PLOS ONE

RESEARCH ARTICLE

Backup transcription factor binding sites protect human genes from mutations in the promoter

Jay C. Brown *

Department of Microbiology, Immunology and Cancer Biology, University of Virginia School of Medicine, Charlottesville, Virginia, United States of America

* jcb2g@virginia.edu

OPEN ACCESS

Citation: Brown JC (2023) Backup transcription factor binding sites protect human genes from mutations in the promoter. PLoS ONE 18(8): e0281569. https://doi.org/10.1371/journal.pone.0281569

Editor: Roberto Mantovani, Università degli Studi di Milano, ITALY

Received: January 25, 2023

Accepted: August 16, 2023

Published: August 31, 2023

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by an award from the University of Virginia. The University of Virginia had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The author has declared that no competing interests exist.

Abstract

This study was designed to test the idea that the regulatory regions of human genes have evolved to be resistant to the effects of mutations in their primary function, the control of gene expression. It is proposed that the transcription factor/transcription factor binding site (TF/TFBS) pair having the greatest effect on control of a gene is the one with the highest abundance among the regulatory elements. Other pairs would have the same effect on gene expression and would predominate in the event of a mutation in the most abundant pair. It is expected that the overall regulatory design proposed here will be highly resistant to mutagenic change that would otherwise affect expression of the gene. The idea was tested beginning with a database of 42 human genes highly specific for expression in brain. For each gene, the five most abundant TF/TFBS pairs were identified and compared in their TFBS occupancy as measured by their ChIP-seq signal. A similar signal was observed and interpreted as evidence that the TF/TFBS pairs can substitute for one another. TF/TFBS pairs were also compared in their ability to substitute for one another in their effect on the level of gene expression. The study of brain specific genes was complemented with the same analysis performed with 31 human liver specific genes. Like the study of brain genes, the liver results supported the view that TF/TFBS pairs in liver specific genes can substitute for one another in the event of mutagenic damage. Finally, the TFBSs in the brain specific and liver specific gene populations were compared with each other with the goal of identifying any brain selective or liver selective TFBSs. Of the 89 TFBSs in the pooled population, 58 were found only in brain specific but not liver specific genes, 8 only in liver specific but not brain specific genes and 23 were found in both brain and liver specific genes. The results were interpreted to emphasize the large number of TFBS in brain specific genes.

Introduction

Investigators continue to be puzzled by the large number of transcription factor binding sites present in the regulatory regions of human genes. Whereas in prokaryotic organisms one or a very few TF binding sites suffice for the regulatory needs of most genes, hundreds are the rule in human gene promoters. For instance, a total of 357 TFBS are annotated in DLGAP3, a human gene of average length. High TFBS numbers are also the rule in the genomes of other eukaryotes. Further, in many human genes distinct TFBS clusters are observed outside the promoter but within the coding region of a gene. Five such clusters are annotated in DLGAP3 [1, 2].

Reasonable suggestions have been advanced to account for the large number of TFBS found in the genomes of eukaryotes. TFs and TFBSs have been proposed, for example, to be involved in specifying the tissue where a gene is expressed, a feature less relevant in prokaryotes [3–6]. Eukaryotic gene expression could be controlled by combinations of TFs in a way that would require an increased number of TFBSs [7–9]. Despite the availability of reasonable suggestions, however, there is currently no consensus about the role of the "extra" TF/TFBSs present in eukaryotic genomes.

The study described here was carried out to test the idea that the large number of TFBSs in eukaryotic promoters and other control elements has evolved to protect the regulatory region from the effect of mutations. Since the interaction of a TFBS with its TF occurs in the promoters of many different genes, it is reasonable to expect that additional protection against mutagenic change may be required for TFBS compared to DNA regions with fewer interaction targets. The hypothesis tested in the present study proposes that the additional TFBS in eukaryotes has evolved to meet the need for the additional protection against mutation in the TF/TFBS contact.

The mechanism proposed for protection depends on the abundance of binding sites for a TF in the promoter. It is suggested that expression of a gene depends to the greatest extent on the TF that binds to the highest abundance TFBS. Lesser effects on transcription would be provided by TFBS with lower abundance in the promoter. This situation would prevail until there is a mutation causing loss of the highest abundance TF/TFBS interaction. In that case the second most abundant TFBS/TF pair would become the most abundant and gene expression would continue unchanged because the new highest abundance TFBS would dominate regulation of transcription. The proposal depends on the idea that transcription regulation depends on the identity of the highest abundance TF/TFBS pair, not on the identity of the individual TF/TFBS pair involved. A variety of different TF/TFBS pairs are proposed to be able dominate regulation of a gene if the pair is the most abundant one in the promoter.

The study described here was designed to test the idea outlined above for control of gene expression. It was reasoned that if control were indeed dominated by the most abundant TF/TFBS pair with the second most abundant as a backup, then the ChIP-seq signal from the most abundant TF/TFBS pair should be numerically similar to that of the second most abundant. The third should resemble the second and so on. Also, a recognizably similar result should be obtained if one plotted gene expression level against TFBS occupancy for the most abundant TFBS/TF pair and the second. If the second most abundant pair has evolved to take the place of the most abundant in the event of a mutation, then the second most abundant pair should have a similar effect on transcription as the first.

The study was performed beginning with a database of 42 human genes each with highly selective expression in the brain. Public databases were used to accumulate the expression level of each gene and the ChIP-seq signal from each of the five most abundant TFBSs. The tests described above were then carried out with all 42 database genes. The ChIP-seq signal from the five most abundant TFBS were compared with each other, and plots of gene expression vs ChIP-seq signal were compared for the three most abundant ChIP-seq pairs. It was expected that resemblance of the ChIP-seq signal and the ChIP-seq/expression results from the database genes would support the idea that the second most abundant rank of TFBS would have the ability to replace the first if the first were lost due to mutation.

A similar study was conducted with a database of human genes with highly selective expression in liver. Information was collected about gene expression level and TFBS occupancy for different abundance ranks as described for the brain database. ChIP-seq signals and plots of expression against ChIP-seq signal were created in the same way. It was anticipated that the outcome of the liver study would clarify whether the brain results are unique to brain or if they would apply more broadly.

Finally, a study was focused on the TFBSs present in the brain database. All TFBSs in the brain population were pooled and compared to the similar pool of TFBSs in the liver group. The results were expected to identify any instances of brain-selective or liver-selective TFBSs.

Materials and methods

Gene databases: TF binding levels and gene expression

All genes were derived from the GRCh38/hg38 version of the human genome. Brain specific genes were accumulated for the study from a combination of locally curated genes and those from the database of Sonawane et al, [3]. A gene was considered to be brain-specific if its expression was 10-fold or greater in brain than in the tissue with the next highest expression level. Focus was on a tissue specific gene population (i.e. brain) to maximize the possibility that conclusions would apply to the level of gene expression and minimize contributions from the potential effects of TFs on tissue distribution. The expression level for each gene was obtained from the GTEx Portal of RNA-seq results found in the UCSC Genome Browser (version hg38 (https://genome.ucsc.edu/)). The database of liver-specific genes was curated using the same procedure described above for brain.

For each gene examined, the TF binding abundance of each TFBS was determined from the ChIP-seq values reported in the ENCODE project database of cis-Regulatory Elements (3 November 2018 version) by way of the UCSC Genome Browser. ChIP-seq values were identified from the promoter and from other TFBS clusters within the gene coding region. No effort was made to include regulatory elements outside the promoter and gene coding region. Table Browser was searched on Regulation followed by TF Clusters, and the output was focused on the "score" entry. The score was summed for each TFBS as there were several (usually ~3–10) instances of a TFBS annotated in the regulatory region of each gene. The highest scoring TFBS and its score were then recorded in the brain-specific or liver-specific database under the heading "ChIP1." Values for the second highest TFBS were recorded under ChIP2 and so on through ChIP5. Note that the TFBS associated with the ChIP1 rank cannot be the same as that for ChIP-seq2, or ChIP-seq 3 and so on. For each gene, each ChIP-seq rank has its own TFBS.

Data analysis

Data were manipulated with RStudio and Excel. Graphic rendering was done with SigmaPlot v14.5.

Results

Experimental strategy

As indicated above, the analyses described here were designed to test the idea that the regulatory regions of human genes have evolved to create redundant layers of TFBS able to substitute for one another in the event of mutagenic damage. Relevant TFBS groups were identified according to the abundance of their occupancy by the cognate TF as determined from the results of ChIP-seq experiments. For each gene, the highest abundance TFBS/TF pair was put into one pool (rank 1), the second most abundant was put in a second pool (rank 2) and so on to rank 5. Each data point therefore has its own TF and its own level of TFBS occupancy (i.e. ChIP-seq signal). If the information described above is to function in a redundant manner to control the level of gene expression, then it is expected that each TFBS/TF rank will: (1) resemble the others in TF occupancy as measured by its ChIP-seq signal; and (2) have a recognizably similar effect on the level of gene expression as measured by the RNA-seq measurement. The above expectations were tested in the studies described below. It was assumed that a positive result would support the existence of redundant layers of regulatory control in human genes.

Effect of TFBS occupancy rank on brain specific gene expression

The effect of TFBS occupancy rank on gene expression was examined with a population of 42 protein coding, human, brain specific genes (see S1 Table). For each gene the table shows the most abundant TFBS together with its ChIP-seq signal (columns labeled "tfbs1" and "ChIP1," respectively). Other columns show the TFBS and ChIP-seq signals for ranks 2–5. Perhaps the easiest way to evaluate the ability of the ChIP-seq results to serve as backups for each other is to compare the ChIP-seq values. The closer the values cluster with each other the better they should be able to substitute for their function. A mixture of ranges was observed in the brain gene population (S1 Table). For instance, a wide range of ChIP-seq values (2231–1000) was observed with ANKRD34C, while narrower ranges were found with others such as ANDRD63 (1691–1327) and GRIN2B (11,057–8550). Clear outlier values were observed with the ChIP-seq signal of some genes (e.g. GRM4).

To summarize the results, outlier values were discarded and the average values for the ChIP-seq signal in adjacent ranks was plotted for the 42 genes in the brain database (Fig 1A). A skewed distribution favoring lower values for the inter-rank ChIP-seq values was observed, a result that favors the view that adjacent TF/TFBS pairs would be well suited to substitute for each other as predicted. The presence of higher inter-rank ChIP-seq values in the distribution (e.g. values greater than 200 or 300) indicates the presence of TF/TFBS pairs less well able to substitute for their neighbors in the event of a mutation.

TFBS occupancy rank and liver specific gene expression

The same analysis described above for brain specific genes was performed with a database of 31 genes with specific expression in liver (S2 Table). The goal was to determine whether analysis of liver specific genes would support the same interpretation as the brain specific ones, namely that there are TF/TFBS pairs in the regulatory regions that can serve as backup for each other in the event of mutations to the DNA. As with the brain specific genes, for each liver specific gene the TF/TFBS pairs with the highest to the fifth highest abundance as judged by ChIP-seq score were accumulated and compared. Similar scores among the five TF/TFBS pairs would favor the idea that the five were well suited to serve as backups for one another.

A mixture of results was observed. In some genes a wide range of ChIP-seq signals was observed, while in others measurements were more similar. AFP is an example of the first group, and AGXT and AHSG are examples of the second. To characterize the overall distribution, an average for ChIP-seq values in adjacent TF/TFBS pairs were computed and plotted (Fig 1B). The results showed a skewing of the distribution toward lower inter-TF/TFBS pairs, a result that favors the ability of nearby TF/TFBS pairs to substitute for one another.

Tests with gene expression: Brain specific genes

A more demanding test of the ability of TF/TFBS ranks to substitute for each other would occur if entire ranks of TF/TFBSs could be compared at once rather than comparing individual

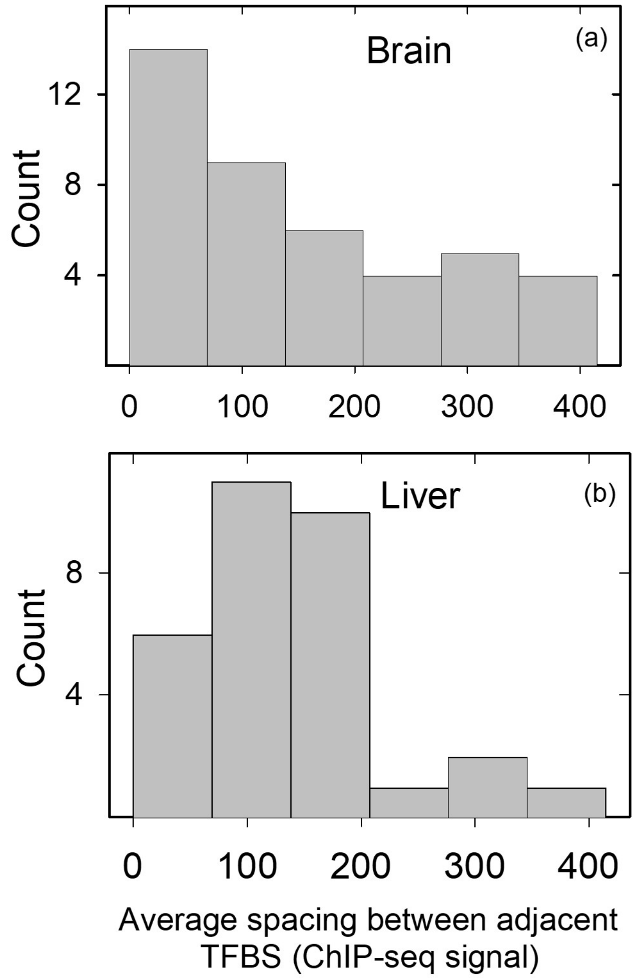

Fig 1. Similarity in the ChIP signal in the five most abundant TFBS. Histogram showing the difference in ChIP-seq signal between adjacent TFBS in the promoters of database brain (a) and liver (b) genes. Data plotted in (a) and (b) are derived from S1 and S2 Tables, respectively.

https://doi.org/10.1371/journal.pone.0281569.g001

TF/TFBS pairs only. This goal was accomplished by involving the gene expression level as measured by GTEx RNA-seq. For each gene in the brain specific database, the value for the expression level was plotted against the value of the ChIP-seq signal for rank 1 TF. This was then repeated for rank 2 and rank 3 TF, and the plots were compared. It was expected that a resemblance among the three plots would suggest the ability of the three ChIP-seq ranks to substitute for each other in regulating the level of gene expression.

Fig 2 shows the results obtained with genes in the brain specific database. Clear similarities were observed among the three ranks in the expression/ChIP-seq signal relationship. For instance, in the lower left of the plot the RTP1, PNMA6F and FGF3 genes are found in similar locations in the three abundance ranks. Near the center of the plot, HTR5A, HCRT and TLX3 are located similarly. CACNG8, TBR1 and CREG2 locations are related in all three plots. The results are interpreted to indicate that the three ChIP-seq abundance ranks have the potential to serve in a redundant fashion to drive expression of the genes to the same level. It is suggested that the small differences observed in ChIP-seq signal for the same gene in separate ranks might produce tolerable differences in expression level if one rank needed to be substituted for another.

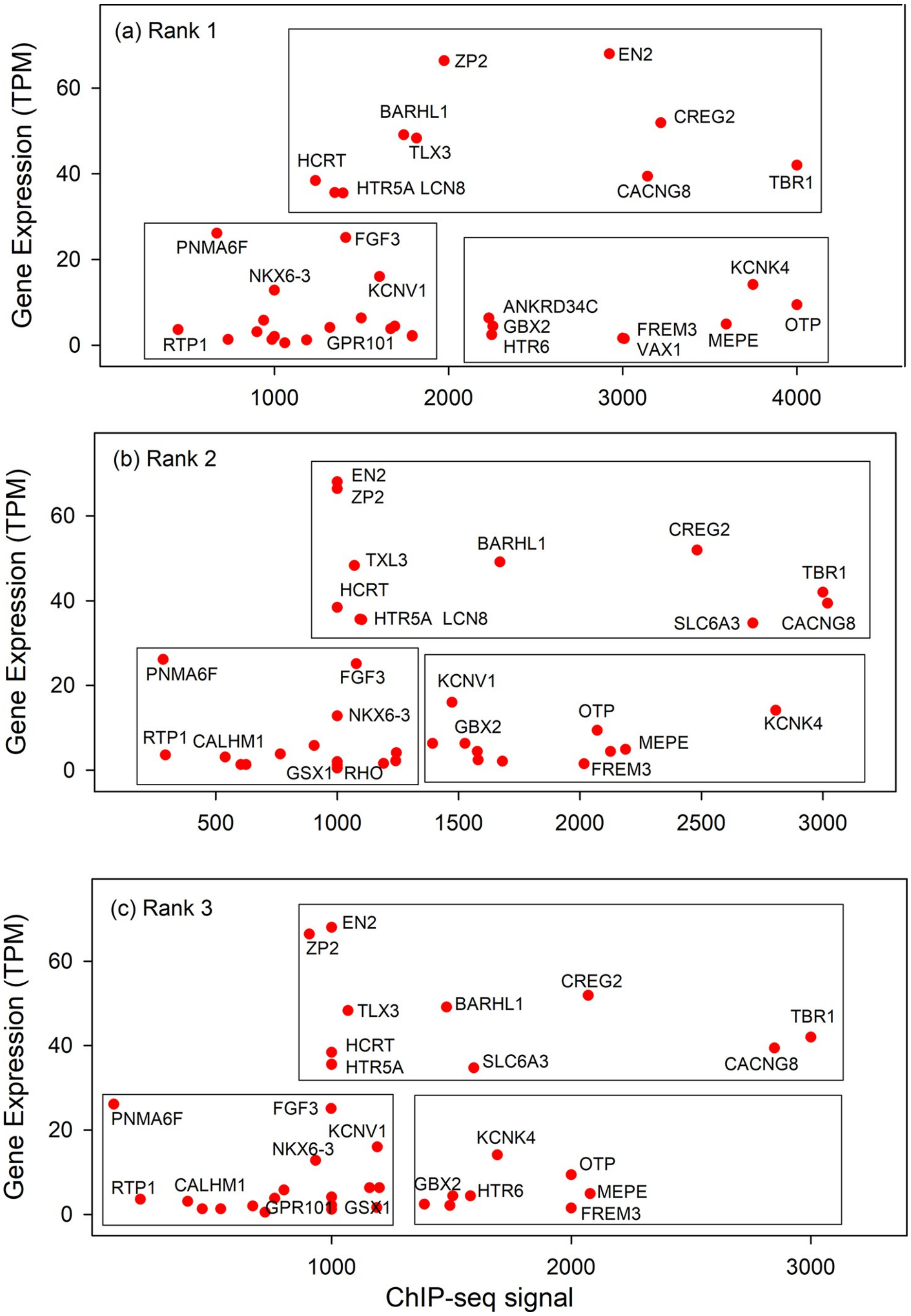

Fig 2. Gene expression/ChIP-seq signal relationship in ranks 1–3 of genes in the brain specific database. Plots for Ranks 1, 2, and 3 are shown in panels (a), (b) and (c), respectively. Numerical values for the data points shown can be found in S1 Table. Boxes are intended to aid in interpretation of the results. They do not contain any experimental results. Note that data points for the same gene are related in the three plots.

https://doi.org/10.1371/journal.pone.0281569.g002

Tests with gene expression: Liver specific genes

To test the generality of the results described above with brain specific genes, a similar analysis was carried out with the database of liver specific genes. The results showed that the same genes were found in similar locations in the three plots (see Fig 3). CFHTR3 and UGT2B10, for instance, are in the high expression part of each plot with CFHR3 in the lower ChIP-seq signal end. An arc of four genes with varying expression levels is found in all three plots near the low ChIP-seq signal level (see SPP2, SERPINA7, F13B and SLC22A10). All plots also have a cluster of four genes at the high end of the ChIP-seq axis (i.e., genes MBL2, INHBC, SLC22A25 and INS-IGF2). The results are interpreted to indicate that the three ChIP-seq abundance ranks can serve in a redundant or backup manner to shield expression of the gene from mutagenic change in the promoter. Further, the results show that the same backup promoter design is found in liver specific as in brain specific human genes.

TFBSs in the promoters of brain and liver specific human genes: All database TFBSs

The information accumulated here about promoter content of TFBSs provides the opportunity to compare the promoters of brain specific and liver specific genes (see S1 and S2 Tables). While previous studies have demonstrated a degree of tissue selectivity to TFBSs, it is rare for such studies to involve the large populations of tissue-specific genes available here [10]. For analysis, TFBSs in the promoters of brain specific genes were assembled in two forms, (1) all the TFBSs shown in S1 Table. That is, the five most abundant TFBSs in each of the 42 brain specific genes or 210 TFBSs in all: and (2) the population of non-redundant TFBSs. Each TFBS is counted only once regardless of how many times it occurs in the overall population. The same two TFBS populations were assembled for the liver-specific genes. Finally, histograms were plotted to show the count of each TFBS in the total population (Figs 4 and 5).

The results for brain specific genes emphasize binding of TFs that use histone methylation to influence gene expression (Fig 4). For instance, of the five most abundant TFBSs three (EZH2, SUZ12 and RNF2) are components of polycomb repressive complex 2 (PRC2), a complex that uses histone methylation to attenuate gene expression [11, 12]. A fourth, KDM4A, also uses histone methylation to regulate transcription [13]. The result suggests an important role for histone methylation in the regulation of brain specific genes.

No similar emphasis on histone methylation was observed with liver-specific genes (Fig 5). Of the seven most abundant TFBSs, three bind TFs that attract transcription machinery by binding to the promoter region (i.e., FOXA1, FOXA2 and HNF4A) [14–16]. Three of the four remaining TFBSs affect transcription of a wide variety of genes. RXRA mediates the effects of retinoids while SP1 and YY1 can activate or repress gene expression depending on other factors [17–19]. Comparing the brain and liver specific gene populations suggests the dominant mechanisms of gene regulation are distinct in the two tissues.

TFBSs in the promoters of brain and liver specific human genes: All unique database TFBSs

Non-redundant TFBS in the brain and liver specific databases were accumulated with the expectation that they might identify TFBSs found in brain-specific genes, but not in liver specific ones. TFBS found in liver specific, but not brain specific genes would also be identified. Scans of S1 and S2 Tables revealed the presence of 81 non-redundant TFBS among the 210 total brain-specific genes and 31 among the 155 liver-specific genes (Table 2). Percentages were 39% in the case of brain specific genes and 20% for liver. Among the 81 non-redundant brain TFBS, 58 (72%) were found in the brain population only and 23 (28%) in both brain and liver populations. Among the 31 non-redundant TFBS in liver, 8 (26%) were found in liver, but not brain and 23 (74%) in both liver and brain.

Two conclusions stand out from analysis of the non-redundant TFBS populations. First is the high proportion of brain-only (i.e. not found in liver) non-redundant TFBS among the

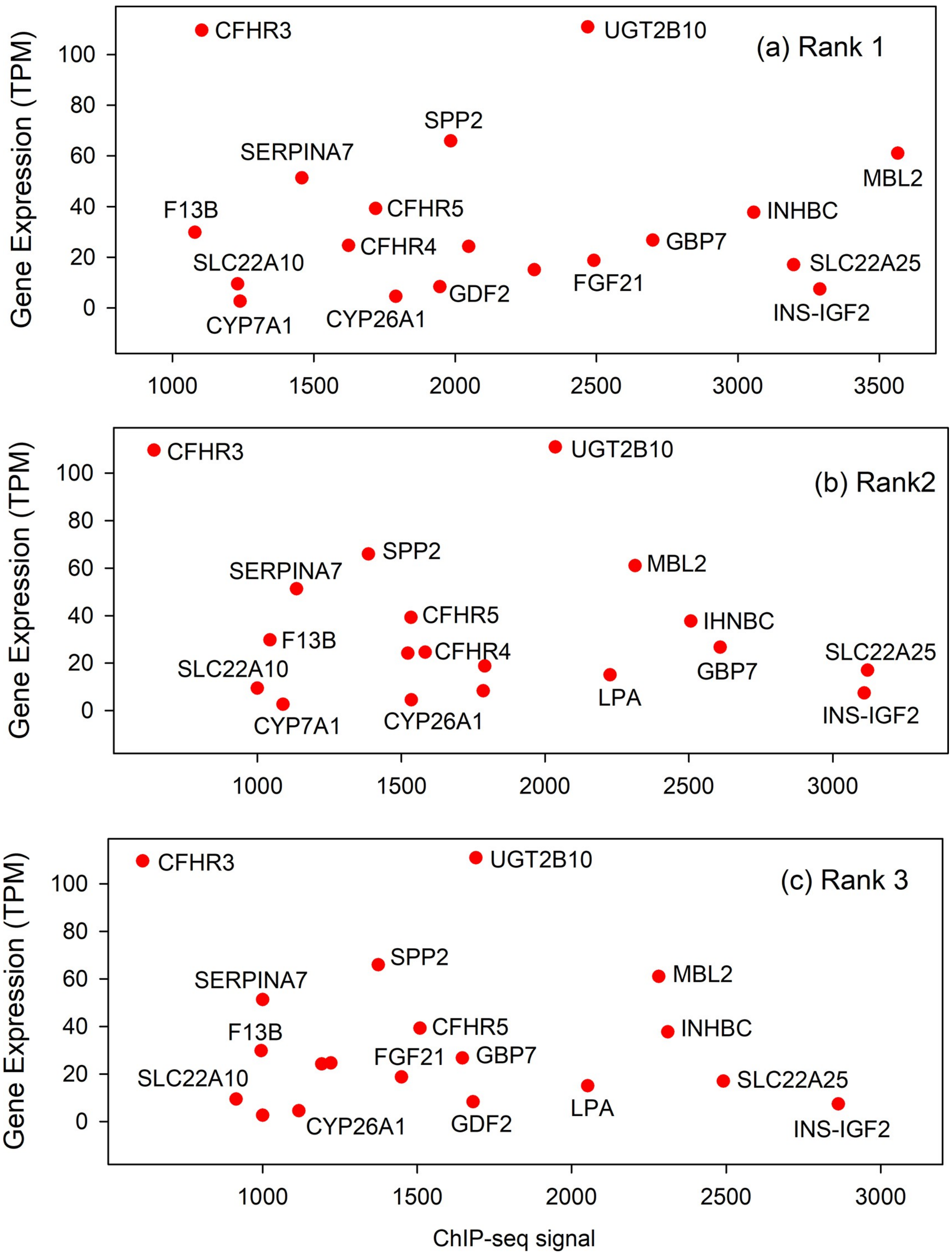

Fig 3. Gene expression/ChIP-seq signal relationship in ranks 1–3 of genes in the liver specific database. Plots for Ranks 1, 2 and 3 are shown in panels (a), (b) and (c), respectively. Numerical values for the data points shown can be found in S2 Table. Note that the location of data points for the same gene are similar in the three plots.

https://doi.org/10.1371/journal.pone.0281569.g003

brain specific genes (58 in the 81 total non-redundant TFBS population). For liver, the comparable figures are 8 of 31 non-redundant TFBS. The high proportion among brain genes may result from a greater need for gene regulatory activity in the brain due perhaps to a greater

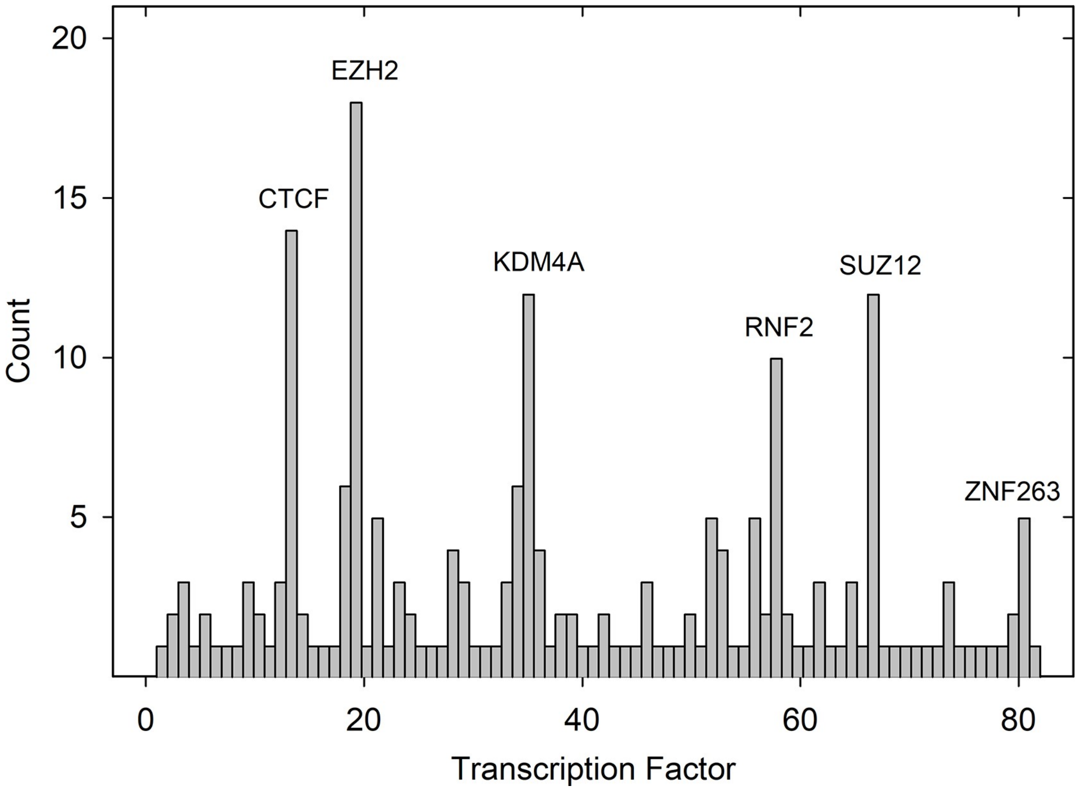

Fig 4. Transcription factor binding site counts in the aggregate of all five brain specific gene ranks. Histogram shows the count of each TFBS in the brain specific gene population (see S1 Table). TFBS in all five ranks are included. TFBS not identified in the figure are found in Table 1. Note that 3 of the 6 TFBS with the highest abundance bind TF subunits of the polycomb repressive complex 2 involved in histone methylation (i.e., EZH2, RNF2 and SUZ12).

https://doi.org/10.1371/journal.pone.0281569.g004

number of functionally distinct regions in the brain or to a more diverse developmental program.

Second is the absence of any evidence for a tissue specific TFBS. All the TFBS found in the brain-specific gene population, for instance, are those that can be readily identified in the promoters of non-brain genes. The same is true for the TFBS identified here in liver specific, but not brain specific genes (Table 2). All can be found abundantly in other human genes. The overall finding indicates that the role of TF in influencing the tissue where a gene is expressed must involve multiple TFs or TFs in combination with other regulatory mechanisms.

Discussion

Backup organization to allow promoter function to co-exist with mutagenic damage

The backup system of promoter organization suggested by the results described here is distinct from other methods humans use to deal with mutagenic change to the genome. The suggested system does not result in prevention of mutations or correction of them once they have occurred. Rather, the system describes a method promoters use to continue to function normally despite the existence of mutations in their own sequences. The results indicate the existence of backup or redundant systems of TFBS that are each able to provide gene regulatory function in the event of mutagenic damage to another. An individual promoter is thereby able

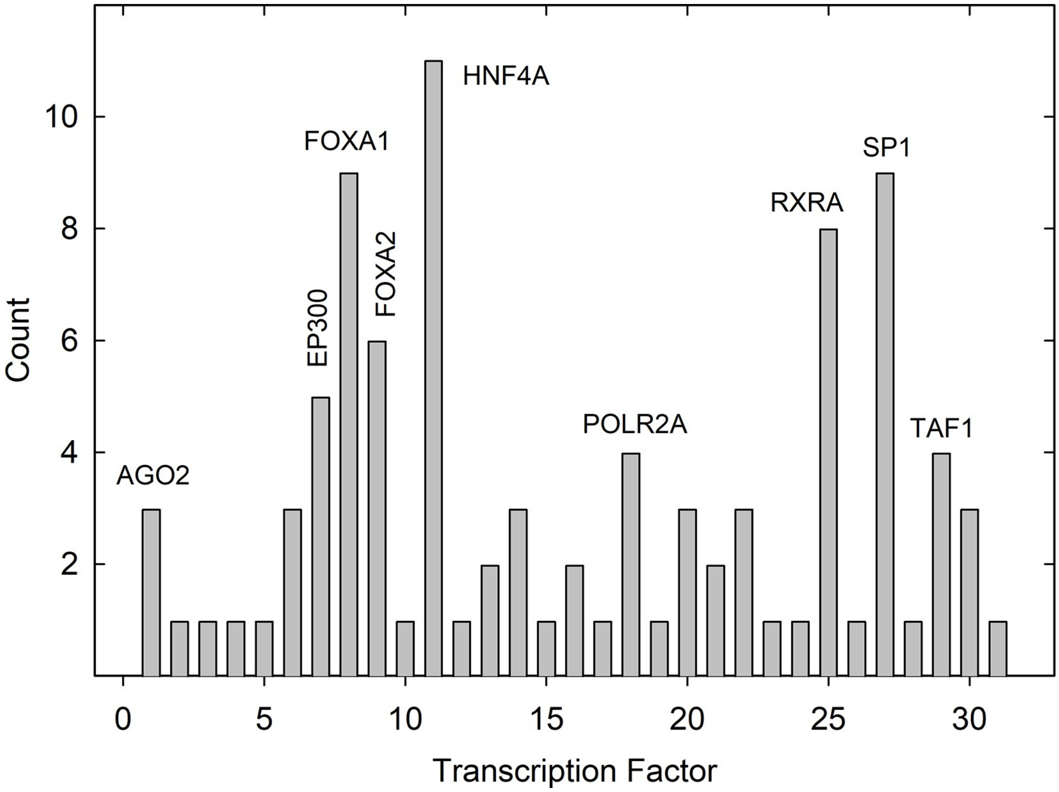

Fig 5. Transcription factor binding site counts in the aggregate of all five liver specific gene ranks. Histogram shows the count of each TFBS in the liver specific gene population (see S2 Table). TFBS in all five ranks ae included. TFBS not identified in the figure are found in Table 1. Note the high abundance of HNF4A binding sites in the liver specific gene population.

https://doi.org/10.1371/journal.pone.0281569.g005

to co-exist with a diverse background of mutagenic events using methods that complement other mechanisms that prevent or correct mutations.

The backup system can be understood as related to other genome-wide features that are able to allow the individual to accommodate potentially disastrous mutagenic events. An example is the way genes encoding proteins that are parts of a single structure or function are distributed widely in the genome rather than being located together in a cluster. The 31 genes encoding distinct protein components of the human nuclear pore complex are an example. The genes are spread among 17 different chromosomes increasing the likelihood that damage to any one gene might be able to be accommodated by the overall structure [20, 21]. This arrangement prevents a mutation in a single chromosome from affecting more than one or a few nuclear pore genes. Other distributed genomic elements that enable humans to co-exist in a mutagenic environment include gene enhancers, homologous genes and exons distributed within the same gene [22–25].

High TFBS content of eukaryotic promoters

The system of backup TFBSs proposed here is consistent with the observation that the promoters of human genes and those of other eukaryotic organisms contain many more TFBSs than those of prokaryotes. The ChIP-seq signal from the most abundant TFBS is in most cases a sum of several individual TFBSs in the promoter. The same is true of the second most abundant TFBS, the third and so on. It is easy to imagine therefore that the number of individual

Table 1. Key to transcription factor binding site numbering in Figs 4 and 5.

No.	Brain (Fig 4)	Liver (Fig 5)
1	AGO2	AGO2
2	ASH2L	ARID3A
3	ATF2	ATF2
4	ATF7	ATF7
5	BCOR	CREM
6	CBFA2T3	CTCF
7	CBX2	EP300
8	CBX8	FOXA1
9	CEBPB	FOXA2
10	CHD1	HDAC2
11	CREB1	HNF4A
12	CTBP1	HNRNPK
13	CTCF	HNRNPL
14	E2F6	JUND
15	EGR1	MAFK
16	EHMT2	MAX
17	EP300	NCOR1
18	EZH2	POLR2A
19	FIPL1	POLR2G
20	FOS	RAD21
21	FOXA1	RBFOX2
22	FOXA2	RBM22
23	GABPA	RBM39
24	GATA2	RUNX3
25	GATA3	RXRA
26	GATAD2B	SMARCA4
27	HDAC1	SP1
28	HDAC2	SRSF4
29	HDAC6	TAF1
30	HNF4A	YY1
31	HNRNPL	ZBTB33
32	IKZF1	
33	JUND	
34	KDM1A	
35	KDM4A	
36	L3MBTL2	
37	MAFF	
38	MAFK	
39	MAX	
40	MGA	
41	MNT	
42	MXI1	
43	MYC	
44	NR2C1	
45	NR2F2	
46	NRF1	
47	PAX5	

(*Continued*)

Table 1. (Continued)

No.	Brain (Fig 4)	Liver (Fig 5)
48	PCBP1	
49	PKNOX1	
50	POLR2A	
51	POLR2G	
52	RAD21	
53	RBBP5	
54	RBFOX2	
55	RBM25	
56	REST	
57	RFX1	
58	RNF2	
59	RUNX3	
60	RXRA	
61	SETDB1	
62	SIN3A	
63	SMARCA4	
64	SMC3	
65	SP1	
66	STAT3	
67	SUZ12	
68	TBP	
69	TRIM28	
70	USF	
71	USF1	
72	XRCC5	
73	YY1	
74	ZBTB33	
75	ZEB1	
76	ZEB2	
77	ZKSCAN1	
78	ZMTM3	
79	ZNF143	
80	ZNF263	
81	ZNF639	

https://doi.org/10.1371/journal.pone.0281569.t001

TFBSs in the aggregate of all abundance classes (i.e. ranks) might reach the high number observed in an entire gene promoter region.

Tissue specific genes

This study was focused on tissue specific genes with the idea that the results might clarify the way tissue specificity is encoded in promoters. The results confirmed a previously noted selectivity of HNF4A for liver genes [26]. A binding site for HNF4A was observed to be among the five most abundant in 19 of the 31 liver specific genes examined here (see S2 Table). Evidence for other tissue specific TFBSs was less convincing. For instance, 21 different TFBS were found among the 42 highest abundance TFBS in the brain specific genes examined here (see S1 Table).

Table 2. Transcription factor binding sites present in the promoters of database brain-specific and liver-specific genes (all five abundance ranks).

No.	Brain only	No.	Brain only (cont.)	No.	Liver only	Both Brain and Liver
1	ASH2L	31	NR2C1	1	ARID3A	AGO2
2	BCOR	32	NR2F2	2	CREM	ATF2
3	CBFA2T3	33	NRF1	3	HNRNPK	ATF7
4	CBX2	34	PAX5	4	NCOR1	CTCF
5	CBX8	35	PCBP1	5	RBM22	EP300
6	CEBPB	36	PKNOX1	6	RBM39	FOXA1
7	CHD1	37	RBBP5	7	SRSF4	FOXA2
8	CREB1	38	RBM25	8	TAF1	HDAC2
9	CTBP1	39	REST	9		HNF4A
10	E2F6	40	RFX1	10		HNRNPL
11	EGR1	41	RNF2	11		JUND
12	EHMT2	42	SETDB1	12		MAFK
13	EZH2	43	SIN3A	13		MAX
14	FIPL1	44	SMC3	14		POLR2A
15	FOS	45	STAT3	15		POLR2G
16	GABPA	46	SUZ12	16		RAD21
17	GATA2	47	TBP	17		RBFOX2
18	GATA3	48	TRIM28	18		RUNX3
19	GATAD2B	49	USF	19		RXRA
20	HDAC1	50	USF1	20		SMARCA4
21	HDAC6	51	XRCC5	21		SP1
22	IKZF1	52	ZEB1	22		YY1
23	KDM1A	53	ZEB2	23		ZBTB33
24	KDM4A	54	ZKSCAN1			
25	L3MBTL2	55	ZMTM3			
26	MAFF	56	ZNF143			
27	MGA	57	ZNF263			
28	MNT	58	ZNF639			
29	MXI1					
30	MYC					

https://doi.org/10.1371/journal.pone.0281569.t002

Overall, the results provide little support for the view that there is a single TF that uniquely determines the tissue location of a brain or liver specific human gene.

Role of TFBS occupancy in the level of gene expression

The results described here provide a test of the idea that the level of TF binding to its TFBS may influence the level of a gene's expression, at least in some cases. Support for the idea would be obtained if the level of TFBS occupancy as measured by the ChIP-seq signal were found to be correlated with the level of gene transcription. Such correlations can be found in both the brain and liver gene populations examined here (see Figs 2 and 3). Among the brain genes an inverse correlation is observed among the genes ZP2, BARHL1, CREG2, CACNG8 and TBR1 in all three abundance ranks plotted. In the liver gene population, a positive correlation is observed relating genes F13B, SERPIN7, SPP2 and UGT2B10 in all three ranks shown. The results suggest that expression of genes in the brain group is repressed by TFBS occupancy in all three TF ranks while expression is activated in the liver gene group. In both cases the results support the view that interaction of TF and TFBS can have an influence on transcription level in the genes involved. A similar TF dose effect on human gene expression has recently been reported for the human SOX9 gene [27].

Supporting information

S1 Table. All human brain-specific genes used in this study (42 genes). [a] Most abundant transcription factor binding site (i.e., rank 1 transcription factor binding site). [b] ChIP-seq signal for rank 1 transcription factor binding site. [c] Gene expression in TPM from GTEx. (DOCX)

S2 Table. All human liver-specific genes used in this study (31 genes). [a] Most abundant transcription factor binding site (i.e., rank 1 transcription factor binding site). [b] ChIP-seq signal for rank 1 transcription factor binding site. [c] Gene expression in TPM from GTEx. (DOCX)

Author Contributions

Conceptualization: Jay C. Brown.

Data curation: Jay C. Brown.

Formal analysis: Jay C. Brown.

Investigation: Jay C. Brown.

Methodology: Jay C. Brown.

Project administration: Jay C. Brown.

Resources: Jay C. Brown.

Validation: Jay C. Brown.

Writing – original draft: Jay C. Brown.

Writing – review & editing: Jay C. Brown.

References

1. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003; 424(6945):147–51. https://doi.org/10.1038/nature01763 PMID: 12853946
2. Hawley DK, McClure WR. Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Res. 1983; 11(8):2237–55. https://doi.org/10.1093/nar/11.8.2237 PMID: 6344016
3. Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, et al. Understanding Tissue-Specific Gene Regulation. Cell Rep. 2017; 21(4):1077–88. https://doi.org/10.1016/j.celrep.2017.10.001 PMID: 29069589
4. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 2007; 39 (6):730–2. https://doi.org/10.1038/ng2047 PMID: 17529977
5. Liu T, Zhang S, Xiang D, Wang Y. Induction of hepatocyte-like cells from mouse embryonic stem cells by lentivirus-mediated constitutive expression of Foxa2/Hnf4a. J Cell Biochem. 2013; 114(11):2531–41. https://doi.org/10.1002/jcb.24604 PMID: 23744720
6. Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics. 2008; 9:271. https://doi.org/10.1186/1471-2105-9-271 PMID: 18541026
7. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38(4):576–89. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432
8. Roudnicky F, Kim BK, Lan Y, Schmucki R, Kuppers V, Christensen K, et al. Identification of a combination of transcription factors that synergistically increases endothelial cell barrier resistance. Sci Rep. 2020; 10(1):3886. https://doi.org/10.1038/s41598-020-60688-x PMID: 32127614

9. Bright AR, van Genesen S, Li Q, Grasso A, Frolich S, van der Sande M, et al. Combinatorial transcription factor activities on open chromatin induce embryonic heterogeneity in vertebrates. EMBO J. 2021; 40(9):e104913. https://doi.org/10.15252/embj.2020104913 PMID: 33555045

10. Papathanasiou P, Perkins AC, Cobb BS, Ferrini R, Sridharan R, Hoyne GF, et al. Widespread failure of hematolymphoid differentiation caused by a recessive niche-filling allele of the Ikaros transcription factor. Immunity. 2003; 19(1):131–44. https://doi.org/10.1016/s1074-7613(03)00168-7 PMID: 12871645

11. Blackledge NP, Rose NR, Klose RJ. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. Nat Rev Mol Cell Biol. 2015; 16(11):643–9. https://doi.org/10.1038/nrm4067 PMID: 26420232

12. Schuettengruber B, Bourbon HM, Di Croce L, Cavalli G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell. 2017; 171(1):34–57. https://doi.org/10.1016/j.cell.2017.08.002 PMID: 28938122

13. Klose RJ, Kallin EM, Zhang Y. JmjC-domain-containing proteins and histone demethylation. Nat Rev Genet. 2006; 7(9):715–27. https://doi.org/10.1038/nrg1945 PMID: 16983801

14. Cirillo LA, Zaret KS. Specific interactions of the wing domains of FOXA1 transcription factor with DNA. J Mol Biol. 2007; 366(3):720–4. https://doi.org/10.1016/j.jmb.2006.11.087 PMID: 17189638

15. Cernilogar FM, Hasenoder S, Wang Z, Scheibner K, Burtscher I, Sterr M, et al. Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. Nucleic Acids Res. 2019; 47 (17):9069–86. https://doi.org/10.1093/nar/gkz627 PMID: 31350899

16. Walesky C, Apte U. Role of hepatocyte nuclear factor 4alpha (HNF4alpha) in cell proliferation and cancer. Gene Expr. 2015; 16(3):101–8.

17. Sharma S, Shen T, Chitranshi N, Gupta V, Basavarajappa D, Sarkar S, et al. Retinoid X Receptor: Cellular and Biochemical Roles of Nuclear Receptor with a Focus on Neuropathological Involvement. Mol Neurobiol. 2022; 59(4):2027–50. https://doi.org/10.1007/s12035-021-02709-y PMID: 35015251

18. Gordon S, Akopyan G, Garban H, Bonavida B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. Oncogene. 2006; 25(8):1125–42. https://doi.org/10.1038/sj.onc.1209080 PMID: 16314846

19. Kaczynski J, Cook T, Urrutia R. Sp1- and Kruppel-like transcription factors. Genome Biol. 2003; 4 (2):206.

20. Cronshaw JM, Krutchinsky AN, Zhang W, Chait BT, Matunis MJ. Proteomic analysis of the mammalian nuclear pore complex. J Cell Biol. 2002; 158(5):915–27. https://doi.org/10.1083/jcb.200206106 PMID: 12196509

21. Hoelz A, Debler EW, Blobel G. The structure of the nuclear pore complex. Annu Rev Biochem. 2011; 80:613–43. https://doi.org/10.1146/annurev-biochem-060109-151030 PMID: 21495847

22. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 2006; 7:29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623 PMID: 16719718

23. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005; 39:309–38. https://doi.org/10.1146/annurev.genet.39.073003.114725 PMID: 16285863

24. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005; 6(5):386–98. https://doi.org/10.1038/nrm1645 PMID: 15956978

25. Lin X, Liu Y, Liu S, Zhu X, Wu L, Zhu Y, et al. Nested epistasis enhancer networks for robust genome regulation. Science. 2022; 377(6610):1077–85. https://doi.org/10.1126/science.abk3512 PMID: 35951677

26. DeLaForest A, Di Furio F, Jing R, Ludwig-Kubinski A, Twaroski K, Urick A, et al. HNF4A Regulates the Formation of Hepatic Progenitor Cells from Human iPSC-Derived Endoderm by Facilitating Efficient Recruitment of RNA Pol II. Genes (Basel). 2018; 10(1). https://doi.org/10.3390/genes10010021 PMID: 30597922

27. Naqvi S, Kim S, Hoskens H, Matthews HS, Spritz RA, Klein OD, et al. Nature Genetics 2023; 55:841–851.