Published in final edited form as:

Neurogenetics. 2018 August ; 19(3): 135–144. doi:10.1007/s10048-018-0546-8.

Toward deciphering the mechanistic role of variations in the Rep1 repeat site in the transcription regulation of *SNCA* gene

A. Afek^{1,2}, L. Tagliafierro^{2,3}, O.C. Glenn^{2,3}, D.B. Lukatsky⁴, R. Gordan^{1,2,5,*}, and O. Chiba-Falek^{2,3,*}

¹Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA

²Center for Genomic and Computational Biology, Duke University Medical Center, Durham, NC 27710, USA

³Department of Neurology, Duke University Medical Center, Durham, NC 27710, USA

⁴Department of Chemistry, Ben-Gurion University of the Negev, Be'er Sheva 8410501 Israel

⁵Department of Computer Science, Duke University, Durham, NC 27705, USA

Abstract

Short structural variants – variants other than single nucleotide polymorphisms – are hypothesized to contribute to many complex diseases, possibly by modulating gene expression. However, the molecular mechanisms by which noncoding short structural variants exert their effects on gene regulation haven't been discovered. Here, we study simple sequence repeats (SSR), a common class of short structural variants. Previously, we showed that repetitive sequences can directly influence the binding of transcription factors to their proximate recognition sites, a mechanism we termed non-consensus binding. In this study, we focus on the SSR termed Rep1, which was associated with Parkinson's disease (PD) and has been implicated in the *cis*-regulation of the PD-risk *SNCA* gene. We show that Rep1 acts via the non-consensus binding mechanism to affect the binding of transcription factors from the GATA and ELK families to their specific sites located right next to the Rep1 repeat. Next, we performed an expression analysis to further our understanding regarding the GATA and ELK family members that are potentially relevant for *SNCA* transcriptional regulation in health and disease. Our analysis indicates a potential role for GATA2, consistent with previous reports. Our study proposes non-consensus transcription factor binding as a potential mechanism through which noncoding repeat variants could exert their pathogenic effects by regulating gene expression.

*Co-corresponding authors: Ornit Chiba-Falek, Dept of Neurology, DUMC Box 2900, Duke University, Durham, North Carolina 27710, USA., Tel: 919 681-8001, Fax: 919 684-6514, o.chibafalek@duke.edu. Raluca Gordan, Depts of Biostatistics & Bioinformatics and Computer Science, DUMC Box 3382, Duke University, Durham, North Carolina 27708, USA., Tel: 919 684-9881, Fax: 919 668-0795, raluca.gordan@duke.edu.

CONFLICT OF INTEREST

The authors declare that they have no competing financial interests.

INTRODUCTION

Simple Sequence Repeats (SSRs; also known as microsatellites or short tandem repeats), including homopolymer stretches, are one of the most polymorphic and abundant repeat classes in the human genome. Recently, there has been increased support for the idea that SSRs, which are largely ignored in large-scale genetic studies, may be involved in many human complex diseases, and may also contribute significantly to variation in gene expression in humans[1–4]. SSRs are thought to affect phenotypes by altering the regulation of gene transcription[5–10], splicing[11], and translation, and it is by these mechanisms that SSRs may play a role in the etiology of human diseases. However, the roles of noncoding SSRs in human complex diseases, and specifically the mechanisms whereby SSRs regulate gene expression and exert their pathogenic effects, have yet to be discovered.

Genetic variations associated with Parkinson's disease (PD)-risk demonstrated associations with *SNCA* (MIM: 163890) expression levels. A prominent example is the Rep1 polymorphisms, a complex SSR associated with PD-risk in diverse populations[12]. Rep1[6] is a polymorphic dinucleotide repeat site ((TC)_nTT(TC)_m(TA)_x(CA)_z where n=10–11, m=9–11, x=7–9, and z=10–13) located ~10 kb upstream of the *SNCA* transcription start site[13, 14]. The length of Rep1 appears to be associated with increased risk of PD[15]. In particular, most association studies, including a large meta-analysis, have shown that the extended alleles of Rep1 confer increased risk to develop late-onset, 'idiopathic' PD, while the shorter allele is protective[12, 16–19]. We investigated the effect of the Rep1 polymorphism on *SNCA* expression, and discovered that Rep1 regulates *SNCA* transcription in human brain tissues[20]. These results have been confirmed using luciferase reporter assay[6, 21] and in a humanized mouse model[22]. The PD-risk Rep1 allele led to increased *SNCA*-mRNA levels, providing further support for the pathogenic effect of *SNCA* overexpression[6, 20, 22]. Retrospectively, the Rep1 site was assessed using the SSV evaluation system and had a high score (Total Impact Score=31) relative to the complete list of SSRs within *SNCA* genomic regions including +/-50kb flanking regions[23]. However, the mechanism underlying the effect of Rep1 variants on *SNCA* transcription remains elusive.

Recently, Afek et al. showed that certain repetitive DNA sequences, when present in the flanking regions of specific transcription factor (TF) binding sites, can significantly affect the intensity of TF-DNA binding through a mechanism termed “non-consensus binding”[24, 25]. In particular, Afek et al. showed that the magnitude of the non-consensus effect can reach 66% of consensus TF-DNA binding, and that non-consensus binding free energy depends on the exact repeat type and on the length of repeated DNA sequence elements[25]. However, the role of the TFs non-consensus binding mechanism in the context of the human genome has not been investigated, and its potential implication in the regulation of human gene expression has yet to be discovered.

Here, we propose that the mechanistic role of Rep1 in the differential transcriptional regulation of the *SNCA* gene may be mediated through non-consensus binding of TFs. Using an *in vitro*, cell-free system for measuring TF-DNA interactions quantitatively and in high throughput[25, 26], we studied whether the variable alleles at the Rep1 site show

differential non-consensus binding of TFs. We focus on two TF families, GATA and ELK, because 1) they have consensus binding sites in very close proximity to the Rep1 region, and 2) they have been reported to be expressed in brain and/or play a role in brain development. We found that variable alleles at the Rep1 locus drive differential binding affinities for both TF families. Using expression analysis we further our understanding regarding the GATA and ELK family members that are potentially relevant to *SNCA* transcriptional regulation in health and disease.

MATERIALS AND METHODS

Human tissue samples

The TF expression analysis of the human tissues consisted of temporal cortices from individuals with three autopsy-confirmed neuropathological diagnoses: (1) Parkinson's disease (PD; N=3), (2) Dementia with Lewy body (DLB; N=3), (3) clinically and neuropathologically normal subjects (N=3); and blood samples from cognitively healthy subjects (N=3). The projects were approved by the Duke Institutional Review Board and followed appropriate ethical protocols, and all experiments were performed in accordance with relevant guidelines and regulations. The brain tissues were obtained through the Layton Aging & Alzheimer's Disease Center at Oregon Health and Science University. Subjects (or their Legally Authorized Representatives) provided informed consent for their data and brain specimens to be used in future research.

RNA extraction and cDNA synthesis

Total RNA was extracted from human temporal cortex, blood samples, and iPSC-derived cells, using TRIzol reagent (Invitrogen) followed by purification with an RNeasy kit (Qiagen) used per the manufacturer's protocol. RNA concentration was determined spectrophotometrically at 260 nm, while the quality of the purification was determined by 260 nm/280 nm ratio that showed values between 1.9 and 2.1, indicating high RNA quality.

cDNA was synthesized using MultiScribe RT enzyme (Applied Biosystems) using the following conditions: 10 min at 25°C and 120 min at 37°C.

Real time PCR

Real-time PCR was used to quantify the levels of the transcription factors (TFs). Briefly, samples were assayed by relative quantitative real-time PCR using TaqMan expression assays and the ABI QuantStudio 7 to determine the level of the message RNA (mRNA) of the TFs relative to mRNAs encoding housekeeping genes. ABI MGB probe and primer set assays (Applied Biosystems) that were used to amplify the target transcription factors and the house-keeping reference controls are listed in (Supplementary Table S1). Each cDNA (10 ng) was amplified in duplicate in at least two independent runs (overall 4 repeats), using TaqMan Universal PCR master mix reagent (Applied Biosystems) and the following conditions: 2 min at 50 °C, 10 min at 95 °C, 40 cycles: 15 sec at 95 °C, and 1 min at 60°C. As a negative control for the specificity of the amplification, we used RNA control samples that were not converted to cDNA (no-RT) and no-cDNA/RNA samples (no-template) in each plate. No amplification product was detected in control reactions. Data were analyzed with a

threshold set in the linear range of amplification. The cycle number at which any particular sample crossed that threshold (Ct) was then used to determine fold difference, whereas the geometric mean of the two control genes served as a reference for normalization. Fold difference was calculated as $2^{-[\text{Ct}(\text{TF-target}) - \text{Ct}(\text{geometric mean of reference})]}$. $\text{Ct} = [\text{Ct}(\text{sample})] - [\text{Ct}(\text{calibrator})]$. The calibrator, a particular RNA sample obtained from a control cortex, was used repeatedly in each plate for normalization within and across runs. The variation of the Ct values among the calibrator replicates was smaller than 10%.

Statistical association analyses

Expression measurements of the different TF-mRNAs in human tissues represent three subjects each repeated four times, and in iPSC-derived cells represent four biological and technical replicates. Statistical significance of differences in expression between the samples for each TF was analyzed by pairwise comparisons using the Tukey-Kramer HSD test (JMP Pro13, SAS).

Recombinant protein expression and purification

Gateway-compatible clones containing the full-length genes for GATA family protein Gata1 and ELK protein Elk1 were purchased from GeneCopoeia and the genes were transferred into pDEST15 using the LR Clonase reaction (Life Technologies). This vector enables the production of N-terminally GST-tagged proteins. Cells were grown in LB broth to an OD₆₀₀ of 0.4 to 0.6 and then expression was induced with IPTG at 30° to 37° C. Pelleted cells were frozen, stored at -20 C, and thawed cells were lysed with lysozyme. GST-tagged protein was purified from the soluble portion of the lysate using GST resin (GE Healthcare) according to manufacturer's instructions.

Design of the DNA library for protein-DNA binding assays

Custom protein-DNA binding experiments based on the protein-binding microarray (PBM) technology [25, 26, 33, 51, 52] can be used to test TF binding to up tens of thousands of custom sequences of up to 36bp long. Here, we designed a custom DNA library based on the Rep1 genomic region (chr4: 90767174-90767251, GRCh37/hg19). Taking 50bp upstream of Rep1 plus 10bp of the repeat region, and showing it on the reverse strand, the genomic sequence is

CACACACACGGAAGGCATCAGATATCTCATGTGTGTATACACATACATATATATA
GG, where the two bold underlined sequences are core binding motifs for GATA and ELK, respectively, and the italic sequence represents the first 10 bases of the Rep1 region. On the reverse strand, the Rep1 region is (TC)_nTT(TC)_m(TA)_x(CA)_z, where n=10–11, m=9–11, x=7–9, and z=10–13. Since probes in custom PBM libraries are limited to 36bp, in order to test the influence of Rep1 repeats on binding of GATA and ELK to DNA we designed sets of probes that contain the GATA and/or ELK sites in their immediate genomic flanks, followed by all possible combinations of repeats based on the Rep1 alleles. Specifically, we designed the following sets of probes:

1. Probes with no specific binding sites, to use as negative controls. These probes contain all possible overlapping 36mers generated from all possible 144 combinations of the Rep1 repeats.
2. “ELK binding site” probes, which contain the ELK core binding site in its immediate genomic context (CGGAAGGCATCAG) preceded by 23bp sequences designed according to all possible combinations of repeats found in Rep1 alleles.
3. “GATA binding site” probes, which contain the GATA core binding site in its immediate genomic context (ATCAGGATATCTCATG) preceded by 21bp sequences designed according to all possible combinations of repeats found in Rep1 alleles.
4. “ELK+GATA binding sites” probes, which contain both the ELK and the GATA core binding sites in their genomic context (CGGAAGGCATCAGATATCTCATG) followed by 13bp sequences designed according to all possible combinations of repeats found in Rep1 alleles.
5. Probes designed from another repeat region potentially relevant for *SNCA* expression, which we call the “Pure CA” region (chr4:90636808-90636841, GRCh37/hg19). A consensus binding site for the GATA family is located in close proximity to this region:

...CACACACACACTTTT**ATATCTTTTC** (10 bases of the CA repeat are shown in italic font, the reverse complement of the GATA core binding site is shown in bold underlined font). Based on this repeat region, we designed probes containing the genomic region upstream of the repeat, followed by CA repeats of different length, and by the GATA site in its immediate flanking sequence.

For all probes, the 36bp custom sequence was followed by a constant 24bp sequence (GTCTTGATTGCTTGACGCTGCTG) complementary to a primer used for primer extension in order to double-strand the DNA (see below).

***In vitro* protein-DNA binding assays**

Custom protein-DNA binding experiments were carried out following the standard protein-binding microarray (PBM) protocol [26, 33] but using the custom DNA library described above. Briefly, we first performed primer extension to obtain double-stranded DNA oligonucleotides on the microarray. Next, each microarray chamber was incubated with a 2% milk blocking solution for 1 h, followed by incubations with a PBS-based protein binding mixture [33] for 1 h (containing either 75nM GATA1 or 65nM ELK1 recombinant protein) and with Alexa488-conjugated anti-GST antibody (1:40 dilution, Invitrogen A-11131) for 1 h. The array was gently washed as previously described [33] and then scanned using a GenePix 4400A scanner (Molecular Devices) at 2.5-μm resolution. All blocking and washing solutions contained 50μM zinc acetate, as recommended for zinc-finger proteins, including GATA [33]. Data were normalized with standard analysis scripts [26, 33]. The normalized data were further processed to compute median values over

replicate spots containing identical DNA sequences. The raw and processed data generated in this study is currently being submitted to GEO.

Data Availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information files), or available from the corresponding author on reasonable request.

RESULTS

Two potential transcription factors binding sites are located in very close proximity to the Rep1 repeat region

We analyzed the genomic region in the neighborhood of Rep1 to identify putative TF binding sites, using the TF binding specificities databases UniPROBE and cisBP[27, 28]. We identified an ELK binding site (CGGAAGGC) located right downstream of the Rep1 region (on the reverse strand, as *SNCA* is encoded on the reverse stand), and a GATA binding site (AGATATCTCA) located 3 nucleotides downstream of the ELK site (on the reverse strand). ELK proteins, which include ELK1, ELK3, and ELK4 in human, are part of the ETS structural family of transcription factors. All three ELK human proteins bind to sequences containing the TTCC/GGAA core motif, and their DNA-binding preferences are indistinguishable according to existing binding data[29]. Thus, we selected one of the ELK proteins, ELK1, to investigate the effects of repeat sequences on ELK binding to DNA. The GATA protein family also contains multiple TFs (GATA1 through GATA6) with indistinguishable DNA-binding specificities. Thus, as a representative of the GATA family, we selected GATA1 to investigate the effects of repeats on its DNA binding affinity. Interestingly, members of both GATA and ELK transcription factor families play important roles in brain development and function[30] and some have even been shown to regulate the expression of the *SNCA* gene[31, 32], albeit through regulatory regions distinct from the Rep1 region studied here. Thus, the proximity of ELK and GATA binding sites to the Rep1 region, together with the known effects of Rep1 alleles on *SNCA* gene expression, led us to hypothesize that the length of different repeat units that comprise the Rep1 region could have a direct influence on ELK and/or GATA binding at this genomic site. We tested this hypothesis using a quantitative, cell-free protein-DNA binding assay.

Briefly, we designed a custom DNA library containing the GATA and ELK binding sites identified in close proximity to the Rep1 region, flanked by Rep1 repeat elements of different length (Methods). We note that the genomic region covering the Rep1 site together with the GATA and ELK binding sites is relatively long (74–90bp for Rep1, plus 19bp for the GATA/ELK binding sites), which makes it difficult to obtaining sensitive TF binding measurements for the different Rep1 alleles. Thus, in order to determine whether the DNA repeat elements in the Rep1 regions have the potential to quantitatively affect TF binding, we tested each pair of consecutive, variable Rep1 repeat elements, *i.e.* $(TC)_n TT(TC)_m$, $(TC)_m (TA)_x$, and $(TA)_x (CA)_z$, for their effects on GATA and/or ELK binding (Methods). We synthesized the DNA oligos on a chip and we used the protein-DNA binding microarray technology (PBM)[26, 33] to quantitatively measure binding of recombinant GATA and

ELK proteins to thousands of sites containing Rep1 repeat elements of different lengths. First, though, we confirmed that GATA and ELK do indeed bind with high specificity to the consensus sites we identified next to the Rep1 region (Fig. 1). We note that also PBM assays measure TF-DNA binding in a cell-free system, previous work from our group and others has demonstrated that this technology can reveal new insights into TF binding mechanism in living cells[34, 35]

Repetitive elements in the Rep1 region have significant effects on GATA and ELK binding to DNA

Next, we analyzed the GATA and ELK binding signals for probes containing repeats of different lengths. Given the relatively short length of the probes that can be used in quantitative binding assays, testing the full Rep1 alleles followed by the ELK and GATA sites is not feasible. Instead, for each repeat element in the Rep1 region $(TC)_n TT (TC)_m (TA)_x (CA)_z$ we tested the effect on GATA and ELK binding to their consensus sites flanked by 1, 2, 3, 4, ... repeated units of each type of repeat (CA, TA, and TC). We preserved the genomic context as much as possible, given the technical limitation that probes can contain at most 36 custom base pairs.

Consistent with our hypothesis, we found that the lengths of the Rep1 repeat components have significant effects on GATA and ELK binding to their consensus binding sites (Fig. 2). For all three repeat types (CA, TA, and TC), longer repeats are significantly correlated with higher binding affinities for both GATA (Fig. 2a) and ELK (Fig. 2b), when the repeats are located right next to the transcription factor binding sites. Overall, the effect of the individual repeat types on TF binding in the context of Rep1 is very complex (Supplementary Fig. 1). The structure of Rep1 is complex and contains different repeat components in tandem. Therefore, in the design of our DNA library, increasing the length of one repeat (*e.g.* $(CA)_z$) is done simultaneously with decreasing the length of another repeat (*e.g.* $(TA)_x$), in order to keep the total length of the oligos constant within the native genomic context. Thus, the measured effects on TF binding could come from either, or both, of the repeat components in each probe. Nevertheless, our analysis of a 'pure' CA polymorphic site showed that increasing the length of a CA repeat in the proximity of a GATA consensus binding site leads to increased GATA binding (discussed below and Fig. 3). The TT component in the Rep1 region also adds to the complexity. As shown in Fig. 2 (lower panels), increasing the length of the $(TC)_m$ component leads to increased GATA and ELK binding. However, as the length of the $(TC)_m$ component increases, the TT component moves further apart from the transcription factor binding site. One could consider the TT component as 'breaking' a long $(TC)_{n+m}$ repeat. In this interpretation, the location of the break, relative to the transcription factor binding site, could be causal for the effect we observed on binding of GATA and ELK to DNA. Additional binding studies and biochemical characterization would be needed to further understand the mechanisms by which these sequence repeats have a direct effect on TF binding.

Given the importance of the CA repeat component in the Rep1 region [21, 22], and to further our understanding of the contribution of CA repeats in their native genomic context, we searched for a 'pure' CA polymorphic site that has a potential TF binding site in its close

proximity. We identified a 'pure' CA repeat variant (rs10586902) nearby a GATA site in the *SNCA* gene noncoding flanking region (chr4:90636808-90636841, GRCh37/hg19), positioned within a potentially regulatory sequence (ENCODE). We tested the effect of the CA repeat length on GATA binding in the genomic context of this 'pure' CA variant. Consistent with our results on the Rep1 region, longer CA repeats in this polymorphic region lead to stronger GATA binding to its consensus site located right next to the repeat variant (Fig. 3). This finding is also consistent with our previous *in vitro* and *in vivo* observation that Rep1 alleles with longer CA components drive higher transcription of *SNCA* gene [21, 22].

Characterization of the expression of GATA and ELK families in brain tissues relevant to PD and DLB

While the DNA-binding specificities of GATA and ELK factors are expected to be identical within their respective protein families, each family member may be expressed at different levels across various tissue and cell types. Thus, we next sought to gain insights into which ELK and GATA transcription factors are potentially relevant for the regulation of *SNCA* transcription in the context of synucleinopathies. We characterized the mRNA expression of TFs of the GATA and ELK families in human brain cortical tissues from healthy and disease donors and in blood samples from healthy subjects. We compared the mRNA expression levels among the different neuropathologies, *i.e.* PD, DLB and normal, and between tissues, *i.e.* normal brain vs. blood (summarized in Supplementary Table S2). Each tested transcription factor showed a distinct expression profile across the analyzed tissues (Fig. 4). GATA1, GATA3 and ELK3 were predominantly expressed in the blood ($p < 0.0001$), and no significant differences were observed among the different neuropathologies. GATA2 and GATA6 expression in normal blood were also higher compared to normal cortex but in a lesser extent ($p = 0.005, 0.09$), while ELK1 levels showed an opposite trend. Overall no significant trends were observed among the cortex tissues from the different diagnoses. However, several TFs showed higher mean levels in the cortex from disease compared to normal, particularly DLB > normal, that warrant investigations using a larger sample size (Fig. 4 and Supplementary Table S2).

GATA2 was the most abundant TF in the human brain cortex samples, ~2–3fold increase across each pathology. Thus, we have prioritized GATA2 for further studies using iPSC-derived neurons from an apparently healthy control (Control), and from a patient with the triplication of the *SNCA* gene (*SNCA*-tri). GATA2 expression was elevated in mature dopaminergic and cholinergic neurons derived from disease and normal iPSC lines compared to neuronal precursor cells (NPCs), but no further changes were observed upon inducing aging (Supplementary Fig. S2a). Consistently, *SNCA*-mRNA expression increased throughout neuronal maturation, *i.e.* NPC < mature neurons. However, although induce aging did not affect GATA2 expression, with the exception of normal cholinergic neuron, the expression of *SNCA*-mRNA slightly further increased in the aged neurons (Supplementary Fig. S2b). Noteworthy, in each stage (NPC, mature neurons, and aged) for both iPSC sources (patient and normal) GATA2 expression were significantly higher ($p < 0.0001$) in the cholinergic lineage compared to the dopaminergic cells (Supplementary Fig. S2). Thus,

GATA2 emerged as the top candidate transcription factor for future follow up studies, as discussed below.

DISCUSSION

Changes in expression levels of specific genes in the brain are associated with late-onset neurodegenerative diseases, and these expression changes are genetically regulated [36–38]. A prominent example is the *SNCA* gene and the contribution of alteration in its expression to the etiology of synucleinopathies [36, 39–42], [38, 43–45]. Therefore, it is highly important to reveal the genetic mechanisms underpinning the differential expression of *SNCA* and other key genes. Transcription is one of the mechanisms that contribute to gene expression and its involved interaction between *cis* elements – DNA sequences – and *trans* factors – transcription factors. The simple sequence repeat (SSR) region Rep1 has been suggested to act as an enhancer element of *SNCA* gene. Here, we identified potential TFs, GATA and ELK families, that bind proximate to Rep1 and whose binding affinities are influenced directly by the Rep1 sequence. Although our results do not exclude the influence of other cellular factors (such as nucleosomes), they do suggest that the role of Rep1 in modulating *SNCA* transcription may be mediated, at least in part, by its effect on GATA/ELK binding via the non-consensus mechanism.

Studies using human brain tissues[20], luciferase reporter assay[6, 21] and humanized mouse model[22] showed the role of Rep1 in the regulation of *SNCA*-mRNA levels. However, these results conflict with the recent findings that showed no effect of Rep1 on *SNCA*-mRNA level using genome edited iPSC-derived neurons [46]. It is possible that difficulties in quantifying the total *SNCA* transcripts levels affected the validity of the reported conclusions. Therefore, we cannot rule out the possibility of a false negative for the *SNCA*-Rep1 finding reported by Soldner et al [46]. Low abundance of relevant TFs in their cellular model system may explain the challenges they encountered to detect an effect.

Noteworthy, other groups have shown that GATA TFs bind to elements within intron 1 [31] and the promoter [32] of *SNCA*, where they have been proposed to play a role as inducers of transcription. Consistently with these previous reports, we showed that, among the GATA family members, GATA2 is highly abundant in human cortex, suggesting that GATA2 plays a role in the regulation of *SNCA* transcription. Thus, we have prioritized GATA2 for further studies and conducted analysis of expression profiles in iPSC derived dopaminergic and cholinergic neurons as *in vitro* models for PD and DLB, respectively [47]. We showed higher levels in cholinergic derived cells, inferring that GATA2 may exert a differential contribution to the etiologies of DLB vs PD that warrants in depth investigations using single cells analyses and model systems.

Previously we showed that Rep1 plays a role in the transcriptional regulation of *SNCA*, and that the variable alleles led to differential expression of the *SNCA* gene [6, 20, 22] [21]. Here we investigated the mechanism underlying the transcription role of Rep1. We dissected the extent of the effect of each component of the Rep1 sequence, and found that the CA component contributes substantially to the non-consensus binding. These results are consistent with our previous finding showing the contribution of the CA length to the levels

of *SNCA*-mRNA in human brains[20], humanized mice[22] and luciferase reporter system[6] [21].

The results of this *in vitro* study suggest that Rep1 variants have the potential to significantly affect binding of TFs to their genomic regulatory sites, via non-consensus binding mechanism, which in turn could regulate *SNCA*-mRNA expression levels. The direct impact of the non-consensus TF binding of Rep1 on *SNCA*-mRNA levels warrants further investigations using *in vivo* model systems. However, such investigations are not trivial. The genomic binding of TFs in the cell is typically tested with chromatin immunoprecipitation (ChIP) assays, however, traditional ChIP assays are not sufficient to directly compare TF binding between samples[48] (in our case, samples with different Rep1 alleles). Furthermore, robust *in vivo* testing will require the generation of multiple *isogenic* cell-lines with the various Rep1 alleles. Therefore, *in vivo* testing of all the variants tested in our cell-free binding assay may not even be feasible. Nevertheless, our comprehensive cell-free assay provided the initial mechanistic evidence, as well as good candidates TF and repeat elements. These results established the foundation for further testing in future follow-up studies using a new developed cell-based system in combination with sensitive ChIP based assays.

In conclusion, using the Rep1 repeat variant upstream of the *SNCA* gene as a prototype, we provided initial support for the proposed general phenomenon that differential non-consensus binding of transcription factors (TFs) to the variable alleles at repeat sequence sites is likely to be responsible, at least in part, for differential expression of the *cis*-regulated genes. In fact, we have initiated follow-up experiments for two additional polymorphic SSR in *SNCA* region that are likely to be functional [23, 49], both of which exhibit strong potential binding sites for neuronal TFs +/- 50bp of the repeat site and are strong candidates for the non-consensus binding effect. Our study serves as a model for future investigations aiming to explore the effect of other noncoding polymorphic repeat sequences on transcriptional regulation in the human brain, in relation to neurodegenerative diseases in aging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Layton Aging & Alzheimer's Disease Center at Oregon Health and Science University for providing us with the brain tissues, and Dr. Randy Woltjer for his assistance in obtaining the required brain samples for the study.

FUNDING

This work was funded in part by the Holland-Trice award (to O.C-F. and R.G.), the National Institutes of Health/ National Institute of Neurological Disorders and Stroke (NIH/NINDS) [R01 NS085011 to O.C-F.], and the National Institutes of Health/National Institute of General Medical Sciences (NIH/NIGMS) [R01 GM117106 to R.G.].

References

- 1Mirkin SM. Expandable DNA repeats and human disease. *Nature*. 2007; 447:932–40. [PubMed: 17581576]
- 2Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nature reviews Genetics*. 2005; 6:729–42.
- 3Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. Genomes Project C. The landscape of human STR variation. *Genome Res*. 2014; 24:1894–904. [PubMed: 25135957]
- 4Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet*. 2016; 48:22–9. [PubMed: 26642241]
- 5Akai J, Kimura A, Hata RI. Transcriptional regulation of the human type I collagen alpha2 (COL1A2) gene by the combination of two dinucleotide repeats. *Gene*. 1999; 239:65–73. [PubMed: 10571035]
- 6Chiba-Falek O, Nussbaum RL. Effect of allelic variation at the NACP-Rep1 repeat upstream of the alpha-synuclein gene (SNCA) on transcription in a cell culture luciferase reporter system. *Hum Mol Genet*. 2001; 10:3101–9. [PubMed: 11751692]
- 7Okladnova O, Syagailo YV, Trinitz M, Stober G, Riederer P, Mossner R, et al. A promoter-associated polymorphic repeat modulates PAX-6 expression in human brain. *Biochemical and biophysical research communications*. 1998; 248:402–5. [PubMed: 9675149]
- 8Peters DG, Kassam A, St Jean PL, Yonas H, Ferrell RE. Functional polymorphism in the matrix metalloproteinase-9 promoter as a potential risk factor for intracranial aneurysm. *Stroke*. 1999; 30:2612–6. [PubMed: 10582986]
- 9Searle S, Blackwell JM. Evidence for a functional repeat polymorphism in the promoter of the human NRAMP1 gene that correlates with autoimmune versus infectious disease susceptibility. *Journal of medical genetics*. 1999; 36:295–9. [PubMed: 10227396]
- 10Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, et al. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett*. 1999; 455:70–4. [PubMed: 10428474]
- 11Hefferon TW, Groman JD, Yurk CE, Cutting GR. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc Natl Acad Sci U S A*. 2004; 101:3504–9. [PubMed: 14993601]
- 12Maraganore DM, de Andrade M, Elbaz A, Farrer MJ, Ioannidis JP, Kruger R, et al. Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease. *Jama*. 2006; 296:661–70. [PubMed: 16896109]
- 13Xia Y, Rohan de Silva HA, Rosi BL, Yamaoka LH, Rimmler JB, Pericak-Vance MA, et al. Genetic studies in Alzheimer's disease with an NACP/alpha-synuclein polymorphism. *Ann Neurol*. 1996; 40:207–15. [PubMed: 8773602]
- 14Touchman JW, Dehejia A, Chiba-Falek O, Cabin DE, Schwartz JR, Orrison BM, et al. Human and mouse alpha-synuclein genes: comparative genomic sequence analysis and identification of a novel gene regulatory element. *Genome Res*. 2001; 11:78–86. [PubMed: 11156617]
- 15Goldman SM, Umbach DM, Kamel F, Tanner CM. Head injury, alpha-synuclein Rep1 and Parkinson's disease: a meta-analytic view of gene-environment interaction. *Eur J Neurol*. 2015; 22:e75. [PubMed: 26060929]
- 16Kay DM, Factor SA, Samii A, Higgins DS, Griffith A, Roberts JW, et al. Genetic association between alpha-synuclein and idiopathic Parkinson's disease. *Am J Med Genet B Neuropsychiatr Genet*. 2008; 147B:1222–30. [PubMed: 18404644]
- 17Farrer M, Maraganore DM, Lockhart P, Singleton A, Lesnick TG, de Andrade M, et al. alpha-Synuclein gene haplotypes are associated with Parkinson's disease. *Hum Mol Genet*. 2001; 10:1847–51. [PubMed: 11532993]
- 18Mizuta I, Nishimura M, Mizuta E, Yamasaki S, Ohta M, Kuno S. Meta-analysis of alpha synuclein/NACP polymorphism in Parkinson's disease in Japan. *J Neurol Neurosurg Psychiatry*. 2002; 73:350. [PubMed: 12185186]

- 19Mellick GD, Maraganore DM, Silburn PA. Australian data and meta-analysis lend support for alpha-synuclein (NACP-Rep1) as a risk factor for Parkinson's disease. *Neurosci Lett*. 2005; 375:112–6. [PubMed: 15670652]
- 20Linnertz C, Saucier L, Ge D, Cronin KD, Burke JR, Browndyke JN, et al. Genetic regulation of alpha-synuclein mRNA expression in various human brain tissues. *PLoS One*. 2009; 4:e7480. [PubMed: 19834617]
- 21Chiba-Falek O, Touchman JW, Nussbaum RL. Functional analysis of intra-allelic variation at NACP-Rep1 in the alpha-synuclein gene. *Hum Genet*. 2003; 113:426–31. [PubMed: 12923682]
- 22Cronin KD, Ge D, Manninger P, Linnertz C, Rossoshek A, Orrison BM, et al. Expansion of the Parkinson disease-associated SNCA-Rep1 allele upregulates human alpha-synuclein in transgenic mouse brain. *Hum Mol Genet*. 2009; 18:3274–85. [PubMed: 19498036]
- 23Saul R, Lutz MW, Burns DK, Roses AD, Chiba-Falek O. The SSV Evaluation System: A Tool to Prioritize Short Structural Variants for Studies of Possible Regulatory and Causal Variants. *Hum Mutat*. 2016; 37:877–83. [PubMed: 27279261]
- 24Afek A, Cohen H, Barber-Zucker S, Gordan R, Lukatsky DB. Nonconsensus Protein Binding to Repetitive DNA Sequence Elements Significantly Affects Eukaryotic Genomes. *PLoS Comput Biol*. 2015; 11:e1004429. [PubMed: 26285121]
- 25Afek A, Schipper JL, Horton J, Gordan R, Lukatsky DB. Protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci U S A*. 2014; 111:17140–5. [PubMed: 25313048]
- 26Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*. 2006; 24:1429–35.
- 27Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic acids research*. 2015; 43:D117–22. [PubMed: 25378322]
- 28Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014; 158:1431–43. [PubMed: 25215497]
- 29Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO journal*. 2010; 29:2147–60. [PubMed: 20517297]
- 30Nardelli J, Thiesson D, Fujiwara Y, Tsai FY, Orkin SH. Expression and genetic interaction of transcription factors GATA-2 and GATA-3 during development of the mouse central nervous system. *Dev Biol*. 1999; 210:305–21. [PubMed: 10357893]
- 31Scherzer CR, Grass JA, Liao Z, Pepivani I, Zheng B, Eklund AC, et al. GATA transcription factors directly regulate the Parkinson's disease-linked gene alpha-synuclein. *Proc Natl Acad Sci U S A*. 2008; 105:10907–12. [PubMed: 18669654]
- 32Brenner S, Wersinger C, Gasser T. Transcriptional regulation of the alpha-synuclein gene in human brain tissue. *Neurosci Lett*. 2015; 599:140–5. [PubMed: 26002080]
- 33Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols*. 2009; 4:393–411. [PubMed: 19265799]
- 34Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*. 2016; 351:1450–4. [PubMed: 27013732]
- 35Gordan R, Hartemink AJ, Bulyk ML. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res*. 2009; 19:2090–100. [PubMed: 19652015]
- 36Chiba-Falek O, Lopez GJ, Nussbaum RL. Levels of alpha-synuclein mRNA in sporadic Parkinson disease patients. *Mov Disord*. 2006; 21:1703–8. [PubMed: 16795004]
- 37Linnertz C, Anderson L, Gottschalk W, Crenshaw D, Lutz MW, Allen J, et al. The cis-regulatory effect of an Alzheimer's disease-associated poly-T locus on expression of TOMM40 and apolipoprotein E genes. *Alzheimer's & dementia: the journal of the Alzheimer's Association*. 2014; 10:541–51.

- 38Linnertz C, Lutz MW, Ervin JF, Allen J, Miller NR, Welsh-Bohmer KA, et al. The genetic contributions of SNCA and LRRK2 genes to Lewy Body pathology in Alzheimer's disease. *Hum Mol Genet.* 2014; 23:4814–21. [PubMed: 24777780]
- 39Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, et al. alpha-Synuclein locus triplication causes Parkinson's disease. *Science.* 2003; 302:841. [PubMed: 14593171]
- 40Farrer M, Kachergus J, Forno L, Lincoln S, Wang DS, Hulihan M, et al. Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications. *Ann Neurol.* 2004; 55:174–9. [PubMed: 14755720]
- 41Miller DW, Hague SM, Clarimon J, Baptista M, Gwinn-Hardy K, Cookson MR, et al. Alpha-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. *Neurology.* 2004; 62:1835–8. [PubMed: 15159488]
- 42Grundemann J, Schlaudraff F, Haeckel O, Liss B. Elevated alpha-synuclein mRNA levels in individual UV-laser-microdissected dopaminergic substantia nigra neurons in idiopathic Parkinson's disease. *Nucleic Acids Res.* 2008; 36:e38. [PubMed: 18332041]
- 43Rockenstein E, Hansen LA, Mallory M, Trojanowski JQ, Galasko D, Masliah E. Altered expression of the synuclein family mRNA in Lewy body and Alzheimer's disease. *Brain research.* 2001; 914:48–56. [PubMed: 11578596]
- 44Asi YT, Simpson JE, Heath PR, Wharton SB, Lees AJ, Revesz T, et al. Alpha-synuclein mRNA expression in oligodendrocytes in MSA. *Glia.* 2014; 62:964–70. [PubMed: 24590631]
- 45Lutz MW, Saul R, Linnertz C, Glenn OC, Roses AD, Chiba-Falek O. A cytosine-thymine (CT)-rich haplotype in intron 4 of SNCA confers risk for Lewy body pathology in Alzheimer's disease and affects SNCA expression. *Alzheimer's & dementia: the journal of the Alzheimer's Association.* 2015; 11:1133–43.
- 46Soldner F, Stelzer Y, Shivalila CS, Abraham BJ, Latourelle JC, Barrasa MI, et al. Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature.* 2016; 533:95–9. [PubMed: 27096366]
- 47Tagliaferro L, Glenn OC, Zamora ME, Beach TG, Woltjer RL, Lutz MW, et al. Genetic analysis of alpha-synuclein 3' untranslated region and its corresponding microRNAs in relation to Parkinson's compared to dementia with Lewy bodies. *Alzheimers Dement.* 2017
- 48Bonhoure N, Bounova G, Bernasconi D, Praz V, Lammers F, Canella D, et al. Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.* 2014; 24:1157–68. [PubMed: 24709819]
- 49Lutz MW, Saul R, Linnertz C, Glenn OC, Roses AD, Chiba-Falek O. A cytosine-thymine (CT)-rich haplotype in intron 4 of SNCA confers risk for Lewy body pathology in Alzheimer's disease and affects SNCA expression. *Alzheimer's & dementia: the journal of the Alzheimer's Association.* 2015
- 50Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{−(Delta Delta C(T))} Method. *Methods (San Diego, Calif.* 2001; 25:402–8.
- 51Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* 2013; 3:1093–104. [PubMed: 23562153]
- 52Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordan R. Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics.* 2013; 29:i117–25. [PubMed: 23812975]

Fig. 1.

GATA1 and ELK1 transcription factors bind with high specificity to their consensus sites in the proximity of the Rep1 repeat region. Boxplots show the distributions of TF binding signals (measured as the natural logarithm of fluorescence intensity) for sets of probes containing either no ELK/GATA binding sites (leftmost boxes), the putative ELK binding site, the putative ELK and GATA binding sites, or the putative GATA binding site in proximity of the Rep1 region (see Method for details on the DNA library design). Left: Probes containing the putative GATA site show significantly higher GATA1 binding signals than probes without the putative binding site, indicative of specific binding. Right: Probes containing the putative ELK site show significantly higher ELK1 binding signals than probes without the putative binding site, indicative of specific binding.

Fig. 2. The length of Rep1 repeats has a significant impact on GATA1 and ELK1 binding to their consensus sites proximal to the Rep1 region $(TC)_nTT(TC)_m(TA)_x(CA)_z$

(a) Increasing the length of the $(CA)_z$, $(TA)_x$, and $(TC)_m$ repeats, when located next to the GATA consensus binding sites, leads to increased GATA1 binding levels. Boxplots shows the natural logarithm of the florescence intensity measured by PBM (Methods). Each box shows the binding signal at 10 replicate DNA spots randomly distributed across the glass surface. The central mark indicates the median of the replicates, and the bottom and top edges of each box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points, excluding outliers. **(b)** Similar to panel (a), but showing the increase in ELK1 binding signal due to increased length of the $(CA)_z$, $(TA)_x$, and $(TC)_m$ repeats, when located next to the ELK consensus binding site.

Fig. 3. The length of the CA repeats in a ‘pure’ CA polymorphic site has a significant effect on GATA1 binding to its consensus site in the proximity of the repeat

Boxplot show the natural logarithm of the fluorescence intensity measured by PBM (Methods). Each box shows the binding signal at 10 replicate DNA spots randomly distributed across the glass surface. The central mark indicates the median of the replicates, and the bottom and top edges of each box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points, excluding outliers.

Fig. 4. Expression profiles of GATA and ELK transcription factors

The expression of GATA1, GATA2, GATA3, GATA6, ELK1, and ELK3 was evaluated in temporal cortex obtained from cognitively normal human brain tissues (N=3), PD patients (N=3), DLB patients (N=3), and blood samples (N=3). Levels of each TF were measured by real-time RT-PCR and calculated relatively to the geometric mean of GAPDH-mRNA and PPIA-mRNA reference controls using the 2^{-CT} method. Each column represents the mean of four biological and technical replicates. The error bars represent the S.E.M.