Published online 4 July 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 14* ***6649–6659***
doi: 10.1093/nar/gkw593

Identification of consensus binding sites clarifies FMRP binding determinants

Bart R. Anderson[1,2,†], Pankaj Chopra[2,3,†], Joshua A. Suhl[2], Stephen T. Warren[2,3,4,*] and Gary J. Bassell[1,5,*]

[1]Department of Cell Biology, Emory University School of Medicine, Atlanta, GA, USA, [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA, [3]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA 30322, USA, [4]Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA and [5]Department of Neurology, Emory University School of Medicine, Atlanta, GA 30322, USA

Received January 26, 2016; Revised June 16, 2016; Accepted June 17, 2016

ABSTRACT

Fragile X mental retardation protein (FMRP) is a multifunctional RNA-binding protein with crucial roles in neuronal development and function. Efforts aimed at elucidating how FMRP target mRNAs are selected have produced divergent sets of target mRNA and putative FMRP-bound motifs, and a clear understanding of FMRP's binding determinants has been lacking. To clarify FMRP's binding to its target mRNAs, we produced a shared dataset of FMRP consensus binding sequences (FCBS), which were reproducibly identified in two published FMRP CLIP sequencing datasets. This comparative dataset revealed that of the various sequence and structural motifs that have been proposed to specify FMRP binding, the short sequence motifs TGGA and GAC were corroborated, and a novel TAY motif was identified. In addition, the distribution of the FCBS set demonstrates that FMRP preferentially binds to the coding region of its targets but also revealed binding along 3′ UTRs in a subset of target mRNAs. Beyond probing these putative motifs, the FCBS dataset of reproducibly identified FMRP binding sites is a valuable tool for investigating FMRP targets and function.

INTRODUCTION

Fragile X Mental Retardation Protein (FMRP) is an important regulator of neuronal translation. Absence or dysfunction of FMRP results in fragile X syndrome (FXS). Although numerous studies have been performed attempting to identify the binding determinants for recognition of target mRNAs by FMRP, no consensus has been reached. However, several binding motifs have been proposed. The most well characterized FMRP binding motif was the first identified, the G-quadruplex (1,2). G-quadruplexes are a structural motif composed of spaced GG dinucleotides that assemble into stacked planar tetrads, with the GG separated by loops of variable length and sequence. FMRP binds to G-quadruplexes through its RGG box domain (3). U-rich RNA was the second proposed motif bound by FMRP, identified by two groups using cDNA-SELEX and yeast 3-hybrid approaches (4,5). Using a SELEX approach, a particular class of RNA pseudoknot structure known as the kissing complex was proposed by Darnell *et al.* as a structural motif bound by FMRP's KH2 domain, although no natural FMRP target RNAs containing this motif have yet been identified (6). A unique structural motif composed of three independent stem-loops was bound by FMRP in the *Sod1* mRNA via the RGG box domain and promoted translation of Sod1, in contrast to FMRP's canonical role as a translational repressor (7). On the basis of CLIP sequencing, Ascano *et al.* identified two short sequences enriched in FMRP-bound RNAs from HEK293 cells: WGGA and ACUK (W = A or T/U, K = G or T/U) (8). Most recently, Ray *et al.* used a technique called RNAcompete to find linear RNA sequences bound by hundreds of RNA binding domains and identified the sequence GAC as bound by the KH domains of FMRP, dFmr and the FMRP paralogs FXR1 and FXR2 (9). Additionally, FMRP has been shown to be bound all along the coding sequence (CDS) in an apparent sequence independent manner (10), which raises key questions on how FMRP binds selectively to a small subset of mRNAs, estimated at 4% (11).

[*]To whom correspondence should be addressed. Gary J. Bassell. Tel: +1 404 727 3772; Fax: +1 404 727 6256; Email: gbassel@emory.edu
Correspondence may also be addressed to Stephen T. Warren. Tel: +1 404 727-5979; Fax: +1 404 727 3949; Email: swarren@emory.edu

[†]These authors contributed equally to the paper as first authors.

Present addresses:
Bart R. Anderson, Exicure, Skokie, IL, USA.
Joshua A. Suhl, Booz Allen Hamilton, Westborough, MA, USA.

Our recent bioinformatics analysis examined proposed FMRP-bound motifs in consensus FMRP target datasets and found that the WGGA sequence, WGGA-clusters consistent with G-quadruplexes and GAC motifs are all enriched in FMRP target genes (12). This analysis assessed FMRP targets at the level of FMRP-bound target genes. To further build on this data, we pursued analysis of FMRP binding at the level of individual binding sites within each target mRNA. Seeking to clarify the binding determinant for FMRP, we compared the binding sites identified in two separate large-scale CLIP studies of FMRP binding (8,10) to identify FMRP consensus binding sequences (FCBS). This shared dataset contains 34 218 binding sites within 3703 genes. We then characterized these FMRP binding sites including their location, sequence and structural motifs, and potential interactions with RNAi, methylation and translation.

MATERIALS AND METHODS

GeneUniverse whole genome meta-gene dataset

Ensembl was used to construct the whole genome dataset. This dataset was limited to protein-coding genes containing both 5′ UTR and 3′ UTR, and for each gene a meta-gene was generated containing all protein-coding exons as well as the 5′ UTR and 3′ UTR. This resulted in a set of 18 325 genes, referred to herein as the GeneUniverse.

Identifying FMRP consensus binding sequences (FCBS)

Sequencing of FMRP-bound mRNAs has been published by two groups, each with two variants. Darnell *et al.* (10) performed HITS-CLIP of endogenous FMRP in mouse brain lysates and polysomes. Ascano *et al.* (8) performed PAR-CLIP from HEK293 cells transfected with isoform 1 and isoform 7 of human FMRP. For each of these studies we obtained the complete set of mapped CLIP reads and combined replicates of both variants to generate a single set of all CLIP reads generated by each group. The genomic coordinates of the Darnell tags were converted from mouse mm9 to human hg19 coordinates using the LiftOver utility available at the UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgLiftOver) (13). This set of human coordinates was then restricted to tags located within exons of protein-coding genes in the GeneUniverse described above, including 5′ UTR, CDS and 3′UTR exons. The length of each tag was normalized by finding the midpoint and keeping 50 nt of exonic sequence on each side, resulting in a set of 101 nt Darnell-derived exonic sequences. The PAR-CLIP protocol identifies a binding peak within each tag and this peak was used as the centering point for the Ascano tags, which were similarly normalized to 101 nt. The two normalized sets were compared and overlapping sites identified. Because multiple tags from one dataset could overlap with a single tag in the other dataset, it was necessary to select a reference set. We used the Ascano tags as the reference set because this dataset contained PAR-CLIP peaks which provided an exact base position of FMRP crosslinking. Although a method to identify crosslinking sites in HITS-CLIP has been published (14), this method relies on a minor fraction (8–20%) of sequence reads which contain reverse transcription-induced mutations and were discarded in original analysis, and therefore these sequence reads were not available for use in our analysis. All of the Ascano-derived tags that were overlapped by at least one Darnell-derived tag were kept as FMRP consensus binding sequences (FCBS). Therefore, each sequence of FCBS is centered on an Ascano PAR-CLIP peak with 50 nt of exonic sequence on either side. There were a total of 34 218 such sites from 3703 genes. The full FCBS set is available as Supplementary Table S1.

We assessed the significance of the overlap in two ways. First, we performed a chi-square test to determine if the number of genes that were common (N = 5881) in the Ascano (N = 9103) and Darnell (N = 8249) datasets was significant. The overlap was highly significant ($P < 2.2 \times 10^{-16}$). Next, we determined if the number of genes having CLIP tags in close proximity in both the Ascano and Darnell datasets was significant. We performed permutation tests with varying window sizes to assess whether the tags from the two datasets were closer to each other than would be expected by chance. We found these to be highly significant (P-values ≤ 0.0001 for window sizes of 50, 100, 250, 500 or 1000 nt). Additional details of permutations given in Supplementary Data. For further analyses, we chose a conservative window size of 101 nt.

Generating background datasets

For permutation analyses, negative sets were generated from mRNA sequences in the GeneUniverse dataset described above. Randomly selected 101nt sequences were normalized to match the FCBS distribution of 5′UTR, CDS and 3′UTR, using the 1st exon and last exon as proxies for 5′UTR and 3′UTR.

For sequence enrichment analyses, the background dataset was a Markov model order 2 of expected oligo frequencies derived from all mRNA sequences in the GeneUniverse. This was generated using the create-background-model tool available from the Regulatory Sequence Analysis Tools (RSAT) suite (http://rsat.ulb.ac.be/rsat/) (15,16).

Assessing kissing complexes

From the complete set of FCBS sites (N = 34 218), the McGenus algorithm (17) was used to predict pseudoknot formation and predicted folds were filtered to include kissing complex pseudoknots. For the permutation, an equal number of random sites were selected from all mRNA sequences in the GeneUniverse, controlling for the distribution within the 5′UTR, CDS and 3′UTR regions. A total of 5000 permutations were done. The permuted P-value is the fraction of times the number of sites predicted to form kissing complexes in the mRNA permuted sets exceeded the number of sites predicted to form kissing complexes in FCBS.

Assessing G-quadruplexes

By using a 101 nt region around each PAR peak rather than restricting to the sequenced read itself, the FCBS re-captures any G-rich sequences that may have been depleted due to the use of RNase T1 in the Ascano PAR-CLIP protocol. Within each FCBS, we looked for two different patterns that could be considered potential G-quadruplex forming motifs (QFM). We searched for QFM with loops of 2–7 nt because although examples of longer loop sequences have been described, published literature suggests that 7 nt is a common maximum number of loop nucleotides (18,19). We chose 2 as the minimum number of loop nucleotides because our initial studies were focused on the putative FMRP binding motif 'WGGA', which in a series of tandem motifs that could form a tetrad G-quadruplex would be 'WGGAWGGAWGGAWGGA'. The underlined 'AW' forced the analysis to contain no <2 loop nucleotides between the double G residues. Additionally, while loops of just one nucleotide have been reported, they appear to be uncommon. Generic QFM were defined as sequences matching the pattern GGN{2,7}GGN{2,7}GGN{2,7}GG, where 'N{2,7}' indicates any sequence 2–7 nt in length. WGGA-QFM were defined as sequences matching the pattern WGGAN{0,5}WGGAN{0,5}WGGAN{0,5}WGGA. The same approach was used to assess the QFM and WGGA-QFM in FCBS + 101 nt downstream sequence. For comparison, a set of random human mRNA sequences was selected, matching the number and length of the FCBS sequences, and the number of these sequences containing QFM and WGGA-QFM was determined. This process was repeated for 50 000 permutations to generate a permutation *P*-value.

Assessing sequence motifs

The percentage of tags containing a given motif and frequency of each in the FCBS set were calculated using the dna-pattern and create-background utilities within the Regulatory Sequence Analysis Tools (RSAT) suite (http://rsat.ulb.ac.be/rsat/) (15,16). To assess whether specific patterns (WGGA, TGGA, etc.) are enriched we performed a permutation analysis. In each permutation, for each of the 34 218 clip tags within the FCBS set, we selected a corresponding random peak position, taking care to select the random position in the same transcript and the same region (i.e. 5′UTR, CDS or 3′UTR) as the FCBS clip tag, to have counts for each motif. We conducted 1000 such permutations. The *P*-value was calculated as the number of times a sequence motif count in the set of random permutations exceeded the count in FCBS divided by the total number of permutations ($N = 1000$).

Unbiased motif enrichment and position enrichment were assessed using MEME-ChIP version 4.9.1 (20). The DREME tool assesses motif enrichment and the CentriMo tool assesses positional enrichment.

Assessing codon usage

To assess whether particular Amino Acids were significantly enriched in FCBS, a permutation analysis was conducted. In each of the 1000 permutations, for each of the 34 218 clip tags a random position was selected in the same transcript, and the number and type of in-frame AA codons within 20 bp of this random position was recorded. The enrichment *P*-value for any AA codon was calculated as the number of times the permuted counts for that AA codon exceeded the counts in the FCBS sequence divided by the total number of permutations ($N = 1000$).

Distance from FMRP sites to miRNA sites

For the 100 miRNAs most highly expressed in HEK293 cells (21), the locations of miRNA seed sites was identified in each FMRP binding site. Analysis of miRNA seed sites included all FCBS sites, without restricting solely to 3′UTR sites. The minimum distance between a miRNA seed site and the FMRP binding site PAR peak was calculated. The distance distribution was compared to distance when the same calculations were performed for a negative control CLIP dataset (c22orf28 (22)).

RESULTS

Identifying consensus FMRP binding sites

To investigate the determinants of FMRP binding we took advantage of two published large-scale CLIP sequencing studies of FMRP target sites. Darnell *et al.* performed HITS-CLIP of endogenous mouse FMRP from brain polysomes in seven CLIP replicates using two protocol variations (10). Ascano *et al.* performed PAR-CLIP of tagged human FMRP transfected into HEK293 cells, using isoform 1 and isoform 7 in independent CLIP experiments (8). Although differences in cell type, species and protocol will inevitably lead to differences even in legitimate FMRP binding locations, we reasoned that locations bound by FMRP in both these datasets represent high-confidence FMRP binding sites. Therefore, we compared the CLIP tags generated by these two groups and identified the locations that were bound by FMRP in both datasets (see Figure 1, Tables 1 and 2, and Supplementary Figure S1). After converting the Darnell CLIP tags to human coordinates, restricting both datasets to exonic locations in protein-coding genes and normalizing all tags to a standard 101 nt length, we found 34 218 sites that were bound by FMRP in both studies, hereafter referred to as FCBS. The full FCBS set is available as Supplemental Table S1. These 34 218 sites represent ∼30% of the exonic tags from each parental dataset. There are 11 471 genes containing a CLIP tag in the Ascano and Darnell datasets. Of these, there are 5881 genes that contain a CLIP tag in both the Ascano and in the Darnell datasets. A chi-square test shows that this overlap is highly significant ($P < 2.2 \times 10^{-16}$). The FCBS identifies specific sites within those genes that were bound in both datasets. With an overlap window size of 101 nt, such sites were located in 3703 genes, representing 41–45% of the genes bound in the parental datasets. Both parental datasets exhibited bias toward the CDS, which was further accentuated in the combined dataset, with ∼73% of FCBS located in CDSs.

Presence and frequency of previously proposed sequence motifs

The fraction of FCBS sequences that contain each previously proposed FMRP binding sequence was assessed and

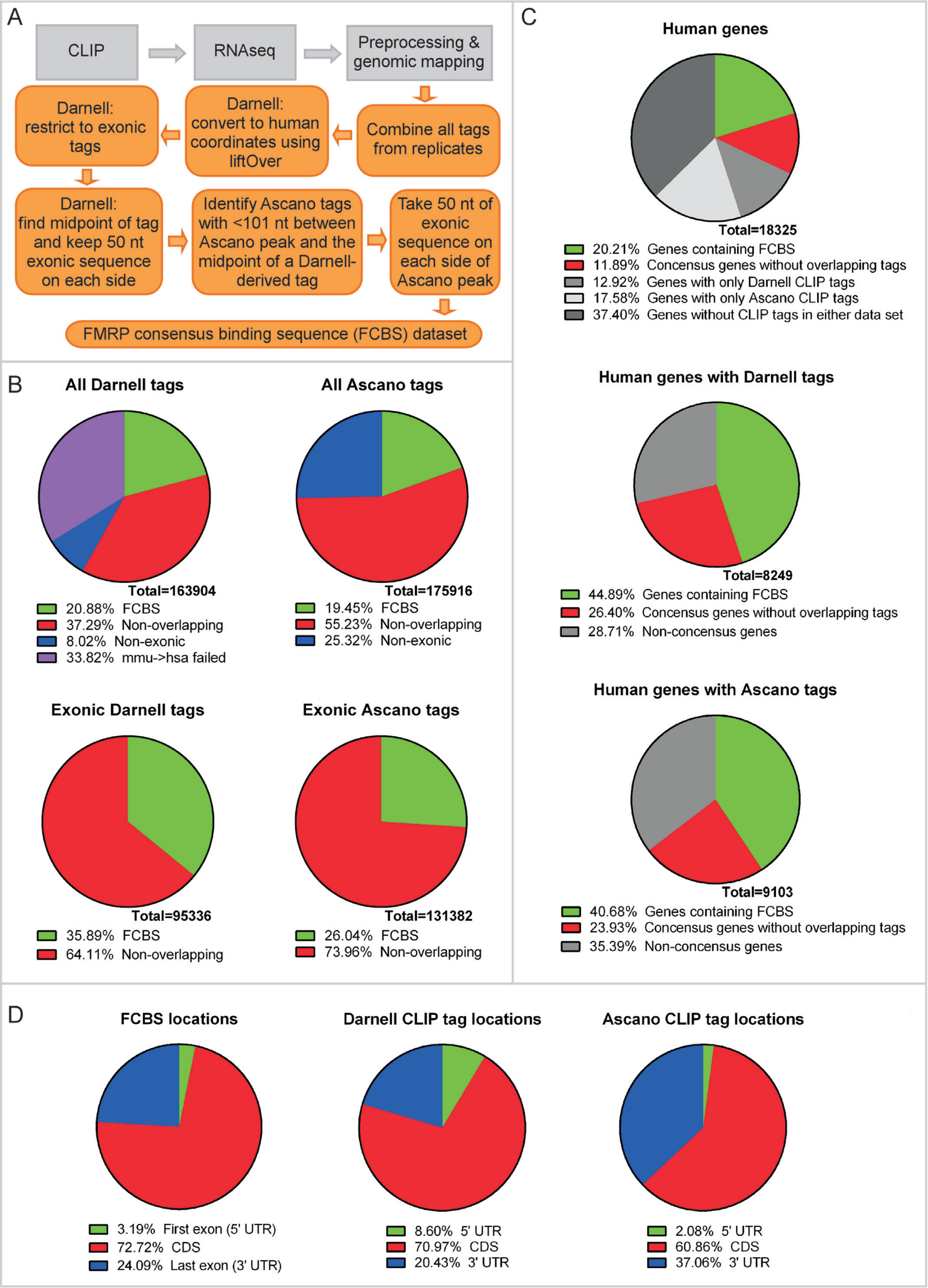


Figure 1. Identification of FMRP consensus binding sequences (FCBS). (**A**) Workflow showing processing steps (orange) to generate FCBS set from parental CLIP datasets (gray). (**B**) Overlap of FMRP-bound CLIP tags from Darnell *et al*. and Ascano *et al*. datasets. (**C**) Genes containing sites bound in Darnell *et al*. and Ascano *et al*. datasets. (**D**) Location distribution of FMRP-bound sites in parental and FCBS sets.

Table 1. FMRP-bound CLIP tags through processing steps to generate FCBS set

	CLIP tags	
	Darnell	Ascano
Total tags	163 904	175 916
Human coordinates	108 478	175 916
Exonic	95 336	131 382
FCBS	>34 218	34 218

Table 2. Genes containing FMRP-bound CLIP tags through processing steps to generate FCBS set

	Genes with CLIP tags	
	Darnell	Ascano
Human coordinates	8249	9103
FCBS	3703	3703

compared to random 101 nt human mRNA sequences (Figure 2A). As any given motif could potentially occur multiple times within one FCBS sequence, the frequency of occurrence for each motif in the FCBS set was also calculated and compared to the frequency of the motifs in random mRNA sequences (Figure 2B). By both measures there was an increase in TGGA, ACTK and GAC motifs and a decrease in U-rich (TTTT) motifs.

Unbiased motif discovery

The presence and frequency of sequence motifs are informative descriptive measures but do not provide statistical information about enrichment. We therefore performed unbiased motif discovery using the MEME suite (20). The DREME tool identifies enriched motifs, and TGGA was the top motif identified (Figure 3). AGGA and TGGT motifs were also among the top 10 motifs identified and an additional 9 motifs also contained GG, for a total of 12 out of 26 motifs identified (see Supplementary Figure S2). The presence of TGGA and AGGA supports the WGGA motif proposed by Ascano *et al*. Furthermore, WGGA and the other GG sequences could contribute to the formation of G-quadruplex structures. Three other motifs among the top ten contained TAY (Y = C or T), a motif that has not been previously indicated as a FMRP-bound sequence. The sequence GAC was identified as part of a larger motif but it was not enriched as an independent motif, although this may be due to the high frequency of this trinucleotide sequence in both the FCBS and negative datasets (see Figure 2). No ACUK-containing motifs were identified.

Position bias of sequence motifs

The CentriMo tool within the MEME suite identifies not whether the abundance of a motif is enriched but rather whether the position of a motif is enriched. This tool was used to look for local enrichment of the motifs identified by DREME as well as a database of RNA motifs bound by RNA-binding domains (RBD) (9). The motifs with the most significant bias in position were TGGA motifs identified by DREME as enriched in abundance (Figure 4A). The remaining of the top ten motifs with significant positional bias were from the Ray *et al*. database of RBD-bound motifs. Notably, four of these ten were GAC motifs that were identified by Ray *et al*. as motifs bound by FMRP and its two paralogs FXR1P and FXR2P. Additionally, two ATG motifs and an AGAGA motif were among the most positionally biased motifs. There were three major patterns of position bias (Figure 4B). Central bias was most strongly exhibited by GAC motifs and was also displayed by ATG motifs. The TAY motifs identified by DREME were similar to a central bias but offset slightly to the 5′ side of center with a depletion on the 3′ side of center. TGGA and related motifs also exhibited a central bias but were bimodal. This bimodal distribution would be consistent with the repeated GG-containing motifs composing a G-quadruplex structure. The third pattern of position bias exhibited was central depletion, which was seen with AGAGA motifs. Although not among the most significantly biased in position, when U-rich sequences were identified by DREME or CentriMo they were also depleted in the center of the FCBS sequence. No ACUK motifs were identified as having significant position enrichment.

Kissing complex motifs

The kissing complex is a form of RNA pseudoknot and RNAs forming kissing complexes were found to bind to a KH domain-containing truncation of FMRP with high affinity (6). We therefore asked whether FCBS sequences are enriched for putative kissing complexes. We used the McGenus algorithm (17) to predict pseudoknot folds that could be formed by each FCBS sequence and then assessed those folds for kissing complex structures similar to those previously found to bind FMRP. We asked whether predicted kissing complexes were enriched in FCBS compared to randomly selected mRNA sequences. This analysis yielded a permuted *P*-value of 0.1426, indicating no significant enrichment of predicted kissing complexes.

With the intent to assess whether they were pulled out in FMRP CLIP experiments, Pubmed was searched for established kissing complexes. However, literature searches did not produce any confirmed kissing complexes in mammalian mRNA.

G-quadruplex motifs

The G-quadruplex structure is the best characterized of the proposed FMRP-bound motifs. Several specific examples have been validated (1,2,23–25). Furthermore, predicted WGGA-containing G-quadruplexes are enriched in consensus FMRP target genes (12). Therefore, we assessed the potential of FCBS to form G-quadruplexes. Within each FCBS, we searched for a generic QFM and compared this to the number of QFM found in randomly selected mRNA sequences of the same length. Because preliminary analyses using several variations of QFM criteria produced similar findings (data not shown), in this analysis we searched on QFM containing 2–7 nt loops because these criteria capture the commonly accepted G-quadruplex formation requirements (18,19), although single nucleotide loops and loops longer than 7 nt have also been characterized (26)

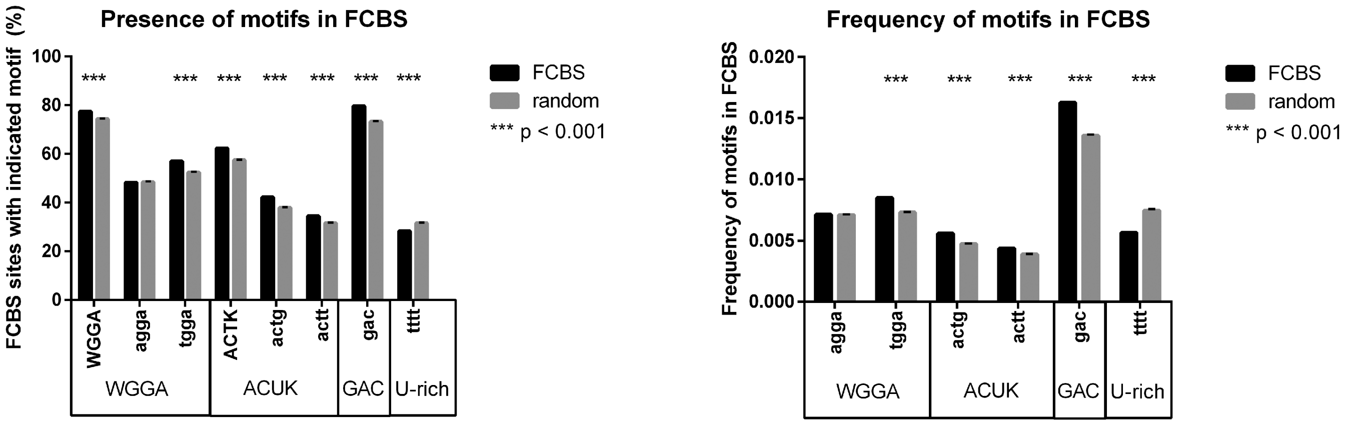


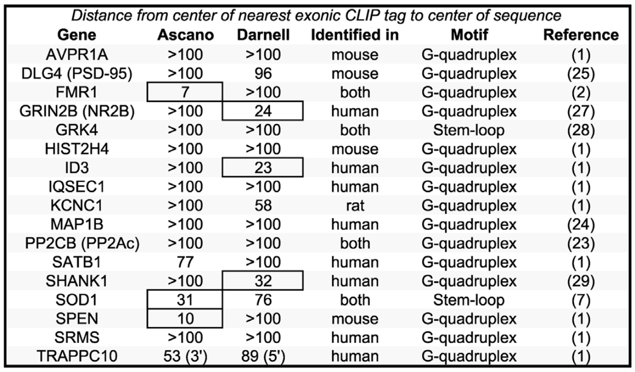
Figure 2. Previously described sequence motifs in FCBS set. Presence of sequence motifs (**A**) and frequency of motif occurrence (**B**) in FCBS set and in 1000 permutations of random 101 nt mRNA sequences.

which would not be captured in our analysis. This analysis found QFM in 12% of FCBS with no statistical enrichment over random mRNA sequences. Because WGGA-containing QFM are enriched in FMRP target genes (12), we also looked for WGGA-QFM in FCBS and again found no statistical enrichment over random mRNA sequences. Considering the possibility that G-quadruplexes were immediately adjacent to FMRP binding sites, rather than directly bound by FMRP, we looked for G-quadruplexes in the FMRP tags plus an additional 101 nt of downstream sequence. The permuted *P*-values of these analyses were 0.9999 for QFM and 0.9045 for WGGA-QFM, indicating that QFM and WGGA-QFM are in fact depleted in FCBS rather than enriched. This suggests that although QFM may be bound by FMRP and is enriched in genes bound by FMRP, this motif alone is not sufficient for FMRP binding *in vivo*, leaving the role of QFM as a binding determinant unresolved.

Examining previously identified FMRP binding sites in FMRP CLIP datasets

A literature search revealed 17 FMRP target mRNAs for which the FMRP-bound sequence has been identified (1,2,7,23–25,27–29). Fifteen of these sites are putative or confirmed G-quadruplexes. The remaining two each form unique stem-loop structures. We looked for evidence of FMRP binding to these sites in the two published FMRP CLIP datasets (8,10). Assuming the footprint of a typical RNA-binding protein is ∼40 nt, we asked whether the center of each site is within 50 nt of the center of any FMRP CLIP tag. To our surprise, of these 17 sites, only six were centered within 50 nt of an FMRP CLIP sequencing tag from either dataset, and none were within 50 nt of a sequencing tag from both sets (Table 3), highlighting the need for a dataset of reproducible FMRP binding sites.

Table 3. Previously identified FMRP-bound sequences in FMRP CLIP datasets



	Distance from center of nearest exonic CLIP tag to center of sequence				
Gene	**Ascano**	**Darnell**	**Identified in**	**Motif**	**Reference**
AVPR1A	>100	>100	mouse	G-quadruplex	(1)
DLG4 (PSD-95)	>100	96	mouse	G-quadruplex	(25)
FMR1	7	>100	both	G-quadruplex	(2)
GRIN2B (NR2B)	>100	24	human	G-quadruplex	(27)
GRK4	>100	>100	both	Stem-loop	(28)
HIST2H4	>100	>100	mouse	G-quadruplex	(1)
ID3	>100	23	human	G-quadruplex	(1)
IQSEC1	>100	>100	human	G-quadruplex	(1)
KCNC1	>100	58	rat	G-quadruplex	(1)
MAP1B	>100	>100	human	G-quadruplex	(24)
PP2CB (PP2Ac)	>100	>100	both	G-quadruplex	(23)
SATB1	77	>100	human	G-quadruplex	(1)
SHANK1	>100	32	human	G-quadruplex	(29)
SOD1	31	76	both	Stem-loop	(7)
SPEN	10	>100	mouse	G-quadruplex	(1)
SRMS	>100	>100	human	G-quadruplex	(1)
TRAPPC10	53 (3')	89 (5')	human	G-quadruplex	(1)

Nucleotide distance from center of nearest exonic CLIP tag to center of previously identified FMRP-bound sequences. Boxes highlight indicate sequences centered within 50 nt of FMRP CLIP tags.

Position of FMRP binding sites along mRNA

We examined where along each mRNA the FCBS sites are located (Figure 5). This analysis revealed that the majority of FCBS are in the coding region of mRNA. The FCBS in the coding region display a slight trend toward increased FMRP binding moving from the 5′ toward the 3′ end of the CDS. Those FCBS located in the 5′UTR are biased toward the 3′ end, near the coding region. Similarly, those FCBS located in the 3′UTR are biased toward the 5′ end, near the coding region. Overall, this data indicates that FMRP binds target mRNAs predominantly in the coding region but also suggest that binding sites along the 3′UTR may contribute to FMRP mediated regulation.

Some studies suggest that FMRP interacts directly with the ribosome (30,31). If FMRP's RNA association is dominated by interactions with the ribosome rather than by direct interaction with mRNAs, then FMRP binding sites should be enriched in the CDS and the binding sites that fall in the 3'UTR would be expected to cluster near the CDS-proximal end of the 3'UTR due to association with ribosomes that have not yet fallen off the mRNA. Although our data demonstrating a bias for FMRP binding predominantly within or near the CDS is consistent with a ribosome-association model of FMRP binding, it can neither provide direct support nor disprove this model.

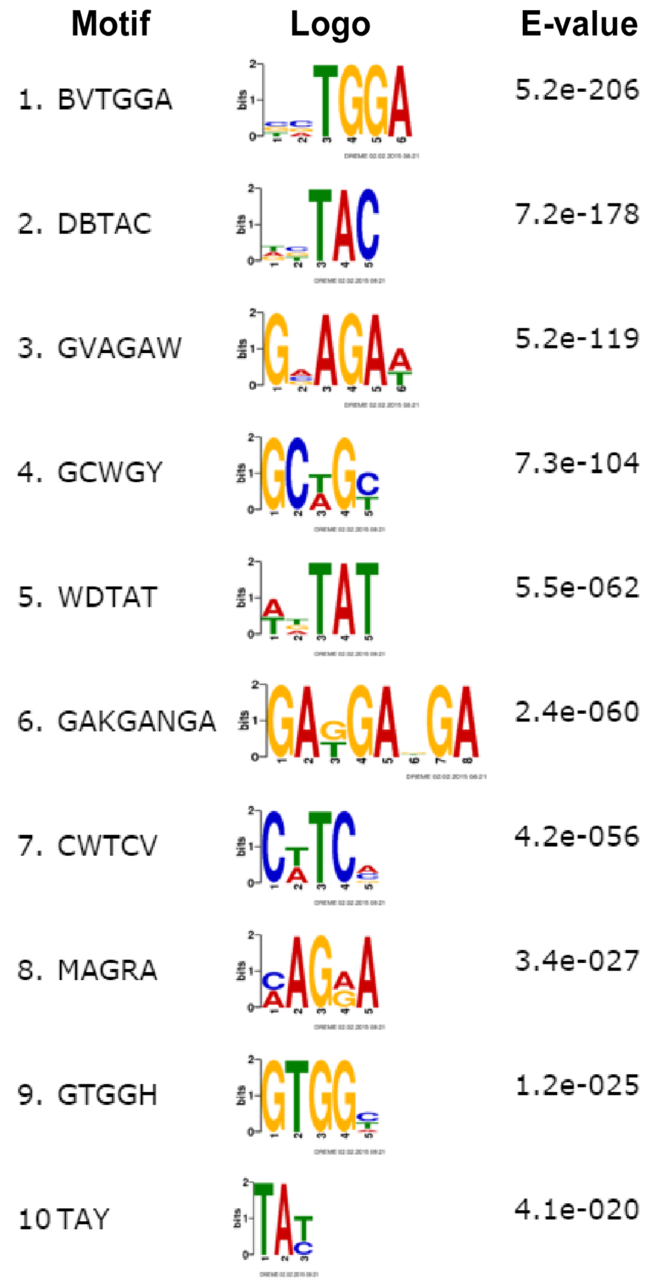


Figure 3. Ten most significantly enriched motifs in FCBS set as identified by DREME tool.

miRNA analysis

Because FMRP has been previously proposed to function cooperatively with miRNA (32–34), we looked for evidence of interaction with miRNAs. First, we looked at miRNA seed sites for the reverse complement of motifs enriched in FMRP overlap tags: GAC, TAY or TGGA. Of 2588 known human miRNAs, 388 (15%) have one of these in their seed. If analysis is restricted to the 100 miRNAs most highly expressed in HEK293 cells (the cell type from which the Ascano dataset was generated), 24 of the top 100 miRNAs have a FMRP-bound motif reverse complement in their seed. After considering that the majority of these miRNAs were closely related from several miRNA families (374a/b, 30a/b/c/d/e, let-7a/b/c/d/e/f/g/i and 196a/b), only nine distinct miRNAs are represented, which does not suggest that these FMRP binding sites represent and enrichment of miRNA binding sites.

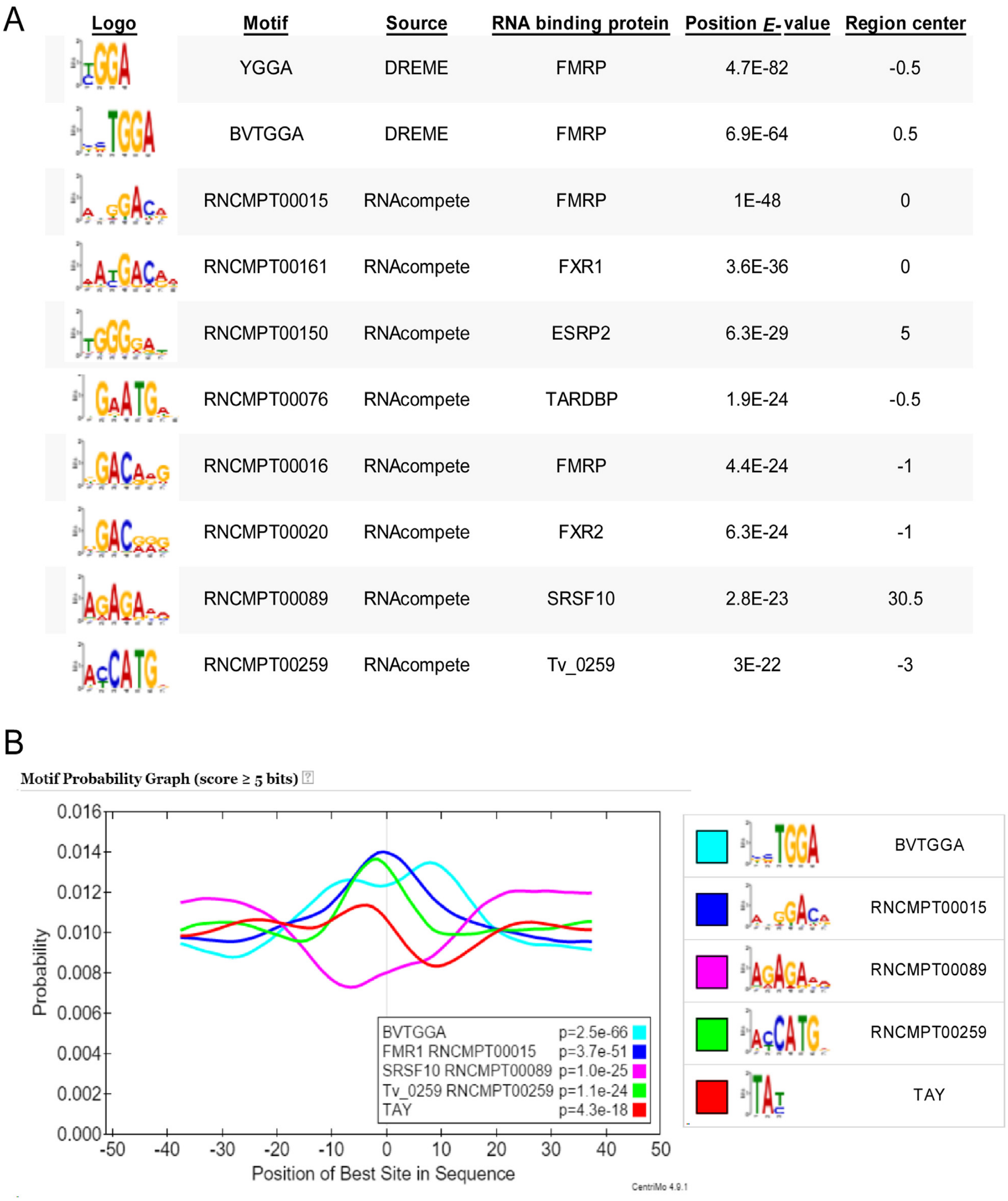
We reasoned that even if the FCBS motifs don't correspond with miRNA seed sites, the FCBS may be enriched near miRNA binding sites. Therefore, we assessed the distance from FCBS sites to miRNA seed sites. For comparison, the same calculations were performed using CLIP tags from c22orf28, a protein with no known or proposed interaction with RNAi. As seen in Figure 6, the FCBS were no closer to miRNA sites than c22orf28 CLIP tags, indicating that FCBS are not enriched near miRNA sites.

Codons

Because the motifs identified in FCBS are short 3–4 nt sequences, we asked whether these sequences correspond to codons for particular amino acids. As shown in Table 4, the five possible codons represented by FMRP-bound motifs encode four amino acids (Asp, Tyr, Trp and Gly). The anticodon tRNAs are not unusually abundant nor rare. It is notable, however, that these FMRP-bound sequences account for all of the Asp, Tyr and Trp tRNAs. We questioned whether these FMRP-bound sequences correspond to amino acids pause sites, but observed no relationship to known pause sites (data not shown). We also found it notable that the TGG codon for Trp is the least abundant of all vertebrate codons and that the TAY codons for Tyr are also low-abundance; in contrast, the GAC codon for Asp is of average abundance. Because these possible codon-binding observations are only relevant if the codons are in-frame, we assessed the in-frame amino acids within the FCBS (Figure 7). For most amino acids, the frequency was only slightly altered in FCBS as compared to random positions within the same transcripts. For three amino acids, there was >15% difference in frequency, each with P-value < 0.001. These were enrichment of Asp (D), which includes the GAC codon, enrichment of Trp (W), encoded by TGG and depletion of Lys (K), which includes the AAA codon. Notably, the enrichment of these in-frame codons correspond to FCBS-enriched motifs GAC and TGGA. Conversely, in-frame Lys/K codons are depleted in FCBS, consistent with the depletion of the AAA in the center of FCBS.

Similarity of FCBS motifs to m6A methylation sites

We also noted that the FMRP-bound GAC motif is similar to the sequence motif that signals m6A methylation. The preferred 5 nt consensus sequence for m6A methylation is RRACH (where H = A,C,U. R = A/G) (35), which therefore includes a central GAC in half of the possible RRAC sequences. We also noted that the binding motifs identified by Ray *et al.* for all Fmr family proteins include RRACH. We therefore asked whether this m6A motif signal is enriched in FCBS. The frequency of both GGAC and AGAC is increased in FCBS (5.3 and 4.4 occurrences per 1000 nt, respectively) as compared to all human mRNA sequences (each at 3.9 occurrences per 1000 nt). Additionally, the RRACH motif is found in 57% of the FCBS, as com-



A

Logo	Motif	Source	RNA binding protein	Position *E*-value	Region center
	YGGA	DREME	FMRP	4.7E-82	-0.5
	BVTGGA	DREME	FMRP	6.9E-64	0.5
	RNCMPT00015	RNAcompete	FMRP	1E-48	0
	RNCMPT00161	RNAcompete	FXR1	3.6E-36	0
	RNCMPT00150	RNAcompete	ESRP2	6.3E-29	5
	RNCMPT00076	RNAcompete	TARDBP	1.9E-24	-0.5
	RNCMPT00016	RNAcompete	FMRP	4.4E-24	-1
	RNCMPT00020	RNAcompete	FXR2	6.3E-24	-1
	RNCMPT00089	RNAcompete	SRSF10	2.8E-23	30.5
	RNCMPT00259	RNAcompete	Tv_0259	3E-22	-3

Figure 4. Motifs with significant position bias in FCBS. (**A**) Ten motifs with the most significant bias in location in FCBS set as identified by CentriMo tool. For motifs originating from RNAcompete (9), the RNA binding protein which binds that motif is indicated. (**B**) Common patterns of position bias within FCBS set as identified by CentriMo tool.

Table 4. Codons corresponding to FCBS-enriched sequence motifs.

Codon	Amino acid	FCBS-enriched sequences	AA category
GAC	Asp / D	one of two codons	acidic
UAU	Tyr / Y	both codons	aromatic
UAC	Tyr / Y	both codons	aromatic
TGG	Trp / W	only codon	aromatic
GGA	Gly / G	one of four codons	small

Notes:
Phe/F is the only aromatic amino acid not encoded by FCBS-enriched sequence motifs. Phe/F codons are UUU and UUC.
TGG (Trp/W) is the least abundant of all vertebrate codons. TAT and TAC (Tyr/Y) are also low-abundance codons.

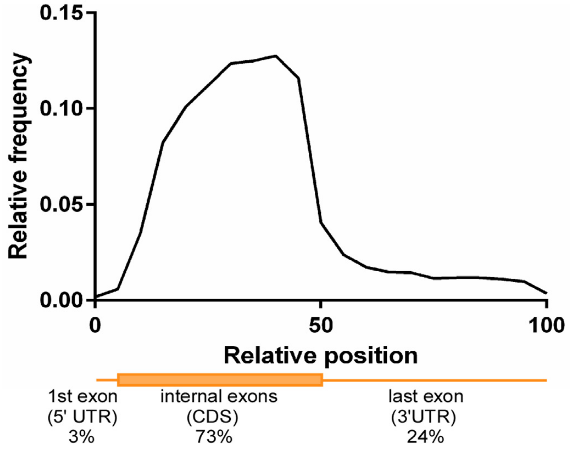


Figure 5. Normalized distribution of FCBS sites along all bound mRNAs, with the percentage of FCBS sites located in each region indicated.

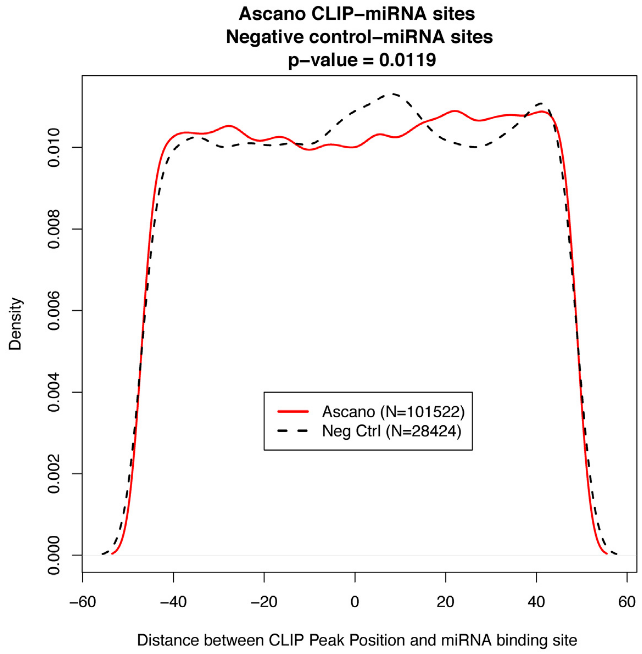


Figure 6. FMRP CLIP tags are not enriched near miRNA seed sites. Using the Ascano *et al.* FMRP CLIP dataset and a negative control CLIP dataset (c22orf28 (22)), the distance was calculated from each PAR peak to the nearest miRNA seed site for the 100 most highly expressed miRNAs in HEK cells.

pared to 49% of randomly generated sequences of the same length.

DISCUSSION

Following numerous reports there remains a lack of clear understanding of what specifies a mRNA or sequence to be bound by FMRP. Although several motifs have been proposed and many sites identified, as well as the proposal of a sequence independent binding mechanism, a lack of reproducibility has hampered progress toward a consensus. A set of high-confidence FMRP-bound sequences is needed to facilitate future studies of FMRP binding and function. Here, we generated a set of FMRP consensus binding sequences consisting of the exonic sequences bound by FMRP from two separate high-throughput sequencing studies. This set consists of 34 218 sites in 3703 genes, representing ∼20% of the sites in each of the two parental datasets. Due to differences in cell type (HEK293 versus whole brain), species (human versus mouse) and methodology (PAR-CLIP versus HITS-CLIP) there will be many legitimate FMRP-bound sequences that are not included in the FCBS set. Our goal was not to create a library of every site that FMRP binds, but rather to create a conservative set of reproducibly bound sites, which will be useful in identifying common features in FMRP-bound sites.

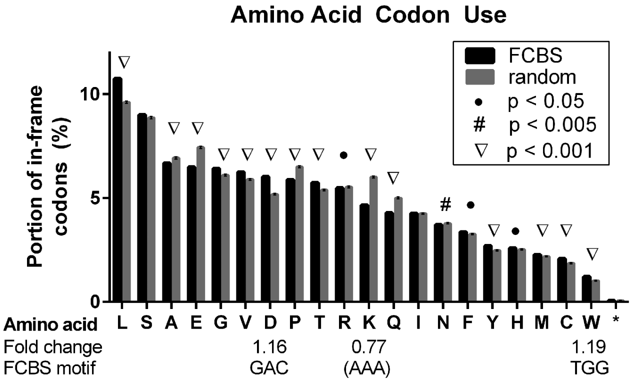


Figure 7. Codon usage in FCBS. For each FCBS site or in 1000 permutations of random positions within the same transcripts, the in-frame codon usage was calculated. Bar graph indicates percent of all in-frame codons. * indicates stop codons. For amino acids with ≥15% difference between FCBS and random positions, the fold change is indicated as well as corresponding motifs enriched or depleted (depletion indicated by parenthesis) in FCBS.

Using this set of FCBS sites, we examined previously proposed FMRP binding sequence motifs. Enriched presence and frequency in FCBS lends support for FMRP binding to TGGA, ACUK and GAC sequences, whereas poly-U was found to be depleted (Figure 2). When using unbiased motif discovery, we again found support for TGGA and GAC motifs, which were enriched and centered in FCBS, whereas ACUK and poly-U sequences were not enriched in FCBS (see Figures 3 and 4, and Supplementary Figure S2). Previous work demonstrated WGGA and GAC enrichment within consensus FMRP targets at the level of the genes (12). Here we extend those findings to the level of consensus-bound target mRNA sequences and demonstrate both enrichment and position bias within the FMRP-bound sequences. During this analysis the sequence TAY was also found to be enriched and centered, adding another trinucleotide sequence motif as a contributing feature in FMRP binding. As FMRP has been demonstrated to interact and regulate translation of specific transcripts cooperatively with RISC (32–34), we assessed miRNA target sites within FCBS but found no evidence for enrichment (Figure 6); the marginal *P*-value is likely due to the large sample size rather than actual enrichment. We similarly examined previously proposed FMRP structural motifs in the FCBS, and found no enrichment of kissing complex nor G-quadruplex motifs. The lack of enriched G-quadruplexes in particular was surprising as there are well-documented examples of FMRP interaction with G-quadruplexes in target mRNAs (1,2,23–25,27). Our data does not indicate that FMRP does not interact with G-quadruplexes in those mRNAs. Rather, the FCBS set indicates that although FMRP may interact with some target mRNAs via G-quadruplexes, they are not a feature in the majority of FMRP-bound sequences and therefore are not the primary feature by which FMRP targets are bound *in vivo*. In addition, it has been suggested that FMRP binds to G-quadruplexes made up of RNA dimers (1,3) or at the junction of G-quadruplexes and RNA duplexes (36), whereas our analysis here examined only intramolecular G-quadruplexes composed of a single linear sequence at the FMRP binding site. Similarly, G-quadruplexes with single nt loops or loops longer than 7 nt would not be captured in our QFM search. Therefore, FMRP-bound G-quadruplexes existing as part of a more complicated structure would not have been identified in our analysis.

The majority of the FCBS sites are located within the coding region of mRNAs. Although both parental datasets showed enrichment within the coding regions of mRNAs, it was possible that the sites common between sets would display a different pattern of enrichment and so we examined the distribution of FCBS sites. Here we found that not only were the majority of FCBS sites located in the coding region, but that this enrichment was larger than in either of the parental sets. The strong bias of FCBS toward the CDS suggests that FMRP functions primarily through interaction with the coding region. It has been proposed that FMRP interacts with ribosomal proteins or with ribosomal RNA (30,31). Although the CDS-biased distribution of FCBS is consistent with a ribosome-interaction model, the FCBS dataset cannot provide insight as to whether FMRP interacts directly with the ribosome, nor distinguish between interactions with ribosomal proteins versus RNA. The presence of UTR FCBS sites (27% of tags) is consistent with published reports of FMRP regulating specific transcripts via binding to the 5′ UTR or 3′ UTR (1,7,23–25,27). Of those FCBS found in the 5′ and 3′ UTR, the number of FCBS is highest near the coding region and decreases moving distally from the CDS (Figure 5). Further work is needed to assess the role of 3′UTR tags, which could play an important role in selectivity of binding.

The 3–4 nt motifs identified in FCBS are insufficiently long to provide specificity for FMRP binding. It is unknown whether the remaining specificity is provided by interaction with RNA structural features, protein–protein interactions, or other factors. For example, we observed a striking similarity between the enriched FCBS motif GAC and the m6A methylation signal sequence RRACH (35,37), and noted that RGACH is located in 57% of the FCBS. Additionally we noted that the FCBS motifs GAC and TGG, when used as codons encode Asp and Trp, which are enriched in FCBS relative to their use in random sequences from the same transcripts (Figure 7). Further studies should be performed to investigate the contribution of these factors to FMRP binding.

Here, we generated the shared FCBS dataset and used it to investigate the reproducibility of previously proposed FMRP binding motifs. Going forward, the FCBS dataset will be invaluable in further investigation of these and other possible determinants of FMRP binding. While the FCBS data reveals the significance of mRNA sequence motifs, including WGGA, GAC and TAY binding motifs, more work is clearly needed to understand additional factors that provide FMRP with selectivity in binding its cohort of mRNAs. One important question is whether selectivity in mRNA binding requires a minimal threshold for the number and/or density of FMRP binding sites per target mRNA molecule. Additionally, more work is needed to understand the potential interplay between 3′UTR and CDS binding sites, which could be important for FMRP-containing RNA transport granules (38). In addition, the role of FMRP binding motifs in translational regulation still remains unclear. By identifying reproducible FMRP binding sites, the FCBS data is a rich resource for all avenues of research on molecular mechanisms of FMRP mediated regulation of mRNA and its dysregulation in fragile x syndrome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health [T32MH087977 to B.R.A., J.A.S., 1F32NS083205 to B.R.A, NS091859 to S.T.W., 1R01MH109026 to G.J.B]. Funding for open access charge: Departmental funds.

Conflict of interest statement. None declared.

REFERENCES

1. Darnell,J.C., Jensen,K.B., Jin,P., Brown,V., Warren,S.T. and Darnell,R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell*, **107**, 489–499.
2. Schaeffer,C., Bardoni,B., Mandel,J.L., Ehresmann,B., Ehresmann,C. and Moine,H. (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *EMBO J.*, **20**, 4803–4813.
3. Ramos,A., Hollingworth,D. and Pastore,A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *RNA*, **9**, 1198–1207.
4. Chen,L., Yun,S.W., Seto,J., Liu,W. and Toth,M. (2003) The fragile X mental retardation protein binds and regulates a novel class of mRNAs containing U rich target sequences. *Neuroscience*, **120**, 1005–1017.
5. Dolzhanskaya,N., Sung,Y.J., Conti,J., Currie,J.R. and Denman,R.B. (2003) The fragile X mental retardation protein interacts with U-rich RNAs in a yeast three-hybrid system. *Biochem. Biophys. Res. Commun.*, **305**, 434–441.
6. Darnell,J.C., Fraser,C.E., Mostovetsky,O., Stefani,G., Jones,T.A., Eddy,S.R. and Darnell,R.B. (2005) Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes Dev.*, **19**, 903–918.
7. Bechara,E.G., Didiot,M.C., Melko,M., Davidovic,L., Bensaid,M., Martin,P., Castets,M., Pognonec,P., Khandjian,E.W., Moine,H. *et al.* (2009) A novel function for fragile X mental retardation protein in translational activation. *PLoS Biol.*, **7**, e16.
8. Ascano,M. Jr, Mukherjee,N., Bandaru,P., Miller,J.B., Nusbaum,J.D., Corcoran,D.L., Langlois,C., Munschauer,M., Dewell,S., Hafner,M. *et al.* (2012) FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, **492**, 382–386.
9. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.

10. Darnell,J.C., Van Driesche,S.J., Zhang,C., Hung,K.Y., Mele,A., Fraser,C.E., Stone,E.F., Chen,C., Fak,J.J., Chi,S.W. *et al.* (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, **146**, 247–261.
11. Brown,V., Jin,P., Ceman,S., Darnell,J.C., O'Donnell,W.T., Tenenbaum,S.A., Jin,X., Feng,Y., Wilkinson,K.D., Keene,J.D. *et al.* (2001) Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell*, **107**, 477–487.
12. Suhl,J.A., Chopra,P., Anderson,B.R., Bassell,G.J. and Warren,S.T. (2014) Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. *Hum. Mol. Genet.*, **23**, 5479–5491.
13. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
14. Zhang,C. and Darnell,R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.*, **29**, 607–614.
15. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
16. Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.*, **39**, W86–W91.
17. Bon,M., Micheletti,C. and Orland,H. (2013) McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.*, **41**, 1895–1900.
18. Stegle,O., Payet,L., Mergny,J.L., MacKay,D.J. and Leon,J.H. (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374–i382.
19. Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
20. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
21. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
22. Baltz,A.G., Munschauer,M., Schwanhausser,B., Vasile,A., Murakawa,Y., Schueler,M., Youngs,N., Penfold-Brown,D., Drew,K., Milek,M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.
23. Castets,M., Schaeffer,C., Bechara,E., Schenck,A., Khandjian,E.W., Luche,S., Moine,H., Rabilloud,T., Mandel,J.L. and Bardoni,B. (2005) FMRP interferes with the Rac1 pathway and controls actin cytoskeleton dynamics in murine fibroblasts. *Hum. Mol. Genet.*, **14**, 835–844.
24. Menon,L., Mader,S.A. and Mihailescu,M.R. (2008) Fragile X mental retardation protein interactions with the microtubule associated protein 1B RNA. *RNA*, **14**, 1644–1655.
25. Stefanovic,S., Bassell,G.J. and Mihailescu,M.R. (2015) G quadruplex RNA structures in PSD-95 mRNA: potential regulators of miR-125a seed binding site accessibility. *RNA*, **21**, 48–60.
26. Guedin,A., Gros,J., Alberti,P. and Mergny,J.L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.
27. Stefanovic,S., DeMarco,B.A., Underwood,A., Williams,K.R., Bassell,G.J. and Mihailescu,M.R. (2015) Fragile X mental retardation protein interactions with a G quadruplex structure in the 3'-untranslated region of NR2B mRNA. *Mol. Biosyst.*, **11**, 3222–3230.
28. Maurin,T., Melko,M., Abekhoukh,S., Khalfallah,O., Davidovic,L., Jarjat,M., D'Antoni,S., Catania,M.V., Moine,H., Bechara,E. *et al.* (2015) The FMRP/GRK4 mRNA interaction uncovers a new mode of binding of the Fragile X mental retardation protein in cerebellum. *Nucleic Acids Res.*, **43**, 8540–8550.
29. Zhang,Y., Gaetano,C.M., Williams,K.R., Bassell,G.J. and Mihailescu,M.R. (2014) FMRP interacts with G-quadruplex structures in the 3'-UTR of its dendritic target Shank1 mRNA. *RNA Biol.*, **11**, 1364–1374.
30. Siomi,M.C., Zhang,Y., Siomi,H. and Dreyfuss,G. (1996) Specific sequences in the fragile X syndrome protein FMR1 and the FXR proteins mediate their binding to 60S ribosomal subunits and the interactions among them. *Mol. Cell. Biol.*, **16**, 3825–3832.
31. Chen,E., Sharma,M.R., Shi,X., Agrawal,R.K. and Joseph,S. (2014) Fragile X mental retardation protein regulates translation by binding directly to the ribosome. *Mol. Cell*, **54**, 407–417.
32. Caudy,A.A., Myers,M., Hannon,G.J. and Hammond,S.M. (2002) Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes Dev.*, **16**, 2491–2496.
33. Ishizuka,A., Siomi,M.C. and Siomi,H. (2002) A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes Dev.*, **16**, 2497–2508.
34. Muddashetty,R.S., Nalavadi,V.C., Gross,C., Yao,X., Xing,L., Laur,O., Warren,S.T. and Bassell,G.J. (2011) Reversible inhibition of PSD-95 mRNA translation by miR-125a, FMRP phosphorylation, and mGluR signaling. *Mol. Cell*, **42**, 673–688.
35. Dominissini,D., Moshitch-Moshkovitz,S., Schwartz,S., Salmon-Divon,M., Ungar,L., Osenberg,S., Cesarkas,K., Jacob-Hirsch,J., Amariglio,N., Kupiec,M. *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.
36. Vasilyev,N., Polonskaia,A., Darnell,J.C., Darnell,R.B., Patel,D.J. and Serganov,A. (2015) Crystal structure reveals specific recognition of a G-quadruplex RNA by a beta-turn in the RGG motif of FMRP. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E5391–E5400.
37. Meyer,K.D., Saletore,Y., Zumbo,P., Elemento,O., Mason,C.E. and Jaffrey,S.R. (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, **149**, 1635–1646.
38. Bassell,G.J. and Warren,S.T. (2008) Fragile X syndrome: loss of local mRNA regulation alters synaptic development and function. *Neuron*, **60**, 201–214.