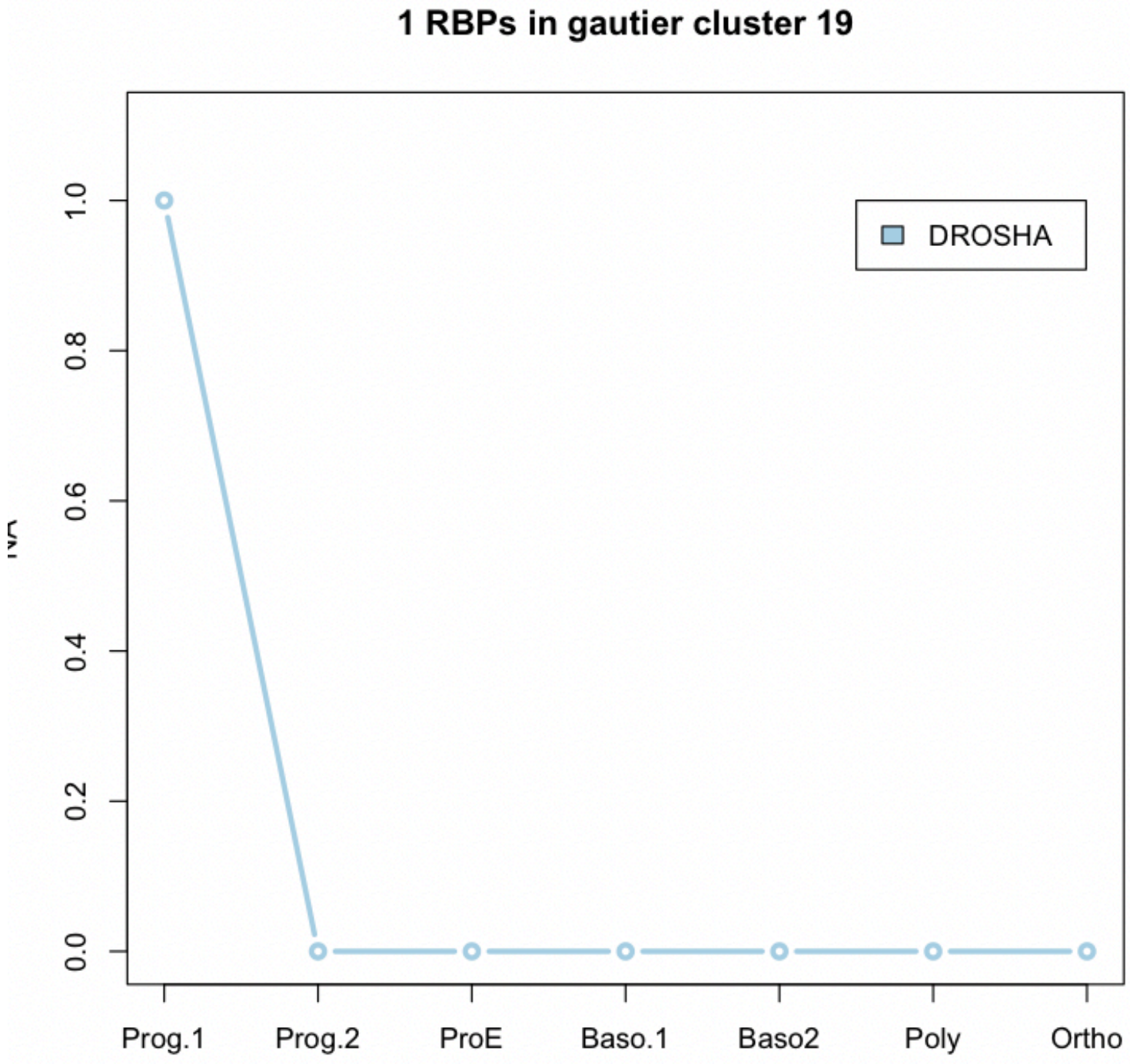

2 RBPs in gautier cluster 30
UNK
YTHDC1
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

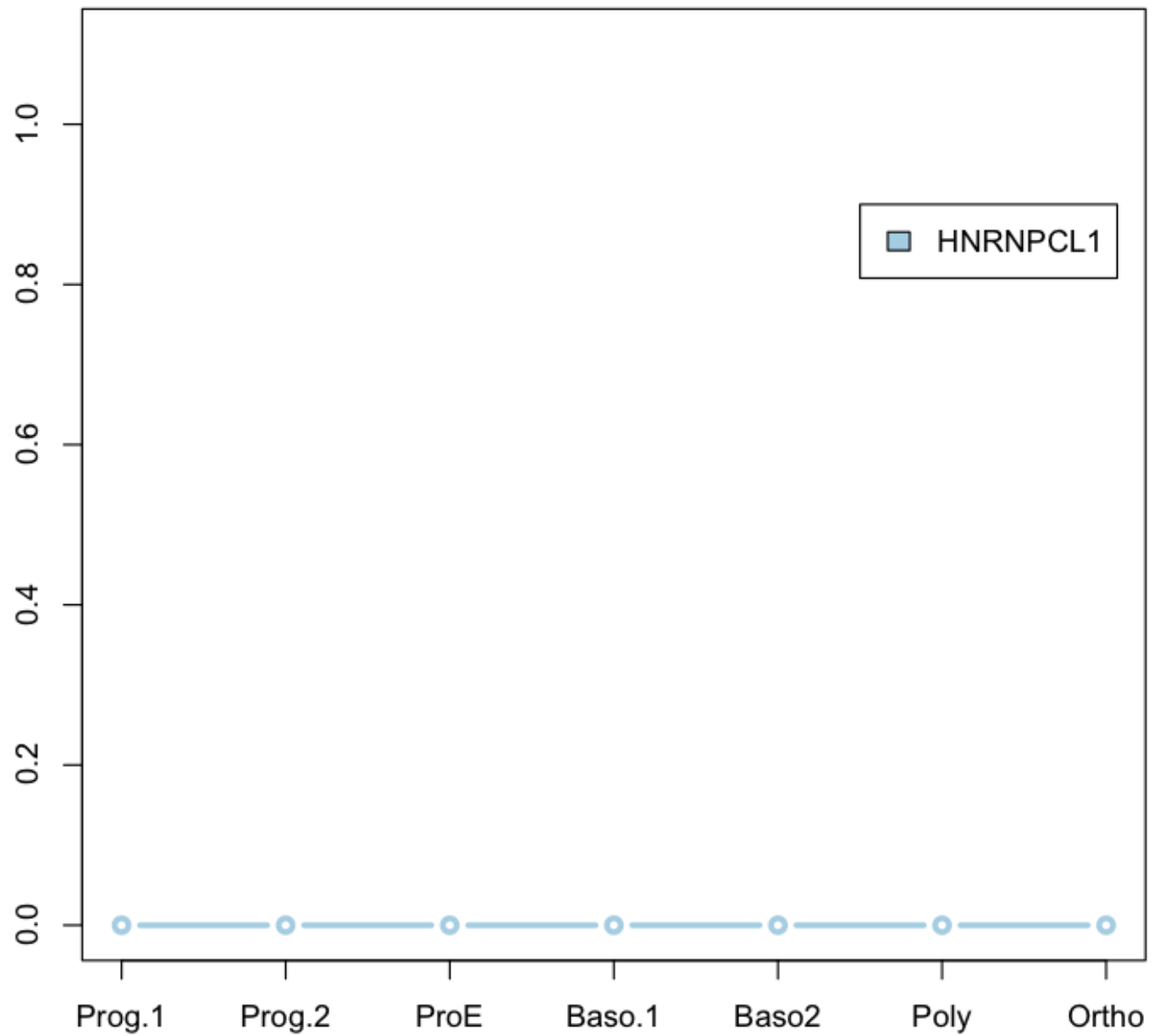

1 RBPs in gautier cluster 1
A1CF
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

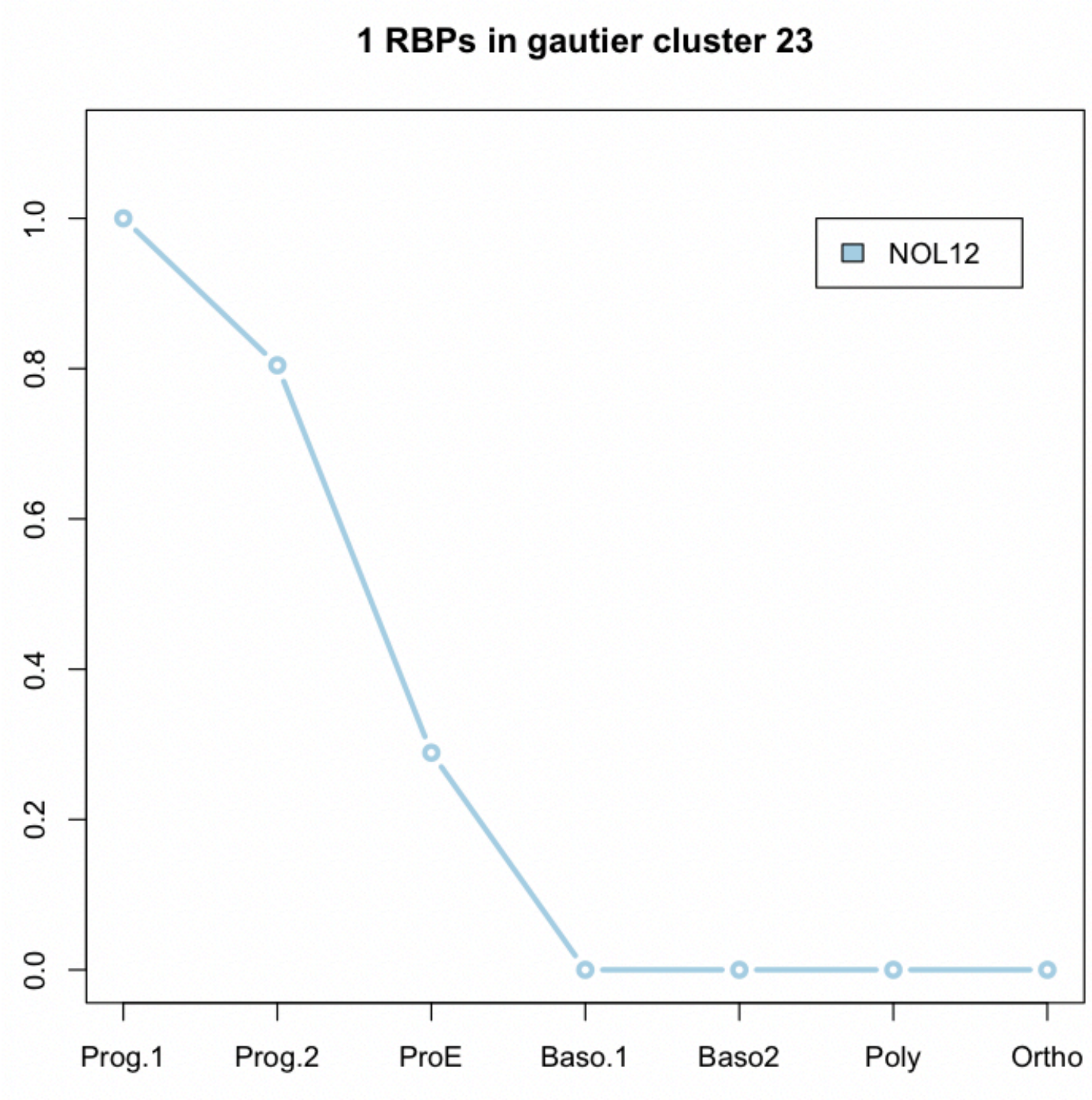

1 RBPs in gautier cluster 5
AGGF1
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

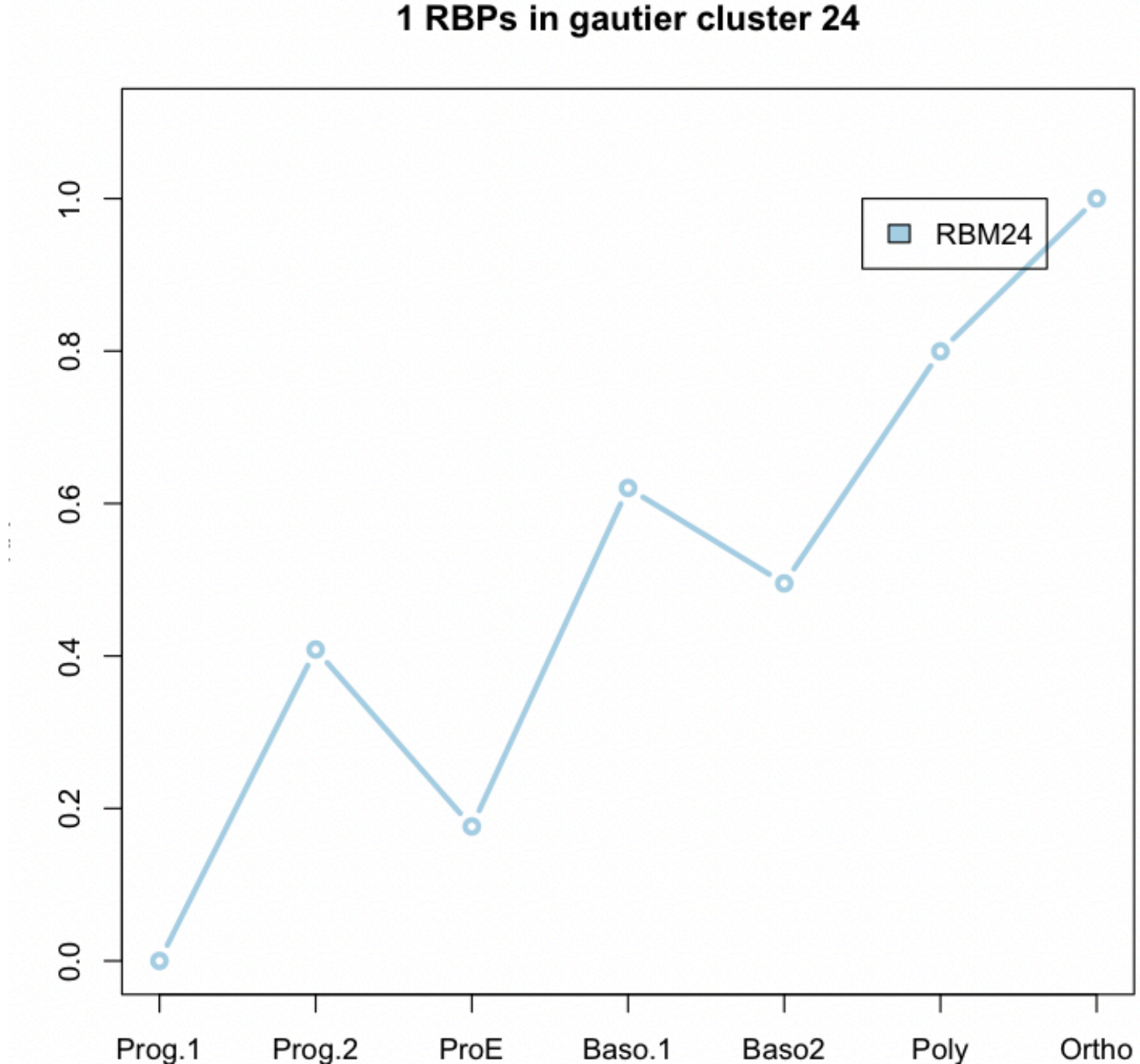

1 RBPs in gautier cluster 14
CPEB4
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

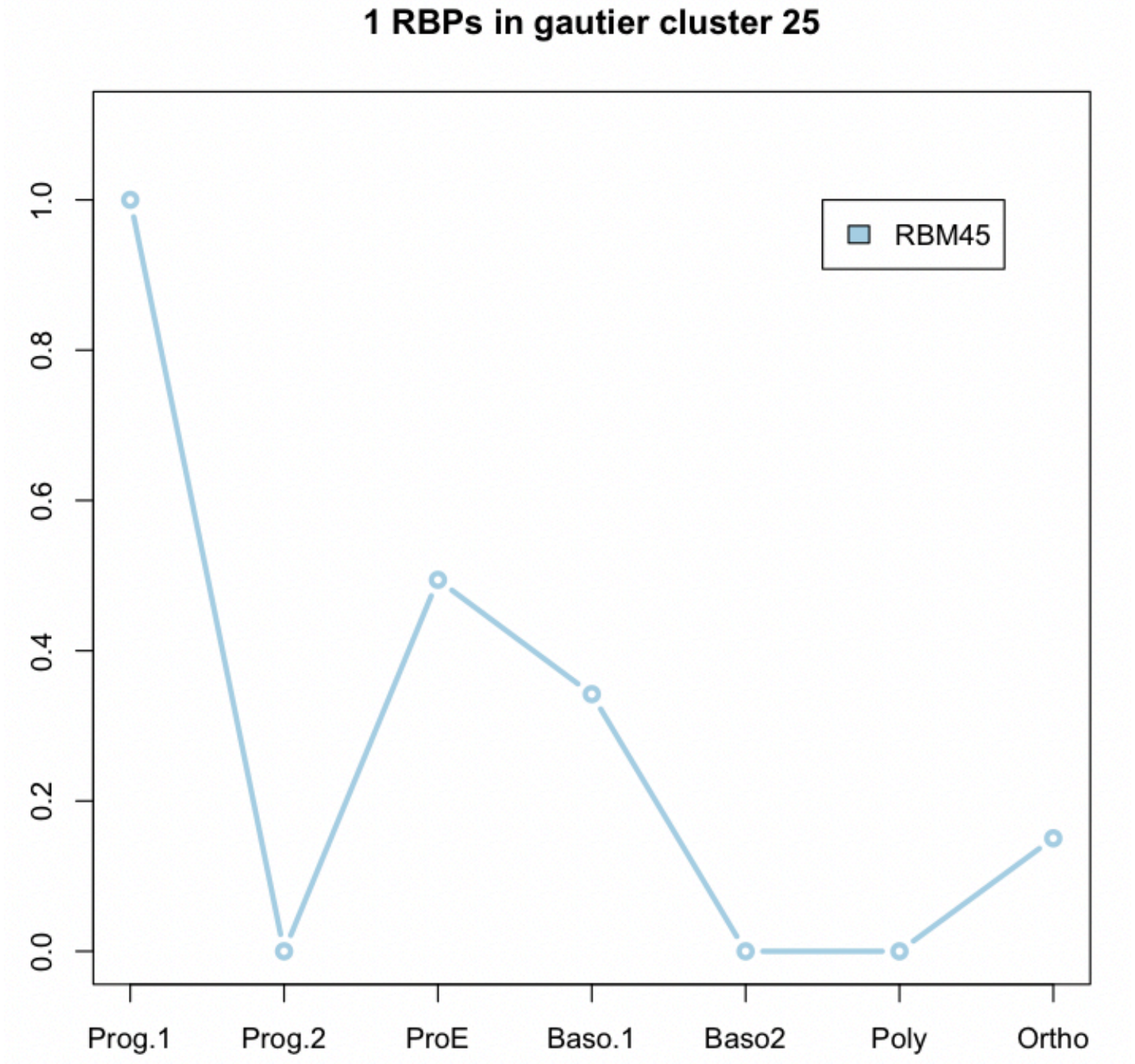

1 RBPs in gautier cluster 16
CSTF2T
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

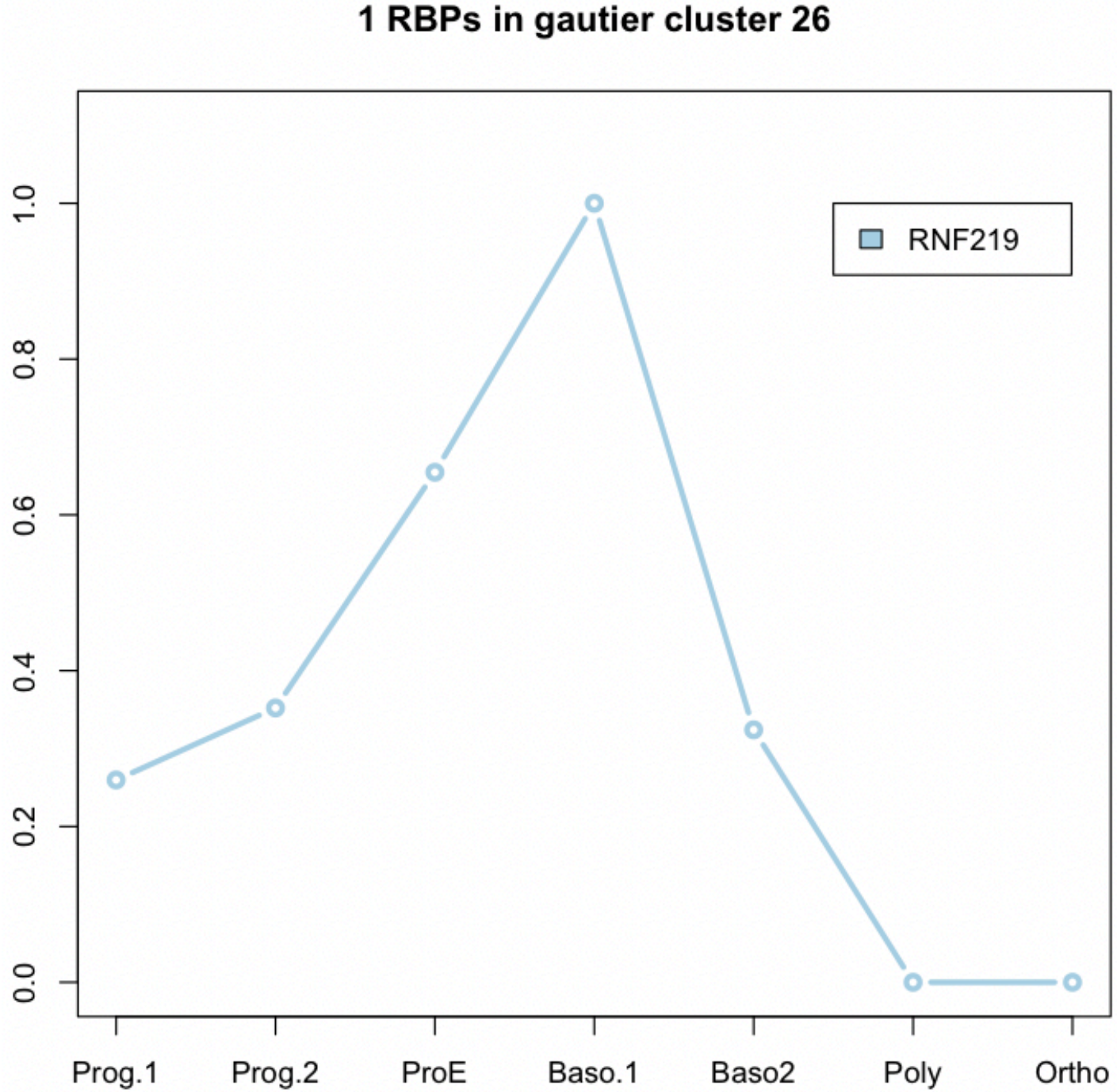

1 RBPs in gautier cluster 19
DROSHA
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

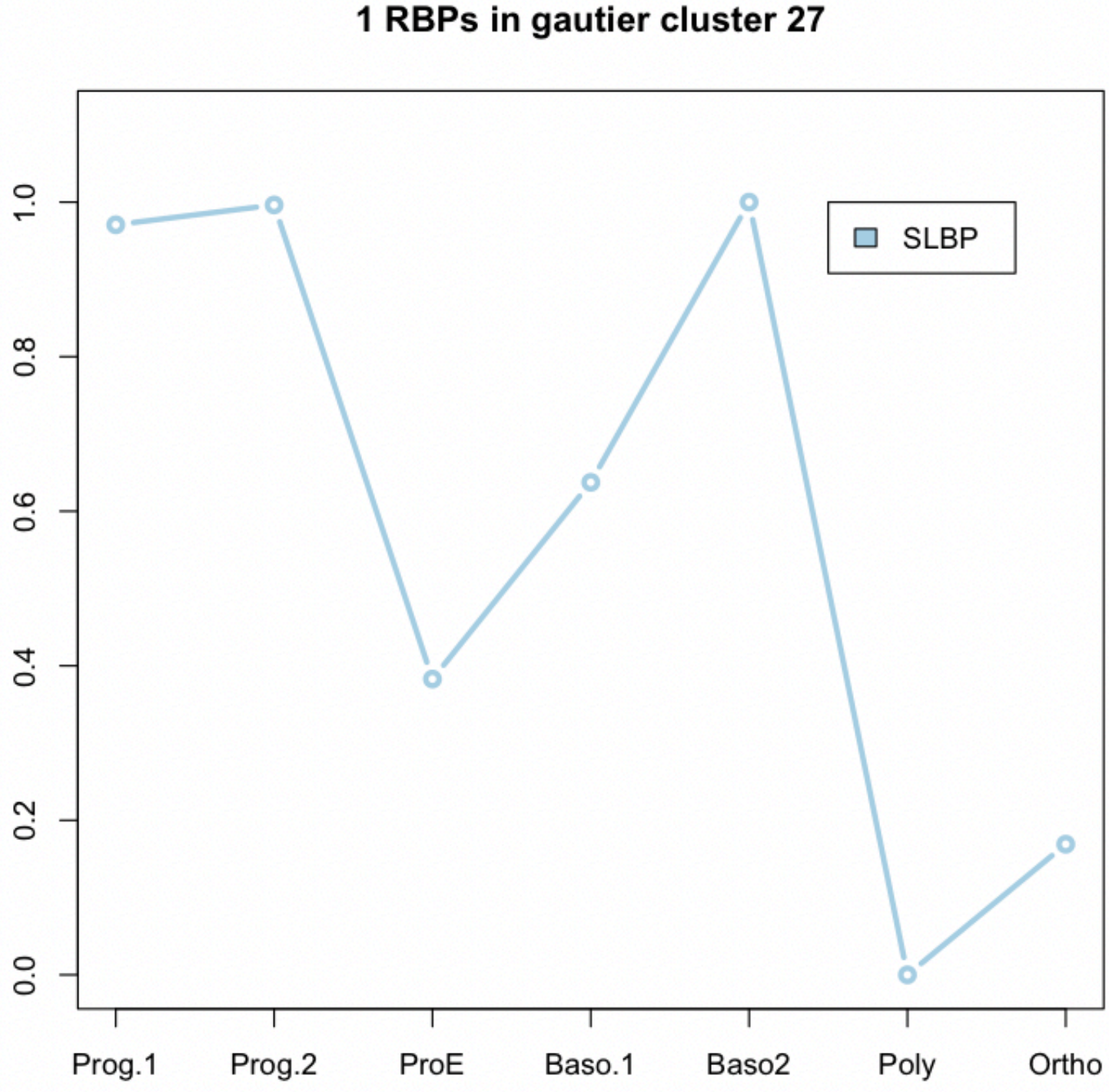

1 RBPs in gautier cluster 22
HNRNPCL1
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

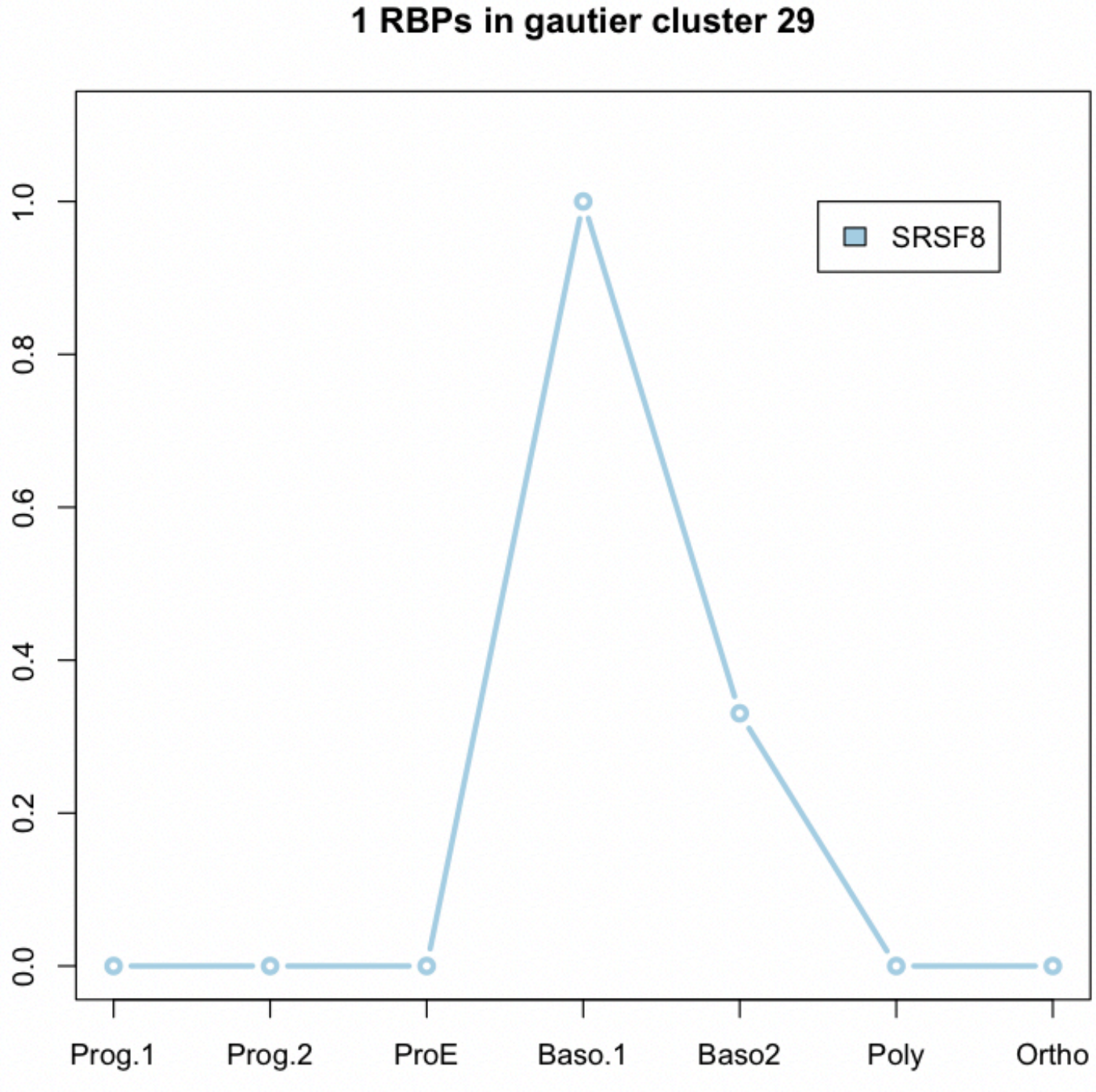

1 RBPs in gautier cluster 23
NOL12
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

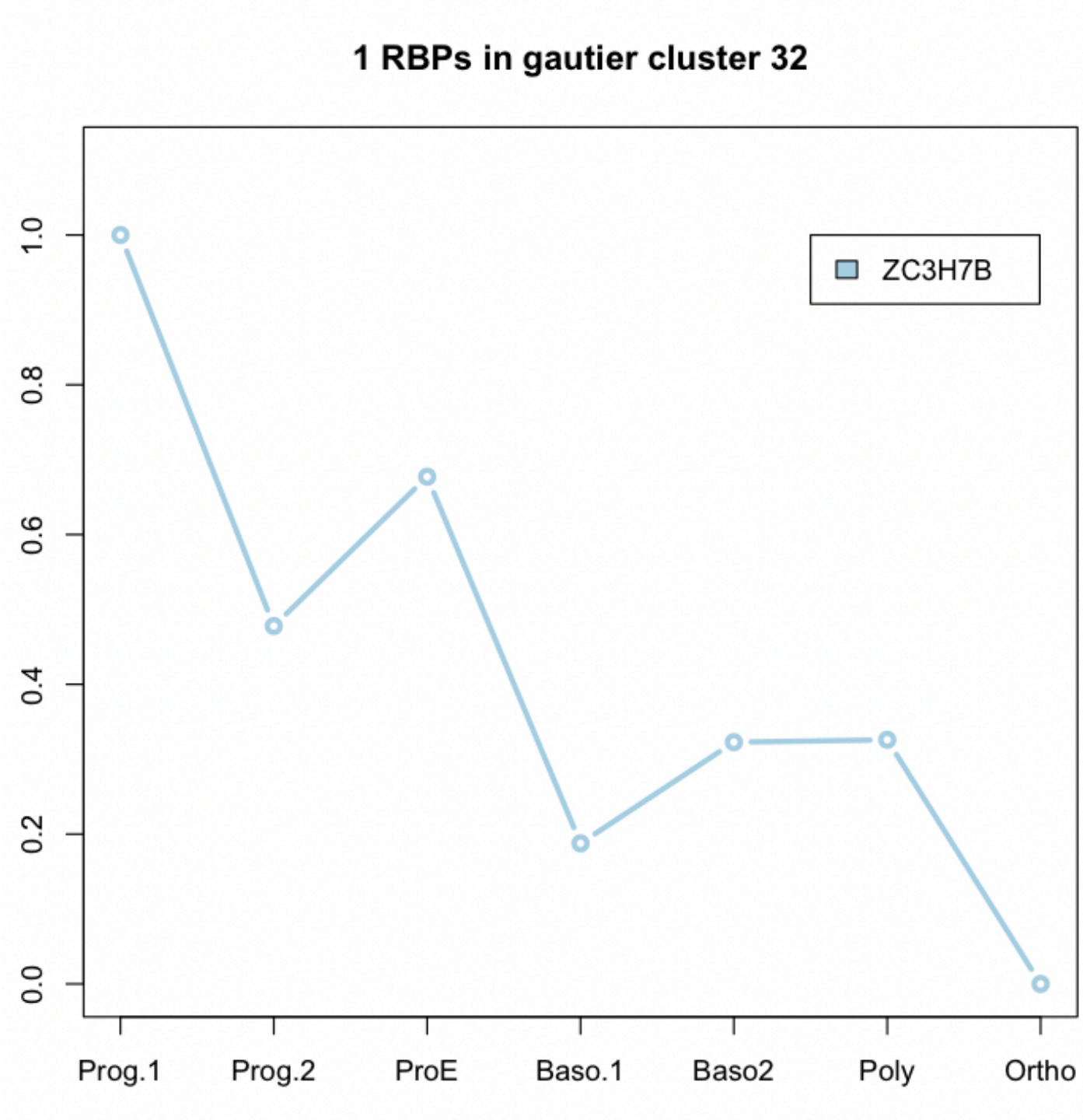

1 RBPs in gautier cluster 24
RBM24
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho

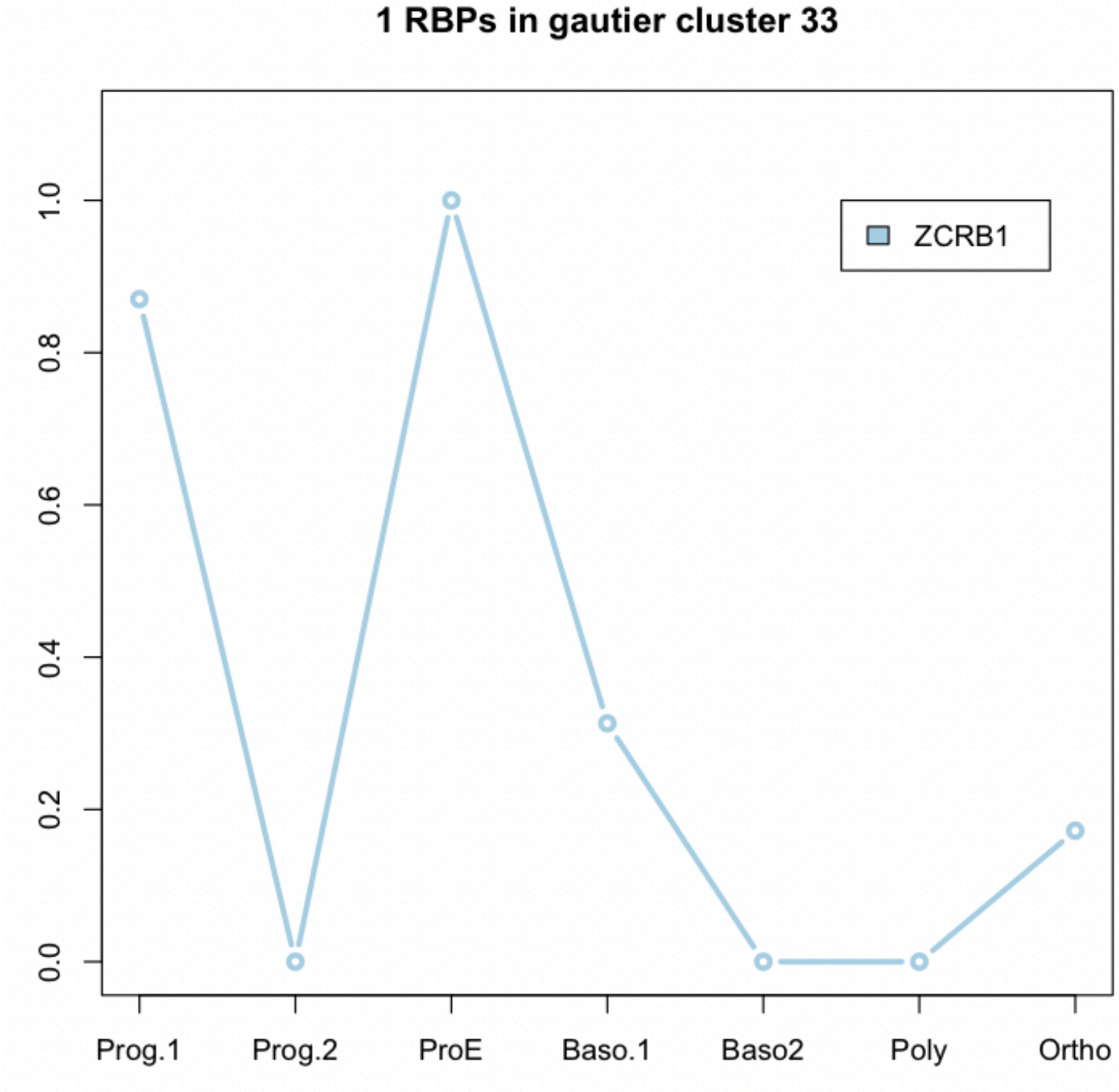

1 RBPs in gautier cluster 25
RBM45
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho


1 RBPs in gautier cluster 26
RNF219
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho


1 RBPs in gautier cluster 27
SLBP
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho


1 RBPs in gautier cluster 29
SRSF8
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho


1 RBPs in gautier cluster 32
ZC3H7B
1.0
0.8
0.6
0.4
0.2
0.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho


1 RBPs in gautier cluster 33
ZCRB1
0.0
0.2
0.4
0.6
0.8
1.0
Prog.1
Prog.2
ProE
Baso.1
Baso2
Poly
Ortho