



Article

# A SINE-VNTR-*Alu* in the LRIG2 Promoter Is Associated with Gene Expression at the Locus

Ashley Hall <sup>1</sup>, Anni K. Moore <sup>2</sup>, Dena G. Hernandez <sup>2</sup>, Kimberley J. Billingsley <sup>2</sup>, Vivien J. Bubb <sup>1</sup>, John P. Quinn <sup>1,\*</sup> and NABEC (North American Brain Expression Consortium) <sup>†</sup>

<sup>1</sup> Department of Pharmacology and Therapeutics, Institute of Systems, Molecular & Integrative Biology, University of Liverpool, Liverpool L69 7BE, UK; ashley.hall@liverpool.ac.uk (A.H.); jillbubb@liverpool.ac.uk (V.J.B.)

<sup>2</sup> Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA; anni.moore@nih.gov (A.K.M.); hernand@mail.nih.gov (D.G.H.); kimberley.billingsley@nih.gov (K.J.B.)

\* Correspondence: jquinn@liverpool.ac.uk

† Membership of the ELANS Study Group is provided in the Acknowledgments.

Received: 2 October 2020; Accepted: 9 November 2020; Published: 11 November 2020

**Abstract:** The hominid SINE-VNTR-*Alu* (SVA) retrotransposons represent a repertoire of genomic variation which could have significant effects on genome function. A human-specific SVA in the promoter region of the gene leucine-rich repeats and immunoglobulin-like domains 2 (*LRIG2*), which we termed SVA\_LRIG2, is a common retrotransposon insertion polymorphism (RIP), defined as an element which is polymorphic for its presence or absence in the genome. We hypothesised that this RIP might be associated with differential levels of expression of *LRIG2*. The RIP genotype of SVA\_LRIG2 was determined in a subset of frontal cortex DNA samples from the North American Brain Expression Consortium (NABEC) cohort and was imputed for a larger set of that cohort. Utilising available frontal cortex total RNA-seq and CpG methylation data for this cohort, we observed that increased allele dosage of SVA\_LRIG2 was non-significantly associated with a decrease in transcription from the region and significantly associated with increased methylation of the CpG probe nearest to SVA\_LRIG2, i.e., SVA\_LRIG2 is a significant methylation quantitative trait loci (mQTL) at the *LRIG2* locus. These data are consistent with SVA\_LRIG2 being a transcriptional regulator, which in part may involve epigenetic modulation.

**Keywords:** SVA; polymorphism; retrotransposon; gene regulation; *LRIG2*; evolution; RIP; retrotransposon insertional polymorphism

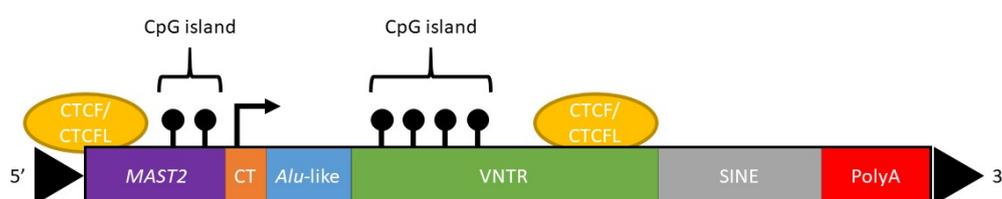
## 1. Introduction

Across many diseases, most disease-associated genetic variation identified by genome-wide association studies and gene candidate studies is located within non-coding DNA. A major constituent of non-coding DNA is sequence derived from retrotransposons, which constitute approximately 45% of the human genome [1]. Retrotransposon sequences propagate throughout the genome via a ‘copy-and-paste’ mechanism, first involving transcription into RNA followed by reverse transcription and insertion of a new copy at a distant genomic site. They are important contributors to genetic diversity via their ability to introduce novel splice sites, promoters and regulatory elements into a locus [2,3], thus facilitating the generation of novel species- and tissue-specific expression patterns in the germline.

Non-long terminal repeat (non-LTR) retrotransposons are the only known contemporarily active retrotransposons in the human genome, of which SINE-VNTR-*Alu* (SVA) elements are the evolutionarily youngest [1]. Present only in hominids, there are approximately 3000 identified in the

reference human genome (Hg38). Considered 5' to 3', SVAs are composed of a CT element (CCCTCTn repeat), an antisense *Alu*-like (another retrotransposon family) sequence, one or two variable number tandem repeat (VNTR) regions, a short interspersed nuclear element (SINE) region derived from the retroviral HERV-K10 element, and a poly-A tail [4] (Figure 1). SVAs do not transpose autonomously but are mobilised in *trans* by the transposition machinery of LINE-1, an autonomous class of non-LTR retrotransposon [5]. Full-length SVAs vary from 1 to 4 kb in size, and are classified A–F in order of evolutionary age based on their SINE region [4]. Additionally, an F1 class of SVA contains a 5' transduction of exon 1 of the *MAST2* gene [6], such that much or all of the CT element is replaced (Figure 1). SVA subfamilies E, F and F1 are human specific, and with the addition of the D subfamily, are transpositionally active within the human genome [4]. It is estimated that one novel SVA insertion event occurs every 916 births, corresponding to  $7 \times 10^6$  distinct insertions worldwide [7]. A genomic site where a retroelement may be present or absent is known as a retrotransposon insertion polymorphism (RIP), and the average human is estimated to possess 56 SVA RIPs that differ from the reference genome [8]. Additionally, SVAs may be polymorphic with regards to the length (repeat copy number) of their CT element, VNTR and poly-A signal components, and may also contain single-nucleotide polymorphisms (SNPs). Accordingly, SVAs are considered the most polymorphic structural variants in the genome [8].

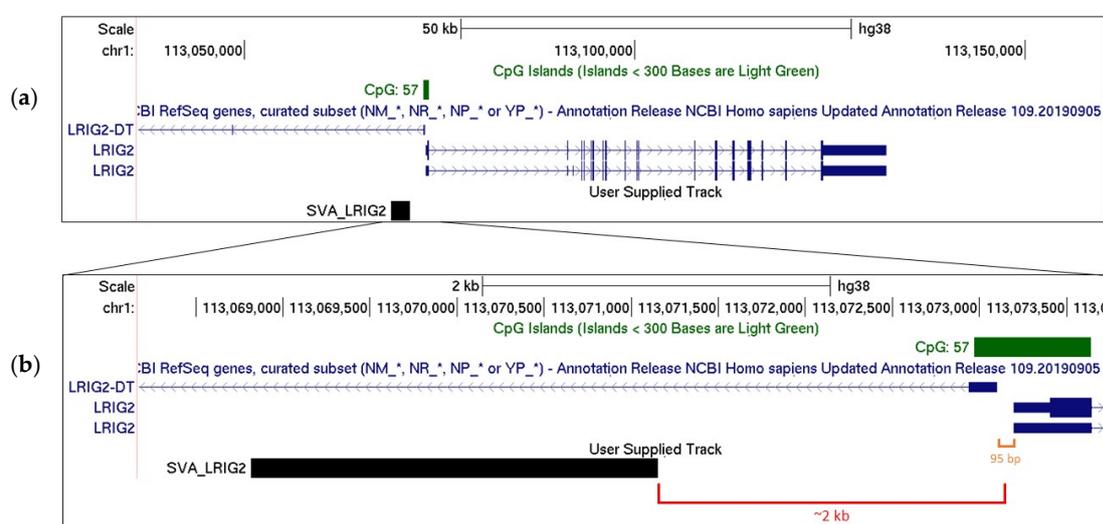
#### SVA F1



**Figure 1.** The canonical structure of the SVA F1 retrotransposon subclass, 5' to 3' displayed left to right. Purple: 5' transduction of exon 1 of *MAST2*. Potential CCCTC-binding factor (CTCF) binding site and CpG island are displayed. Orange: The (CCCTCT)<sub>n</sub> hexamer repeat, which may be variable in copy number. Note, much of this sequence has been replaced by the *MAST2* transduction. Blue: The *Alu*-like region, which is composed of two antisense *Alu* regions separated by a small intervening sequence. Green: The VNTR region which may be one or two regions of 35–50 bp tandem repeats, and may reach GC content of 70%. A CpG island is pictured, with black lollipops signifying 5-methylcytosine residues. A potential binding site for CTCF is shown. Grey: The SINE region is derived from the 3' LTR of the retroviral HERV-K10 element. Red: The canonical polyadenylated tail can act as a signal for transcription termination but may be bypassed by RNA polymerase, and is thought to be important for SVA retrotransposition. Black arrowheads indicate target site duplication regions resulting from insertion.

On chromosome 1, there is a 2.4 kb long (full-length) SVA F1 positioned approximately 2 kb upstream from the transcriptional start site of leucine-rich repeats and immunoglobulin-like domains 2 (*LRIG2*) (Figure 2), a protein which is involved in epidermal growth factor signalling. The *LRIG2* protein is ubiquitously expressed but deleterious mutations are associated with urofacial syndrome [9,10], a disorder characterised by improper nerve control of the face and bladder [11]. *LRIG2* shares its bidirectional promoter region with a non-coding divergent transcript, *LRIG2-DT*, whose expression could be similarly regulated [12–14] (Figure 2). The SVA proximal to this *LRIG2* locus (herein SVA\_*LRIG2*) is a RIP, meaning that it transposed relatively evolutionarily recently, and may represent an insertion event involved in human-specific modulation of gene expression. Indeed, SVA RIPs and length polymorphisms may have profound effects on gene regulation and genomic architecture. Over 60% of SVAs in the reference genome are found in genes or within 10 kb of a gene despite this representing only approximately 1% of the human genome [15], suggesting preferential

insertion into actively transcribed regions [4]. The GC content of an SVA is approximately 60% and may exceed 70% within the VNTR region, satisfying the definition of a CpG island [16]. SVAs have been shown to be methylated in somatic tissue, with evolutionarily older classes of SVA generally showing greater methylation [17]. DNA methylation is known to be capable of spreading laterally along DNA from CpG islands [18], and mouse B1 SINE retrotransposons have been shown to induce transcriptional repression by DNA methylation spreading [19]. An SVA insertion may therefore influence expression of nearby genes by altering local chromatin structure or recruiting factors depending on both the SVA's primary sequence and epigenetic marks [20,21]. Indeed, the 5' *MAST2* exon 1 transduction associated with SVA F1 elements is defined as a CpG island and has been demonstrated to influence transcription in human germline cells [22]. SVAs may also influence chromatin structure by recruiting CCCTC-binding factor (CTCF), a master regulator of 3D chromatin structure [23]. It has been demonstrated that CTCF can bind the SVA VNTR region in vitro, while the germline-expressed paralog CTCF like (CTCFL, also referred to as BORIS) can bind the VNTR region in vivo [21]. Interestingly, CTCFL has also been shown to bind immediately upstream of the SVA F1 subfamily [21].



**Figure 2.** (a) The *LRIG2* locus as shown on UCSC genome browser hg38, displaying two validated *LRIG2* isoforms and first exon of *LRIG2-DT* from the RefSeq genes curated subset, and the *LRIG2* promoter-associated CpG island. *SVA\_LRIG2* position is displayed. (b) Regions flanking *SVA\_LRIG2* are displayed, with distance to *LRIG2* locus transcriptional start sites highlighted in red. Distance between *LRIG2* and *LRIG2-DT* transcriptional start sites highlighted in orange.

Reporter gene studies have demonstrated that gene expression was differently modulated by different SVA components and the full-length SVA both in vitro and in vivo [15,24]. Indeed, genomic VNTRs have previously been shown to represent regions of stimulus responsive histone marks [25]. It should come as no surprise that insertion of SVAs into open reading frames is directly causative of at least 12 diseases due to transcriptional interference [26]. It is interesting to note that in X-linked Dystonia-Parkinsonism not only does a disease-specific SVA insertion appear causative of disease, but that the length of the CT hexamer repeat strongly correlates inversely with age of onset of the disease [27,28].

It is clear that SVAs represent non-coding DNA elements that have the potential to exert potent regulatory influences on nearby gene expression, and that RIPs or length polymorphisms between individuals may contribute to interpersonal differences in expression patterns. However, most work on SVAs thus far focus on their properties in vitro or in disease; there is currently a scarcity of data on the functional influence that SVAs may have in normal gene regulation in humans in situ. *SVA\_LRIG2* represents an opportunity to study the influence of an SVA insertion in situ as it is a RIP,

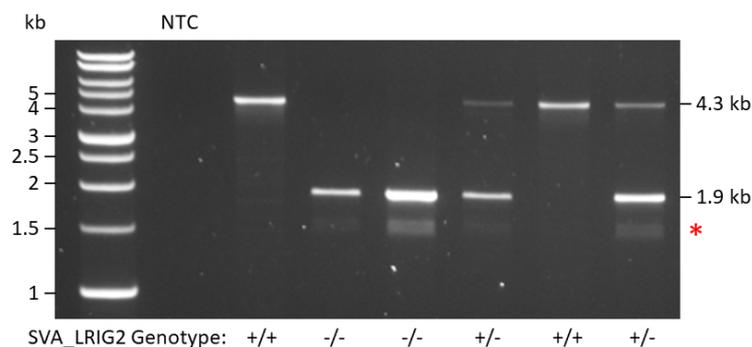
meaning that DNA samples from the general populace represent a naturally occurring resource to address the influence of presence or absence of an SVA on genome structure or gene regulation. Indeed, the Database of RIPs in Humans (dbRIP) assigns an allele frequency of 0.422 to the presence of SVA\_LRIG2 (dbRIP accession RIP3000013) [29], meaning that presence and absence of the SVA should be observed almost equally often. Furthermore, SVA\_LRIG2 has a GC content of approximately 70% and contains 170 CpG dinucleotides, representing a potentially substantial CpG island. By contrast, the CpG island associated with the *LRIG2* promoter region, identified in the UCSC genome browser, is 669 bp long, approximately 65% GC content and contains 57 CpGs (Figure 2). We might postulate that this CpG island would be sensitive to methylation changes induced by the much larger one represented by SVA\_LRIG2. To investigate the potential for this RIP to exert regulatory properties at the *LRIG2* locus we genotyped SVA\_LRIG2 in DNA samples from the North American Brain Expression Consortium (NABEC), a cohort of neurologically normal frontal cortex samples from the Database of Genotypes and Phenotypes (dbGaP) for which molecular datasets including genome-wide genotyping, total RNA sequencing and CpG methylation were available.

## 2. Results

### 2.1. SVA\_LRIG2 Genotypes Were PCR Validated or Imputed for 288 Individuals

Although the NABEC cohort has undergone whole-genome sequencing (WGS), repetitive DNA elements such as retrotransposons are routinely filtered out due to their inherent difficulties in mapping back to the reference genome [30]. It was therefore necessary to genotype SVA\_LRIG2 in the available samples from the NABEC cohort. Ninety-six brain DNA samples underwent PCR with forward and reverse primers that annealed 1.1 kb upstream and 0.8 kb downstream of genomic SVA\_LRIG2, respectively. The inclusion of these flanking regions yielded a 1.9 kb PCR product when SVA\_LRIG2 was absent from the locus and an approximately 4.3 kb PCR product when the SVA was present (Figure 3). This production of two distinctly sized PCR fragments allowed SVA RIP genotype (presence or absence) to be readily determined following agarose gel electrophoresis and visualisation. A representative gel is shown in Figure 3. Of the 96 samples genotyped, 14 had no copies of SVA\_LRIG2 present at the locus (-/-), 43 had 1 copy present (+/-, i.e., the SVA was present on a single chromosome) and 39 had 2 copies present (+/+, i.e., the SVA was present on both chromosomes) (Table 1).

Although repetitive elements are not typically included in WGS processing and genotype calling, it is assumed that these variants may occur within haplotype blocks shared with variants through linkage disequilibrium. Therefore, the genotypes of these typically uncalled variants of interest may be inferred through ‘proxy’ (or ‘tagging’) SNPs that are in linkage disequilibrium [31]. In this way the genotype of SVA\_LRIG2 can be imputed in NABEC individuals whose DNA was not available in our lab. We observed that SVA\_LRIG2 exhibited length polymorphism with 4 VNTR alleles and developed proxy SNPs for each of them in order to fully capture the genetic diversity of SVA\_LRIG2. This produced a total of 29 proxy SNPs, and the SNP with highest linkage disequilibrium  $r^2$  and  $D'$  value for each VNTR allele was taken forward (see Materials and Methods). VNTR allele 4 was excluded after filtering for standard genotype missingness. The chosen proxy SNPs recapitulated the 96 PCR-validated genotypes with 97.4% accuracy (187 out of 192 alleles, data not shown). We then grouped these alleles into SVA\_LRIG2 RIP genotypes: -/-, +/- and +/+. When added to the existing validated genotyping, this yielded 49 SVA\_LRIG2 -/- genotypes, 154 SVA\_LRIG2 +/- and 85 SVA\_LRIG2 +/+, for a total of 288 individuals (Table 1).



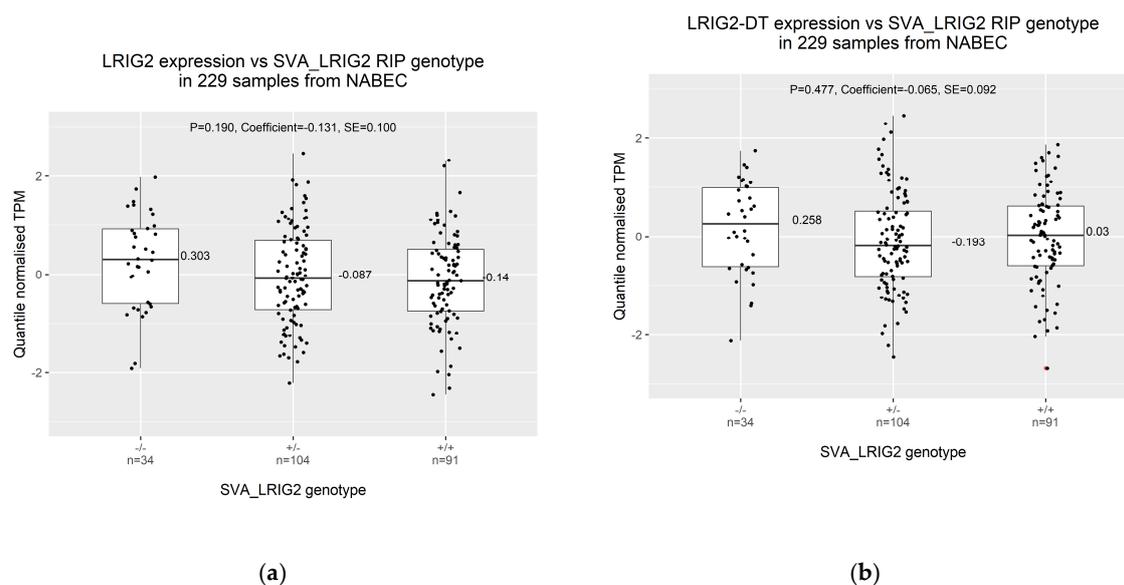
**Figure 3.** Representative image of NABEC DNA samples following PCR with primers targeting regions flanking SVA\_LRIG2. A 1.9 kb PCR product corresponds to SVA\_LRIG2 being absent from the locus and a 4.3 kb PCR product corresponds to SVA\_LRIG2 presence. Sample genotypes are displayed along the bottom. NTC = Non-template control. Red asterisk indicates formation of non-specific PCR products of approximately 1.5 kb.

**Table 1.** SVA\_LRIG2 RIP genotype counts in NABEC samples.

SVA_LRIG2 Genotype	Validated Counts	Imputed Counts	Total Counts
-/-	14	35	49
+/-	43	111	154
+/+	39	46	85
Total	96	192	288

## 2.2. Increased Allele Dosage of SVA\_LRIG2 Is Non-Significantly Associated with Reduced Transcription from the LRIG2 Locus

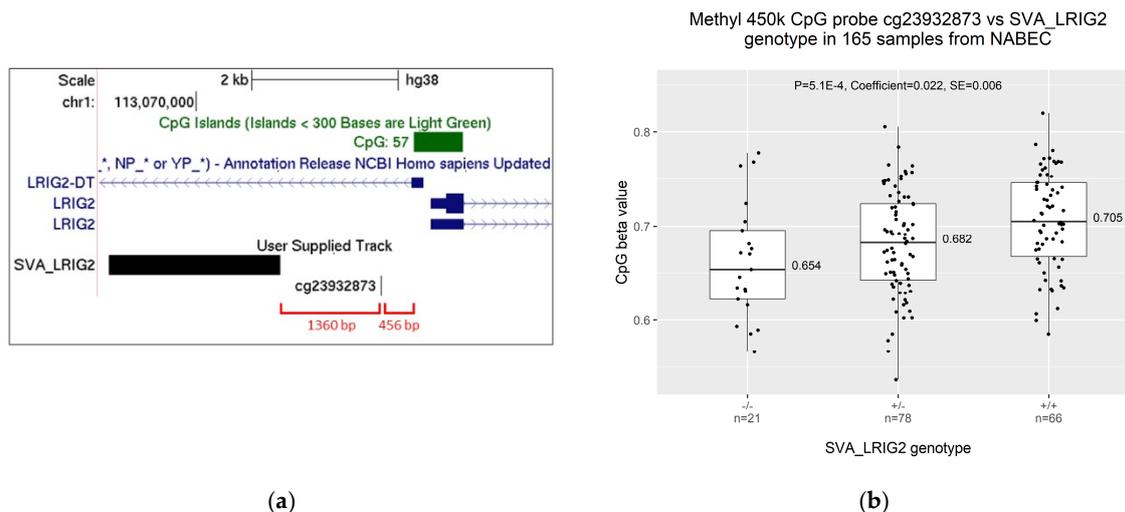
Having determined SVA\_LRIG2 RIP genotypes experimentally or computationally, we assessed whether these genotypes correlated with transcription at the *LRIG2* locus. Of the 288 individuals in NABEC with validated or imputed SVA\_LRIG2 genotypes, frontal cortex total RNA-seq data were available for 229. Expression values, expressed as quantile normalised transcripts per kilobase million (TPM), were extracted for *LRIG2* and *LRIG2-DT*. These expression values were stratified on the basis of SVA\_LRIG2 RIP genotype, resulting in 34 individuals of the genotype SVA\_LRIG2 -/-, 104 that were SVA\_LRIG2 +/- and 91 that were SVA\_LRIG2 +/+. Compared to those with no copies of SVA\_LRIG2 (-/-), individuals with the RIP genotype SVA\_LRIG2 +/- displayed a median expression of *LRIG2* that was 0.390 standard deviations lower and those with SVA\_LRIG2 +/+ were 0.443 lower in the quantile normalised data (Figure 4a). Similarly, median expression of *LRIG2-DT* was 0.451 standard deviations lower in SVA\_LRIG2 +/- individuals and 0.228 lower in SVA\_LRIG2 +/+ individuals compared to those of genotype -/- (Figure 4b). A simple linear regression was used to assess whether SVA\_LRIG2 was an expression quantitative trait loci in neurologically normal brain, i.e., whether increasing allele dosage of SVA\_LRIG2 correlated with differential expression of *LRIG2* or *LRIG2-DT* in the NABEC RNA-seq data. The linear regression model included the known covariates age, gender, ethnicity, originating brain bank and RNA integrity number. Alpha significance level was set at 2.94E-3 using Bonferroni correction for multiple comparisons (see Section 4). Although we found that SVA\_LRIG2 RIP genotype was not significantly associated with expression of *LRIG2* or *LRIG2-DT* (Figure 4,  $p = 0.190$  and  $p = 0.477$ , respectively), our model indicated a negative trend between SVA\_LRIG2 allele dosage and expression of both transcripts (Figure 4, negative coefficient values).



**Figure 4.** Frontal cortex total RNA-seq data for *LRIG2* (ENSG00000198799.12) and *LRIG2-DT* (ENSG00000238198.2) in 229 NABEC individuals grouped by SVA\_LRIG2 RIP genotype. A total of 96 genotypes were PCR validated and 145 were imputed for a total of 241 genotypes. (a) *LRIG2* expression. (b) *LRIG2-DT* expression. RNA-seq data expressed as quantile normalised transcripts per kilobase million (TPM). Standard deviations from the mean of the normalised data are displayed on the y-axis. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE). Outlier value is marked in red.

### 2.3. Increased Allele Dosage of SVA\_LRIG2 Is Associated with Increased Methylation of the Nearest 450K Methylation Probe, cg23932873

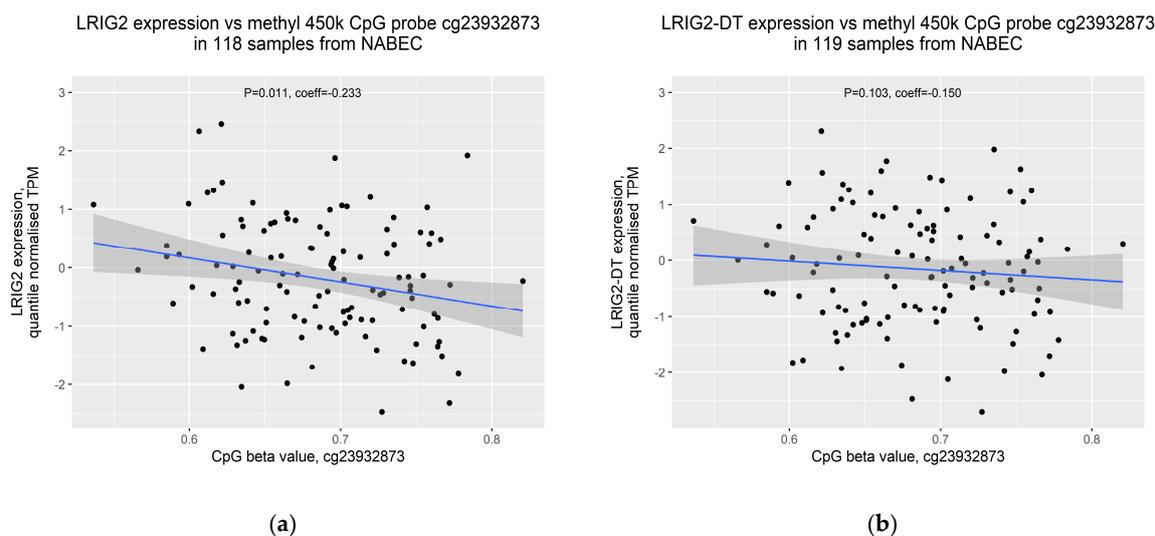
SVAs may simplistically be thought of as mobile CpG islands and DNA methylation is known to spread along adjacent DNA [16,18]. Therefore, we determined whether SVA\_LRIG2 RIP genotypes might be associated with differences in methylation at the locus. Frontal cortex 450K CpG methylation data were available for 165 of the 288 individuals in NABEC with validated or imputed SVA\_LRIG2 genotypes. Methylation data were again stratified on the basis of SVA\_LRIG2 RIP genotype, resulting in 21 individuals of the genotype SVA\_LRIG2  $-/-$ , 78 that were SVA\_LRIG2  $+/-$  and 66 that were SVA\_LRIG2  $+/+$ . Publicly available ENCODE Methylation 450K Bead Array data list 15 CpG methylation probes in a 20 kb window around SVA\_LRIG2, of which 13 are less than 3 kb away from the *LRIG2* promoter region (probes listed in Materials and Methods). As with the expression data, linear regression was used to determine whether SVA\_LRIG2 was a methylation QTL (mQTL) for these 15 probes, i.e., whether increasing allele dosage of SVA\_LRIG2 correlated with changes in CpG methylation. The linear regression model included the known covariates age, gender, ethnicity and originating brain bank. After Bonferroni correction for multiple comparisons (see Section 4), only the CpG dinucleotide probe nearest to SVA\_LRIG2, cg23932873, was found to pass the adjusted alpha level of  $2.94E-3$  when methylation was correlated against SVA\_LRIG2 genotype (Figure 5a). Compared to individuals with the genotype SVA\_LRIG2  $-/-$ , the median proportion of cg23932873 that was methylated was found to be 0.028 (2.8%) higher in those with the genotype  $+/-$  and 0.055 (5.1%) higher in those with SVA\_LRIG2 genotype  $+/+$  (Figure 5b). Our linear regression model provided a p value of  $5.1E-4$  and coefficient of 0.022, indicating that there is a significant association between SVA\_LRIG2 allele dosage and increasing methylation of cg23932873. In other words, SVA\_LRIG2 is a significant mQTL for a CpG dinucleotide at the *LRIG2* locus.



**Figure 5.** (a) The *LRIG2* promoter region as displayed on UCSC genome browser hg38, with two validated *LRIG2* isoforms and first exon of *LRIG2-DT* shown by the RefSeq genes curated subset. The position of *SVA\_LRIG2* is displayed in black. The position of *cg23932873*, the CpG probe closest to *SVA\_LRIG2*, is shown in black and the *LRIG2* promoter-associated CpG island is shown in green. The distances in base pairs from *cg23932873* to *SVA\_LRIG2* and the CpG island are shown in red. (b) Frontal cortex CpG 450K methylation data for probe *cg23932873* in 184 NABEC individuals grouped by *SVA\_LRIG2* RIP genotype. A total of 96 genotypes were PCR validated and 88 were imputed for a total of 184 genotypes.

#### 2.4. Increased Methylation of *cg23932873* Is Weakly Correlated with Decreased Expression of *LRIG2*

DNA methylation is known to generally repress gene expression through mechanisms including impairment of transcription factor binding and recruitment of proteins that confer repressive histone modifications [20]. Therefore, having observed that increased *SVA\_LRIG2* allele dosage is associated with decreased transcription from the *LRIG2* locus and increased methylation proximal to *SVA\_LRIG2*, we hypothesised that the two may be correlated. Both RNA-seq data for *LRIG2* and CpG methylation data were available for 118 samples in the NABEC cohort after removal of outliers, while 119 samples were available for *LRIG2-DT* expression. CpG beta values for *cg23932873* were plotted against normalised TPM values for *LRIG2* or *LRIG2-DT* and tested with Pearson correlation coefficient. We observed a weak yet significant inverse correlation between *cg23932873* methylation and *LRIG2* expression (Figure 6a), and no correlation between *cg23932873* methylation and *LRIG2-DT* expression (Figure 6b).



**Figure 6.** Expression of transcripts from the *LRIG2* locus versus CpG 450K methylation data for probe cg23932873 in frontal cortex of 129 NABEC individuals. (a) *LRIG2* (ENSG00000198799.12). (b) *LRIG2-DT* (ENSG00000238198.2). Blue line indicates trend line; dark grey zone indicates 95% confidence interval. Pearson correlation coefficient and corresponding p values displayed.

### 3. Discussion

Despite transposable elements making up nearly half of the human genome, they remain relatively unexplored as sources of important genomic variation [1], as their repetitive nature makes them difficult to map to the human genome. The evolutionarily youngest SVA retrotransposons may be particularly important for human health and evolution since they have introduced hominid-specific elements with regulatory potential to specific genomic loci, a process that is ongoing for SVA subclasses D–F1 [32]. It should be noted that SVA regulatory influences need not be dramatic to be biologically important in individuals of certain genetic backgrounds—for example, in the study of genetically complex disease, it is generally accepted that disease can result from the cumulative effect of many low-contribution variants [33]. Here we have investigated how presence or absence of an SVA retrotransposon 2 kb upstream of a divergent promoter region might influence regulation of transcription. SVA\_*LRIG2* is a relatively common RIP, meaning that the general populace represents a ready-made resource of the potential effects of SVA\_*LRIG2* presence vs absence at the *LRIG2* locus. The RIP genotype of SVA\_*LRIG2* was determined in the NABEC cohort using PCR and imputation and correlated with both gene expression and methylation at the nearby *LRIG2* and *LRIG2-DT* promoter region.

We observed a negative correlation between increased SVA\_*LRIG2* allele dosage and transcription of both *LRIG2* and the region's divergent transcript *LRIG2-DT* in our linear regression models, but this did not achieve statistical significance. However, we note that this is not unexpected in a relatively small study of a heterogeneous population. Furthermore, the frontal cortex samples represent a tissue of heterogeneous cell types including neurons and glial cells, which may act to convolute expression and methylation patterns. Additionally, in this study, the SVA\_*LRIG2* genotype  $-/-$  is made up of a considerably fewer samples than genotypes  $+/-$  or  $+/+$ , and we speculate that expression differences may achieve statistical significance with an increased sample/cohort size. We postulate that this may be the case for *LRIG2* expression in particular, which displayed the clearest association between SVA\_*LRIG2* allele dosage and transcription. Despite falling short of statistical significance, these observations suggest that insertion of SVA\_*LRIG2* near to the *LRIG2* promoter region may be associated with decreased expression of transcripts from the locus, particularly *LRIG2*. That the two transcripts were similarly affected supports previous models in which divergent transcripts were co-regulated [12–14], and suggests a largely orientation-independent model for transcriptional modulation by SVA\_*LRIG2* at this locus.

When assessing the 15 CpG methylation probes in the region surrounding SVA\_LRIG2, only the closest probe, cg23932873, was found to be a significant mQTL. Specifically, we found a strong positive association between SVA\_LRIG2 allele dosage and levels of methylation of cg23932873 in the frontal cortex of individuals from the NABEC cohort. This indicates that SVA\_LRIG2 may induce increased methylation of cg23932873, potentially via lateral spread of methylation from the GC- and CpG-rich SVA [16,18]. Furthermore, when expression from the *LRIG2* locus was compared directly to methylation of cg23932873 we observed a weak but significant inverse correlation, suggesting that increased methylation at cg23932873 induced by SVA\_LRIG2 might contribute to the observed decrease in expression from the *LRIG2* locus. It has been demonstrated that methylated CpG dinucleotides are recognised by proteins associated with transcriptional repressor complexes [34], and that DNA methyltransferases can cooperate with enzymes that add methylation or remove acetylation at histones [35–37]. In this way, increased methylation at cg23932873 resulting from SVA\_LRIG2 might contribute to formation of a transcriptionally repressive environment at the *LRIG2* locus, leading to the modest modulation of *LRIG2* and *LRIG2-DT* expression we have observed here.

We note that transcriptionally repressive effects of SVA\_LRIG2 may also be mediated by other properties of SVA F1s that are not necessarily mutually exclusive to increasing local methylation. For example, binding of the chromatin structural regulator CTCF/CTCFL to SVA\_LRIG2, either within the VNTR or upstream of the *MAST2* transduction [21], may form novel chromatin loops that preclude interactions between the *LRIG2/LRIG2-DT* promoter region and enhancer elements [23]. However, only expression and methylation datasets were available in the NABEC cohort, meaning that analysis of the influence of SVA\_LRIG2 on 3D chromatin structure is currently beyond the scope of this study. We also note that *LRIG2* and *LRIG2-DT* may be differently affected, particularly at the post-transcriptional level, by SVA\_LRIG2 owing to its genomic context; SVA\_LRIG2 is 2 kb upstream of the promoter region of both transcripts, but is within the first intron of *LRIG2-DT*. It has been demonstrated that intronic SVAs are associated with intron retention in mRNA [27,38], which may trigger various nuclear degradation mechanisms and result in decreased overall transcript levels [39].

Taken together, these results suggest that recent SVA insertion events may have the potential to fine-tune the expression of nearby transcripts, and that one potential mechanism for this is via modulation of methylation at the promoter region. These in situ observations expand upon previous work demonstrating the in vitro regulatory effects of SVAs [15,24]. This is pertinent when we consider that over 60% of SVAs are found in genes or within 10 kb of a gene [15], and may therefore influence a wide variety of biological processes. Given that the data presented here were derived from the NABEC cohort, our findings are specifically applicable to neurologically normal frontal cortex. Our observations therefore complement previous hypotheses that SVAs may be important sources of genomic variation in the central nervous system. As hominid-specific elements, SVAs may be implicated in the development of higher cognitive function and ongoing human-specific genome evolution [32]. It is also conceivable that transposable elements, particularly SVAs, which are currently filtered out of short-read WGS data represent an under-appreciated source of variation that may be relevant to genetically complex processes. This may be especially true for very recent, human-specific insertion events and human-specific physiology and evolution. We note that short-read sequencing technology produces reads of up to 600 bases, with inherent difficulties in mapping repetitive elements back to the genome. By contrast, ‘long-read’ sequencing regularly generates reads of over 10 kb [40], meaning large or repetitive structural variants such as retrotransposons may be readily detected in WGS. Indeed, recent advances and decreasing costs of long-read sequencing technologies should soon make its implementation routine [41]. In light of the work presented here, we expect that improved detection of retrotransposons in WGS projects will inevitably lead to greater emphasis on their roles as important structural variants in the immediate future of human genomics research.

#### 4. Materials and Methods

DNA samples from the NABEC cohort (dbGaP accession phs001300.v1.p1) underwent PCR with 0.02 U/ $\mu$ l KOD Hotstart polymerase (Merck 71086, Darmstadt, Germany) and 1.5 mM MgSO<sub>4</sub>, 0.2 mM

dNTPs, 1 M betaine, and 0.3  $\mu$ M of each primer. To amplify SVA\_LRIG2 and its flanking regions, the forward primer used was 5'-TCCACGGACTTCCTAGAACG-3' and the reverse primer used was 5'-CGAGATGGCGGCAGTACC-3'. Thermal cycling conditions were 95 °C for 2 min, followed by 35 cycles of 95 °C for 20 s, 58 °C for 10 s and 70 °C for 1 min 50 s. Samples then underwent agarose gel electrophoresis at 110 V for 90 min and were visualised on a BioDoc-It™ UV transilluminator (Analytik Jena: UVP, Upland, CA, US).

In order to accurately capture the full genetic diversity of SVA\_LRIG2, proxy SNPs were generated using the '--show-tags' option in PLINK v1.90 [42] for each of the 4 SVA\_LRIG2 VNTR region length variants observed in the 96 NABEC individuals validated in our lab (genotyping data not shown). For each variant, the proxy SNP with the highest  $r^2$  and  $D'$  values was selected: VNTR 1 proxy rs114767321,  $r^2 = 1$ ,  $D' = 1$ ; VNTR 2 proxy rs183751190,  $r^2 = 1$ ,  $D' = 1$ ; VNTR 3 proxy rs12744009,  $r^2 = 0.894$ ,  $D' = 1$ . Proxy SNPs for VNTR variant 4 were excluded after filtering for genotype missingness ( $<0.1$ ). The remaining proxy SNPs for SVA\_LRIG2 variants were then grouped into RIP genotypes: -/-, +/- or +/+.

Generation of the NABEC frontal cortex total RNA-seq data and CpG methylation data has been previously described in detail by Gibbs et al. [43]. Data relating to individuals under the age of 15 were excluded to minimize developmental effects in the data. Expression data were extracted for LRIG2 and LRIG2-DT with Ensembl accession numbers ENSG00000198799.12 and ENSG00000238198.2, respectively. CpG data samples with sex mismatch were excluded. We noted that Illumina's Infinium MethylationEPIC manifest file listed cg23932873 as containing the SNP rs61818193 with MAF = 0.5, which would be a potential confounder of this association analysis. However, rs61818193 does not have a frequency listed on dbSNP and was absent in pre-filtered WGS data in the NABEC cohort. Therefore, we concluded it was acceptable to proceed with use of this methylation probe as an mQTL readout.

The CpG probes analysed at the *LRIG2* locus were: cg13503476, cg23932873, cg22598841, cg16709384, cg26091510, cg04139429, cg23175215, cg17310611, cg24448849, cg15031996, cg10983720, cg14912723, cg21504385, cg09332974, and cg23961141.

Linear regression analyses were performed in RStudio Version 1.2.1335 (Boston, MA, USA) using the 'lm' and 'summary' functions, where test statistics follow a Student's t distribution. When assessing effects of SVA\_LRIG2 genotype in RNA-seq and CpG methylation data for NABEC individuals, a total of 17 tests were performed: 2 RNA transcripts (*LRIG2* and *LRIG2-DT*) and 15 CpG probes were examined. The 0.05 alpha significance level for these linear regression analyses was therefore adjusted using Bonferroni correction for multiple tests:  $0.05/17 = 2.94E-3$ .

**Author Contributions:** Conceptualisation, A.H., V.J.B. and J.P.Q.; methodology, A.H., A.K.M., D.G.H., K.J.B., V.J.B., J.P.Q. and NABEC; formal analysis, A.H. and NABEC; data interpretation A.H., A.K.M., D.G.H., K.J.B., V.J.B., J.P.Q. and NABEC; writing—original draft preparation, A.H.; writing—review and editing, A.H., A.K.M., D.G.H., K.J.B., V.J.B., J.P.Q. and NABEC; funding acquisition, A.H., V.J.B. and J.P.Q. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Wellcome Trust PhD Programme, grant number 215192/Z/19/Z.

**Acknowledgments:** We thank members of the North American Brain Expression Consortium (NABEC) for providing DNA samples derived from brain tissue and corresponding total RNA-seq and CpG methylation data. Brain tissue for the NABEC cohort was obtained from the Baltimore Longitudinal Study on Aging at the Johns Hopkins School of Medicine, the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, the Banner Sun Health Research Institute Brain and Body Donation Program, and from the University of Kentucky Alzheimer's Disease Center Brain Bank. This research was supported in part by the Intramural research Program of the NIH, National institute on Aging.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

NABEC	North American Brain Expression Consortium
SVA	SINE-VNTR- <i>Alu</i>
RIP	Retrotransposon insertion polymorphism
LTR	Long terminal repeat
Hg38	Human genome assembly GRCh38
VNTR	Variable number tandem repeat
SINE	Short interspersed nuclear element
SNP	Single-nucleotide polymorphism
CTCF	CCCTC-binding factor
CTCFL	CTCF like
LRIG2	Leucine-rich repeats and immunoglobulin-like domains 2
dbRIP	Database of RIPs in humans
dbGaP	Database of Genotypes and Phenotypes
WGS	Whole-genome sequencing
TPM	Transcripts per kilobase million
mQTL	Methylation quantitative trait loci
SE	Standard error

## References

- Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921, doi:10.1038/35057062.
- Erwin, J.A.; Marchetto, M.C.; Gage, F.H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.* **2014**, *15*, 497–506, doi:10.1038/nrn3730.
- Cordaux, R.; Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **2009**, *10*, 691–703, doi:10.1038/nrg2640.
- Wang, H.; Xing, J.; Grover, D.; Hedges, D.J.; Han, K.; Walker, J.A.; Batzer, M.A. SVA Elements: A Hominid-specific Retroposon Family. *J. Mol. Biol.* **2005**, *354*, 994–1007, doi:10.1016/j.jmb.2005.09.085.
- Hancks, D.C.; Goodier, J.L.; Mandal, P.K.; Cheung, L.E.; Kazazian, H.H., Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **2011**, *20*, 3386–3400, doi:10.1093/hmg/ddr245.
- Bantysh, O.B.; Buzdin, A.A. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochemistry (Moscow)* **2009**, *74*, 1393–1399.
- Xing, J.; Zhang, Y.; Han, K.; Salem, A.H.; Sen, S.K.; Huff, C.D.; Zhou, Q.; Kirkness, E.F.; Levy, S.; Batzer, M.A.; et al. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res.* **2009**, *19*, 1516–1526, doi:10.1101/gr.091827.109.
- Bennett, E.A.; Coleman, L.E.; Tsui, C.; Pittard, W.S.; Devine, S.E. Natural Genetic Variation Caused by Transposable Elements in Humans. *Genetics* **2004**, *168*, 933–951, doi:10.1534/genetics.104.031757.
- Stuart, H.M.; Roberts, N.A.; Burgu, B.; Daly, S.B.; Urquhart, J.E.; Bhaskar, S.; Dickerson, J.E.; Mermerkaya, M.; Silay, M.S.; Lewis, M.A.; et al. LRIG2 Mutations Cause Urofacial Syndrome. *Am. J. Hum. Gen.* **2013**, *92*, 259–264, doi:10.1016/j.ajhg.2012.12.002.
- Holmlund, C.; Nilsson, J.; Guo, D.; Starefeldt, A.; Golovleva, I.; Henriksson, R.; Hedman, H. Characterization and tissue-specific expression of human LRIG2. *Gene* **2004**, *332*, 35–43, doi:10.1016/j.gene.2004.02.002.
- Newman, W.G.; Woolf, A.S. Urofacial Syndrome. In *GeneReviews*<sup>®</sup>; Adam, M.P., Ardinger, H.H., Pagon, R.A., Wallace, S.E., Bean, L.J.H., Stephens, K., Amemiya, A., Eds.; University of Washington: Seattle, WA, USA, 1993.
- Haberle, V.; Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 621–637, doi:10.1038/s41580-018-0028-8.
- Seila, A.C.; Calabrese, J.M.; Levine, S.S.; Yeo, G.W.; Rahl, P.B.; Flynn, R.A.; Young, R.A.; Sharp, P.A. Divergent transcription from active promoters. *Science* **2008**, *322*, 1849–1851, doi:10.1126/science.1162253.
- Core, L.J.; Martins, A.L.; Danko, C.G.; Waters, C.T.; Siepel, A.; Lis, J.T. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **2014**, *46*, 1311–1320, doi:10.1038/ng.3142.
- Savage, A.L.; Bubb, V.J.; Breen, G.; Quinn, J.P. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol. Biol.* **2013**, *13*, 101, doi:10.1186/1471-2148-13-101.

16. Strichman-Almashanu, L.Z.; Lee, R.S.; Onyango, P.O.; Perlman, E.; Flam, F.; Frieman, M.B.; Feinberg, A.P. A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res.* **2002**, *12*, 543–554, doi:10.1101/gr.224102.
17. Szpakowski, S.; Sun, X.; Lage, J.M.; Dyer, A.; Rubinstein, J.; Kowalski, D.; Sasaki, C.; Costa, J.; Lizardi, P.M. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene* **2009**, *448*, 151–167, doi:10.1016/J.GENE.2009.08.006.
18. Turker, M.S. Gene silencing in mammalian cells and the spread of DNA methylation. *Oncogene* **2002**, *21*, 5388–5393, doi:10.1038/sj.onc.1205599.
19. Yates, P.A.; Burman, R.; Simpson, J.; Ponomoreva, O.N.; Thayer, M.J.; Turker, M.S. Silencing of mouse Aprt is a gradual process in differentiated cells. *Mol. Cell Biol.* **2003**, *23*, 4461–4470, doi:10.1128/mcb.23.13.4461-4470.2003.
20. Moore, L.D.; Le, T.; Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **2013**, *38*, 23–38, doi:10.1038/npp.2012.112.
21. Pugacheva, E.M.; Teplyakov, E.; Wu, Q.; Li, J.; Chen, C.; Meng, C.; Liu, J.; Robinson, S.; Loukinov, D.; Boukaba, A.; et al. The cancer-associated CTCFL/BORIS protein targets multiple classes of genomic repeats, with a distinct binding and functional preference for humanoid-specific SVA transposable elements. *Epigenet. Chromatin* **2016**, *9*, 35, doi:10.1186/s13072-016-0084-2.
22. Zabolotneva, A.A.; Bantysh, O.; Suntsova, M.V.; Efimova, N.; Malakhova, G.V.; Schumann, G.G.; Gayfullin, N.M.; Buzdin, A.A. Transcriptional regulation of human-specific SVA1retrotransposons by cis-regulatory MAST2 sequences. *Gene* **2012**, *505*, 128–136, doi:10.1016/j.gene.2012.05.016.
23. Wang, D.C.; Wang, W.; Zhang, L.; Wang, X. A tour of 3D genome with a focus on CTCF. *Semin. Cell Dev. Biol.* **2018**, *90*, 4–11, doi:10.1016/J.SEMCDB.2018.07.020.
24. Savage, A.L.; Wilm, T.P.; Khursheed, K.; Shatunov, A.; Morrison, K.E.; Shaw, P.J.; Shaw, C.E.; Smith, B.; Breen, G.; Al-Chalabi, A.; et al. An Evaluation of a SVA Retrotransposon in the FUS Promoter as a Transcriptional Regulator and Its Association to ALS. *PLoS ONE* **2014**, *9*, e90833, doi:10.1371/journal.pone.0090833.
25. Vasilioi, S.A.; Ali, F.R.; Haddley, K.; Cardoso, M.C.; Bubb, V.J.; Quinn, J.P. The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro. *Addict. Biol.* **2012**, *17*, 156–170, doi:10.1111/j.1369-1600.2010.00288.x.
26. Kaer, K.; Speek, M. Retroelements in human disease. *Gene* **2013**, *518*, 231–241, doi:10.1016/j.gene.2013.01.008.
27. Aneichyk, T.; Hendriks, W.T.; Yadav, R.; Shin, D.; Gao, D.; Vaine, C.A.; Collins, R.L.; Domingo, A.; Currall, B.; Stortchevoi, A.; et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **2018**, *172*, 897–909, doi:10.1016/J.CELL.2018.02.011.
28. Bragg, D.C.; Mangkalaphiban, K.; Vaine, C.A.; Kulkarni, N.J.; Shin, D.; Yadav, R.; Dhakal, J.; Ton, M.-L.; Cheng, A.; Russo, C.T.; et al. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in <i>TAF1</i>. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E11020–E11028, doi:10.1073/pnas.1712526114.
29. Wang, J.; Song, L.; Grover, D.; Azrak, S.; Batzer, M.A.; Liang, P. dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.* **2006**, *27*, 323–329, doi:10.1002/humu.20307.
30. De Koning, A.P.J.; Gu, W.; Castoe, T.A.; Batzer, M.A.; Pollock, D.D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet.* **2011**, *7*, e1002384, doi:10.1371/journal.pgen.1002384.
31. Halldorsson, B.V.; Istrail, S.; De La Vega, F.M. Optimal selection of SNP markers for disease association studies. *Hum. Hered.* **2004**, *58*, 190–202, doi:10.1159/000083546.
32. Vasieva, O.; Cetiner, S.; Savage, A.; Schumann, G.G.; Bubb, V.J.; Quinn, J.P. Potential Impact of Primate-Specific SVA Retrotransposons During the Evolution of Human Cognitive Function. *Trends Evol. Biol.* **2017**, *6*, doi:10.4081/ni.2017.6514.
33. Lohmueller, K.E.; Pearce, C.L.; Pike, M.; Lander, E.S.; Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **2003**, *33*, 177–182, doi:10.1038/ng1071.
34. Ng, H.H.; Zhang, Y.; Hendrich, B.; Johnson, C.A.; Turner, B.M.; Erdjument-Bromage, H.; Tempst, P.; Reinberg, D.; Bird, A. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat. Genet.* **1999**, *23*, 58–61, doi:10.1038/12659.
35. Geiman, T.M.; Sankpal, U.T.; Robertson, A.K.; Zhao, Y.; Zhao, Y.; Robertson, K.D. DNMT3B interacts with hSNF2H chromatin remodeling enzyme, HDACs 1 and 2, and components of the histone methylation system. *Biochem. Biophys. Res. Commun.* **2004**, *318*, 544–555, doi:10.1016/j.bbrc.2004.04.058.

36. Fuks, F.; Burgers, W.A.; Brehm, A.; Hughes-Davies, L.; Kouzarides, T. DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat. Genet.* **2000**, *24*, 88–91, doi:10.1038/71750.
37. Fuks, F.; Hurd, P.J.; Deplus, R.; Kouzarides, T. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Res.* **2003**, *31*, 2305–2312, doi:10.1093/nar/gkg332.
38. Stacey, S.N.; Kehr, B.; Gudmundsson, J.; Zink, F.; Jonasdottir, A.; Gudjonsson, S.A.; Sigurdsson, A.; Halldorsson, B.V.; Agnarsson, B.A.; Benediktsdottir, K.R.; et al. Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. *Hum. Mol. Genet.* **2016**, *25*, 1008–1018, doi:10.1093/hmg/ddv622.
39. Jacob, A.G.; Smith, C.W.J. Intron retention as a component of regulated gene expression programs. *Hum. Genet.* **2017**, *136*, 1043–1057, doi:10.1007/s00439-017-1791-x.
40. Pollard, M.O.; Gurdasani, D.; Mentzer, A.J.; Porter, T.; Sandhu, M.S. Long reads: Their purpose and place. *Hum. Mol. Genet.* **2018**, *27*, R234–R241, doi:10.1093/hmg/ddy177.
41. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 30, doi:10.1186/s13059-020-1935-5.
42. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, 7, doi:10.1186/s13742-015-0047-8.
43. Gibbs, J.R.; van der Brug, M.P.; Hernandez, D.G.; Traynor, B.J.; Nalls, M.A.; Lai, S.L.; Arepalli, S.; Dillman, A.; Rafferty, I.P.; Troncoso, J.; et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **2010**, *6*, e1000952, doi:10.1371/journal.pgen.1000952.

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).