Diverse eukaryotic CGG binding proteins produced by independent domestications of hAT transposons

Isaac Yellan¹, Ally W.H. Yang², and Timothy R. Hughes¹,²,*

¹Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada
²Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada

*Corresponding author: t.hughes@utoronto.ca
The human transcription factor (TF) CGGBP1 (“CGG Binding Protein”) is conserved only in amniotes, and is believed to derive from the zf-BED and Hermes transposase DNA-binding domains (DBDs) of a hAT DNA transposon. Here, we show that sequence-specific DNA binding proteins with this bipartite domain structure have resulted from dozens of independent hAT domestications in different eukaryotic lineages. CGGBP1s display a wide range of sequence specificity, usually including preferences for CGG or CGC trinucleotides, while some bind AT-rich motifs. The CGGBP1s are almost entirely non-syntenic, and their protein sequences, DNA binding motifs, and patterns of presence or absence in genomes are uncharacteristic of ancestry via speciation. At least eight CGGBP1s in the coelacanth *Latimeria chalumnae* bind distinct motifs, and the expression of the corresponding genes varies considerably across tissues, suggesting tissue-restricted function.
INTRODUCTION

Transposons can contribute to the evolution of regulatory networks by providing novel DNA, and in some cases, such as the hAT family of DNA transposons, being co-opted as sequence-specific DNA binding proteins (DBPs), such as transcription factors (TFs) (Feschotte and Pritham 2007; Atkinson 2015). The hAT N-terminal region encodes a BED zinc finger (zf-BED) and the Hermes transposon DNA binding and protein dimerization domain, which are involved in sequence specific DNA binding to both subterminal repeat sequences and inverted terminal repeats (Hickman, et al. 2005; Hickman, et al. 2014). The ZBED family of vertebrate TFs is an example of hAT domestication, having arisen from two different exaptations (Hayward, et al. 2013). The human gene CGGBP1 provides another example; it was first identified in a survey of proteins that bind to the Fragile X (CGG)n repeat expansion (Deissler, et al. 1996; Deissler, et al. 1997) and bears remote homology to the N-terminal region of hATs (Singh and Westermark 2015).

Transposons often evolve quickly (Khan, et al. 2006; Feschotte and Pritham 2007; Kofler, et al. 2015; Arkhipova 2017), and their domain models, which are based on amino acid (AA) sequence alignments, may not generalize across all instances of a transposon family. The same is presumably true of host proteins derived from transposons. Indeed, no significant domains are detected in CGGBP1 by Pfam (El-Gebali, et al. 2019) and SMART (Letunic and Bork 2018) databases searches, even though CGGBP1 contains an apparent zf-BED and Hermes DNA binding domain (DBD). The Panther (Mi, et al. 2019) and Interpro (Mitchell, et al. 2019) databases do contain a CGGBP1 superfamily definition, however. Intriguingly, this domain model identifies high-scoring host sequences (“CGGBPs”) in several bony fish, the sea lamprey Petromyzon marinus, as well as several insects and fungi. The CGGBP1 domain model also identifies 29/824 hATs in the RepeatMasker database, mainly in the hAT-19 subclass (13/16), suggesting that the CGGBP1 domain model corresponds only to specific hAT subtypes. There is considerable overlap between species containing hAT-19 RepeatMasker hits, and those with host genes encoding CGGBPs (Table S1).
Strikingly, the genome of the coelacanth *Latimeria chalumnae* includes 62 CGGBP genes, with 47 containing the full zf-BED and Hermes DBDs (Amemiya, et al. 2013; Yates, et al. 2019).

These CGGBP proteins are largely unstudied and undocumented, except for mammalian CGGBP1, and their presence in various databases. Here, we present several lines of evidence indicating that the non-amniote CGGBPs represent a previously undocumented class of *bona fide* DBPs that have arisen many times through independent *hAT* integrations.

**CGGBPs possess diverse DNA sequence preferences**

Determining what sequences DBPs bind is often a starting point in their characterization (Lambert, et al. 2018). We queried the sequence specificity of 101 host CGGBPs (from Ensembl (Yates, et al. 2019)) and five CGGBP-like *hAT* proteins, using two independent universal Protein Binding Microarray (PBM) designs (Berger, et al. 2006). The two arrays (“ME” and “HK”) (Badis, et al. 2008) each contain all possible DNA 10-mers, but each is embedded within a completely different 35-mer relative to the other array. The CGGBPs were selected on the basis of sequence diversity (Figure S1). As the vast majority contain introns and/or are truncated (< 200 AAs), they are likely to be transcribed as mRNA and are not functional *hAT*s (Figure S3). PBM data were processed to obtain estimates of significance of binding to 8-base sequences (represented by multiple spots on the array), known as 8-mer “E-scores” (Berger, et al. 2006) as well as motifs that summarize the top scoring 8-mers (Ray, et al. 2013).

Eighteen of the CGGBP proteins, and four *hAT* proteins (including variants of *hAT-19_Crp*) displayed clear sequence preferences, i.e. a set of high-scoring 8-mers with related sequences, that were shared between the two array designs and yielded a clear motif. These 22 proteins all had relatively long N-terminal extensions (and thus a longer linker between the GST tag and the CGGBP domain), and most contained both Cys and both His residues that complex zinc in the zf-BED, a leucine two positions N-terminal to the zf-BED, and two prolines that are widely conserved in the Hermes domain,
suggesting that these residues are required for DNA sequence specificity. **Figure S2** contains an alignment of all CGGBPs tested, with key residues annotated. Conditioned on having five of the seven conserved residues and a 14AA minimum linker, 36% (22/61) of the constructs were successful, consistent with previous results from many bona fide DBD classes (Weirauch, et al. 2014).

**Figures 1** and **2** summarize the data for these 22 proteins. In **Figure 1**, the scores obtained for individual 8-mers are clustered. The groups formed therefore reflect only the DNA binding data, irrespective of the protein sequences or the genomes in which they are found. In **Figure 2**, the protein phylogram reflects the AA replacement rate, with DNA binding motifs shown as sequence logos. The groups of closely related CGGBPs in **Figure 2** therefore presumably reflect recent shared ancestry.

Several observations support the accuracy of the PBM data. Human CGGBP1 had highest preference to CGGCGG-containing sequences, reflecting how it was initially identified (Deissler, et al. 1997). In addition, the DNA binding motifs for all four hAT-encoded CGGBPs have high-scoring matches within subterminal regions of the corresponding Repbase model (**Figure S3**). Also, in several cases, there are clear similarities between DNA binding motifs obtained from CGGBPs with related protein sequences, even from distantly related genomes. The coelacanth gene Coel_14934 and the coelacanth hAT-13_LCh transposon encode similar protein sequences, and yielded very similar DNA binding motifs, suggesting that Coel_14934 may be a recently co-opted hAT-13_LCh copy. Sea lamprey, which is >515MYa distant (Kumar, et al. 2017) also contains a related CGGBP with similar sequence specificity. A similar relationship is found between the CGGBPs from kanglang fish and climbing perch, which diverged >200MYa, and the CGGBPs from spiny chromis and the killifish *Austrofundulus limnaeus*, which diverged >80MYa.

Most CGGBPs possess unique DNA sequence specificity (**Figures 1** and **2**): the 8-mer preferences from the ME and HK arrays for each CGGBP correlate more highly with each other than to any other protein (**Figure 1**). The only clear exceptions are the human and chicken CGGBP1 orthologs, and the three single-residue variants of
Most of the motifs contain a CGG trinucleotide, but several of the coelacanth CGGBP s in fact prefer GCGC-containing sequences, while AT-rich motifs were obtained from hAT transposons in the insects Rhodnius prolixus (kissing bug) and Locusta migratoria (migratory locust), and the one CGGBP-containing protein in the fungus Rhizophagus irregularis (Figure 2). Thus, the CGGBP s display diverse sequence specificity signatures.

Similarity among CGGBP DNA sequence specificity and AA sequence often deviate from what would be expected from the species tree (Figures 1 and 2). To explore this issue more closely, we examined orthology relationships among the CGGBP s.

**Lack of synteny among non-amniote CGGBP-containing proteins indicates distinct hAT domestications**

To ask whether the CGGBP s are orthologs (i.e. related by common ancestry in the host genome), we examined the homology of nearby genes at their chromosomal locations, because common descent is often reflected in shared ordering of genes (synteny). In bony fish, for example, gene order is largely preserved since the whole genome duplication ~300 MYa (Jaillon, et al. 2004). In contrast, few of the CGGBP s are found at syntenic positions (based on the presence of at least two nearby homologs) (Figure 3). Three coelacanth proteins are adjacent on the chromosome (shown in Figure 3; all others are on different scaffolds), and these proteins have similar sequences, consistent with tandem duplication following initial integration. Two of these three yielded motifs (Figure 2), with no high-scoring 8-mers in common (Figure 1), suggesting that diversification in DNA sequence specificity may have promoted retention following duplication.

Lack of synteny is consistent with the CGGBP s originating from distinct hAT integration events. Several other observations are consistent with this notion. One is that the intron/exon structures of the CGGBP s differ from each other (Figure S4). Another is that CGGBP s show a spotty distribution, absent from many (and perhaps most) fish and insect genomes, and apparently most fungi, in a pattern that does not strictly reflect
established clades (Table S1). Likewise, as noted above, protein sequence relationships among CGGBPs often deviate from the species tree, as do CGGBP DNA sequence specificities (Figures 1 and 2). Coelacanth CGGBPs, for example, are distributed throughout trees of both CGGBP AA sequence similarity (Figure 2) and DNA binding similarity (Figure 1), suggesting they may derive from different hAT subtypes. In another example, CGGBPs of the spiny chromis and the killifish are similar to each other, and bind very similar motifs, but their AA sequences are more related to CGGBPs from kissing bug and migratory locust than to those of other fish, suggesting they may have originated from horizontal transfer of related hATs. Likewise, the CGGBP from yellow fever mosquito is more closely related to crocodile hAT-19_Crp than it is to the hAT-19 proteins we analyzed from other insects, and these two proteins also have similar DNA sequence specificity.

**Varied expression of coelacanth CGGBPs suggests tissue-restricted function**

Collectively, these observations indicate that many independent domestications of the hAT CGGBP domain have spawned a relatively large family of DBPs with distinct sequence preferences, which are present in a variety of fish and insects and some fungi. In particular, the coelacanth CGGBPs have highly disparate sequences, relative to other CGGBPs, and we speculate that they originate from different hAT-19 families, or other related hATs. It is known that DNA transposons have been recently active in coelacanth (Naville, et al. 2014), and indeed, relatively few of the 62 CGGBPs present in *L. chalumnae* can be found in the transcriptome shotgun assembly of the other extant coelacanth species, *L. menadoensis*, which diverged ~22 MYa from *L. chalumnae* (Amemiya, et al. 2013) (we identified only eight CGGBPs using reciprocal BLAST and scanning with the Panther HMM).

Lack of orthology presents a challenge in examining evolutionary pressures on the CGGBPs. The fact that the non-amniote CGGBPs are found entirely in non-model species also complicates functional analysis. Nonetheless, the *Latimeria chalumnae* CGGBPs are often expressed in different tissues from each other (Figure S5).
suggesting that their functions may be tissue restricted in addition to being differentiated by distinct binding specificities.

Finally, the fact that the CGGBP domain model represents only a small subset of hATs, mainly the hAT-19 subclass, raises two intriguing possibilities. First, the observation that both the CGGBPs and hAT-19 are found mainly among a few disparate classes of eukaryotes suggests horizontal movement of hATs between very different branches, even though these classes (e.g. Rhizophagus and bony fish) inhabit very different environments. Second, the narrow coverage of the CGGBP domain model among hATs (29/824) raises the possibility that domain models corresponding to the same regions of other hAT subclasses may uncover many additional classes of host DBPs.

**Methods**

We designed and built expression constructs for the 114 CGGBP proteins to include the entire domain plus 50 flanking AAs on both sides (unless the end of the protein was reached), fused to an N-terminal GST tag and driven by a T7 promoter. Three proteins with unknown AA residues (X) were run as all possible variants consistent with corresponding codon sequences. PBM data were generated as described (Weirauch, et al. 2014), with motifs derived using Top10AlignZ. Orthologs, gene family members, and synteny were determined through manual examination of Ensembl and NCBI gene tracks, with homology called according to Ensembl Compara, NCBI orthology predictions for A. limnaeus, or the NCBI-annotated identity of inferred Kanglang Fish genes.

**Data availability**

Raw and normalized PBM data is deposited on GEO under accession [GSE157085](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157085). Motifs and PBM construct sequences are available at [http://hugheslab.ccbr.utoronto.ca/supplementary-data/CGGBP1](http://hugheslab.ccbr.utoronto.ca/supplementary-data/CGGBP1).

**Acknowledgements**
This work was supported by the National Science and Engineering Research Council of Canada (NSERC) (RGPIN-2016-05643) and the Canadian Institutes of Health Research (FDN-148403). IY held a CIHR CGS-M award. We are grateful to Arttu Jolma for helpful conversations, Debashish Ray and Rozita Razavi for critical evaluation of the manuscript, and Mihai Albu for website development and IT support.
FIGURE LEGENDS

Figure 1. DNA 8-mer preferences of CGGBPs. Heatmap shows E-scores for 8-mers enriched (E ≥ 0.45) in at least one PBM assay. Cell values are scaled E-scores (10^{(E+10)^{-3}}, then normalized between 0 and 100), and are grouped by hierarchical agglomerative clustering with average linkage. Green labels represent coelacanth; blue labels represent hAT transposases. Dendrograms were produced through complete-linkage hierarchical clustering of Pearson distances. Silhouettes were obtained from PhyloPic, or created using Public Domain or Creative Commons images.

Figure 2 – Phylogram of CGGBPs with DNA binding motif displayed. Proteins with DNA binding preferences on PBMs were aligned with Mafft E-INSi v7.310 (Katoh and Standley 2013) and any positions with gaps in greater than 20% of the sequences were trimmed using TrimAI. The unrooted maximum likelihood (ML) tree was determined using RAxML-NG (Kozlov, et al. 2019), using the LG substitution matrix, ML estimated invariant site proportions, and a 4-category gamma model of rate heterogeneity (LG+I+G4), and 300 bootstrap trees. The substitution model was selected according to maximum AIC as determined by ModelTest-NG. Motifs were produced using Top10AlignZ (Weirauch, et al. 2014).

Figure 3 – Synteny of CGGBP-containing loci. Orthologous loci are displayed as boxes along a bold line, and relationships between homologous genes indicated by thin connecting lines. All species with a sequence-specific CGGBP are included, as well as two fish with non-DNA binding proteins for comparison. Gene order was obtained from Ensembl or NCBI gene tracks. Non-homologous intervening genes have been removed, with two diagonal lines indicating such gaps, and compared to orthologous loci in other species. For species with multiple paralogous CGGBPs, the name/accession of the visualized CGGBP is given. An asterisk is shown next to the name of species for which the CGGBP loci displays no synteny, but which do contain conserved genes. Species tree from TimeTree (Kumar, et al. 2017).
REFERENCES

Amemiya CT, Alfoldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311-316.

Arkhipova IR. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA 8:19.

Atkinson PW. 2015. hAT Transposable Elements. Microbiol Spectr 3.

Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol Cell 32:878-887.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24:1429-1435.

Deissler H, Behn-Krappa A, Doerfler W. 1996. Purification of nuclear proteins from human HeLa cells that bind specifically to the unstable tandem repeat (CGG)n in the human FMR1 gene. J Biol Chem 271:4327-4334.

Deissler H, Wilm M, Genc B, Schmitz B, Ternes T, Naumann F, Mann M, Doerfler W. 1997. Rapid protein sequencing by tandem mass spectrometry and cDNA cloning of p20-CGGBP. A novel protein that binds to the unstable triplet repeat 5'-d(CGG)n-3' in the human FMR1 gene. J Biol Chem 272:16761-16768.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427-D432.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331-368.

Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJ. 2009. A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. Cell 138:314-327.

Hayward A, Ghazal A, Andersson G, Andersson L, Jern P. 2013. ZBED evolution: repeated utilization of DNA transposons as regulators of diverse host functions. PLoS ONE 8:e59940.

Hickman AB, Ewis HE, Li X, Knapp JA, Laver T, Doss AL, Tolun G, Steven AC, Grishaev A, Bax A, et al. 2014. Structural basis of hAT transposon end recognition by Hermes, an octamer DNA transposase from Musca domestica. Cell 158:353-367.

Hickman AB, Perez ZN, Zhou L, Musingarimi P, Ghirlando R, Hinshaw JE, Craig NL, Dyda F. 2005. Molecular architecture of a eukaryotic DNA transposase. Nat Struct Mol Biol 12:715-721.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ouzou-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946-957.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16:78-87.
Kofler R, Nolte V, Schlotterer C. 2015. Tempo and Mode of Transposable Element Activity in Drosophila. PLoS Genet 11:e1005406.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453-4455.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34:1812-1819.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. Cell 175:598-599.

Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46:D493-D496.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47:D419-D426.

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, et al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351-D360.

Naville M, Chalopin D, Voilf JN. 2014. Interspecies insertion polymorphism reveals recent activity of transposable elements in extant coelacanths. PLoS ONE 9:e114382.

Nikaido M, Noguchi H, Nishihara H, Toyoda A, Suzuki Y, Kajitani R, Suzuki H, Okuno M, Aibara M, Ntagungu BP, et al. 2013. Coelacanth genomes reveal signatures for evolutionary transition from water to land. Genome Res 23:1740-1748.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, 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.

Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, Jolma A, Zhong G, Guo H, Kanagalingam T, et al. 2016. Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res 26:1742-1752.

Singh U, Westermark B. 2015. CGGBP1---an indispensable protein with ubiquitous cytoprotective functions. Ups J Med Sci 120:219-232.

Stormo GD, Fields DS. 1998. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23:109-113.

Tan G, Lenhard B. 2016. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics 32:1555-1556.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drew P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158:1431-1443.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2019. Ensembl 2020. Nucleic Acids Res. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430:2237-2243.
Figure 1

870 8-mers enriched in ≥ 1 assay (E ≥ 0.45)

Normalized E-score

Scaled

Assay

PBM

Gilthead seabream

Sparus aurata

Human

Homo sapiens

Chicken

Gallus gallus

Coelacanth

Latimeria chalumnae

Migratory locust

Locusta migratoria

Rhizophagus irregularis

Climbing perch

Anabas testudineus

Kanglang fish

Anabas grahami

Yellow fever mosquito

Aedes aegypti

Saltwater crocodile

Crocodylus porosus

Sea lamprey

Petromyzon marinus

Killifish sp

Australendulus limnaeus

Spiny chromis

Acanthochromis polyacanthus
0.4 aa replacements/site
**Figure 3**

Potential CGGBP Orthologs

| Species       | CGGBP Locus | Ortholog or Gene Family Member |
|---------------|-------------|-------------------------------|
| Human         |             |                               |
| Chicken       |             |                               |
| Climbing Perch |             |                               |
| Orange Clownfish |             |                               |
| Reedfish      |             |                               |
| Orange Chromis |             |                               |
| Spiny Chromis |             |                               |
| Gilthead Seabream |         |                               |
| Austrofundulus Limnaeus | |                               |
| Rhizophagus   |             |                               |
| Kanglang Fish |             |                               |
| Lamprey       |             |                               |

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab007/6129264 by guest on 10 February 2021