The impact of repetitive DNA and its guardian proteins on the evolution of neuronal gene regulatory networks

Farmiloe, G.

Publication date
2023

Citation for published version (APA):
The role of primate-specific ZNF519 in CGG-repeat mediated gene regulation

Grace Farmiloe¹, Ilan Kotlar¹, Elise van Bree¹, Frank Jacobs¹$  

¹University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam Neuroscience, Amsterdam, the Netherlands  

$ Corresponding author
Abstract

Short tandem repeats (STRs) are defined as a short DNA motif that is repeated many times, each repeat directly adjacent to the other. STRs are highly variable and ubiquitous in our genomes, they are often found in the vicinity of genes and they have been shown to have influence on gene expression. STRs have also been implicated in the processes of speciation and evolution. Their variability, the speed at which they can change and their species specificity makes them prime tools for evolution to work with. In particular CGG repeats are associated with neurological disorders in humans, implying they may have an essential role in developing or maintaining a healthy brain. How exactly STRs are controlled, and how STRs regulate gene expression is not well understood. In this study we explore the possibility that the influence of STRs on nearby gene expression is controlled by KRAB zinc finger proteins. We identify the primate-specific ZNF519 as a potentially important regulator of CGG STRs, and find that ZNF519 and the STRs it binds may have contributed to the evolution of gene expression networks underlying human brain developmental processes.
Introduction

Short tandem repeats (STRs) consist of small (1-6bp), repeating nucleotide sequence units of variable length (Tautz 1993). They are ubiquitous in our genomes, and there are estimated to be as many as 1 million STR loci in our genomes. STRs are known to be highly polymorphic between individuals (Weber and Wong 1993; Brinkmann et al. 1998; Li et al. 2002; Legendre et al. 2007). Recent advances in high resolution genetic profiling has revealed that earlier studies have underestimated the extent of this variation and that many reference STRs are under-represented (Fungtammasan et al. 2015; Annear et al. 2021). A clear view of STR variation in the human population is important because many STRs have been ascribed gene-regulatory roles (Gymrek et al. 2016; Chen et al. 2016; Fotsing et al. 2019; Jakubosky et al. 2020). In one study, 28,000 STRs were identified as associated with the expression of nearby genes and a subgroup of these have been implicated in complex traits (Fotsing et al. 2019). In a comparison between humans and non-human primates, genes that contained tandem repeats (TRs) in their 3’ UTRs, introns, and exons showed higher expression divergence than genes without a TR and genes with polymorphic STRs showed the highest expression divergence (Sonay et al. 2015). This association suggests that STRs have played a role in evolutionary differences in gene expression and potentially even speciation.

STRs are known to differ widely between species, and a number of STR-characteristics were found to be specific to the human genome; STRs found in core promoter sequences are longer in humans than other primates and a large number of STRs are expanded in humans (Ohadi et al. 2012; Namdar-Aligoodarzi et al. 2015; Sulovari et al. 2019). In primates a large number of STRs are lineage specific and a large proportion of STRs shared with older lineages are expanded since diverging from mouse, especially in 5’UTR regions (Namdar-Aligoodarzi et al. 2015; Sulovari et al. 2019). Because the 5’UTR has important regulatory roles, this suggests that species-specific STR variations could account for species-specific characteristics of gene expression.

The regulatory influence of STRs on gene expression is most evident in so-called repeat expansion disorders. In most healthy individuals STRs are present in the genome within a certain range of repeat lengths. However in some individuals an STR is abnormally expanded and depending on the locus where the STR is located, this can give rise to a repeat expansion disorder. The pathogenic mechanisms of these disorders are varied and depend on the nucleotide sequence of the repeat unit itself and the genes that are affected; Molecular phenotypes include the production of toxic proteins and complete transcriptional silencing of genes.
The role of primate-specific ZNF519 in CGG-repeat mediated gene regulation

(Malik et al. 2021). It has been shown that when a repeat becomes longer it also becomes more unstable, allowing for even larger expansions in length in future generations or even in somatic cells during embryonic development (Mirkin 2007; Kevtun and McMurray 2008; McMurray 2010). This leads to the question of how much STR expansion is tolerated by our genomes and what factors are present in our cells to restrict this. From an evolutionary point of view, it would be interesting to assess which factors are involved in the differential propensity or vulnerability of STRs to expand, a process that may account for many of the species-specific characteristics of gene expression. The molecular mechanisms by which STR lengths can change include replication strand slippage and double strand break repair (Kristitch and Mirkin 2020). Species-specific adaptations in any of the many factors involved in these complex processes could account for the species-specific characteristics of STRs, including the propensity or vulnerability for expansion.

It was recently established that STRs can be bound by members of the Kruppel associated box (KRAB) zinc finger protein (KZNFs) family (Jacobs et al. 2014; Imbeault et al. 2017; Haring et al. 2021). KZNFs are a very large family of DNA binding proteins characterised by a large number of species-specific genes and gene-structural adaptations (Emerson and Thomas 2009; Thomas and Schneider 2011; Ecco et al. 2017; Imbeault et al. 2017; Helleboid et al. 2019). Given the established regulatory roles of KZNFs at other repetitive DNA loci, it is possible that STR-binding KZNFs play a regulatory role in STR-containing gene promoters. In this analysis we present the hypothesis that variations in STR length and STR mediated gene regulation are mediated through KZNFs. We explore the possibility of the existence of a KZNF as a primate-specific factor which interacts with and modulates STRs and the potential role that these interactions may have played in species-specific characteristics of STR expansions and neuronal gene regulation.
Chapter 5

Results

Multiple KZNFs interact with various simple repeat families

The regulatory effect of specific KZNFs and certain classes of TEs has been well documented (Thomas and Schneider 2011; Jacobs et al. 2014; Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017; Helleboid et al. 2019). Recent ChIP data on a large number of KZNFs has shown that the individual roles of members of this gene family may be more complex and widespread than initially thought (Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017). A number of KZNFs have been shown to bind to gene promoters as well as TEs and some KZNFs do not seem to bind to TEs at all (Imbeault et al. 2017; Farmiloe et al. 2020). In our study of promoter binding KZNFs we made the observation that a number of KZNFs bind to a large proportion of simple tandem repeats (STRs), especially CG-rich repetitive regions found in gene promoters. We focused on the gene-promoter binding KZNFs which bind to >100 promoters, 31 in total, and further analysed their repetitive DNA binding patterns. The KZNFs within this selection could be grouped into three categories, 1: KZNFs interacting primarily with genomic and promoter-embedded TEs in the absence of STRs (18 KZNFs), 2: KZNFs interacting with both TEs and promoter-embedded STRs and (8 KZNFs), 3: KZNFs interacting with promoter-embedded STRs (5 KZNFs) (Figure 1). Only KZNFs which fell in categories 2 and 3 were relevant to our investigation, leaving 13 candidate KZNFs for STR regulation. As repetitive elements are abundant in the human genome, we assessed KZNF binding to control regions 500bp downstream of the KZNF summits and as an extra control, transcription start sites. These controls exhibited a clearly different pattern of repeat binding, suggesting the recognition of STRs is specific to the KZNFs analysed.

Some KZNFs interact with disease-associated STRs

Some genetic variation involving expanded STRs can give rise to diseases. To assess whether any of the KZNFs that bind to STRs interact with disease-associated loci, we overlapped KZNF binding with 17 common repeat expansion disorder (RED) loci (Table 1). We found robust binding by 5 KZNFs on, or next to at least one of the pathogenic repeats in genes associated with 5 REDs (ATXN7, HTT, RUNX2, FMR1, TMEM185A) (Figure 2a). Each of the 5 KZNFs (ZNF93, ZNF202, ZNF468, ZNF519, ZNF534) belong to category 2, and have a mixed binding pattern of STRs and TEs except ZNF202 which predominantly binds to simple repeats and is in category 3 (Figure 2b). The ZNF202 peaks were found at all 5 repeats, however, the peaks seem comparatively low and broad when compared to the
other KZNFs, suggesting low subtype specificity of ZNF202 to STRs in general. ZNF519 has strong, defined peaks with summits that precisely overlap with each pathogenic repeat. These repeats were all either CAG or CGG, the most common trinucleotide motifs associated with REDs. In the analysed data ZNF519 appears to have a preference for the trinucleotide repeats and does not show binding at repeats with a longer motif such as the CNBP tetranucleotide, the C9orf72 hexanucleotide, and the CSTB dodecanucleotide.

![Figure 1: Repeat binding patterns of ZNFs](image)

**Figure 1: Repeat binding patterns of ZNFs** *(left)* Pie charts showing the average pattern of repeat binding after analysis of 31 promoter binding KZNFs. Each of the 31 KZNFs falls into one of three groups: TE favouring: n=18, Mixed binding pattern: n=8, Simple repeat favouring: n=5. *(right)* Showing an example of the location where members of each group might bind in specific, recognised classes of transposable elements and gene promoters.

We also examined KAPI (KRAB associated protein 1) ChIP data at these KZNF sites *(Figure 2a).* KAPI is recruited by KZNFs and influences heterochromatin formation. The presence of both KZNF and KAPI signal at these loci (here seen as loci where the peaks overlap) suggests that the KZNF is playing a repressive role at the loci by recruiting KAPI. A KAPI peak often coincides with the ZNF519 peak for the CAG repeats, as is noticeable for HTT and RUNX2 but KAPI signal is
not present at the sites of ZNF519 binding to CGG repeats in for example FMR1 and TMEM185A (Figure 2a). It is possible that in some of these loci KZNFs are recruiting KAP1 and establish a repressive chromatin structure. At loci lacking a KAP1 signal there may be an alternate function being served by the KZNF at that particular locus.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Product</th>
<th>Disorder</th>
<th>Motif</th>
<th>Repeat size</th>
<th>Pathogenic repeat size</th>
<th>Primary pathogenic mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMR1</td>
<td>FMRP</td>
<td>Fragile X syndrome</td>
<td>CGG</td>
<td>6–53</td>
<td>&gt;200</td>
<td>Gene silencing</td>
</tr>
<tr>
<td>FMR1</td>
<td>FMRP</td>
<td>Fragile X associated tremor/ataxia syndrome</td>
<td>CGG</td>
<td>6–53</td>
<td>55–200</td>
<td>Increased transcription, transcript toxicity</td>
</tr>
<tr>
<td>FXN</td>
<td>Frataxin</td>
<td>Friedreich’s ataxia</td>
<td>GAA</td>
<td>7–34</td>
<td>&gt;100</td>
<td>Reduced frataxin levels30</td>
</tr>
<tr>
<td>HTT</td>
<td>Huntingtin</td>
<td>Huntington’s disease</td>
<td>CAG</td>
<td>10–35</td>
<td>&gt;96</td>
<td>Toxic gain/change of function</td>
</tr>
<tr>
<td>ATXN1</td>
<td>Ataxin 1</td>
<td>Spinocerebellar ataxia 1</td>
<td>CAG</td>
<td>6–35</td>
<td>49–88</td>
<td>Toxic gain of function31</td>
</tr>
<tr>
<td>ATXN2</td>
<td>Ataxin 2</td>
<td>Spinocerebellar ataxia 2</td>
<td>CAG</td>
<td>14–32</td>
<td>33–77</td>
<td>Toxic gain of function32</td>
</tr>
<tr>
<td>ATXN3</td>
<td>Ataxin 3</td>
<td>Spinocerebellar ataxia 3</td>
<td>CAG</td>
<td>12–40</td>
<td>55–86</td>
<td>Toxic gain of function32</td>
</tr>
<tr>
<td>CACNA1A</td>
<td>CACNA1A</td>
<td>Spinocerebellar ataxia 6</td>
<td>CAG</td>
<td>4–18</td>
<td>21–30</td>
<td>Toxic gain of function33</td>
</tr>
<tr>
<td>ATXN7</td>
<td>Ataxin 7</td>
<td>Spinocerebellar ataxia 7</td>
<td>CAG</td>
<td>7–17</td>
<td>38–120</td>
<td>Toxic gain of function34</td>
</tr>
<tr>
<td>AFF2</td>
<td>AF4/FMR2 family member 2</td>
<td>FRAXE mental retardation</td>
<td>CCG</td>
<td>6–35</td>
<td>&gt;200</td>
<td>Gene silencing12</td>
</tr>
<tr>
<td>AFF3</td>
<td>AF4/FMR2 family member 3</td>
<td>FRA2E mental retardation</td>
<td>CCG</td>
<td>3–20</td>
<td>&gt;300</td>
<td>Gene silencing21</td>
</tr>
<tr>
<td>TMEM185A</td>
<td>Transmembrane protein 185A</td>
<td>FRA2X mental retardation</td>
<td>CCG</td>
<td>6–29</td>
<td>&lt;900</td>
<td>Gene silencing24</td>
</tr>
<tr>
<td>DMPK</td>
<td>Myotonic dystrophy protein kinase</td>
<td>Myotonic dystrophy 1</td>
<td>CTG</td>
<td>5–37</td>
<td>&gt;50</td>
<td>Toxic gain of function35</td>
</tr>
<tr>
<td>RUNX2</td>
<td>Run-related transcription factor 2</td>
<td>Cleidocranial dysplasia</td>
<td>CAG</td>
<td>~17</td>
<td>&gt;27</td>
<td>Loss of function36</td>
</tr>
<tr>
<td>CNBP</td>
<td>ZNF9</td>
<td>Myotonic dystrophy 2</td>
<td>CCTG</td>
<td>22–33</td>
<td>&gt;75</td>
<td>Toxic gain of function35</td>
</tr>
<tr>
<td>C9orf72</td>
<td>Guanine nucleotide exchange C9orf72</td>
<td>Amyotrophic lateral sclerosis/frontotemporal dementia</td>
<td>GGGGCC</td>
<td>&lt;30</td>
<td>&gt;31</td>
<td>Toxic RNA gain of function37</td>
</tr>
<tr>
<td>CSTB</td>
<td>Cystatin-B</td>
<td>Unverricht–Lundberg disease</td>
<td>CCCCQC</td>
<td>2–3</td>
<td>&gt;60</td>
<td>Reduced protein levels7</td>
</tr>
</tbody>
</table>

**Table 1: Repeat expansion disorders**
The role of primate-specific ZNF519 in CGG-repeat mediated gene regulation

**Figure 2: KZNF binding at simple repeats associated with repeat expansion disorders**

a, Showing simple repeats in gene promoter regions associated with repeat expansion disorders, the pathogenic repeat is shown in red. KZNF binding shown below the repeats and genes. ZNF ChIP tracks scaled 0-300 (left) and 0-50 (right), KAP1 track scaled 0-50 for all regions. b, Specific repeat class binding of high confidence KZNF summits for each of the KZNFs with a peak around the pathogenic repeats above, number of STRs bound labelled in each pie chart.

ZNF519 binds to disease-associated STRs including fragile sites

The predicted binding motif for ZNF519 is highly enriched for G/C sequences (Figure 3a). Indeed this prediction is supported by our observations that over 75% of the simple repeats bound by ZNF519 are GC rich or GC composed (Figure 3b); together, this accounts for approximately 25% of all ZNF519 summits. In addition, ZNF519 was previously shown (Farmiloe et al, submitted) to bind to a large number of gene promoters which are also typically GC rich (Akan and Deloukas 2008). Notably, ZNF519’s GC binding preferences also include chromosomal fragile sites (Table 2): Folate-sensitive fragile sites are regions of the chromosome prone to gaps or breaks and are associated with expanded CGG-repeats. The majority of these fragile sites are located in or nearby gene promoters and a number of them have been implicated in REDs or other rare disorders (Table 2). The most well-known fragile site is the CGG repeat in the
Chapter 5

promoter of *FMR1* on chromosome X which, when expanded, causes fragile X syndrome. We identified a total of 11 fragile sites from the literature (Knight et al. 1993; Jones et al. 1995; Shaw et al. 2002; Sarafidou et al. 2004; Metsu et al. 2014b, 2014a) of which 7 are associated with rare diseases (Table 2). For 8 fragile sites, we observed a strong *ZNF519* peak, directly overlapping with the fragile site-associated CGG repeats (Figure 3c). Two fragile sites were excluded because they lacked a clear association with a nearby gene. Next to *ZNF519*, a number of other KZNFs also bind in the vicinity of individual fragile site CGG-associated genes although no consistent binding pattern was observed across all fragile sites as is observed for *ZNF519*.

ENCODE histone modification data suggests that all regions except FRA2A (*AFF3*) and FRAXE (*AFF2*) are active promoters; the layered H3K4me data from ENCODE shows a clear peak around the *ZNF519* binding sites at these loci. The lack of layered H3K4me3 at *AFF2* and *AFF3* could be a result of the cells used to generate the data as these genes are highly expressed in the brain and the ENCODE data does not include a neuronal cell line. Similar to KZNF binding sites in loci associated with REDs, no clear *KAP1* binding was observed across all 9 fragile sites. Because *KAP1* closely associates with repressive H3K9me3 marks, the absence of *KAP1* binding is in line with the presence of activating H3K4me3 on these sites. It should be noted that the data on KZNF binding and epigenetic marks is based on ‘healthy’ cells that are expected to contain un-expanded fragile sites. It is unclear whether these binding profiles differ in the case of an expanded repeat. Overall, despite binding of *ZNF519* to the CGG repeats within fragile sites, the role that *ZNF519* could be playing at these loci in both their non-pathogenic and expanded forms remains elusive.

<table>
<thead>
<tr>
<th>Fragile site</th>
<th>Chromosome band</th>
<th>Gene</th>
<th>Motif</th>
<th>Pathology</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRA2A</td>
<td>2q11.2</td>
<td>AFF3</td>
<td>CGG</td>
<td>FRA2A mental retardation 21</td>
</tr>
<tr>
<td>FRA7A</td>
<td>7p11.2</td>
<td>ZNF713</td>
<td>CGG</td>
<td>FRA7A autism spectrum disorder 22</td>
</tr>
<tr>
<td>FRA10A</td>
<td>10q23.3</td>
<td>FRA10AC1</td>
<td>CGG</td>
<td>-</td>
</tr>
<tr>
<td>FRA10B*</td>
<td>10q25.2</td>
<td>-</td>
<td>AT-rich</td>
<td></td>
</tr>
<tr>
<td>FRA11A</td>
<td>11q13.3</td>
<td>C11orf80</td>
<td>CGG</td>
<td>-</td>
</tr>
<tr>
<td>FRA11B</td>
<td>11q23.3</td>
<td>CBL2</td>
<td>CGG</td>
<td>Jacobsen syndrome 19</td>
</tr>
<tr>
<td>FRA12A</td>
<td>12q13.1</td>
<td>DIP2B</td>
<td>CGG</td>
<td>FRA12A mental retardation</td>
</tr>
<tr>
<td>FRA16A*</td>
<td>16p13.11</td>
<td>-</td>
<td>CGG</td>
<td>-</td>
</tr>
<tr>
<td>FRAXA</td>
<td>Xq27.3</td>
<td>FMR1</td>
<td>CGG</td>
<td>FXS</td>
</tr>
<tr>
<td>FRAXE</td>
<td>Xq28</td>
<td>AFF2</td>
<td>CGG</td>
<td>FRAXE mental retardation 12</td>
</tr>
<tr>
<td>FRAXF</td>
<td>Xq28</td>
<td>TEMEM185A</td>
<td>CGG</td>
<td>FRAXF mental retardation 12</td>
</tr>
</tbody>
</table>

**Table 2: List of fragile sites**

*FRA16A and FRA10B do not have a gene in their vicinity and were excluded from further analysis*
The role of primate-specific ZNF519 in CGG-repeat mediated gene regulation

Figure 3: ZNF519 peaks at fragile sites

**a**. ZNF519 predicted binding motif. **b**. Breakdown of simple repeats bound by ZNF519 showing a large proportion are CG enriched. **c**. ZNF519 peaks and summits at fragile sites and their associated genes, expanded repeats are shown with a red outline. Also included are tracks showing ENCODE layered H3K4me3 signal, ChIP signal from KZNFs mentioned in figure 2 and KAP1 signal. All chip tracks scaled 0-50.
ZNF519 binds a large number of CGG repeats across the genome and could play a role in their stability

Based on our observations of ZNF519 binding at GC rich regions and CGG repeats prone to expansion, we further analysed ZNF519 binding on a genomic level. Given the potential underrepresentation of CGG repeats in the reference genome, a new list of CGG repeats was generated using the tandem repeat finder (Benson 1999); the new list contained 6384 CGG repeats. 3661 (~57%) CGG repeats were found to overlap with a ZNF519 peak (Figure 4a) and the unbound CGG repeats show a clear absence of ZNF519 binding (Figure 4b). Around 40% of the ZNF519-bound CGG repeats were found in gene promoter regions, defined as -5000bp and +1000bp of a transcription start site. A gene ontology analysis for these genes showed significant tissue enrichment in the brain (Figure 4c). Analyses of ZNF519 binding sites in general also showed an enrichment for brain related genes (Farmiloe et al, submitted).

A link between CGG repeats and neuronal genes was also recently explored in a study that investigated polymorphism levels for over 6000 CGG repeats genome-wide across 544 healthy individuals (Annear et al. 2021). They found that 410 out of 1400 genes implicated in neurological disease were associated with polymorphic CGG repeats. This implies that pathogenic repeat expansions could play a role in more neurological diseases than are currently documented in the literature. Of the 410 genes from this list, 195 (48%) also have a promoter CGG/ZNF519 binding site. Based on this data, we investigated whether ZNF519 binding may be associated with the CGG repeat length stability at those loci. The cohort polymorphism scores for CGG repeats which overlapped with ZNF519 binding sites were compared with the polymorphism scores from CGG repeats with no ZNF519 binding. The ZNF519-bound CGG repeats showed increased levels of polymorphism compared to those without ZNF519 binding (Figure 4d). Whereas this suggests that ZNF519 binding associates with the length stability of ZNF519-bound CGG repeats, the mechanism of involvement of ZNF519 remains elusive. This implies that CGG repeats with a ZNF519 binding site are more unstable and more likely to expand or contract. In addition to the increased levels of polymorphism we also observed a higher proportion of 5’UTR regions in the fraction of ZNF519 bound CGG repeats (Figure 4e) suggesting gene-regulatory implications of the polymorphisms seen at those CGG repeat-loci.
The role of primate-specific ZNF519 in CGG-repeat mediated gene regulation

Figure 4: Genome wide overlap of CGG repeats and ZNF519

a, b ChIP density plots showing ZNF519 binding at CGG repeats which overlap with ZNF519 peaks (n=3661) (a) and CGG repeats which do not overlap with ZNF519 peaks (n=2723) (b). c, Results from Gene ontology analysis of promoter CGG repeats with a ZNF519 peak. d, Collective levels of cohort polymorphism in CGG repeats which do not (white) and do (grey) overlap with a ZNF519 peak (without ZNF519 n=X, with ZNF519 n=2439), **** = P<.0001, Wilcoxon rank sum test with continuity correction. e, Percent of CGG repeats without ZNF519 peak (white) and with ZNF519 peak (grey) and their genomic annotations.
ZNF519 has a repressive effect at bound, CGG-associated genes

To further investigate whether the observed binding of ZNF519 to CGG repeats in neuronal promoters has an effect on neuronal gene expression, we analysed cortical tissues derived from human ESCs where ZNF519 was genetically deleted (ZNF519-KO) ZNF519-KO cells and wild type (WT) controls were directed into cortical organoids, which were harvested after 5 weeks in culture and analysed by RNA-seq. We first investigated the effect of loss of ZNF519 on the 195 genes implicated in neurological diseases that contain a ZNF519 bound CGG repeat. A total of 14 genes were differentially expressed (adj p<0.01) (Figure 5a; Table 3), and all of them were upregulated in the ZNF519-KO, further supporting the repressive ability of ZNF519 at some genes important for neuronal development and maintenance. The biggest difference was observed for a gene called KCNK9, which has a ZNF519-bound CGG repeat in the promoter region and is upregulated in the ZNF519 KO (Figure 5b). Whereas this indicates that some individual genes like KCNK9 and FOXP1 may be under control of ZNF519 through a CGG repeat in their promoter, no overall significant difference was observed between the groups of all promoter-CGG containing genes with or without ZNF519 binding (Figure 5c).

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Base.mean</th>
<th>Log2 FC</th>
<th>P adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOXP1</td>
<td>2658.6</td>
<td>0.380</td>
<td>1.18E-14</td>
</tr>
<tr>
<td>KCNK9</td>
<td>298.7</td>
<td>0.662</td>
<td>6.21E-13</td>
</tr>
<tr>
<td>GNAO1</td>
<td>5919.9</td>
<td>0.186</td>
<td>2.81E-06</td>
</tr>
<tr>
<td>RYR2</td>
<td>912.7</td>
<td>0.285</td>
<td>0.000178</td>
</tr>
<tr>
<td>PPP1CB</td>
<td>3392.4</td>
<td>0.170</td>
<td>0.0004708</td>
</tr>
<tr>
<td>ODC1</td>
<td>5618.7</td>
<td>0.151</td>
<td>0.0004899</td>
</tr>
<tr>
<td>SOX4</td>
<td>25160.4</td>
<td>0.129</td>
<td>0.0005061</td>
</tr>
<tr>
<td>PGAP1</td>
<td>7079.5</td>
<td>0.135</td>
<td>0.0009039</td>
</tr>
<tr>
<td>MAP2K1</td>
<td>360.6</td>
<td>0.319</td>
<td>0.0024367</td>
</tr>
<tr>
<td>SLC4A4</td>
<td>631.7</td>
<td>0.264</td>
<td>0.0026575</td>
</tr>
<tr>
<td>USP9X</td>
<td>4768.6</td>
<td>0.133</td>
<td>0.0043488</td>
</tr>
<tr>
<td>RARB</td>
<td>199.3</td>
<td>0.339</td>
<td>0.0067473</td>
</tr>
<tr>
<td>FRAS1</td>
<td>850.3</td>
<td>0.231</td>
<td>0.0074015</td>
</tr>
<tr>
<td>VAMP2</td>
<td>2267.1</td>
<td>0.156</td>
<td>0.0094974</td>
</tr>
</tbody>
</table>

Table 3: Differentially expressed genes in ZNF519 KO
Differentially expressed genes involved in neurological disorders after ZNF519 KO in cortical organoids, genes with an adjusted p-value <0.01 shown, expression levels from DESeq2 analysis shown under ‘Base.mean’, Log2 fold change shown under ‘Log2 FC’.
We next performed RNA-seq analysis on RNA isolated from ZNF519 overexpressing HEK293 cells and investigated the effect of ZNF519 on the bulk expression of all promoter-CGG genes with or without ZNF519 binding. We found a significant decrease in the expression of genes with a ZNF519-bound CGG-containing promoter when compared to the genes without a ZNF519 binding site in ZNF519 overexpression conditions (Figure 5d). This suggests that ZNF519 could be playing a repressive role on the promoters of these genes, at least under the conditions of over expression. There are two possible explanations for the different results of the transcriptome analysis of ZNF519-KO organoids and ZNF519 overexpression: First, the analyses were done in two different cell types, and secondly, endogenous levels of ZNF519 may be low which reduces the effect of ZNF519 ablation, while ZNF519 overexpression may result in abnormally high ectopic levels of ZNF519. While the endogenous regulatory influence of ZNF519 remains elusive, our data suggests that in tissues with high levels of endogenous ZNF519, its binding to CGG containing promoters may affect gene expression. Given the recent emergence of ZNF519 in primate genomes and the CGG binding characteristics that seems specific for ZNF519, our data reveals ZNF519 as a potential player in CGG-mediated gene regulation, but functional follow up studies will be important to elucidate the mechanism and relevance of its involvement.
Figure 5: Effect of ZNF519 on the expression of CGG-associated genes

**a,** Volcano plot showing Log2 fold change and significance from DESeq2 analysis of CGG-associated genes involved in neurodevelopment in a comparison between ZNF519 knockout and wild type day 35 cortical organoids. Points in light blue represent genes that pass the threshold of fold change >1.5 and p-value <0.01, dark blue points did not pass this threshold. **b,** RNA-seq signal and locus representation of KCNK9. **c, d,** boxplots comparing the log2 fold change of expressed (baseMean >10), genes with a promoter ZNF519 peak and CGG repeat (grey) with genes with a promoter CGG repeat but no ZNF519 peak (white) in **c** ZNF519 KO of 5 weeks old (bound n=1273, unbound n=1836), **d** ZNF519 overexpression (bound n=1235, unbound n=1687), **=**P<0.01, Wilcoxon rank sum test with continuity correction. Red line shows 95% CI of 10,000 times bootstrapped median of a set of unbound genes with the same sample size as the target genes. Individual data points not shown.
Discussion

Our analyses show a clear preference of ZNF519 binding at CGG repeats and C/G rich sequences. We show that ZNF519 is able to bind to a large number of C/G based repeats, recognises CGG containing fragile sites, and 5 repeats associated with well documented REDs. The high proportion of ZNF519-bound CGG repeats in gene promoters, and specifically their enrichment in genes associated with neural disease suggests that ZNF519 could have a function at these loci and influence neuronal gene expression. Whereas we show functional data that support a direct role for ZNF519 in gene expression, further work is required to truly elucidate the mechanism of its involvement.

Our analysis also revealed that ZNF519 binding is mildly associated with the presence of length-polymorphisms of CGGs in neurological disorder associated genes. In that sense, the extent of ZNF519’s regulatory influence may differ between individuals depending on the individual’s repeat size. This also raises the question of what the role is for ZNF519 at CGG repeats that are expanded to pathology-associated sizes. It is known that expansion of STRs such as CGG can trigger silencing (He and Todd 2011), a process which could involve ZNF519.

There are many questions to be answered about short tandem repeats and why a high level of variability is tolerated in genomes despite the potentially pathogenic effects of repeat expansions on gene expression. It is possible that the situation is one of antagonistic pleiotropy (Williams 1957), where highly variable CGG repeats which have the potential to become pathogenic are tolerated in the genome because their variability conveys an advantage. CGG repeats themselves are sources of variation located mainly in gene promoters which could allow for rapid evolution and tweaking of gene expression. ZNF519 could have an important role in this evolutionary tweaking process, but further analyses are needed to fully understand this relationship between ZNF519, CGG repeats and variation in repeat length.
Chapter 5

Materials and methods

Analysis of ChIP-seq data and KZNF binding sites
MACS2 peaks from ChIP-Seq data generated by Imbeault et al. (GEO accession GSE78099), Schmitges et al. (GEO accession GSE76496), and Najafabadi et al (GEO accession GSE52523) for 203 KZNFs and their binding sites were used. The data was visualised using the UCSC genome browser and a track hub was produced. Manual observation was used to identify individual KZNFs with binding peaks on the tandem repeat involved in the pathological expansion of 17 genes associated with well-known neurological disorders, and an additional 6 associated with chromosomal fragile sites.

ChIP-Seq data processing
ChIP-seq data from Imbeault et al (2017) was reprocessed to generate bigWig files for further visualisation and MACS2 summits for analyses. The public US Galaxy server (Afgan et al. 2018), usegalaxy.org) was used for processing data. Reads were trimmed and adaptor sequences removed using trimmomatic (Bolger et al. 2014) version 0.36.5 for paired-end reads (ILLUMINA CLIP TruSeq3 paired-end), cutting if average per base quality in a 4-base sliding window was below 20, the CROP setting was set to 100 bases. Trimmed reads were mapped to hg19 with Bowtie2 using the very-sensitive-end-to-end setting (Langmead and Salzberg 2012). Bigwig files were generated with the bamCoverage tool without any normalisation or scaling (Ramírez et al. 2016). MACS2 was used to call peaks and summits using the default settings (Zhang et al. 2008).

Simple repeat analysis
MACS2 summit files for the 51 KZNFs which bind >50 gene promoters were filtered for a signal value >10. The summit coordinates were then intersected using the UCSC table browser (Karolchik et al. 2004) with the Repeatmasker dataset to give an output of all repeats intersecting with KZNF summits (Smit et al.). These data were processed in excel to generate pie charts showing the breakdown of bound repeats. Based on this breakdown, KZNFs were divided into three groups. To make ‘archetypal’ pie charts, averages were taken across each of the three groups.

CGG list generation
A list of CGG repeats was generated using tandem repeat finder (TRF) (Benson 1999). The genome assembly GRCh38/hg38 was downloaded from the UCSC browser in fasta format. TRF was run on these files with parameters: MatchScore: 2, MismatchScore: 5, IndelScore: 17, PM: 80, PI:10, Minscore: 24, Maxperiod: 3.
The resulting VCF was parsed with a custom awk script and loaded in R. The output was further filtered for all reading frames of CGG: (CCG|CGC|CGG|GCC|GCG|GGC) and all alternative chromosomes were filtered out. The TRF output was finally transformed in BED format.

**ZNF519, CGG, promoter intersection**
ZNF519 peaks published in Imbeault (2017) were intersected with TRF generated CGG repeat coordinates using the UCSC table browser to generate a list of CGG repeats which overlapped with a ZNF519 peak. The list of CGG repeats was also intersected with promoter regions generated by taking 5000bp upstream and 1000bp downstream of transcription start sites from the hg19 ensembl gtf file downloaded from the UCSC genome browser to generate a list of CGG repeats in promoter regions with and without a ZNF519 summit. For the polymorphism analysis ZNF519 MACS peaks generated by Imbeault (2017) were intersected with the list of CGG repeats generated by Annear et al (2021). ChIP density plots were generated with the list of CGGs with and without a ZNF519 summit using computeMatrix (Galaxy Version 3.1.2.0.0) and plotHeatmap (Galaxy Version 3.1.2.0.1) (Ramírez et al., 2016).

**CGG polymorphism analysis**
CGG start locations and polymorphism levels were downloaded from Annear et al (2021). CGG start locations were then intersected with ZNF519 peaks from Imbeault et al (2017). This generated a list of CGGs with and without an overlap with a ZNF519 peak. These cohort polymorphism scores from these lists were used to generate a boxplot in R using the ggplot2 package (Wickham 2009). A Wilcoxon rank sum test with continuity correction was performed to test significance.

**Neurological disorders analysis**
The list of 410 genes associated with neurological disease was taken from Annear et al (2021) and intersected with the list of genes containing a ZNF519-bound promoter CGG repeat. The p-value and log2 fold change from the DESeq2 differential analysis for the resulting 195 genes were plotted on a volcano plot using ggplot2 in R (Wickham 2009).

**Gene expression analysis**
RNA sequencing data were obtained from 5 week old (35 days) cortical organoids, generated from a ZNF519KO and wild type cell line and from a ZNF519 OE experiment in HEK293 cells described in Chapter 3. Previously generated DESeq2 results tables were used for this analysis. Gene lists from the intersection described previously were used for comparison of genes with promoter CGG
Chapter 5

repeats with and without ZNF519 peaks and genes with promoter CGG repeats vs genes without.

Log2FC of bound and unbound genes was compared using R (R Core Team 2019) and visualised using ggplot2 (Wickham, 2009). The number of bound and unbound genes expressed differed greatly and so the 95% confidence interval (CI) of the median log2FC of unbound genes was calculated by 10,000 times bootstrapping the median of a random set of unbound genes with a similar sample size as the bound genes. For comparison of log2 fold change of expressed target genes vs non-target genes, a Wilcoxon rank sum test with continuity correction was performed.
References


Chapter 5


The role of primate-specific ZNF519 in CGG-repeat mediated gene regulation


