UNIVERSITY OF
LIVERPOOL

Hominid retrotransposons as a modulator of genomic function

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy

By Abigail Lucy Savage

October 2013

Acknowledgements

I am very grateful to my supervisors Prof. John Quinn and Dr Jill Bubb for all their help over the three years of my PhD. I would like to acknowledge our collaborators at King’s College London and St Mary’s Hospital Manchester for their assistance and access to clinical samples, in particular Dr Gerome Breen for his advice regarding statistical analysis and the use of genome browser software. I would also like to thank the Anatomy department for funding my PhD. Also a special mention to all members of white block past and present for making it a great place to work and learn; I made some fabulous friends and a very special housemate for which I am very grateful. A final thank you to my family, Tony and all my friends.

Abstract

Transposable elements constitute 45% of the human genome contributing to our evolution, creating new exons, structural variation and influencing the regulation of transcription. SINE-VNTR-Alus (SVAs) are a hominid specific retrotransposon that are still actively retrotransposing in the human genome today. The structure and sequence of SVAs, in particular their variable number tandem repeat (VNTR) domain, suggest their potential for influencing the regulation of gene expression through binding of transcription factors, differential methylation patterns and formation of secondary structures along with potential for genetic variation between individuals. This project has identified novel regulatory domains and genetic variation within elements belonging to a hominid specific group of retrotransposons. A global analysis undertaken of their distribution identified their preference for genic regions over gene deserts and their insertion into functional regions of the genome such as promoters and introns. An in depth analysis of two SVA insertions, one upstream of the FUS gene and another upstream of the PARK7 gene, demonstrated the ability of SVAs to affect reporter expression *in vitro* and *in vivo*. Both of these SVAs were identified as polymorphic in their central VNTR regions and the PARK7 SVA also demonstrated different copy numbers of repeats it its 5’ CCCTCT domain. Analysis of the PARK7 SVA insertion and gene in cell lines indicated the SVA is not epigenetically silenced, as dogma might suggest to suppress retrotransposition, but present in a transcriptionally active region of the genome. There is increasing evidence for loss of silencing of retrotransposons including within the human brain which would allow for greater influence of potential transcriptional properties embedded within SVAs impacting on genomic function.

Contents

Chapter 1

General Introduction

The regulation of gene expression is a complex process, from the sequence of the DNA itself in promoters, enhancers and repressors to the epigenetic modifications and the way that DNA is packaged. Genetic variation in the human genome between individuals allows for differential modulation of gene expression and response to environmental cues, leading to potential phenotypic differences and predisposition to disease. The sequencing of the genome of multiple species and human individuals has led to a greater understanding of the components of the genome and potential sources of genetic variation. The aim of this study is to identify novel domains involved in the regulation of gene expression and determine whether the genetic variation encompassed within such regions could contribute to functional differences between individuals.

1.1 Composition of the human genome

The human genome is over 3 billion base pairs in length with less than 2% coding for proteins, however the exact number of genes is still unknown (Clamp et al. 2007). Much of the human genome is composed of repeated sequences that fall into several categories: interspersed repeats derived from transposons, duplicated genes, simple sequence repeats, large blocks of duplicated sections of the genome and blocks of tandemly repeated sequences at sites such as telomeres (Lander et al. 2001). The non-coding part of the genome, initially thought of as ‘junk’, at the very least harbours important functional domains including promoters, enhancers and repressors to control and fine tune the levels of gene expression and non-coding RNAs which have structural and regulatory roles.

The greatest degree of conservation across species is within exons, especially those of genes with roles in development to maintain key functions of the proteins they encode. Evolutionary conserved regions (ECRs) of the human genome (other than exons) can indicate domains involved in other functions such as gene regulation, as it is hypothesised their sequence has been maintained through evolution for a purpose (MacKenzie and Quinn 2004; Visel et al. 2007). Many of these ECRs are highly conserved and their ability to affect gene expression has been demonstrated *in vitro* and *in vivo* (Prabhakar et al. 2006; Bonello et al. 2011; Paredes et al. 2011). Proteins are highly conserved for specific function but it is also the amount, tissue specificity and the modulation of gene expression in response to the environment that can account for phenotypic differences. Therefore important regulatory domains may not only be located in the ECRs but regions containing genetic variation that may in part account for the differences between species and individuals of the same species. These polymorphic regulatory domains may come in different forms ranging from repetitive sequences such as variable number tandem repeats (VNTRs) to transposable elements inserted into our genome. Regulatory domains may be located within introns or even hundreds of kilobases away from the start of transcription (Visel et al. 2009a; Visel et al. 2009b). Identifying potential regulatory domains in a vast amount of non-coding DNA can prove difficult, but the wealth of data now freely available for the genomes of many species within genome browsers can aid this search. Not only can the homology and the sequence of a region indicate potential function but the information held in these databases, such as the state of chromatin and transcription factor binding across different cell lines, can help narrow down these potential functional domains. Much of this data is contained within the UCSC genome browser (http://genome.ucsc.edu/index.html), including the encyclopaedia of DNA elements (ENCODE) data. ENCODE is freely available data on the functional elements of the human genome which is an international collaboration between research groups funded by the National Human Genome Research Institute (NHGRI) (http://www.genome.gov/10005107) and is a very useful tool in analysing potential functional regions within the genome.

1.2 Genetic variation within the human genome

Genetic variation within the genome includes single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and VNTRs. SNPs, the most common form of genetic variation, are changes at a single nucleotide base and must occur at a greater frequency than 1% of the population to be considered polymorphic rather than a mutation. SNPs are found throughout the human genome with a recent study of 1092 human genomes identifying 38 million SNPs (Abecasis et al. 2012). SNPs within coding exons may introduce a change at the amino acid level (non synonymous SNPs) and can therefore influence the function of a protein. SNPs in non-coding regions may alter features such as transcription factor binding or splicing for example. SNPs can be used to map the inheritance of blocks of DNA known as haplotypes and the International HapMap project (http://hapmap.ncbi.nlm.nih.gov/) was established to create a map of the haplotypes in the human genome and common genetic variants across different populations. Many multifactorial diseases such as heart disease, cancer, diabetes and neurological degenerative diseases have both an environmental and genetic component. Diseases that involve neurological degeneration include Multple Sclerosis and Parkinson’s Disease but should be differentiated from non-pathological cognitive decline that is associated with normal ageing. It is the small differences in the genetics of an individual that may reveal the predisposition of one individual compared to another for a specific disorder. Different methodologies have been used to analyse genetic variants to identify risk alleles for a specifc disease and include genome wide association studies (GWAS) and candidate gene studies. GWAS have used SNPs to map susceptibility loci in the human genome correlated with disease (Wagner 2013). The SNPs in these identified regions may be directly influencing a genetic predisposition to a disease or it could be another element of genetic variation inherited in the same haplotype playing a role. Candidate gene studies involve the analysis of genes that have been previously linked to a particular disease and have contributed greatly to the study of risk variants, however are limited by prior knowledge of the disease processes.

Another type of genetic variation linked to a predisposition for specific disease and a role in gene regulation are VNTRs, which are repeated sequences of nucleotides that vary in copy number between individuals. There are 600,000 candidate VNTRs within the human genome, many located in functional positions such as intron-exon splicing junctions or promoters where the repetitive nature of the VNTR can provide multiple transcription factor binding sites (Breen et al. 2008; Sawaya et al. 2013). The number of alleles of a specific VNTR can vary, for example there are two alleles of a VNTR located in the promoter of the serotonin transporter gene (SLC6A4) and 9 alleles identified of a VNTR located in the 3’UTR of the dopamine transporter gene (SLC6A3) (Haddley et al. 2008). VNTRs have demonstrated the ability to direct gene expression in a tissue and allele specific manner *in vitro* and *in vivo* (MacKenzie and Quinn 1999b; Roberts et al. 2007; Vasiliou et al. 2012) and have been linked to genetic predisposition to disease

(Anguelova et al. 2003; Herman et al. 2005; Guindalini et al. 2006; Munafo and Johnstone 2008).

Two VNTRs located within the SLC6A4, one in the promoter and the other in intron 2, have been widely studied for their functional properties in gene regulation and association with disease due to the role of the serotonin transporter in modulating the levels of serotonin in the synaptic cleft between neurons and therefore serotonin signalling. The VNTR located within the promoter of the SLC6A4 gene has two alleles consisting of either 14 or 16 copies of a 22-23bp repeat (LPR) and the VNTR within intron 2 has three alleles of 9, 10 or 12 copies of a 16-17bp repeat (Stin2). A variety of techniques have been employed to analyse the transcriptional properties of these VNTRs including chromatin immunoprecipitation (ChIP), transient and stable transfections of reporter gene plasmids and generation of transgenic mice (MacKenzie and Quinn 1999b; Ali et al. 2010; Vasiliou et al. 2012). ChIP was used to identify the binding of specific transcription factors, such as CCCTC binding-factor (CTCF), methyl CpG binding protein 2 (MeCP2) and Y box binding protein 1 (YB-1), across the two VNTRs with differential binding observed across the alleles of the VNTRs in response to challenge indicating expression of SLC6A4 could be modulated in an allelic dependant manner *in vitro* (Vasiliou et al. 2012). The VNTRs of the SLC6A4 gene demonstrated the ability to differentially modulate expression in a reporter gene model *in vitro* and this expression can be modulated by the transcription factor CTCF (Ali et al. 2010). Two of the alleles of the Stin2 VNTR (10 and 12) were used to generate transgenic mice where the alleles displayed differential regulatory properties in a region of the hindbrain known to express the SLC6A4 gene at that time in development (MacKenzie and Quinn 1999b). Both of these VNTRs have been analysed in a variety of cohorts for associations with neurological and psychiatric diseases such as bi and uni polar depression, schizophrenia and migraine, the aim being to identify ‘at risk’ alleles of these VNTRs for a specific disorder (Haddley et al. 2012). There have been associations found in some cohorts but these have not replicated in others; which may be due to ethnic differences, the need to consider more than one genetic variant in the pathway to identify stronger links, environmental factors, the parameters used to measure specific traits, publication bias or a lack of sufficient size to reach statistical significance. The VNTRs of the SLC6A4 are examples of the functional properties embedded within VNTRs, including the allelic dependant modulation of gene expression *in vitro, in vivo* and in response to challenge, and identifies methods that can be employed to determine the gene regulatory potential within specific loci in the genome.

Mobile DNA or transposable elements (TEs) are another source of genetic variation between species and potentially individuals of the same species if still actively transposing. An element called a SINE-VNTR-Alu (SVA) is a hominid specific TE (containing a VNTR domain) and it is the potential functional role, the genetic variation they contribute to the human genome, and association with disease that is to be addressed throughout this project.

1.3 Transposable elements

Transposable elements were first identified by Barbara McClintock in 1950 while studying gene regulation in maize (Mc 1950). There are two classes of TEs: class I or retrotransposable elements that move within the genome through a ‘copy and paste’ mechanism and class II elements or DNA transposons that move through a ‘cut and paste’ mechanism. Retrotransposable elements are mobilised through a RNA intermediate that is reverse transcribed and it is this cDNA ‘copy’ that is inserted back into the host genome at a different loci to the source element increasing their numbers present in the host genome. DNA transposons encode a transposase that removes or ‘cuts’ the transposon from its locus in the host genome and inserts it at a different site. The percentage of the genome that consists of TEs differs between species as does the type of TEs present (Chenais et al. 2012) with nearly half of the human genome consisting of TEs. DNA transposons constitute approximately 3% of the human genome and are no longer active (Lander et al. 2001). The retrotransposable elements can be further subdivided into two main groups: long terminal repeats (LTR) retrotransposons and non-LTR retrotransposons. The non-LTR retrotransposons contain the only known currently active TEs in the human genome and include long interspersed elements (LINEs), short interspersed elements (SINEs), SINE-VNTR-Alus (SVAs) and processed pseudogenes (PP). The structure of the different types of TEs are shown in Figure 1.1 (Beck et al. 2011).

The cell type in which new insertions of TEs occur will determine if they are passed onto the next generation (Figure 1.2) (Muotri et al. 2007). New insertions into primordial germ cells or very early in development in germ cell progenitors will be passed onto following generations. Insertions into other cell types within early development will not be heritable but will contribute to somatic mosaicism of the individual. A study carried out by Kano et al used transgenic mouse and rat models with human and mouse LINE-1 (L1) elements to determine when and where L1 retrotransposition was occurring in development. It was identified that L1 retrotransposition occurred more frequently in embryonic development compared to within germ cells, estimated at approximately 1 out of 50 to 1 out of 500 somatic cells compared to 1 out of 1000 sperm, resulting in fewer heritable insertions and greater somatic mosaicism (Kano et al. 2009).

This text box is where the unabridged thesis included the following third party copyrighted material:

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics* **12**: 187-215.
Figure 1

Figure 1.1: Structure of the different types of transposable elements in mammals. (Beck et al. 2011)

This text box is where the unabridged thesis included the following third party copyrighted material:

Muotri AR, Marchetto MC, Coufal NG, Gage FH. 2007. The necessary junk: new functions for transposable elements. *Human molecular genetics* **16 Spec No. 2**: R159-167.
Figure 3

Figure 1.2: Consequences of retrotransposition in different cell types.. (Muotri et al. 2007)

TEs have often been referred to as parasitic or selfish DNA indicating no benefit to the host genome; however in the genome of nearly all living organisms TEs of one class or another have been identified. TEs have extensively colonised the human genome throughout evolution, even with measures to prevent their transposition in place, there are still active elements altering our genome today. TEs have had a negative impact through disease causing insertions; however there is evidence for their positive contribution to our genomes adaptability during evolution through germ line insertions creating new heritable genes and gene regulation pathways. The host genome has learnt to tolerate or even co-evolved alongside TEs incorporating them into genomic processes; for example Alu elements are involved in driving expression of a cluster of miRNA on chromosome 19 via RNA polymerase III and new evidence indicating that Alu sequences (found in >5% of human UTRs) themselves are targeted by miRNA and could therefore be involved in global post transcriptional regulation of gene expression (Muotri et al. 2007).

1.3.1 Retrotransposons

Retrotransposons replicate through a RNA intermediate increasing their number in the host genome and expanding it. The mechanism through which this retrotransposition occurs differs between the types of element but the enzyme reverse transcriptase is key to this process, creating a DNA copy of the RNA transcribed by the host cell. Autonomous retrotransposons encode for the proteins that are required for their own retrotransposition, in contrast non-autonomous retrotransposons do not encode proteins and need to ‘hijack’ the proteins coded for by the autonomous elements for their retrotransposition. The most successful of these elements in the human genome are the non-LTRs.

1.3.1.1 Long terminal repeat retrotransposons

Long terminal repeats (LTR) are an autonomous retrotransposon named due to their characteristic direct repeats at their 5’ and 3’ ends that are around 300-1000bp long and contain promoter sequences that regulate transcription of the element. Human endogenous retroviruses (HERVs) are members of the LTR family and constitute 8% of the human genome (Lander et al. 2001) and a full length element is approximately 9.5kb in length (Shin et al. 2013). HERVs are previous retroviral infections that entered the germ line of the host, became unable to reinfect and remained within the host genome (Goodier and Kazazian 2008) and therefore HERVs share many characteristics of retroviruses including the proteins they code for and their life cycle. HERVs are divided into three classes (I, II, III) by their similarities to genera of retroviruses and then further subdivided by a primer binding site for a specific tRNA molecule (e.g. HERV-W, HERV-K) when reverse transcription is initiated (Griffiths 2001). The genes of the HERV sequence (*gag*, *pro*, *pol* and *env*) encode for proteins with properties such as a protease, reverse transcriptase, integrase and also structural proteins, however there are many inactivating mutations found within their sequences and these open reading frames (ORFs) no longer code for functional proteins (Kim 2012). There are polymorphic members of the HERV-K elements in the human population in terms of their absence or presence at specific loci, indicating they have been actively retrotransposing since the human-chimpanzee divergence (Belshaw et al. 2005; Shin et al. 2013). L1, Alus and SVAs are known to be actively retrotransposing today due the identification of *de novo* insertions within specific individuals. There have been no *de novo* insertions of HERVs identified therefore it is uncertain whether these elements are still active or have undergone a recent cessation in their retrotransposition capability.

1.3.1.2 Non-LTRs retrotransposons

1.3.1.2.1 Long interspersed elements

Long interspersed elements (LINEs) are the only autonomous non-LTR retrotransposon in the human genome. The majority of these are LINE-1 (L1) elements with approximately 500,000 copies constituting 18% of the human genome (Goodier and Kazazian 2008). A full length L1 element is 6kb in size with two ORFs (Scott et al. 1987) with both ORF encoded proteins required for retrotransposition (Moran et al. 1996). ORF1 encodes for a 40kDa protein (ORF1p) that binds to nucleic acids (single stranded preferentially) (Hohjoh and Singer 1997) and ORF2 encodes for a 150kDa protein (ORF2p) with reverse transcriptase and endonuclease functions (Mathias et al. 1991; Feng et al. 1996). The L1 encoded proteins demonstrate a *cis* preference for their encoding RNA over L1 RNA from retrotransposition deficient elements and cellular mRNAs, to ensure functioning L1 RNA is more likely to be inserted into the host genome (Wei et al. 2001). Retrotransposition of the L1 elements occurs through a process called target primed reverse transcription (TPRT) and this is summarised in Figure 1.3 (Babushok and Kazazian 2007). The L1 RNA is transcribed by RNA polymerase II which is regulated by a promoter within the 5'UTR of the L1 (Minakami et al. 1992) and then exported into the cytoplasm of the cell where translation of the ORF1 and ORF2 proteins occurs. The ORF1p, ORF2p and L1 RNA form a ribonucleoprotein complex (L1 RNP) which is transported back in to the nucleus. The ORF2p with its endonuclease activity nicks the bottom strand of the host's DNA at the consensus sequence 5'TTTTAA 3' at the TA site. This exposes a 3' hydroxyl group that the ORF2p uses as a primer to reverse transcribe the L1 RNA. The ORF2p then nicks the top strand of the host DNA and the newly reverse transcribed cDNA of the L1 is integrated into the host genome. The complementary strand of DNA is then synthesised. TPRT can result in target site duplications (TSDs), 5’ truncations, 3’transductions and internal rearrangement and inversions.

This text box is where the unabridged thesis included the following third party copyrighted material:

Babushok DV, Kazazian HH, Jr. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Human mutation* **28**(6): 527-539.
Figure 3

Figure 1.3: Model of L1 integration reaction. (Babushok and Kazazian 2007)

More than 99.9% of the L1s in the human genome are no longer active due to mutations in their ORFs or rearrangements in their structure such as inversions and truncations (Lander et al. 2001). Brouha et al identified 90 L1 elements with intact ORFs from the 2001 working draft of the haploid human genome sequence and assayed 82 of these elements for their retrotransposition capabilities, they found 40 of which were active in the cell culture retrotransposition assay. Of these elements six were highly active and were responsible for 84% of the retrotransposition capability assayed. This led to the prediction that there are 80-100 L1 elements that are retrotransposition competent in a given human genome with smaller number of 'hot' elements that are responsible for the majority of retrotransposition in the human population (Brouha et al. 2003). L1 elements have not only expanded the human genome through their own proliferation but also mobilise non-autonomous retrotransposons including SINEs, SVAs and PPs.

1.3.1.2.2 Short interspersed elements

Over 120 families of short interspersed elements (SINEs) have been identified in eukaryotic genomes originating from cellular RNA sequences transcribed by RNA polymerase III and are on average 150-300bp long (Kramerov and Vassetzky 2011). The most successful SINE to populate the human genome is the primate specific Alu element with more than 1 million copies (Lander et al. 2001) and many different subfamilies that have been actively expanding our genome for the past 65 million years (Batzer and Deininger 2002). These elements were named Alu due the presence of the AluI restriction enzyme site in their sequence (Houck et al. 1979). Alus are 300bp long and their sequence originates from a processed 7SL RNA gene (Ullu and Tschudi 1984). Alus contain an internal RNA polymerase III promoter to regulate their transcription and are retrotransposed by the L1 encoded proteins (Batzer and Deininger 2002).

1.3.1.2.3 Processed pseudogenes

Processed pseudogenes (PP) or retropseudogenes are RNAs that have been reverse transcribed and inserted back into the genome by L1 retrotransposition machinery (Pavlicek et al. 2006), they do not contain introns and have a 3’ poly A tail (Ding et al. 2006). There are an estimated 11,000 PPs in the human genome (Beck et al. 2011). PPs are mostly transcriptionally silent as the sequence of the gene retrotransposed does not include promoter and regulatory sequences, however there is evidence of some PPs being transcribed with one study estimating this number at 4-6% (Harrison et al. 2005). PPs are no longer under selective pressure and the vast majority accumulate mutations and do not code for proteins. There are examples of PPs that have maintained their protein coding ability and are transcribed and translated into a functional protein and these are called retrogenes (Ding et al. 2006) with an estimated 120 of these retrogenes in the human genome (Vinckenbosch et al. 2006).

1.3.1.2.4 SINE-VNTR-Alus

SINE-VNTR-Alus (SVAs) are the youngest of the retrotransposable elements in the human genome and are hominid specific. They are the most recent to be identified and least widely studied but will be the focus of this thesis.

Historically SVAs were originally identified as a sequence derived from part of the envelope *(env*) gene and a 3'LTR from the HERV-K10 endogenous retrovirus with a poly A-tail and a GC-rich tandem repeat directly upstream and were named SINE-R elements (Ono et al. 1987; Zhu et al. 1992). The *env* gene of a HERV codes for envelope surface and transmembrane proteins of the retrovirus. It was later shown that in the C2 gene, the GC-rich tandem repeat of the SINE-R element was a VNTR. This composite element was termed a SINE-VNTR-Alu (SVA) when further analysis of its components revealed the Alu-like sequences adjacent to the VNTR (Shen et al. 1994). Thus typically SVAs consist of a hexamer repeat (CCCTCT), an Alu-like sequence, a GC-rich VNTR, a SINE and a poly A-tail, however a proportion of the SVAs contain two central GC-rich VNTRs as opposed to one (Figure 1.1). SVAs vary in length from 700-4000bp with 63% of SVA insertions in the human genome full length: containing all five domains of a canonical element (Wang et al. 2005). A precursor of the VNTR domain found within the SVAs is present within the rhesus macaque genome, many of these precursor elements are also present in the human genome suggesting they were retrotransposing prior to the divergence of the old world monkeys and the hominoids (Han et al. 2007). The precursor sequence was termed SVA2 and contains a GC-rich VNTR, a unique 3' sequence and a poly A tail with 40 copies identified in the rhesus macaque genome (Hancks and Kazazian 2010).

SVAs are divided into subtypes (A-F) by the SINE region and their age estimated at 13.56 million years (Myrs) old for the oldest subtype (A) and 3.18Myrs old for the youngest subtype (F) (Wang et al. 2005). A seventh subtype has been identified that contains a 5' transduction of the sequence from the first exon of the MAST2 gene and associated CpG island and has been referred to as either CpG-

SVA, MAST2 SVA or SVA F1 (Bantysh and Buzdin 2009; Damert et al. 2009; Hancks et al. 2009). The sequence of the MAST2 loci that has been incorporated into the F1 structure has been shown to act as a positive regulator of transcription in a reporter gene construct when transfected into human germ cells and is thought to have contributed to the success of the subtype in its retrotransposition (Zabolotneva et al. 2012). The percentage of SVAs for each of the subtypes in the human genome is shown in figure 1.4. Subtype D is by far the largest consisting of 44% of all SVAs with the most recent F1s the smallest group at 3%. Subtypes E, F and F1 are human specific as are some members of SVA subtype D with a total of 864 SVA insertions since the human-chimpanzee divergence ~6 million years ago (Mills et al. 2006). After the human-chimpanzee divergence the SVAs continued to expand within the chimpanzee genome as well as the human creating chimpanzee specific insertions including a subtype unique to chimpanzees called SVA PtA (Wang et al. 2005).

Actively retrotransposing elements cause inter-individual variation between humans with elements being polymorphic for their absence or presence, SVAs included. This has been analysed for a group of human specific SVAs which estimated that 37.5 % of SVA Es and 27.6% of SVA Fs were polymorphic for their presence in the genome (Wang et al. 2005) and the average human is estimated to have 56 SVA absence/presence polymorphisms (Bennett et al. 2004). The frequency of this presence or absence of specific SVAs located in HLA genes has been shown to be variable between groups with different ethnic origins (Kulski et al. 2010).

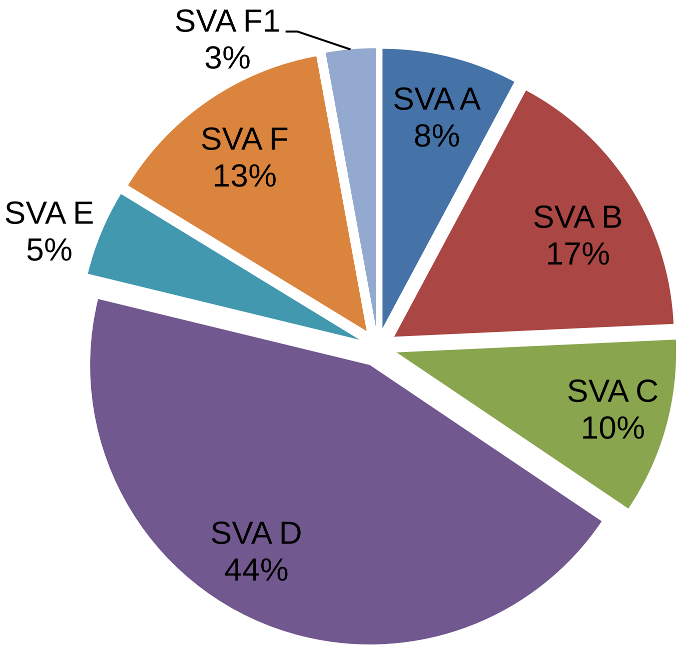


Figure 1.4: The proportion of each SVA subtype in the Hg19 according to UCSC. The SVA subtypes A-F are defined by their SINE region and SVA F1 is subtype F with addition of sequence from the exon 1 of the MAST2 gene.

The site of SVA insertions show the hallmarks of LINE-1 mediated retrotransposition such as insertion at a consensus L1 endonuclease recognition motif (5'TTTTAA 3'), poly A-tails, inversions and rearrangements, target site duplications, truncations and transductions (Hancks and Kazazian 2010). The mobilisation by the L1 protein machinery was validated by two separate studies (Hancks et al. 2011; Raiz et al. 2012) and their retrotransposition rate is estimated at 1 in every 916 births (Xing et al. 2009). To demonstrate the methods employed to analyse the ability of SVAs to retrotranspose the cell line model used by Hancks et al will be outlined. A 'passenger' plasmid was generated containing the sequence of the SVA to be tested for its retrotransposition capability marked with a neomycin retrotransposition indicator cassette. The cassette contains a SV40 promoter and neomycin resistance gene (*neo*) on the opposite strand to the SVA (antisense). The *neo* gene contains an intron on the sense strand interrupting the gene. Therefore the *neo* gene will only confer resistance to the host cell once it has undergone transcription, splicing, reverse transcription and integration into the genome and the ORF is restored. This cassette can be used to show if retrotransposition of the SVA has occurred in the cell line by selecting for cells with the resistance gene and the retrotransposition frequency can be calculated. The passenger plasmid is co-transfected into the cell line (in this study HeLa cells) with a 'driver' plasmid which is a highly active unmarked L1 element. The driver plasmid can be modified, for example to introduce mutations in the ORFs to determine the requirements for SVA retrotransposition in greater detail. In the studies by Hancks et al and Raiz et al, SVAs from subtypes D, E and F1 were shown to be retrotransposition competent in multiple cell lines but at differing frequencies with the ORF2p essential for retrotransposition but the requirement of ORF1p was variable. The retrotransposition capability of the SVA D tested by Hancks et al was ORFp1 independent whereas the SVA F1 required ORFp1. L1 retrotransposition requires both of its ORF encoded proteins but Alus, like SVAs mobilised in *trans* by the L1 machinery, do not require the ORFp1 (Dewannieux et al. 2003). The L1 encoded proteins show a *cis* preference for their encoded RNA therefore some non-autonomous elements may have evolved to require only ORFp2 in an attempt to increase their success.

The regulation of transcription of the SVA mRNA is yet to be fully defined unlike the regulation of L1 and Alu elements. A recent study to determine the nature of SVA retrotransposition revealed that no individual domain of an SVA is fundamental for this to occur, but each domain differentially affected the rate at which retrotransposition can take place in the human osteosarcoma cell line (U2OS) (Hancks et al. 2012). Removal of the CCCTCT repeat or the central VNTR reduced the retrotransposition activity of the SVA by 75% and 79% respectively however the retrotransposition activity of the CCCTCT repeat and Alu-like sequence alone was nearly double that of the whole SVA. All domains of the SVA are dispensable to a certain degree and retrotransposition of the sequence will still occur in the cell line model used by Hancks et al (outlined above). The CCCTCT repeat and the Alu-like sequence of a SVA F were shown to have some promoter activity when cloned into a promoter-less vector and transfected in to the Tera-1 cell line (Zabolotneva et al. 2012). This study also showed that the acquisition of the sequence of the MAST2 CpG island enhanced the transcriptional activity of the F1 subtype showing that the incorporation of an external regulatory sequence contributed to the success of this recent SVA subtype.

SVAs can influence the genomic location of their insertion through mechanisms such as alternative splicing, exon shuffling, formation of secondary structure, recombination events and generation of differentially methylated regions (Hancks and Kazazian 2010).

1.3.2 The impact of transposable elements on the human genome

Transposable elements, despite long being thought of as ‘junk’ DNA, have influenced the human genome during its evolution through mechanisms such as insertional mutagenesis, recombination events, exonisation and modulation of gene expression (Muotri et al. 2007; Goodier and Kazazian 2008). Several classes of TEs are actively retrotransposing creating human specific traits within our genome and even between human individuals, whether harmful or beneficial, these differences can impact on our phenotype. There is a vast array of literature focused on the L1 and Alu retrotransposable elements and their impact on the human genome as they have been the most widely studied. The focus of this thesis is the SVA retrotransposons and therefore will be discussed here in greater detail.

Insertions of the actively retrotransposing non-LTRs into coding or intronic regions affecting transcriptional regulation and processing of genes have been associated with diseases, including haemophilia, Duchenne muscular dystrophy, cystic fibrosis and several cancers (Hancks and Kazazian 2012; Kaer and Speek 2013). To date 8 SVA insertions have been associated with disease (Hancks and Kazazian 2012; Kaer and Speek 2013) through a variety of mechanisms including exon skipping and decreased mRNA production. Table 1.1 adapted from Hancks and Kazazian 2012 summarises the disease causing insertions that are outlined in detail in the following paragraphs.

The first human disease identified to be caused by a SVA insertion was Fukuyama-type congenital muscular dystrophy (FCMD), which is one of the most common autosomal recessive diseases in Japan (0.7-1.2 per 10,000 births). A SVA insertion in the 3'UTR of the fukutin gene is found in 87% of individuals with FCMD and is rarely found outside of the Japanese population (Kobayashi et al. 1998; Watanabe et al. 2005). The insertion of the SVA causes abnormal splicing creating a new splice site removing the original stop codon and creating an additional exon coded for by the sequence of the SVA. The protein therefore lacks 38 amino acids located at the C-terminal of the normal fukutin protein and contains an additional 129 amino acids encoded by the SVA causing mislocalisation of the protein in mammalian cells (Taniguchi-Ikeda et al. 2011). The authors were able to prevent the pathogenic exon trapping of the SVA in cells from FCMD patients and model mice using antisense oligonucleotides returning normal protein levels to approximately 40%.

Abnormal mRNA levels caused by a SVA insertion, in this case within an intron of the TAF1 gene, were linked to X-linked dystonia-parkinsonism (XDP). This resulted in a tissue specific reduction of mRNA of the RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa (TAF1) gene in the caudate nucleus of the patient’s brain and was linked with hypermethylation of the SVA within the same region (Makino et al. 2007). In addition to the reduction of TAF1 mRNA there was also a decrease in the expression of the dopamine receptor D2 gene in the caudate nucleus. The insertion of a SVA either into intron 1 of the LDLRAP1 gene or into exon 3 of the PNPLA2 gene cause autosomal recessive hypercholesterolemia (ARH) and neutral lipid storage disease with subclinical myopathy (NLSDM) respectively, through a decrease in mRNA production with levels undetectable in the assays performed (Wilund et al. 2002; Akman et al. 2010). Further analysis revealed abnormally spliced transcripts of these genes that were then predicted to be subjected to nonsense mediated mRNA decay (NMD) (Taniguchi-Ikeda et al. 2011). In hereditary eliptocytosis and pyropoikiolcytosis (HE and HPP) the insertion of a SVA into exon 5 of the α-spectrin gene resulted in the skipping of this exon and production of a truncated dysfunctional protein (Hassoun et al. 1994). A case of Lynch Syndrome, an autosomal dominant predisposition to cancer, was linked to the production of a mutant transcript which was degraded by NMD caused by an intronic SVA insertion into the PMS2 gene (van der Klift et al. 2012). The authors discuss the difficulties encountered when identifying large heterozygous insertions in genomic DNA due to the preferential amplification of the wild type allele. These technical difficulties are likely to be masking the actual number of this type of disease causing insertion in autosomal dominant disorders.

A detailed study of seven patients with X-linked agammaglobulinemia (XLA) and large scale genomic alterations in the BTK gene identified one patient with an insertion of a fragment of a SVA in exon 9 of the BKT gene causing skipping of that exon (Rohrer et al. 1999). A further study by Conley et al identified another XLA patient with an Alu insertion at the same site as the previously mentioned SVA insertion indicating this site is vulnerable to retrotransposons insertions (Conley et al. 2005). Finally an identical SVA insertion accompanied by a 14kb deletion including the entire HLA-A gene has been linked to leukaemia in three Japanese families (Takasu et al. 2007). Although these three families appear to be unrelated they originate from the same area of Japan suggesting this SVA insertion event may have occurred in a single ancestral individual. The authors compare this SVA insertion and 14kb deletion to the insertion of a L1 element accompanied by a 46kb deletion in the PDHX gene resulting in pyruvate dehydrogenase complex deficiency (Mine et al. 2007), suggesting a similar mechanism is occurring in both of these insertions.

Disease	Gene	SVA Subtype	Size (kb)	Loci of insertion	Affect of Insertion	Reference
FCMD	FKTN	E	3	3'UTR	Alternative splicing/exon trapping	Kobayashi et al 1998 Wantanbe et al 2005 Taniguchi-Ikeda et al 2011
XDP	TAF-1	F	2.6	Intron	Tissue specific mRNA reduction	Makino et al 2007
ARH	LDRAP1	E	2.6	Intron	Alternative splicing	Wilund et al 2002 Taniguchi-Ikeda et al 2011
NLSDM	PNPLA2	E	1.8	Exon	Alternative splicing	Akman et al 2010 Taniguchi-Ikeda et al 2011
HE and HPP	SPTA1	E	0.63	Exon	Exon skipping	Hassoun et al 1994
Lynch Syndrome	PMS2	F	2.2	Intron	Alternative splicing	van der Klift et al 2012
XLA	BTK	-	0.25	Exon	Exon skipping	Rohrer et al 1999 Conley et al 2005
Leukaemia	HLA-A	F1	2	-	14kb deletion	Takasu et al 2007

Table 1.1: The eight SVA insertions that have been associated with disease. Information taken from (Hancks and Kazazian 2012) and cited articles in the table. Diseases: FCMD - Fukuyama-type congenital muscular dystrophy, XDP - X-linked dystonia-parkinsonism, ARH - Autosomal Recessive Hypercholesterolemia, XLA – X-linked agammaglobulinemia, HE – hereditary eliptocytosis, HPP – hereditary pyropoikiolcytosis, NLSDM – neutral lipid storage disease with subclinical myopathy.

Instances of both germ line and somatic insertions of non-LTRs into known cancer related genes have been associated with several types of cancer including breast, ovarian, colon and leukaemia (Chenais 2013). TEs are linked to the development of cancer however the environment of the cancer cells can lead to the activation of these elements and may contribute to the progression of the disease through their mutagenic properties. It has been demonstrated that retrotransposons, in particular the primate specific elements, undergo a loss of methylation in the tumour compared to normal tissue (Szpakowski et al. 2009). This loss of epigenetic silencing of retrotransposons could enable their retrotransposition or open up their regulatory properties impacting on the structure of the genome and gene transcription of the tumour cells. Analysis of somatic retrotransposition in five types of cancer revealed 194 somatic insertions preferentially into genes that are commonly mutated in the tumour (Lee et al. 2012a). The expression of L1 and HERV transcripts have been identified in both breast and ovarian cancers (Bratthauer et al. 1994; Wang-Johanning et al. 2001; Menendez et al. 2004; Wang-Johanning et al. 2007) and high levels of RNA of retrotransposons have also been detected in tumour microvesicles which could potentially contribute to further genomic instability if delivered to other cells (Balaj et al. 2011).

The human specific NANOGP8 processed pseudogene that originated from the retrotransposition of the parent gene NANOG, a transcription factor involved in maintaining pluripotency, inserted into the SINE region of a SVA A already present within in the human genome and demonstrated oncogenic properties (Fairbanks et al. 2012). Experimental evidence has yet to link the regulation of NANOGP8 expression to its site of insertion into the SVA A but the authors hypothesise the regulatory properties of the SVA, in particular the LTR of the SINE, may play a role in the tissue specific expression of this tumourigenic processed pseudogene. NANOGP8 was transcribed in several cancer lines including human osteosarcoma (OS732), human hepatoma (HepG2) and human breast adenocarcinoma (MCF-7) and all cancer tissues tested (uterine, breast and urinary bladder) and protein expression was confirmed in the OS732 cell line (Zhang et al. 2006) .

Somatic retrotransposition has been identified in cancer cells and linked to the tumourigenic process; however there is growing evidence for somatic retrotransposition occurring in neuronal and replicative senescent cells. L1 retrotransposition has been shown to be possible in non-dividing human somatic cells, neural progenitor cells and neuronal cells *in vitro* and *in vivo* (Muotri et al. 2005; Kubo et al. 2006; Coufal et al. 2009). It was also demonstrated that L1 retrotransposition occurs in the adult human brain (Coufal et al. 2009; Evrony et al. 2012). A high-throughput analysis of somatic retrotransposition identified 7743 L1, 13692 Alu and 1350 SVA putative somatic insertions across the brains of three individuals with these insertions occurring at a higher frequency in protein coding genes expressed in the brain (Baillie et al. 2011). Replicative senescent cells have also shown global epigenetic changes of more open chromatin of retrotransposons, in particular the evolutionary recent retrotransposable elements, and associated increase in the presence of their RNA and retrotransposition (De Cecco et al. 2013). This process of somatic mosaicism through retrotransposition could introduce genetic variability between individual cells. Depending on the site of insertion the affect of this process could be 1) neutral, 2) positive as may provide novel and increased variation in transcriptional control for the cell or could be 3) detrimental through mutation and contribute to the aging process and even neurological disease.

Mutations although considered negative in many cases of disease have been a source of genetic variation throughout the evolution of a species and allow the development of new traits and adaptations to changing environments. TEs provide a mechanism for generating genetic variation within a genome adding to the potential adaptability and evolution of the species they are present in (Kazazian 2004; Cordaux and Batzer 2009; Kim et al. 2012). There are 47 genes identified in the human genome that are derived from the sequence of TEs (Lander et al. 2001). For example the syncytin gene is derived from the *env* gene of a HERV-W and is involved in human placental morphogenesis (Mi et al. 2000).

Chimpanzees are the closest living related species to humans and since the publication of the human and chimpanzee reference genomes comparisons of the TEs content of the two have been studied. Since the divergence of humans and chimpanzees there have been an additional 10719 insertions of TEs; 7786 in humans and 2933 in chimpanzees showing there has been a higher rate of transposition in humans than chimpanzees (Mills et al. 2006). Of these insertions 2642 in humans and 990 in chimpanzees were within genes defined as 3kb upstream and 0.5kb downstream of RefSeq genes. These differences in transposon insertions could have in some part influenced speciation and differences in gene expression between the two species. For SVAs there was an additional 864 insertions in the human genome and 396 in the chimpanzee genome. An analysis of the human and chimpanzee genomes revealed that 46537bp had been deleted from the human genome through the processes of SVA insertion mediated deletions and SVA recombination associated deletions (Lee et al. 2012b). A study looking at regions that were highly conserved between chimpanzees and other mammals but deleted in humans revealed tissue specific regulatory domains had been included in these deletions and were hypothesis to be responsible for the loss or gain of traits contributing to the evolution of humans (McLean et al. 2011).

Transduction events during retrotransposition can result in flanking sequence of a SVA being transcribed and retrotransposed along with the SVA duplicating sections of the genome and integration at a different locus; 10% of SVA insertions have transduced sequence at their 3' end (Wang et al. 2005). 5' transduction occurs when an upstream promoter may be used to transcribe the SVA mRNA; the process that created the subtype F1. 3' transductions occur when the RNA polymerase II bypasses the weak polyadenylation signal of the SVA and uses another polyadenylation signal downstream of the SVA. These processes provide mechanisms for creation of new exons or even duplication of genes. Approximately 53kb of genomic sequence has been duplicated by 143 different SVA mediated transduction events including the duplication of the entire solute carrier family 35, member G5 (SLC35G5/AMAC) gene three times with at least two of the SVA transduced genes expressed in humans (Xing et al. 2006). Gene duplication is an important mechanism in the evolution of a species and the generation of new genes. The second gene is no longer under selective pressures to maintain its current function and can therefore undergo mutation with the potential development of a protein with a new function. The genome can evolve with less risk as the function of genes already present can be maintained.

TEs are a source of regulatory elements providing promoters (sense and antisense), binding sites for transcription factors, donor and acceptor splice sites and polyadenylation signals that could affect gene expression (Rebollo et al. 2012). Retrotransposons play an important role in the transcriptome of mammalian cells with retrotransposons located at the 5' of protein coding regions functioning as alternative promoters and that retrotransposon derived transcriptional start sites (TSS) are generally tissue specific and associate with gene dense regions (Faulkner et al. 2009).

SVAs although they have no characterised promoter, would provide due to their repetitive nature, multiple sites for methylation, transcription factor binding and the formation of secondary DNA structures such as G-quadruplex (G4 see below) that could influence gene transcription. SVAs contain large domains of repetitive DNA (VNTRs) similar in copy number and size of individual repeats, that have been found to direct differential tissue specific and stimulus inducible gene expression in many genes and the copy number of those repeats have been correlated to disease predisposition as discussed previously in section 1.2. Due to the young age of the SVAs they still share many similarities even across subtypes therefore they could respond to similar stimuli throughout the genome to give a concerted response to the environment. SVAs can also cause alternative splicing and exon skipping resulting in differential transcripts of a gene as documented by disease causing insertions (Hancks and Kazazian 2012; Kaer and Speek 2013). The poly A-tail present at the 3' end of a canonical SVA insertion if located on the same strand as a gene could affect the transcriptional machinery causing pausing or termination of transcription. The SINE region of the SVA contains LTR sequences from the HERV-K10 which are known to contain regulatory domains and have been hypothesised to be involved in the expression of the human specific processed pseudogene NANOGP8 and the duplicated AMAC genes (Xing et al. 2006; Fairbanks et al. 2012).

The sequence of SVAs is highly GC rich, approximately 60%, with the central VNTR having a GC content of above 70% (Wang et al. 2005) and contain many potential sites of methylation, CG dinucleotides or CpGs. CpGs can be modified by the addition of a methyl group covalently to one of the carbons within the cytosine nucleotide. The number of CpGs located within the human genome is under represented, one fifth of the expected number, due to the spontaneous deanimation of the methylcytosine to thymine (Lander et al. 2001). Throughout the human genome there are regions of high CpG content relative to the rest of the genome and these are called CpG islands (Gardiner-Garden and Frommer 1987). CpG islands are located generally at the 5' and 3' ends of genes and are associated with promoters in particular with the promoters of genes that are widely expressed (Gardiner-Garden and Frommer 1987; Larsen et al. 1992). CpG islands are involved in gene regulation, genomic imprinting and X-chromosome inactivation with hypermethylation of CpG islands associated with stable repression of transcription (Bird 2002; Reik 2007). TEs, including SVAs, are targeted for methylation to prevent their retrotransposition and potential detrimental effects associated with their insertions. SVAs could therefore potentially act as CpG islands at the site of their insertion influencing the neighbouring genomic locus, which could include the repression of expression of nearby genes. SVAs share characteristics that define CpG islands which have been outlined for a handful of these elements in chapter 5.4.6 of this thesis.

The nature of the sequence contained within SVAs also shows the potential for formation of secondary structures such as cruciforms and G4 DNA (Hancks and Kazazian 2010). Cruciform formation requires perfect or imperfect inverted repeats of 6 or more bases, like those seen in the central VNTR of the SVAs, and are involved in processes such as DNA replication and gene regulation (Brazda et al. 2011).

G4 DNA is a secondary structure formed in guanine-rich sequences and is abundant in promoter regions (Huppert and Balasubramanian 2007; Zhao et al. 2007). G4 structures are hypothesised to interfere with replication of DNA and a host of regulatory functions including gene expression, genome stability and telomerase activity (Fletcher et al. 1998; Huppert and Balasubramanian 2005; De and Michor 2011; Clark et al. 2012). Sequences with potential to form G4 are located in the promoters of several genes such as c-MYC and their ability to decrease transcription has been demonstrated (Siddiqui-Jain et al. 2002; Cogoi and Xodo 2006; Membrino et al. 2011). The proto-oncogene c-MYC has been shown to be regulated by G4 formation in the nuclease hypersensitivity region III_1 (NHE III_1) located -142 to -115bp upstream of its promoter1 which regulates up to 90% of the c-MYC transcription (Gonzalez and Hurley 2010). Figure 1.5 depicts the model proposed for the regulation of c-MYC via transcription factors bound and G4 formation to demonstrate how this type of secondary structure is involved in transcription. The expression of c-MYC is activated when the factors Sp1 or hnRNPK and CNBP are bound to the promoter region. Sp1 binding sites are frequently found in regions where sequences with G4 potential are located due the G-rich nature of Sp1 binding motifs. The other factors, hnRNPK and CNBP, are single stranded DNA binding proteins that will bind preferentially to the pyridimine and purine strands respectively. When these factors are not bound, G4 DNA and i-motifs (a type of secondary structure that forms in c-rich regions) are able to form and the expression of c-MYC is repressed. The nature of the sequence of SVAs provides the potential for the formation of G4 DNA which could be involved in the regulation of nearby genes in a similar process as outlined for the c-MYC gene. The CCCTCT hexamer repeat of the SVA has G4 potential on the opposite strand which is similar to the equence found in the NHE III$_1$ of the c-MYC gene. The potential of SVAs to form G4 DNA is analysed in detail in chapter 5.4.5 and the sequence within SVA that have the potential to form G4 DNA is shown in Figure 4.2B and 6.8.

This text box is where the unabridged thesis included the following third party copyrighted material:

Gonzalez V, Hurley LH. 2010. The c-MYC NHE III(1): function and regulation. *Annual review of pharmacology and toxicology* **50**: 111-129. Figure 3

Figure 1.5: Model of transcriptional regulation at the NHE III$_1$ of the c-MYC gene. (Gonzalez and Hurley 2010)

1.4 General aims

The properties of SVAs, in particular the repetitive nature of their sequence, and their potential influence on the regulation of gene expression is the focus of this thesis. The relatively small number of these elements enables a global analysis of their distribution in relation to genomic features of interest including genes and regulatory domains. The VNTR domains of these elements provide additional genetic variation, and therefore potential functional variation between individuals. The structure and functional properties of specific SVA elements will be analysed in an attempt to understand how these elements may be influencing the human genome and the link they may pose with disease predisposition.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Commonly used solutions

TBE buffer 5X - Tris Base 108g, Boric acid 55g, EDTA 5.84g, distilled water up to 2L.

LB Broth - 25g/L in distilled water (Fluka Analytical).

LB Agar - 40g/L in distilled water (Fluka Analytical).

NZY$^+$ Broth – 10g of NZ amine, 5g yeast extract and 5g NaCL up to 1L of deionised water and pH adjusted to 7.5 using NaOH. The solution was autoclaved and then following filter sterilisation, supplements were added prior to use: 12.5ml of 1M $MgCl_2$, 12.5ml of $MgSO_4$ and 10ml of 2M glucose.

2.1.2 Solutions used in Chromatin Immunoprecipitation

Cell lysis buffer – 50mM Hepes-KOH pH7.5, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40**,** 0.25% Triton X-100.

Nuclear lysis buffer – 10mM Tris-HCl, pH8.0, 200mM NaCl, 1mM EDTA, 0.5mM EGTA.

ChIP dilution/sonication buffer – 16.7mM Tris–HCl, pH 8.1, 167mM NaCl, 1.1% Triton X-100, 0.01% SDS, 1.2mM EDTA.

Low salt wash buffer – 20mM Tris–HCl, pH 8.1, 150mM NaCl, 0.1% SDS, 1% Triton X-100, 2mM EDTA.

High salt wash buffer – 20mM Tris–HCl, pH 8.1, 500mM NaCl, 0.1% SDS, 1% Triton X-100, 2mM EDTA.

LiCl wash buffer - 10mM Tris–HCl, pH 8.1, 250mM LiCl, 1% Igepal, 1% sodium deoxycholate and 1mM EDTA.

TE buffer – 10mM Tris–HCl, pH 8.0, 1 mM EDTA.

Elution buffer - 50mM Tris–HCl pH 8, 1mM EDTA, 1% SDS, 50mM $NaHCO_3$.

2.1.3 Sources of cell lines

SK-N-AS – human neuroblastoma cell line, CRL-2137 from the European Collection of Cell Culture (ECACC).

MCF-7 – human breast adenocarcinoma cell line which was provided to the lab by collaborators: Prof Rudland and Prof Palmeri.

JAr - human placental choriocarcinoma cell line, HTB 144 from the ECACC.

2.1.4 Cell culture media

2.1.4.1 Complete media for SK-N-AS cell line

Dulbecco’s Modified Eagles medium with 4500mg glucose/L (Sigma D5796) supplemented with 1% (v/v) non essential amino acid solution (Sigma), 100 units per ml of penicillin and 0.1mg/ml of streptomycin and 10% (v/v) foetal bovine serum (Sigma).

2.1.4.2 Complete media for MCF-7 cell line

Dulbecco's Modified Eagles medium with 4500mg glucose/L (Sigma D5796) supplemented with 100 units per ml of penicillin and 0.1mg/ml of streptomycin and 10% (v/v) foetal bovine serum (Sigma).

2.1.4.3 Complete stripped media for MCF-7 cell line

Dulbecco's Modified Eagles medium with 4500mg glucose/L and phenol red free (Sigma D1145) supplemented with 100 units per ml of penicillin and 0.1mg/ml of streptomycin, L-Glutamine final concentration 2mM (Sigma) and 5% (v/v) charcoal stripped foetal bovine serum (Sigma).

2.1.4.4 Freezing media

90% foetal bovine serum (Sigma), 10% DMSO (Sigma).

2.2 Methods

2.2.1 Bioinformatic and in silico analysis of SVA distribution and structure

Several genome browsers and software available freely on the internet were used in the analysis of the global distribution of SVAs. These are listed below with the version used:

UCSC genome browser Hg19 (http://genome.ucsc.edu/index.html)

Galaxy (http://galaxyproject.org/)

NCBI HuRefChr37.3 (http://www.ncbi.nlm.nih.gov/)

Quadparser (G-quadruplex prediction software) (Wong et al. 2010)

For detailed methods using these programs see methods section of chapter 4 (sections 4.3.1-7).

2.2.2 Cell culture

2.2.2.1 Culturing of SK-N-AS and MCF-7 cell lines

SK-N-AS cells (human neuroblastoma cell line) were grown in media outlined in 2.1.4.1 and MCF-7 cells (human breast adenocarcinoma cell line) were grown in media outlined in 2.1.4.2 both in T175 flasks and when 70-80% confluent they were passaged into new T175 flasks. To passage cells the media was removed from the flask and the cells were washed with 10ml sterile PBS. 5ml of 1x trypsin (Sigma) was added, washed over the cell and then removed. The flask was placed in the incubator at 37^{o}C for 3 minutes until the cells began to detach from the surface. The cells were washed away from the surface of the flask using 10mls of appropriate media and 1-2mls (approximately 1-2.4 million cells depending on cell type) of which was then placed in a new T175 flask with 40mls of media for that cell line.

The cell lines were tested for mycoplasm every six months using MycoAlert Mycoplasma Detection kit (Lonza) to prevent infection.

2.2.2.2 Cell counts with a haemocytometer

To determine the number of cells per ml of media a cell count was completed using a haemocytometer. A T175 flask of cells at approximately 70% confluency was passaged as in methods 2.2.2.1 up to when the cells were washed down with

10mls of media. Prior to use the haemocytometer and coverslip were washed with ethanol. On the centre of the counting surface of the haemocytometer there are 25 squares (5x5) bounded by three parallel lines each containing 25 smaller squares (5x5). The coverslip was placed onto the counting surface of the haemocytometer and 20µl of the media containing the cells was introduced under the coverslip. The counting surface was visualised under a light microscope on the 10x objective and the number of cells within the 25 larger squares bounded by three parallel lines were counted. Any cells on the top or left hand borders of the 25 squares were included in the count where as cells on the bottom or right hand borders were excluded. This area corresponds to 0.1mm^3 therefore the number of cells was multiplied by 10^4 (10000) to give the number of cells in 1cm^3 which is the equivalent of 1ml. This gave the number of cells per ml of media.

2.2.2.3 Freezing cells for storage in liquid nitrogen

For long term storage of cell lines the cells were frozen in freezing media (2.1.4.4) in liquid nitrogen. The cells were grown in T175 flasks until 70-80% confluent and then passaged as in 2.2.2.1 but the cells were washed from the surface of the flask using 10mls of freezing media (2.1.3.4).The freezing media containing cells were split across cryovials with 1.8mls in each. The cryovials were then placed into a Mr Frosty with isopropanol at -80^{0}C for 24 hours. The cryovials were then transferred to liquid nitrogen.

2.2.3 Analysis of endogenous gene expression

2.2.3.1 Extraction of total RNA from cultured cells

Total RNA was extracted using TRIzol reagent (Invitrogen). SK-N-AS and MCF-7 cells were plated out into 6 well plates (450,000 and 400,000 cells per well respectively) and left for 24hrs. The media from each well was removed and 1ml of TRIzol was added per 10cm^2 and pipetted up and down to lyse the cells. 1ml of the lysed cells was added to a microcentrifuge tube and incubated for 5 minutes at room temperature. 0.2ml of chloroform (per 1ml of TRIzol reagent) was added to each sample, shaken by hand for 15 seconds and then incubated at room temperature for 2-3 minutes. The samples were then centrifuged at 12,000 g for 15 minutes at 4^{o}C. After centrifugation the mixture had separated into three layers: a colourless aqueous upper layer containing the RNA, a middle interphase layer and a lower red organic layer containing the DNA and protein. The upper colourless layer was carefully removed and transferred to a new microcentrifuge tube (approximately 500µl). 0.5ml of 100% molecular grade isopropanol (per 1ml of TRIzol reagent) was added to each sample and incubated for 10 minutes at room temperature and then centrifuged at 12,000 g for 10 minutes at 4^{o}C.

The supernatant was removed leaving behind the pellet of RNA. This was then washed with 1ml of 75% ethanol (per 1ml of TRIzol reagent used in initial step). Once the ethanol had been added to the pellet it was vortexed and centrifuged at 7500 *g* for 5 minutes at 4^{o}C. The supernatant was removed and the pellet was air dried for 5 to 10 minutes. The pellet was resuspended in 20µl of nuclease free water and heated on a heat block at 55^{o}C for 10-15 minutes. RNA was quantified using a Nanodrop 8000. The Nanodrop was set to measuring RNA and calibrated with nuclease free water (the dilutant the RNA was resuspended in) and then 1.5µl of each sample was loaded on the pedestal. The amount of UV light absorbed at 260nm by nucleic acids is dependent on their concentration. The Nanodrop measures the optical density of the RNA and then calculates its concentration (an OD_{260nm} of 1 equals an RNA concentration of 40µg/ml). The Nanodrop was also used to assess the quality of the RNA by looking at the 260/280 and 260/230 ratios where expected values of high quality RNA are ~2.0 and 2.0-2.2 respectively. The RNA was then stored at -80^{o}C.

2.2.3.2 First strand synthesis of cDNA from total RNA

cDNA was synthesised from the total RNA extracted in 2.2.3.1 using the GoScript Reverse Transcription System (Promega) following the recommended protocol from the manufacturer that can convert up to 5µg of RNA in each reaction. The same amount of RNA was used in the reverse transcriptase reaction for each sample. The following components were combined into a PCR tube:

RNA (up to 5µg)	Xµl
Random Primers (0.5µg/reaction)	1µl
Nuclease free water	Yµl
Final volume	5µl

The mixture was denatured at 70^{o}C for 5 minutes to and then cooled on ice. The following reverse transcription mix was added to the RNA, random primers and nuclease free water reaction so total volume was 20 µl:

Component	Volume	Final Concentration
Nuclease free water (to a final volume of 15µl)	Xµl	
GoScript 5X reaction buffer	4µl	1X
$MgCl_2$ (25mM)	4µl	5mM
PCR nucleotide mix (10mM of each dNTP)	1µl	0.5mM
Recombinant RNasin Rinonuclease inhibitor (40U/µl)	0.5µl	1U/µl
GoScript Reverse Transcriptase	1µl	

The combined reaction mixes were incubated at 25^{o}C for 5 minutes for the primers to anneal and then incubated at 42^{o}C for 60 minutes (extension step). The reverse transcriptase is inactivated by heating the reaction to 70^{o}C for 15 minutes. The cDNA was then stored at -20^{o}C.

2.2.3.3 Amplification of cDNA for detection of mRNA expression

2.2.3.3.1 Primer design

Primers were designed using Primer3 (http://frodo.wi.mit.edu/). The primers were designed in separate exons to distinguish product from any amplicons arising from contaminating gDNA acting as template. Firstly the complete gDNA of the gene of interest was taken from UCSC genome browser (Hg19) and the introns deleted. This sequence was then used with the Primer3 software. The potential primers were inserted into the In-silico PCR tool of the UCSC genome browser to ensure they were specific to the area of interest and were no other potential PCR products. Primers were synthesised by Eurofins (MWG primers).

2.2.3.3.2 Polymerase chain reaction

GoTaq Flexi DNA Polymerase (Promega) was used in Polymerase chain reaction (PCR) for the amplification of cDNA targets in analysis of gene expression using the cDNA generated in 2.2.3.2 as template.

GoTaq Flexi DNA Polymerase master mix (per reaction)

Component	Volume	Final Concentration
5X Green GoTaq Flexi buffer	5µl	1X
$MgCl_2$ (25mM)	4µl	4mM
PCR nucleotide mix (10mM of each dNTP)	1µl	0.4mM
Upstream Primer (20µM)	0.5µl	0.4µM
Downstream Primer (20µM)	0.5µl	0.4µM
GoTaq DNA polymerase (5u/µl)	0.25µl	0.05U/µl
Nuclease free water	Xµl	
cDNA template	Yµl	
Final Volume	25µl	

The PCR was performed in a thermocycler: QB-96 (Quanta Biotech) or peqSTAR 2X (peqlab). For the cycle conditions for each primer set used see Table A1 in the appendix. The annealing temperature of each primer set was optimised using a gradient PCR. The amount of cDNA template used again varied between primer sets depending on the abundance of the target for amplification.

2.2.3.4 Agarose gel Electrophoresis

PCR products were analysed using agarose gel electrophoresis, which allows the separation of DNA fragments by mass as they migrate through the agarose gel with the application of an electric field. The appropriate amount of agarose (Bioline) was dissolved in 0.5X TBE with the addition of a nucleic acid stain intercalating of either 1µl of GelRed per 10ml (Biotium 10000X) or 0.5µl per 10ml of ethidium bromide (Sigma 10mg/ml). The percentage of the gel used was determined by the size of the fragments to be run on the gel. The gel was poured into the appropriate sized casting tray and combs inserted to set at room temperature. The gels were placed into horizontal gel tanks with 0.5X TBE and if ethidium bromide was used as nucleic acid stain 5µl of ethidium bromide was added for every 1L of buffer. The samples were loaded into the wells of the gel. If no loading dye was already present in the samples (GoTaq Flexi buffer contains dye) loading dye (Promega 6X) was added at 1µl per 5µl of sample. For sizing of the fragments on the gel a DNA ladder, either 100bp (Promega) or 1kb (Promega) depending on the size of the fragments expected, was loaded in at least one of the wells. The voltage (standard is 5V/cm) and time for which the gel was run was dependant on the percentage of the gel and fragment size. The DNA was then visualised using a UV transillumintor (BioDoc-it Imaging System) and an image was captured.

2.2.4 Methods for cloning

2.2.4.1 Primer design for amplification of targets for cloning

The genomic sequence of the locus to be cloned was obtained from UCSC genome browser (Hg19) with flanking sequence. Primer3 (http://frodo.wi.mit.edu/)

was used to design the primers with the specific sequence required in the PCR product indicated using brackets to ensure the primers would be located in the flank of the target sequence. Primers containing restriction sites for directional ‘sticky end’ cloning were designed with approximately 15bp of sequence specific to the region to be amplified, the restriction enzyme site sequence at the 5’end of the primer and then six random bases 5’ of the restriction enzyme site sequence to increase efficiency of digestion. All primers designed were screened using *in-silico* PCR tool from UCSC genome browser to ensure they produced one specific product. Templates used were gDNA from Jar cell line (source of cell lines 2.1.3 and gDNA extraction method 2.2.7) and gDNA from the CEU HapMap cohort (Utah residents with Northern and Western European ancestry from the CEPH collection) which is commercially available DNA and was provided by our collaborator Gerome Breen.

2.2.4.2 Amplification of fragments for cloning using PCR

The proof reading enzyme KOD Hot Start Polymerase (Novagen) that produces blunt ended products was used to amplify the appropriate fragments for cloning.

KOD Hot Start master mix (per reaction):

Component	Volume	Final Concentration
10X buffer for KOD Hot start polymerase	5µl	1X
$MgSO_4$ (25mM)	3µl	1.5mM
dNTPs (2mM of each)	5µl	0.2mM of each dNTP
Upstream Primer (10µM)	1.5µl	0.3µM
Downstream Primer (10µM)	1.5µl	0.3µM

KOD hot start DNA polymerase (5u/µl)	1µl	0.02U/µl
Nuclease free water	Xµl	
gDNA template	Yµl	
Final Volume	50µl	

The products were analysed by agarose gel electrophoresis (2.2.3.4). For details of each primer set and methods used for the different fragments to be cloned see 2.2.5.1 and 2.2.5.2 and Table A1 in the appendix.

2.2.4.3 Restriction enzyme digests

Restriction enzyme digests were used either to create specific nucleic acid overhangs for ligation or as a diagnosis tool for determining the presence and/or orientation of inserts. For restriction enzymes from Promega the following reaction components was used:

Nuclease free water	Xµl
10X Buffer	2µl
Acetylated BSA (10µg/µl)	0.2µl
DNA (1µg)	Yµl
Restriction Enzyme (10u/ µl)	0.5µl
Final volume	20µl

The buffer recommended for each enzyme’s optimum activity was used. The digestion was incubated for 1-4 hours at the appropriate temperature for the enzyme activity and run on an agarose gel to visualise the fragment size (2.2.3.4).

Restriction enzymes used from New England Biolabs were used with the following components:

Nuclease free water	Xµl
10X Buffer	5µl
BSA (10mg/ml)	0.5µl
DNA (1µg)	Yµl
Restriction Enzyme (10u/µl)	0.5µl
Final volume	50µl

The digestion was then incubated at the appropriate temperature for 1-4 hours and run on an agarose gel to visualise the fragment size (2.2.3.4).

2.2.4.4 Extraction of DNA fragments from agarose gels

DNA fragments run on an agarose gel from either a PCR or restriction enzyme digestion were extracted using the QIAquick Gel Extraction Kit (Qiagen) and the manufacturers guidelines provided with the kit were followed. The gel extraction kit can be used for purification of dsDNA fragments from 70bp-10kb removing primers, nucleotides, enzymes, salts, agarose and other impurities for use in downstream applications such as ligation with the appropriate vector. Briefly the fragment of DNA of interest is cut from the agarose gel, which is completely dissolved in the appropriate buffer, purified using a column system and wash buffers and recovered from the column using an elution buffer. The eluate is then used as required in downstream applications or stored at -20^{o}C.

2.2.4.5 Ligation of DNA fragments into pCR-Blunt intermediate vector

For cloning of PCR fragments into an intermediate vector the Zero Blunt PCR Cloning Kit (Invitrogen) was used. The amount of PCR product required was calculated using the following equation:

Xng insert = (10 x Ybp PCR product) X (25ng linearised pCR-Blunt)/3500bp pCR-Blunt

3 ratios from 10:1 to 100:1 of insert:vector were used in the following ligation reaction:

pCR-Blunt (25ng)	1µl
Blunt PCR product	1 to 5µl
10X Ligation Buffer (with ATP)	1µl
Sterile water	Xµl
T4 DNA Ligase (4 units/µl)	1µl
Final volume	10µl

The ligation reaction was incubated at 16^{o}C for 1 hour and then used in the transformation of chemically competent cells 2.2.4.7.1 or stored at -20^{o}C.

2.2.4.6 Ligation of DNA fragments into reporter gene pGL3 vectors

The cloning of inserts into pGL3 vectors was completed using vector and insert with complementary overhangs generated after restriction enzyme digest. The ligations were generally carried out at a molar ratio of vector to insert of 1:3. The amount of insert required was calculated using the following equation:

Insert(ng) = Vector (ng) X size of insert(kb)/size of vector(kb)

Vector DNA	Xµl
Insert DNA	Yµl
10X Ligation Buffer	1µl
Nuclease free water	Zµl
T4 DNA Ligase	0.1-1units
Final volume	10µl

The ligation reaction was incubated at room temperature for 3 hours and then used in the transformation of chemically competent cells 2.2.4.7.1 or 2.2.4.7.2 or stored at -20^{o}C.

2.2.4.7 Transformation of chemically competent cells

2.2.4.7.1 Transformation of DH5α competent *E.coli*

Chemically competent DH5α *E.coli* (Invitrogen) were defrosted on wet ice and divided into 50µl aliquots in 1.5ml microcentrifuge tubes, any aliquots not required were snap frozen in an ethanol dry ice bath and placed at -80^{o}C for storage. 5 µl of each ligation mixture from 2.2.4.5 or 2.2.4.6 was added to 50µl of competent DH5α cells, tapped gently to mix and then incubated on ice for 30mins. The transformation mixture was then heat shocked at 37^{o}C for 20 seconds and placed on ice for 2 minutes. 950µl of pre-warmed LB broth (2.1.1) was added to the transformation mixture and placed in a shaking incubator at an angle to allow for increased gaseous exchange as an aerobe at 37^{o}C and 225rpm for 1hr. 200µl of the transformation mixture was spread onto LBagar plates that have been warmed in the incubator previously and contained appropriate amount of specific antibiotic (75µg/ml ampicillin or 25µg/ml kanamyacin). The plates were then incubated at 37^{o}C overnight. The remainder of the transformation mixture was stored at 4^{o}C for future use if required.

2.2.4.7.2 Transformation of XL10-Gold Ultracompetant cells

A 14ml round bottom BD Falcon polypropylene tube was chilled on ice per transformation reaction and NZY$^+$ broth (2.1.1) was warmed to 42^{o}C. 100µl aliquots of XL-10 Gold ultracompetant cells were then thawed on ice and once defrosted added to the round bottom falcon tubes. 4µl of β-mercaptoethanol was added to the cells, swirled and then incubated on ice for 10 minutes with the cells being swirled every two minutes. 2µl of ligation mixture was added to the cells and then incubated on ice for 30 minutes. The cells were then heated pulsed for 30 seconds at 42^{o}C in a water bath and incubated on ice for 2 minutes. 900µl of pre-warmed NZY$^+$ was added to the cells and placed in a shaking incubator at 225rpm and 37^{o}C for one hour. 200µl of the transformation mixture was spread onto agar plates containing 75µg/ml of ampicillin that had been pre-warmed. These were incubated overnight at 37^{o}C.

2.2.4.8 Miniprep: preparation and purification of plasmid DNA with low yield

For extraction of low yield plasmid DNA from transformed bacteria QIAprep spin Miniprep Kit (Qiagen) was used and the manufacturer’s guidelines followed. These types of extractions were used to screen colonies grown from the transformation of chemically competent bacteria with ligation reactions for determining if an insert of the correct size/orientation is present. Colonies were picked from the agar plates on which the transformation mixture was spread (2.2.4.7.1 and 2.2.4.7.2) and grown in 5ml of LB broth with the appropriate antibiotic overnight to expand the numbers of bacteria containing the plasmid of interest. This culture media could then be used to create a bacterial pellet to extract the plasmid using the miniprep kit. Briefly the miniprep kit system involves the lysis of the bacteria to release the plasmid and subsequent neutralisation of this reaction. The cellular debris is removed and the plasmid DNA is purified using a column system and then eluted. The plasmid was then used in downstream applications such as restriction enzymes digests or stored at -20^{o}C.

Glycerol stocks of transformed bacteria were made for long term storage. 1.4ml of the overnight culture was transferred to a microcentrifuge tube and pelleted by centrifugation at 8000rpm for 3 minutes at room temperature. The supernatant was removed and the pellet resuspended in 0.5ml of sterile 15% glycerol (v/v in LB broth) and transferred to a cryovial. This was then immediately frozen at -80^{o}C.

2.2.4.9 Maxiprep: preparation and purification of plasmid DNA with high yield

A Plasmid Maxi Kit (Qiagen) was used to purify high yields of plasmid DNA from transformed bacteria following the manufacturer's guidelines. This type of plasmid extraction was used to produce a high yield of plasmid DNA with a greater purity than that of a miniprep for use in downstream applications such as *in vitro* reporter gene assays. A small scraping of the glycerol stock was grown in 3ml of LB broth with the appropriate antibiotic for 6-8hrs of which 200µl was then cultured in

100ml of LB broth with the appropriate antibiotic to generate a sufficient quantity of bacteria for extraction of the plasmid. Briefly the bacteria was pelleted and resuspended for lysis. Once the cellular lysis had occurred to release the plasmid the reaction was neutralised and the cellular debris was removed. The plasmid was purified through column by gravity flow, washed and eluted. Isopropanol was used to precipitate the DNA from the eluate. Once the DNA was pelleted it was washed with 70% ethanol to remove excess salt. The ethanol is removed and the pellet is air dried before being resuspended in 300-500µl of EB buffer. The plasmid DNA was quantified using a Nanodrop 8000 and then stored at -20^{o}C.

2.2.4.10 Sequencing

Samples such as plasmids with inserts cloned in and PCR products were sent for sequencing to either Dundee DNA Sequencing and Service or Source Bioscience Life Sciences. The samples and primers were supplied as required by the companies.

2.2.5 Generation of reporter gene constructs

2.2.5.1 Cloning of PCR products of different fragments and alleles of the PARK7 SVA into an intermediate vector

The fragments of interest of the PARK7 SVA and its alleles were amplified using KOD hot start polymerase (2.2.4.2) and ligated into the multiple cloning site of pCR-Blunt intermediate vector (2.2.4.5).

The ligation reactions were transformed into chemically competent DH5α *E.coli* (2.2.4.7.1). During the transformation process of the plasmid containing the whole SVA it is postulated that a recombination event occurred resulting in a truncated insert comprised of the 5’ end of the SVA and a small portion of the SINE region. This was used as a template in order to amplify a fragment of the SVA containing the hexamer repeat, alu-like sequence and the first 10 repeats of the tandem repeat that otherwise would have not been possible due to the repetitive nature of the sequence. The plasmids that appeared to contain the correct insert were sequenced (2.2.4.10). Methods used in this process are described in detail in section 2.2.4 and table 2.1 summaries the information for each construct generated.

Name	Primers for amplification of insert	Template for PCR	Restriction enzyme used to determine orientation of insert	Orientation of insert
Whole SVA I	For 5’ggctttttgataacccctga 3’ Rev 5’tttcggatcacaggcatgagc 3’	gDNA JAr	Bgl1	Rev
SVAΔSINE I	For 5’ggctttttgataacccctga 3’ Rev 5’ ccgcctttctattccacaaa 3’	gDNA JAr	Bgl1	For
TR/VNTR I	For 5’ctcagtgctcaatggtgcc 3’ Rev 5’ ccgcctttctattccacaaa 3’	Whole SVA amplicon	PFIfI	For
Truncated SVA I	For 5’ggctttttgataacccctga 3’ Rev 5’gacggggcggttgcc 3’	Truncation of whole SVA plasmid in pCR blunt	Bgl1	For
Allele 1 I	For 5’ggctttttgataacccctga 3’ Rev 5’gcaaggcttagcttggacag 3’	gDNA of an individual from the CEU HapMap	Bgl1	For
Allele 2 I	For 5’ggctttttgataacccctga 3’ Rev 5’gcaaggcttagcttggacag 3’	gDNA of an individual from the CEU HapMap	Bgl1	For
Allele 3 I	For 5’ggctttttgataacccctga 3’ Rev 5’gcaaggcttagcttggacag 3’	gDNA of an individual from the CEU HapMap	Bgl1	For
Allele 4 I	For 5’ggctttttgataacccctga 3’ Rev 5’gcaaggcttagcttggacag 3’	gDNA of an individual from the CEU HapMap	Bgl1	For

Table 2.1: Plasmids generated using fragments cloned into the intermediate vector pCR-Blunt. This table contains information on the generation of intermediate vectors containing various fragments of the PARK7 SVA and its alleles for use in downstream cloning or required for sequencing. For fragment sizes of PCR products see Table A1 of the Appendix.

2.2.5.2 Cloning of PCR products of different fragments and alleles of the PARK7 SVA into reporter gene vectors

The pCR-Blunt intermediate vectors containing the correct fragments (determined by sequencing) of the PARK7 SVA: whole SVA I, SVAΔSINE I, TR/VNTR I and truncated SVA I were used to clone the fragments into pGL3P in the forward and reverse orientation. The whole SVA I was also used to clone the SVA into the pGL3B vector in the forward and reverse orientation. The intermediate vectors containing the different alleles of the PARK7 SVA were used as template to amplify a shorter fragment removing the 3' flanking sequence. A larger fragment of the alleles of the PARK7 SVA was required to amplify all four alleles from gDNA therefore when being cloned into a reporter gene vector the additional 3' flanking region needed to be removed. These fragments of the different alleles with minimal flanking sequence were cloned into the pGL3P vector using restriction enzymes sites introduced by the PCR primers. All the ligation reactions except that of allele 3 were transformed into DH5α *E.coli* (2.2.4.6.1), however due to problems encountered, the ligation of allele 3 into pGL3P was transformed into XL-10 ultracompetant cells (2.2.4.7.2). The plasmids that appeared to contain the correct insert were sequenced (2.2.4.10). All the methods used in the cloning process are described in section 2.2.4 and table 2.2 summaries the information for each construct generated.

For sufficient quantity and purity for downstream applications the reporter gene constructs were extracted from the transformed DH5α *E.coli* or XL-Gold ultracompetant cells using a maxiprep kit (2.2.4.9).

Name	Vector	Orientation	RE sites of insertion	RE used to digest intermediate plasmid	Template for digest or amplification	Primers for amplification	Digest to detect presence of insert
Whole SVA	pGL3P	For	NheI BglII	BamHI XbaI	Whole SVA I	-	Acc65I NcoI
Whole SVA	pGL3P	Rev	Acc65I XhoI	Acc65I XhoI	Whole SVA I	-	Acc65I XhoI
SVAΔSINE	pGL3P	For	Acc65I XhoI	Acc65I XhoI	SVAΔSINE I	-	Acc65I XhoI
SVAΔSINE	pGL3P	Rev	NheI BglII	BamHI XbaI	SVAΔSINE I	-	Acc65I NcoI
TR/VNTR	pGL3P	For	Acc65I XhoI	Acc65I XhoI	TR/VNTR I	-	Acc65I XhoI
TR/VNTR	pGL3P	Rev	NheI BglII	BamHI XbaI	TR/VNTR I	-	Acc65I NcoI
Truncated SVA	pGL3P	For	Acc65I XhoI	Acc65I XhoI	Truncated SVA I	-	Acc65I XhoI
Truncated SVA	pGL3P	Rev	NheI BglII	BamHI XbaI	Truncated SVA I	-	Acc65I NcoI
Whole SVA	pGL3B	For	NheI BglII	BamHI XbaI	Whole SVA I	-	Acc65I NcoI
Whole SVA	pGL3B	Rev	Acc65I XhoI	Acc65I XhoI	Whole SVA I	-	Acc65I XhoI
Allele 1	pGL3P	For	Acc65I XhoI	-	Allele 1 I	For 5’tgtaggtaccggctttttgataaccc3’ Rev 5’gtaactcgagtttcggatcacaggc3’	Acc65I XhoI
Allele 2	pGL3P	For	Acc65I XhoI	-	Allele 2 I	For 5’tgtaggtaccggctttttgataaccc3’ Rev 5’gtaactcgagtttcggatcacaggc3’	Acc65I XhoI
Allele 3	pGL3P	For	Acc65I XhoI	-	Allele 3 I	For 5’tgtaggtaccggctttttgataaccc3’ Rev 5’gtaactcgagtttcggatcacaggc3’	Acc65I XhoI
Allele 4	pGL3P	For	Acc65I XhoI	-	Allele 4 I	For 5’tgtaggtaccggctttttgataaccc3’ Rev 5’gtaactcgagtttcggatcacaggc3’	Acc65I XhoI

Table 2.2: The reporter gene constructs generated for use in *in vitro* luciferase assays. This table shows information regarding the backbones used and the origin of the inserts for generating reporter gene constructs containing different sized fragments of the PARK7 SVA and its different alleles.

2.2.6 Analysis of reporter gene expression

2.2.6.1 Transient transfection of reporter gene constructs into SK-N-AS and MCF-7 cell lines

Turbofect (Thermo Scientific) was used to transfect SK-N-AS and MCF-7 cell lines following the manufacturer’s guidelines. The cells were counted (2.2.2.2) and were then plated into 24 well plates at the following concentrations 24 hours prior to transfection: SK-N-AS 120,000 cells per well and MCF-7 100,000 cells per well with 1ml of media. For each transfection/well 1µg of the test reporter gene plasmid, 10ng of the internal control and 2µl of Turbofect were combined in a total volume of 100µl of serum free media, vortexed and then incubated for 20 minutes at room temperature. 100µl of the transfection mixture (1/10 of the volume of media) was added per well and after 4 hours the media was changed to remove the transfection mixture reducing cell death.

2.2.6.2 Cell lysis

48 hours after the cells had been transfected with the reporter gene constructs the cells were lysed in preparation for the Dual Luciferase Reporter Assay (Promega). Prior to the addition of the passive lysis buffer (PLB) the media was removed and the cells washed with PBS. 100µl of 1X PLB was added to each well and the 24 well plates were placed on a rocking platform for 15 minutes. 20µl of the cell lysate was transferred to an opaque 96 well plate in preparation for the luciferase assay.

2.2.6.3 Measuring levels of reporter gene using Dual Luciferase Assay

The appropriate amount of luciferase assay reagent II (LARII) and Stop and Glo reagent was prepared for the number of measurements required and allowed to reach room temperature. The opaque 96 well plate containing the cell lysate was placed into a Glomax 96 Microplate Luminometer (Promega). The luminometer has two injectors and therefore can automatically dispense both the LARII and the Stop and Glo one after the other at programmed intervals. The injectors were flushed with distilled water, 70% ethanol, distilled water and then air to thoroughly clean them. The injectors are then primed with the reagents (LARII in injector 1 and Stop and Glo in injector 2) before the promega dual luciferase program is run, which measures the bioluminescence from the reaction catalysed by the firefly and renilla luciferase enzymes. The LARII is added first to measure the bioluminescence produced by the reaction catalysed by the firefly luciferase protein and then the Stop and Glo quenches this reaction and is used to measure the bioluminescence from the reaction catalysed by the renilla luciferase protein.

Using the measurements from the activity of the two reporter gene constructs that were co-transfected the activity of the constructs across the different wells can be accurately compared as the internal control reduces experimental variability caused by differences in transfection efficiencies.

2.2.7 Extraction of gDNA from cell lines

Genomic DNA was extracted from the cell lines available in the lab using the QIAamp DNA mini kit (Qiagen) following the manufacturer’s guidelines. The cells were trypsinised from a confluent T175 flask and counted. The recommended number of cells ($5x10^6$) were transferred to a 1.5ml microcentrifuge tube and centrifuged for 5 minutes at 300g. The supernatant was removed and the cell pellet resuspended in 200µl of PBS. For the extraction from the MCF-7 cell line a range of the number of cells was used due to their reported unusual karyotype. The samples were lysed using enzymes and then passed through a column where the DNA binds to the membrane. The impurities were removed with wash steps and then the gDNA is eluted. The quantity and quality of the gDNA was analysed using a Nanodrop 8000 and then stored at -20^{o}C.

2.2.8 Genotyping of SVAs

Several SVAs were amplified throughout this project and the methods are outlined in the methods sections of the relevant results chapters (3.3.2, 5.3.5, 6.3.5, 6.3.8-10).

2.2.9 Chromatin Immunoprecipitation (ChIP)

2.2.9.1 Harvesting the cell pellet

SK-N-AS and MCF-7 cells were grown in T175 flasks under basal conditions and when they reached 70-80% confluency were harvested for ChIP. The number of cells required to provide enough chromatin for ChIP is specific to each cell line. There are approximately 12 million SK-N-AS cells in a T175 flask which provides sufficient chromatin for ~10 immunoprecipitations using 10µg and a T175 flask of

MCF-7 cells contains approximately 15 million cells and would provide sufficient chromatin for ~15 immunoprecipitations using 10µg.

To crosslink the proteins to the DNA the cells were fixed with 1% formaldehyde (v/v) and incubated at room temperature for 8-10 minutes occasionally swirling the mixture. The reaction was quenched by added glycine to a final concentration of 0.125M and incubated for 5 minutes at room temperature. Each flask of cells was then washed twice with 10mls of ice cold PBS supplemented with 10µl of 200X protease inhibitor cocktail (PIC) (Calbiochem) and 100µl of 0.1M of phenylmethanesulfonyl fluoride (PMSF) (Sigma). 5ml of PBS supplemented with 10µl of 0.1M PMSF was added to each flask and the cells were scraped from the surface of the flask using a cell scraper and collected into a 15ml falcon tube. The cells were pelleted by centrifugation at 1500rpm for 10 minutes at 4^{o}C and the supernatant was removed using an aspirator. The cell pellet can be frozen at this stage of the protocol at -80^{o}C.

2.2.9.2 Cell and nuclear lysis

The pellet was taken directly from 2.2.9.1 and resuspended in 5ml of cell lysis buffer with the addition of 5µl of 200X PIC and incubated for 10 minutes at 4^{o}C on a rotating wheel. The mixture was centrifuged at 3500rpm for 5 minutes at 4^{o}C and then the supernatant was removed. The pellet was resuspended in 5ml of nuclear lysis buffer with the addition of 5µl of 200X PIC and incubated for 10 minutes at 4^{o}C on a rotating wheel. The mixture was centrifuged at 3500rpm for 5 minutes at 4^{o}C and then the supernatant was removed and the pellet resuspended in 1.5ml of sonication buffer.

2.2.9.3 Sonication of chromatin

It is recommended to sonicate fresh chromatin that has not been stored at -80°C as this can affect the shearing process as frozen cell pellets or frozen chromatin are more stable. The Biorupter (Diagenode) was used to sonicate the chromatin from step 2.2.9.2. The chromatin extracted fresh from the SK-N-AS cells was sheared using 30 rounds of 30s on and 30s off on the high setting. The chromatin extracted fresh from the MCF-7 cells was sheared initially using 40 rounds of 30s on and 30s off on the high setting. The sheared chromatin from the MCF-7 cells when analysed (2.2.9.4) had not been sheared enough. Therefore the chromatin from the MCF-7 cell line, that had then been frozen, was sonicated again for another 40 rounds of 30s on and 30s off on the high setting.

2.2.9.4 Estimating the quantity and fragment size of sheared DNA

50µl of sheared chromatin was removed and used to estimate the quantity and fragment size of the DNA. 50µl of nuclease free water was added to the sheared chromatin along with 6µl of 5M NaCl and 2µl of RNase one (Promega). The mixture was vortexed and then incubated at 37°C for 30 minutes. 2µl of proteinase K at 20mg/ml (Sigma) was added to the RNase digested mixture, vortexed and then incubated at 65°C for 2 hours to reverse the cross links. The DNA was then purified using the MiniElute reaction cleanup kit (Qiagen) following the manufacturer's guidelines. The DNA is eluted from the purifying columns in a volume of 10µl and then quantified using a Nanodrop 8000. The remaining DNA was run on a 1% agarose gel (2.2.3.4) with both a 100bp and 1kb ladder to determine the size of the fragments of sheared DNA.

2.2.9.5 Immunoprecipitation of cross-linked protein–DNA interactions

Antibodies for specific histone marks and transcription factors of interest were used in the immmunoprecipitation of the protein-DNA complexes to determine the factors bound to the DNA when the cells were harvested. 100µl of the sheared chromatin was removed for use as the positive control in the PCR analysis. 10µg of sheared chromatin was transferred to a new microcentrifuge tube for each immunoprecipitation to be carried out and made up to 250µl using dilution buffer. The primary antibody was added to the sheared chromatin and incubated on a rotating wheel at 4^{o}C overnight. One tube of sheared chromatin had no primary antibody added to act as a control for non-specific of binding of DNA to the dynabeads. The details of the antibodies and the amount used in displayed in Table A2 in the appendix.

50µl of magnetic dynabeads (Invitrogen 30mg/ml) were transferred to a 1.5ml microcentrifuge tube per immunoprecipitation and washed with 1ml of dilution buffer twice. The second wash was for 2 hours on a rotating wheel at 4^{o}C. The beads were then captured using a magnetic rack and the supernatant removed. The beads were resuspended using 50µl of the antibody chromatin mix and then added to the rest of the chromatin antibody mix and incubated on a rotating wheel at 4^{o}C for 1 hour. The Dynabead and protein-DNA complexes were captured using a magnetic rack and the supernatant removed. The Dynabeads with the bound DNA were washed to remove non-specific DNA and protein bound to the beads. The washes were performed on a rotating wheel at 4^{o}C for 3-5 minutes each with 1ml of the following buffers: low salt wash buffer, high salt wash buffer, LiCl wash buffer and TE buffer.

2.2.9.6 Elution of DNA and reversal of cross links

To elute the immune complex from the beads the bead-immune complex was captured using the magnetic rack and 100ul of elution buffer with a final concentration of 50µg/ml of proteinase K was added. This was mixed at 62°C for 2 hours to release the protein bound DNA and reverse the cross links. The Dynabeads were then captured and the supernatant removed and transferred to a new tube and incubated at 95°C for 10 minutes to denature the protein and inactivate the proteinase K. The DNA was then cleaned up using the MiniElute enzyme reaction cleanup kit (2.2.9.4), quantified using a Nanodrop 8000 and stored at -20°C.

2.2.9.7 Analysis of precipitated DNA using polymerase chain reaction

The regions of interest, the PARK7 gene promoters and 5' of the SVA, were amplified using GoTaq flexi polymerase (2.2.3.4.2) and the conditions for each primer set are listed in Table A1 of the appendix. 5ng of precipitated DNA was used as template for the PCR. A primer set for a gene desert as a negative control for the ChIP assay and the sequence and conditions of the primer set are shown in Table A1 of the appendix. The PCR products were run on 1.2% agarose gels for analysis (2.2.3.4).

Chapter 3

A SVA retrotransposon upstream of the FUS gene can function as a regulatory domain and its implications in ALS

3.1 Introduction

VNTRs have been shown by the lab and others to be important in regulating gene expression and have been associated with disease (section 1.2). One of the projects in the lab at this time was analysing the potential regulatory domains of the fused in sarcoma (FUS) gene important for its association as a candidate gene for amyotrophic lateral sclerosis (ALS). Therefore analysis of the FUS gene locus was undertaken to determine areas with potential regulatory function and genetic variation.

FUS is found on chromosome 16p11.2 and is a RNA/DNA binding protein involved in the regulation of RNA processing (Verma 2011). Mutations within the exons of the FUS gene identified it as a causative gene in some cases of ALS; a fatal disease caused by the degeneration of motor neurons in the brain and spinal cord with death occurring 3-5 years after the onset of symptoms. The incidence of ALS each year in Europe is 2.7 per 100,000 people over the age of 18 occurring more frequently in males than females at a ratio of 1.3:1 (Logroscino et al. 2010) with a lifetime prevelance of 1 in 400 (Al-Chalabi et al. 2010). Familial ALS (FALS) accounts for approximately 5% of all cases of ALS with the rest showing no family history and are considered sporadic (SALS) (Byrne et al. 2011). The heritability of ALS was estimated at 40-45% when the concordance within ALS patients and their parents was analysed inlcuding familial and sporadic forms of the disease and a twin study estimated the heritability of SALS at 61% (Al-Chalabi et al. 2010; Wingo et al. 2011).There have been many studies into the causes of ALS with the identification of 50-60% of genetic mutations in known causative genes of FALS cases and 11% of SALS (Lattante et al. 2012), however there is much still unknown about this disease. The number of ALS cases attributed to mutations in the FUS gene is small;

FUS mutations are present but rare in SALS at around 1% (Corrado et al. 2010; Chio et al. 2011; Lai et al. 2011; Sproviero et al. 2012) and found in only 3-5% of FALS (Kwiatkowski et al. 2009; Vance et al. 2009). Cytoplasmic FUS positive inclusions have been identified in the neurons of individuals with FALS caused by FUS mutations (Kwiatkowski et al. 2009; Vance et al. 2009) but have also been found in the spinal anterior horn neurons of SALS patients without FUS mutations, and in non-superoxide dismutase 1 (SOD1) familial ALS (Deng et al. 2010).

The identification of causative genes for a specific disease, in this case FUS and ALS, can provide insight into pathways involved in the development and progression of the disease and potential regions for genetic associations of common variants within a population. FUS is ubiquitously expressed but the regulation of the gene, in particular in response to challenge, may be an important factor in the initiation or progression of disease processes. Although it is difficult to accurately predict the regulatory domains for a particular gene it has been demonstrated that important domains for gene regulation can occur in both ECRs and VNTRs as discussed in chapter 1 (sections 1.1 and 1.2). The searches for potential areas involved in transcriptional regulation can be aided by the utilisation of the ENCODE data searching for the presence of potential transcription factor binding sites, active histones or DNase 1 hypersensitivity clusters (Doolittle 2013; Kavanagh et al. 2013a) and genome browsers such as UCSC (http://genome.ucsc.edu/index.html). The locus of the FUS gene was analysed using the UCSC genome browser for repetitive regions that may act as regulatory domains, respond to environmental cues and be polymorphic as seen for several of these types of elements throughout the genome. Within 10kb of the TSS of the FUS gene a large repetitive region was identified, which was part of a composite retrotransposable element called a SVA

(SVAs are discussed in detail in sections 1.3.1.2.4 and 1.3.2), Figure 3.1. The potential regulatory properties and association with ALS of this primate specific retrotransposon was addressed.

3.2 Aims

- Identify potential novel regulatory and polymorphic domains of the FUS gene.
- Further analysis of the genetic variation of this SVA element identified in commercial gDNA (personal communication from Thomas Wilm) in a CEU HapMap cohort.
- Assess the genetic variants of the FUS SVA in SALS and matched control cohort to analyse any potential associations of the SVA alleles with disease.
- Determine if there is a tagging SNP for the FUS SVA variants to assist in the analysis of much larger disease cohorts.

3.3 Methods

3.3.1 Bioinformatic analysis of the FUS gene locus

The FUS gene locus was analysed using the UCSC genome browser (http://genome.ucsc.edu/index.html) to determine repetitive sequences in the region that could potentially act as regulatory domains.

3.3.2 Genotyping the TR/VNTR of the FUS SVA

3.3.2.1 Genotyping the TR/VNTR of the FUS SVA in a CEU HapMap cohort

The TR/VNTR of the SVA upstream of the FUS gene was amplified in the CEU HapMap cohort (Utah residents with Northern and Western European ancestry from the CEPH collection) using 1ng of genomic DNA as template (for primers and cycling conditions see Table A1 in Appendix). Go Taq Flexi Polymerase (Promega) was used under standard conditions with the addition of betaine (Sigma) at a final concentration of 1M. For details of the composition of the master mix for Go Taq Flexi Polymerase see 2.2.3.3.2. The products were run on a 1.2% agarose gel (2.2.3.4). To determine if the FUS SVA locus is in Hardy-Weinberg equilibrium (HWE) in the CEU HapMap population the allele frequencies (Figure 3.3A) were used in conjuction with the following equation $p^2+2pq+q^2=1$ to calculate the expected genotypes for this particular locus. The expected values were compared to the observed values using the Chi squared test to determine if the two sets of values were significantly different to each other. If the expected and the observed values are not significantly different to each other ($p>0.05$) the locus can be said to be in HWE.

3.3.2.2 Genotyping the TR/VNTR of the FUS SVA in a SALS and matched controls cohort

The PCR for genotyping the TR/VNTR of the FUS SVA in the SALS and control samples was completed at the laboratory in the MRC Social Genetic and Developmental Psychiatry Research Centre at the Institute of Psychiatry due to the location of the samples, therefore the protocol differed to previous method in 3.3.2.1. The same primers as in 3.3.2.1 were used but the SALS and matched control samples were amplified using Taq Polymerase with the FailSafe 2XD buffer (Cambio) following their recommended protocol with 5ng of gDNA as template. The PCR products were run on a 1.2% agarose gel (2.2.3.4). The SALS and controls matched for ethnicity and age were obtained from King's College London MND DNA Bank All participants gave ethically approved written consent to participate in the study, which was approved by the South London and Maudsley Ethics Committee (reference 222/02). HWE analaysis was carried out for the SALS and control cohorts, see 3.3.2.1 for further details.

3.3.3 Statistical analysis of the genotype data from the SALS and matched controls

The statistical significance of the results from this study were analysed using CLUMP (Sham and Curtis 1995). The CLUMP program was developed as a tool to analyse case-control studies of genetic variants to determine if an allele occurs more frequently in one group or the other.

3.3.4 Identification of a tagging SNP for the FUS SVA alleles

The genotyping results of the individuals of the CEU HapMap cohort could be used in conjunction with publicly available SNP data to identify a tagging SNP for the two variants of the FUS SVA. The genotype for the FUS SVA was determined for the individuals within the CEU HapMap cohort as in method section 3.3.2 and the alleles were named long and short. The long and short genotype of the individuals had to be converted to 'SNPs' so they could be uploaded into the Haploview software (downloaded from the Broad Institute web page http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/downloads) for analysis. Two 'SNPs' were generated one for each allele. FUS SVA SNP 1 corresponded to the long allele and FUS SVA SNP 2 corresponded to the short allele. These 'SNPs' were given genotypes based on the genotype of the individual, see below:

FUS SVA SNP 1	FUS SVA SNP 2
AA=if no long allele present	AA=if no short allele present
AC=if one long allele present	AC=if one short allele present
CC=if two long alleles present	CC=if two short alleles present

This system would therefore distinguish a single allele of the SVA present in a halplotype block using the genotype of the two 'SNPs' that were created; AC=short allele and CA-=long allele.

Genotype data of the known SNPs 200kb either side of the FUS SVA for the CEU HapMap individuals was downloaded from the International HapMap database (http://hapmap.ncbi.nlm.nih.gov/) using release 28. This was copied into an excel spreadsheet and the ‘SNPs’ generated for the FUS SVA genotypes were inserted at the loci corresponding to the actual FUS SVA. The SNP genotypes were then converted from ACGT to numbers as follows 1=A, 2=C, 3=G and 4=T. This data was then used to generate two files: a ped file containing the ID of the individuals and the SNP genotypes and an info file containing the names/rs numbers for the SNPs and their chromosomal loci. These two files were uploaded to the software Haploview where the ped file was used as the data file and the info file as the locus information file. Once uploaded the ‘Run Tagger’ option was selected with the r^2 threshold set to >0.8 for software to identify SNPs in linkage disequilibrium with each other. This generated a list of SNPs that are tagging other SNPs and haplotype blocks to demonstrate how the region of the genome is inherited. The lists of SNPs in linkage disequilibrium with each other were then searched for the SNPs corresponding to the SVA alleles to determine if the SVA is being tagged and inherited in a set haplotype.

3.4 Results

3.4.1 A SVA D is located upstream of the FUS gene

The UCSC genome browser was used to analsye the FUS gene locus. The UCSC genome browser provides a vast array of data and is an extremely useful tool when analysing the genome for potential functional elements. It includes information such as the structure of genes, the location of CpG islands, repetitive regions of DNA, areas of active chromatin and the binding of transcriptions factors to name but a small number of the types of data accessible through this browser.

Analysis of the loci of the FUS gene identified a large GC–rich repetitive region (boxed in red in Figure 3.1A) within 10kb upstream of the FUS gene TSS. The repeat was flanked by DNase 1 hypersensitivity clusters and transcription factor binding sites according to the ENCODE data indicating this region may be active. Similar repeats were searched for using the Blat tool on the UCSC genome and the results contained many homologous regions.

The repetitive region was actually part of a larger composite element called a SVA (boxed in blue in Figure 3.1B). This SVA is a member of the subtype D and is found in humans and chimpanzees only. The structure of this SVA is shown in Figure 3.2A; it differs to the canonical SVA structure as it does not contain a CCCTCT hexamer repeat at the 5' end and a poly A-tail after the SINE. The central repetitive region consisits of a tandem repeat (TR) and a VNTR with two alleles (named long and short) which were identified and discussed in sections 3.4.2 and 3.4.4. The sequence of the TR/VNTR of the SVA is shown in Figure 3.2B. The extra repeat present in the long allele of the SVA is underlined. Many SVAs contain the potential for G-quadruplex (G4) which is a type of secondary structure and the sequence responsible for this potential is in italics. The function of G4 DNA and how this is predicted is dicussed in chapter 1 section 1.3.2 and chapter 4 section 4.4.5 in detail. The SVA present in the chimpanzee genome in the UCSC genome browser (Chimp Feb. 2011) has a much larger central VNTR (873bp) compared to that of the human (447bp) and the sequence of the chimpanzee VNTR is shown in Figure 3.2C. The polymorphic nature of the chimpanzee VNTR has not been analysed as chimpanzee gDNA was unavailable therefore it is not known whether this SVA is variable in the chimpanzee as well as the human.

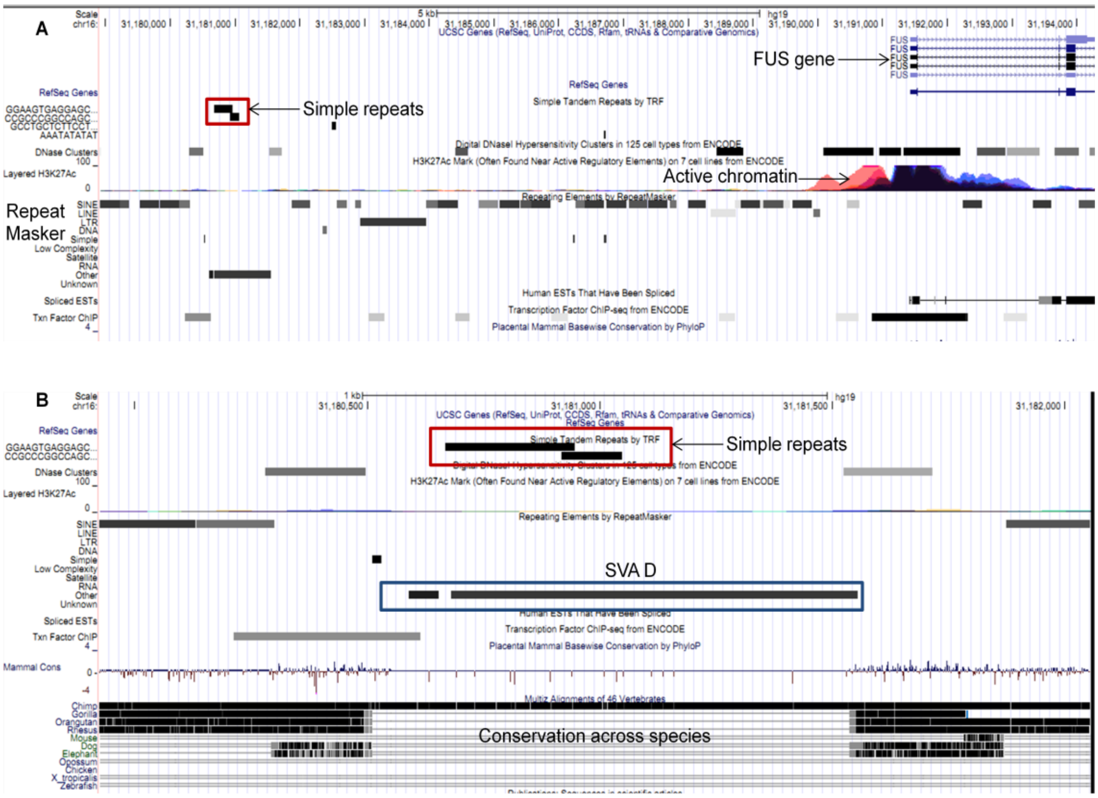
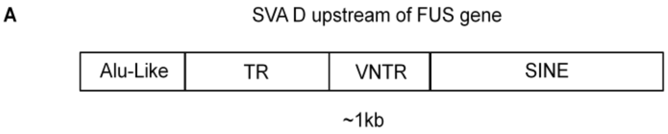


Figure 3.1: Image from UCSC genome browser of the region upstream of the FUS gene. In the UCSC genome browser the names of the tracks on display are found on the left hand side of the image with the scale and chromosomal loci located at the top. The tracks to be displayed in the browser are chosen by the user, in this case from the top of both images A and B the tracks are: UCSC genes, simple repeats, DNase clusters, H3K27Ac histone mark (often found near regulatory elements), repeat masker, ESTs (expressed sequence tags), transcription factor binding and the bottom track in image B shows the mammalian conservation and homology shared with specific species across this region.

A – This screen shot shows the different transcripts of the FUS gene at the top right hand side with DNase hypersensitivity clusters and peaks of active chromatin at the TSS of the gene. The different types of repetitive elements such as LINEs, SINEs and LTRs found in the region are shown by the repeat masker track. There is a large region of repetitive DNA boxed in red within 10kb of the start of the FUS gene identified by the simple repeat track.

B – This is a zoomed in screen shot of the large repetitive region upstream of the FUS gene. The region of repetitive DNA identified in 3.1A is part of a larger composite element called a SVA boxed in blue. This particular SVA is only found in humans and chimpanzees shown by the conservation track at the bottom of the image and has DNase hypersensitivity clusters either side.



B Sequence of human TR/VNTR of the FUS SVA D

TAGGAAGTGAGGAGCGCCTCTTCCCCGCCGCCATCCCATC
TAGGAAGTGAGGAGCGTCTCTGCCTGGCCGCCCATCGTC
TGAGATGTGGGGAGCGCCTCTGCCCCGCCGCCCCGTC
TGGGAGGTGAGGAGCGTCTCTGCCCGGCCGCCCCTTC
TGAGAAGTGAGGAGACCCTCCGCCAGGCAACCGCCCCGTC
TGAGAAGTGAGGAGCCCCTCCGCCCGGCTGCCACCCCGTC
TGGGAAGTGAGGAGCGTCTCCGCCCGGCAGCCACCCCGTC

CGGGAGGGAGGTGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
C***GGGAGGGAGGTGGGGGGGG***TCAGTCCCCCGCCCGGCCAGCCGCCCCGTC
CGGGAGGTGAGGGGCACCTCTGCCCGGCCGCCCCTAC
TGGGAAGTGAGGAGCCCCTCTGCCCGGCCACCACCCCGTC

C Sequence of Chimpanzee VNTR of the FUS SVA D

AGTGAGGAGCGCCTCTTCCCGGCCGCCATCCCATC
TAGGAAGTGAGGAGCGTCTCTGCCCGGCCGCCCATCGTC
TGAGATGTGGGGAGCGCCTCTGCCCCGCCGCCCCGTC
TGGGATGTGAGGAGCGCCTCGGCCTGGCCGCGACCCCGTC
TGGGAGGTGAGGAGCGTCTCTGCCCAGCCGCCCCGTC
TGAGAAGTGAGGAGACCCTCCGCCAGGCAACCGCCCCGTC
TGAGAAGTGAGGAGCCCCTCCGCCCGGCAGCCGCCCCGTC
TGAGAAGTGAGGAGCCCCTCCGCCCGGCTGCCACCTCGTC
TGGGAAGTGAGGAGCGTCTCCGCCCGGCAGCCACCCCGTC

CAGGAGGGAGGTGGGGGTCAGCCCCTGCCAGGCCAGCCGCCCCGTC
CGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCTGTC
CGGGAGGTGAGGGGTGCCTCTGCCCGGCCGCCCCTAC
TGGGAAGTGAGGAGCCCCTCTGCCCGGCCAGCCGCCCCATC
CGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCTGTC
CGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
CGGGAGGTGAGGGGTGCCTCTGCCCGGCCGCCCCTAC
TGGGAAGTGAGGAGCCCCTCTGCCCGGCCAGCCGCCCCATC
CGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
CGGGAGGGAGGTGGGGGGGTCAGCCCCCCCGCCCGGCCAGCCGCCCCGTC
TGGGAGGTGAGGGGCGCCTCTGCCCGGCCGCCCCTAC
TGGGAAGTGAGGAGCCCCTCTGCCCGGCCACCACCCCATC

Figure 3.2: Structure and sequence of the SVA located upstream of the FUS gene.

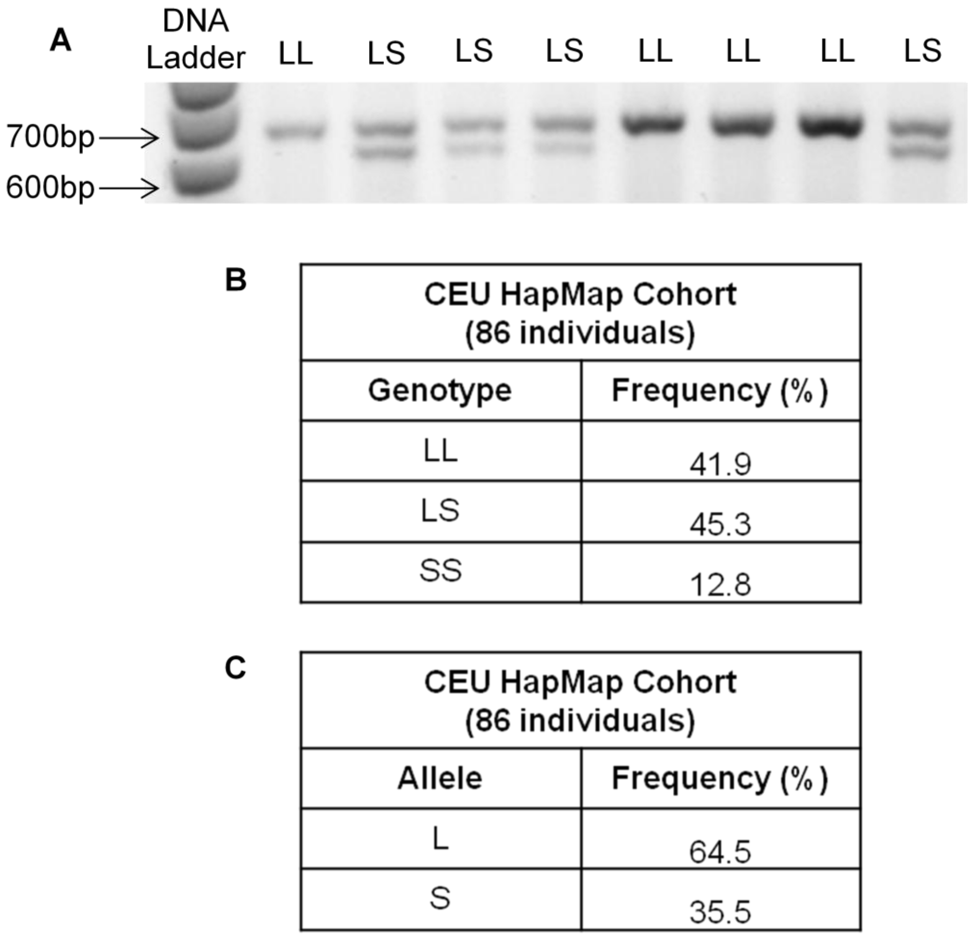
A - A schematic showing the components contributing to the structure of the SVA D located upstream of the FUS gene. It contains an Alu-like sequence, a TR, VNTR and a SINE. This particular SVA is missing the CCCTCT hexamer repeat seen at the 5’ end of a canonical SVA.

B - Sequence of the central repetitive region of the human SVA upstream of the FUS gene. There are 7 copies of a 37-40bp repeat in the TR and 3-4 copies of a 37-50bp repeat within the VNTR (for identification of the two alleles named long and short see section 3.4.2). The repeat underlined is the additional copy found in the long allele and is absent in the short allele of the SVA (sequence in UCSC corresponds to long allele). The sequence in italics within this repeat has G4 potential predicted by the Quadparser software; therefore a small percentage of the SVA long allele (1.8%) has G4 potential whereas the less common short allele does not. G4 is discussed in further detail in chapter 4.

C – The sequence of the central VNTR of the SVA upstream of the FUS gene in the chimpanzee. The VNTR of this SVA is much larger than compared to the human. The variability in the number of repeats of this VNTR has not been analysed therefore this has not been divided into a TR and VNTR as in the human.

3.4.2 Two alleles of the FUS SVA are present in a CEU HapMap cohort

During the cloning of the FUS SVA for use in a reporter gene construct (3.4.3) from commercial gDNA (Promega) two alleles of this SVA were identified (personal communication from Thomas Wilm). The frequency of the genetic variation within this region was then analysed further in the CEU HapMap cohort (Utah residents with Northern and Western European ancestry from the CEPH collection). The primer set used for genotyping targeted the central TR/VNTR. There were also two alleles identified within this cohort with PCR products 665bp in length for the long allele and 615bp for the short allele, which are shown in the example gel image in Figure 3.3A. The long allele is the most common and the frequencies of the different genotypes in the 86 individuals are as follows; LL-41.9%, LS-45.3% and SS-12.8% (Figure 3.3B). The frequency of on the long and short alleles, 64.5% and 35.5% respectively, in the CEU HapMap cohort are shown in Figure 3.3C. Analysis revealed that FUS SVA locus is in HWE within this population (p=0.99). The conditions for the amplification of SVAs require optimisation, sometimes proving to be a difficult region of DNA to amplify, and the conditions can vary between each element as discussed further in chapter 6.



CEU HapMap Cohort (86 individuals)	
Genotype	Frequency (%)
LL	41.9
LS	45.3
SS	12.8

CEU HapMap Cohort (86 individuals)	
Allele	Frequency (%)
L	64.5
S	35.5

Figure 3.3: Genotype and allele frequencies of samples from the CEU HapMap cohort for the SVA upstream of the FUS gene. A - An example image of the PCR products of the alleles of the TR/VNTR of the FUS SVA (L=665bp and S=615bp). The TR/VNTR of the SVA was amplified by PCR and the product run on a 1.2% agarose gel to allow sufficient separation of the alleles and calls were made on the genotype for that individual. B - The table shows the frequency of the genotype for the CEU HapMap cohort. (LL=homozygous long, LS= heterozygous long and short, SS=homozygous short). C – The table shows the allele frequencies of the long and short alleles of the FUS SVA in the CEU HapMap cohort. The locus is in HWE: p=0.99.

3.4.3 The FUS SVA and the VNTR within its structure can act as a regulatory domain *in vitro* and *in vivo*

Functional anaylsis of the SVA, its central VNTR component and genetic variants (long and short alleles) was undertaken in the lab by Thomas Wilm and Kejhal Khursheed. This data has been compiled into a manuscript titled 'An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS' along with data generated on the genetic variants in a SALS cohort discussed in section 3.4.4 of this chapter. The data of the *in vitro* and *in vivo* experiments has been summarised here to show the functional properties of the SVA.

Reporter gene constructs were cloned into the pGL3P vector including both variants of the SVA (L-SVA and S-SVA), and the isolated central TR/VNTR (L-TR/VNTR and S-TR/VNTR). The activity of the constructs was measured in the human neuroblastoma cell line SK-N-AS (Figure 3.4B). Significant differences were observed in the levels of reporter gene expression supported by the complete SVA or the TR/VNTR compared to the minimal SV40 promoter of the pGL3P vector alone (S-SVA $P<0.05$, L-SVA $P<0.05$, S-TR/VNTR $P<0.001$ and L-TR/VNTR $P<0.05$). Both alleles of the complete SVA repressed reporter gene expression while both alleles of the TR/VNTR were activators in this cell line, demonstrating that the SVA may contain multiple and distinct regulatory domains, one of which is a dominant repressor in SK-N-AS cells. When comparing the two different length TR/VNTR constructs no significant difference to the amount of reporter gene activity observed was noted, however there was a small but significant difference in the levels of reporter gene expression when these variants were contained within the complete

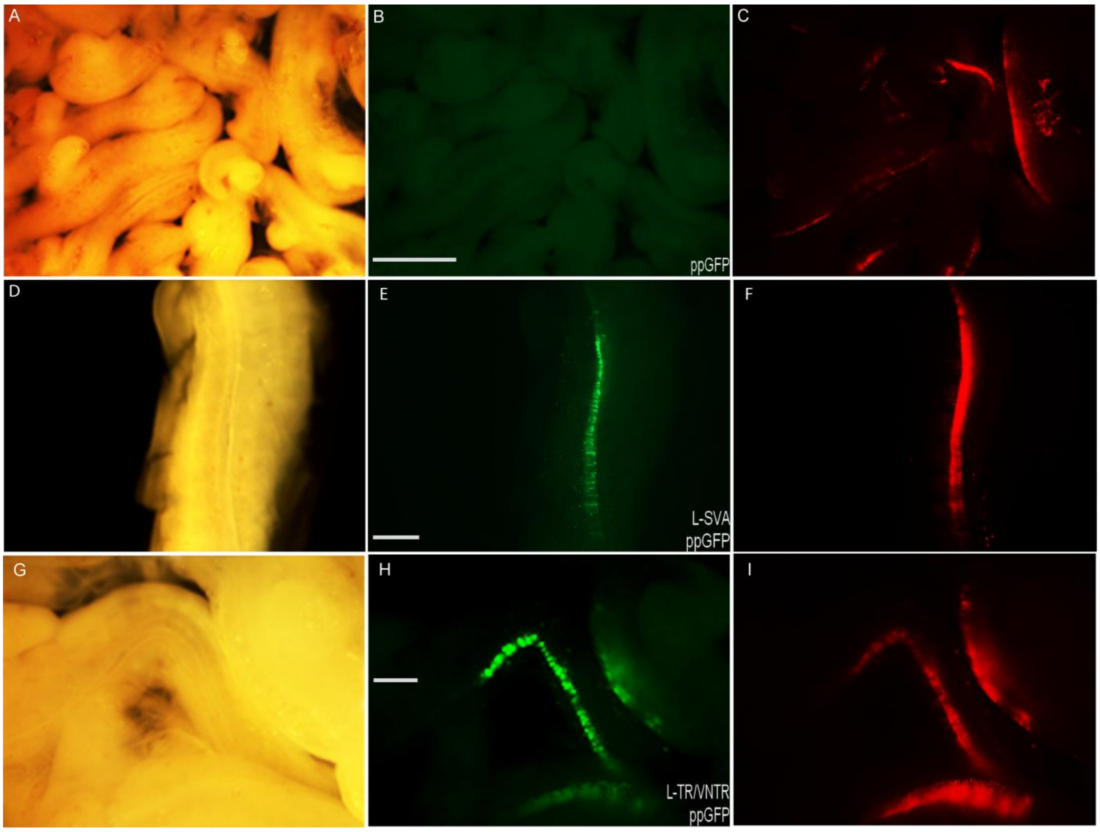
SVA sequence ($P<0.05$). In both the SVA and TR/VNTR constructs it was the long variant that showed lower activity when compared to the short.

The SVA and TR/VNTR (long allele) domains as used above in the SK-N-AS cell line were inserted into a reporter gene vector that had been developed previously in the lab to visualise activity via hrGFP in the chick embryo model. Briefly the reporter vector phrGFP contained the proximal human FUS promoter –160 of the major transcriptional start site to +84 upstream of hrGFP, the L-TR/VNTR and L-SVA sequences were inserted immediately upstream of the promoter sequence.

The test plasmid was injected into the neural tube and then transfected into the cells by electroporation. The reporter gene construct was co-injected with an internal control, the tomato reporter plasmid directed by the chick β-actin promoter; the latter acts as an internal control for cells which were successfully transfected. In this manner the activity and tissue specificity of the L-SVA and L-TR/VNTR reporter could be addressed and the internal control indicated the efficiency of the transfection. The series of FUS reporter gene constructs were injected into the developing embryo at embryonic stage 14HH and monitored at stage 22HH. The proximal FUS promoter alone did not support sufficient reporter gene expression to be observed in this assay (Figure 3.A panel B). Both the L-SVA and L-TR/VNTR reporter gene constructs supported expression; which was readily observed in the neural tube of the chick embryo (Figure 3.A panel E and H respectively). The activation properties of the L-TR/VNTR was consistent in the cell line model and the chick embryo model, however the complete L-SVA showed the repressive qualities in the cell line but acted as an activator in the chick embryo model enhancing expression over that of the minimal FUS promoter.

.

A



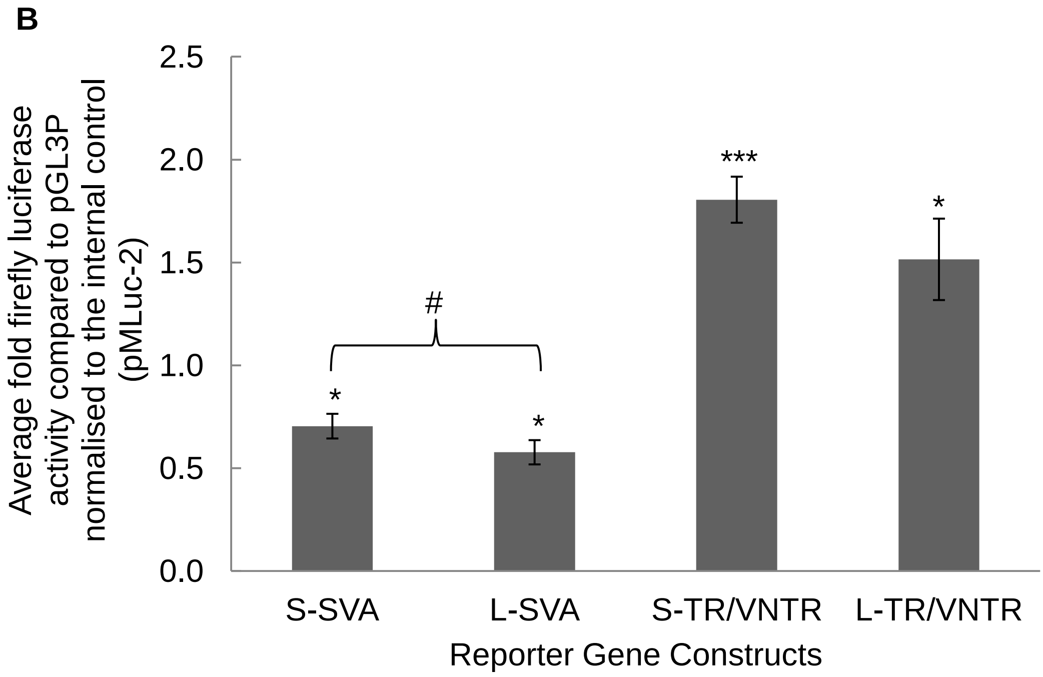


Figure 3.4: The FUS SVA and VNTR act as a functional regulatory domain *in vitro* and *in vivo*. This image has been compiled from data produced from two other members of the lab (TPW and KK) and is present within the manuscript (Savage et al) which has been submitted for publication.

A - Chick embryos were electroporated with either a FUS proximal promoter GFP (ppGFP) reporter construct (Panels A-C), L-SVA ppGFP-reporter (Panels D-F) or L-TR/VNTR ppGFP-reporter (Panels G-I) at stage 14HH and GFP expression analysed 48 hours later (stage 22HH). Expression could not be detected in the neural tube from the FUS proximal promoter sequences alone (panel B), however when either the L-SVA (panel E) or L-TR/VNTR (panel H) sequences were included, GFP reporter gene expression could readily be seen. Panels A, D & G show the corresponding brightfield images. Panels C,F &I show the identical fields taken with a red filter to demonstrate the extent of successful electroporation of the neural tube using a control tomato marker expression plasmid. Scale bar in B is 2mm and in E & H is 1mm.

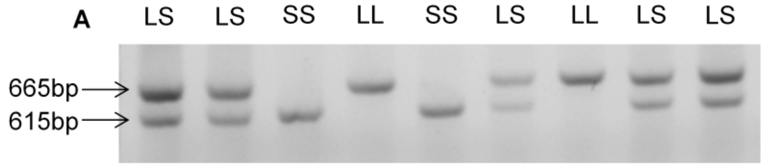
B – Reporter gene constructs containing each allele of the FUS SVA and TR/VNTR (long and short) were transfected into the neuroblastoma cell line, SK-N-AS. The fold values of each construct compared to pGL3P normalised to the internal control (pMLuc-2) to account for differences in transfection efficiency are displayed. One tailed t-test was used to measure the significance of fold activity of the FUS SVA and TR/VNTR over the minimal promoter of the pGL3P vector *P<0.05, ***P<0.001, and to compare the activity of the alleles of the SVA and the TR/VNTR to each other # P<0.05 N=7.

3.4.4 The FUS SVA does not show association with SALS

The genetic variation of the FUS SVA was analysed in a cohort of 241 individuals with SALS and a cohort of 228 matched controls to determine if there was an association of a specific variant of the SVA with SALS. The same primer set was used as in the genotyping of the CEU HapMap cohort and an example gel image of the PCR fragments in the SALS and controls cohort in shown in Figure 3.5A. The sequence of the two alleles of the TR/VNTR of the FUS SVA was confirmed by the sequencing of both an L and S allele after gel purification (2.2.4.4 and 2.2.5.10). This demonstrated that the L allele corresponded to the sequence found in the UCSC browser for the TR/VNTR of this SVA element. The following genotype frequencies were observed in the SALS cohort 45.6% LL, 39% LS and 15.4% SS and 46.9% LL, 42.1% LS and 11.0% SS in the matched controls (Figure 3.5B). In addition the allele frequencies for these cohorts are shown in Figure 3.5C. The genotype frequencies of the two cohorts were analysed for significant differences using CLUMP (Sham and Curtis 1995). Although there was a small difference of 4.4% between the frequencies of SS individuals in the SALS cohort compared to the matched controls these were found not to be significant when analysed. This small difference may be a false positive as the SS frequency in the CEU HapMap cohort is 12.8% (Figure 3.3B) which is in between the two values. The p-values from the T1 and T4 tests from the CLUMP analysis were P=0.36 and P=0.33 respectively (Sham and Curtis 1995).

HWE anlaysis was carried out for the FUS SVA locus in the SALS and control cohorts with both showing to be in HWE, p=0.09 and p=0.88 respectively. However there is a notable difference in the p values of the two cohorts; the SALS cohort had a much lower p value close to being statistically significant indicating there is a trend towards the FUS SVA locus being out of HWE within this group.

This locus was also shown to be in HWE in the CEU HapMap cohort with a p value of 0.99, indicating that out of the three populations the genotype frequencies in the SALS cohort showed the greatest difference to the expected based on allele frequency.



B

Sporadic ALS (241 individuals)	
Genotype	Frequency (%)
LL	45.6
LS	39.0
SS	15.4
Controls (228 individuals)	
LL	46.9
LS	42.1
SS	11.0

C

Sporadic ALS (241 individuals)	
Allele	Frequency (%)
L	65.1
S	34.9
Controls (228 individuals)	
L	68.0
S	32.0

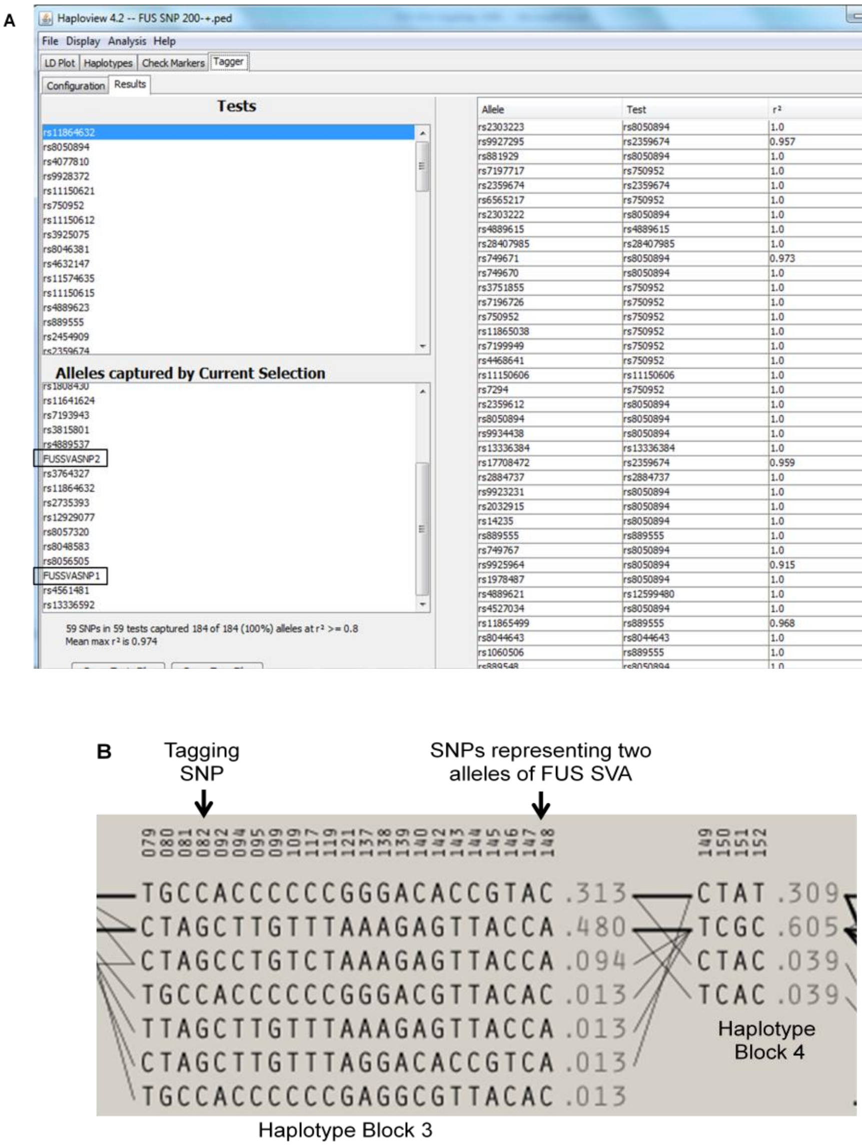
Figure 3.5: Genotype and allele frequencies of SVA located upstream of the FUS gene in Sporadic ALS and matched control cohort A - Example image of the two alleles of the TR/VNTR of the FUS SVA run on a 1.2% agarose gel after amplification using PCR. The genotype of each individual was determined. One of each allele was gel extracted and sequenced and corresponded to the previously sequenced and cloned alleles. (L=665bp and S=615bp) B - Table showing the percentage of each genotype in the SALS patients (241 individuals) and the matched controls (228 individuals). C – Table showing the allele frequencies for the FUS SVA in the SALS and control cohorts. The FUS locus is in HWE in both the SALS and control cohorts, p=0.09 and p=0.88 repectively.

3.4.5 Multiple tagging SNPs identified for the FUS SVA alleles including rs11864632

Although the analysis of the SVA by PCR analysis of the VNTR polymorphism did not show significant statistical association with the SALS cohort I could only address a limited population using this method. I therefore set out to identify tagging SNPs so that such SNPs could be addressed in not only larger cohorts of SALS but other conditions in which FUS is implicated. Tagging SNPs were identified for the alleles of FUS SVA using the genotype data generated for the individuals of the CEU HapMap cohort (3.3.2 and 3.4.2) and the freely avaiable SNP genotype data for this cohort in the International HapMap database. The screen shot in Figure 3.6A shows the SNPs tagged or in linkage disequilibrium with the SNP rs11864632 which includes the SNPs created to represent the alleles of the FUS SVA. The haplotype block (3) in which the SNPs representing the alleles of the SVA are inherited is shown in figure 3.6B which includes rs11864632 along with 20 other SNPs. A genotype of C at rs11864632 is inherited along with the short allele of the FUS SVA (represented at SNP 147 and 148 with a genotype of AC) and a G at rs11864632 is inherited with the long allele of the FUS SVA (represented at SNP 147 and 148 with a genotype of CA). This was validated by addressing the genotype of SNP rs11864632 for each individual in the HapMap that had been genotyped for the FUS SVA and checking they correlated as predicted by the haploview analysis.

The 20 other SNPs that were tagged by rs11864632 with an r^2>0.8, 18 of which were inherited in haplotype block 3 and 2 in haplotype block 4, were analysed for the correlation of their genotype with the genotype of the FUS SVA. Seven also showed the ability to predict the genotype of the FUS SVA as accurately as rs11864632 (100%) with rest being able to predict >90% of the time within the

HapMap individuls genotyped for the FUS SVA. Therefore several SNPs have been shown to be in strong linkage disequilibrium with the alleles of the FUS SVA and could be used as tagging SNPs. This would remove the requirement for the labour intensive PCR and gel electrophoresis method for genotyping the FUS SVA as the SNP genotype could be used to determine alleles of the SVA present. This will constitute future experiments in this area and extrapolated to other appropriate SVAs to determine if they represent a genetic risk based on polymorphism present for particular disease as discussed in chaper 6.



A
Haploview 4.2 -- FUS SNP 200-+.ped
File Display Analysis Help
Key
LD Plot
Haplotypes
Check Markers
Tagger
Configuration
Results

Tests

rs11864632
rs8050894
rs4077810
rs9928372
rs11150621
rs750952
rs11150612
rs3925075
rs8046381
rs4632147
rs11574635
rs11150615
rs4889623
rs889555
rs2454909
rs2359674

Alleles captured by Current Selection

rs1808430
rs11641624
rs7193943
rs3815801
rs4889537
FUSSVASNP2
rs3764327
rs11864632
rs2735393
rs12929077
rs8057320
rs8048583
rs8056505
FUSSVASNP1
rs4561481
rs13336592

59 SNPs in 59 tests captured 184 of 184 (100%) alleles at r^2 >= 0.8
Mean max r^2 is 0.974

Allele	Test	r^2
rs2303223	rs8050894	1.0
rs9927295	rs2359674	0.957
rs881929	rs8050894	1.0
rs7197717	rs750952	1.0
rs2359674	rs2359674	1.0
rs6565217	rs750952	1.0
rs2303222	rs8050894	1.0
rs4889615	rs4889615	1.0
rs28407985	rs28407985	1.0
rs749671	rs8050894	0.973
rs749670	rs8050894	1.0
rs3751855	rs750952	1.0
rs7196726	rs750952	1.0
rs750952	rs750952	1.0
rs11865038	rs750952	1.0
rs7199949	rs750952	1.0
rs4468641	rs750952	1.0
rs11150606	rs11150606	1.0
rs7294	rs750952	1.0
rs2359612	rs8050894	1.0
rs8050894	rs8050894	1.0
rs9934438	rs8050894	1.0
rs13336384	rs13336384	1.0
rs17708472	rs2359674	0.959
rs2884737	rs2884737	1.0
rs9923231	rs8050894	1.0
rs2032915	rs8050894	1.0
rs14235	rs8050894	1.0
rs889555	rs889555	1.0
rs749767	rs8050894	1.0
rs9925964	rs8050894	0.915
rs1978487	rs8050894	1.0
rs4889621	rs12599480	1.0
rs4527034	rs8050894	1.0
rs11865499	rs889555	0.968
rs8044643	rs8044643	1.0
rs1060506	rs889555	1.0
rs889548	rs8050894	1.0

TGCCACCCCCCGGGACACCGTAC .313
CTAGCTTGTTTAAAGAGTTACCA .480
CTAGCCTGTCTAAAGAGTTACCA .094
TGCCACCCCCCGGGACGTTACAC .013
TTAGCTTGTTTAAAGAGTTACCA .013
CTAGCTTGTTTAGGACACCGTCA .013
TGCCACCCCCCGAGGCGTTACAC .013

CTAT .309
TCGC .605
CTAC .039
TCAC .039

Figure 3.6: rs11864632 identified as a tagging SNP for the SVA upstream of the FUS gene. A – Screen shot of the haploview software results from running the tagger analysis. The SNP rs11864632 is highlighted in the top left box is tagging the SNPs in the bottom left box including the SNPs corresponding to the FUS SVA alleles. B – Haplotype block of the region containing the FUS SVA and its tagging SNP. Genotypes of the SNPs are shown. 82 represents the tagging SNP rs11864632 with this loci either containing a G or C. 147 represents the long allele of the FUS SVA and 148 represents the short allele of the FUS SVA (AC=short allele, CA=long allele). A genotype of C at the rs11864632 SNP tags for the short allele of the FUS SVA and a G tags for the long allele. (r^2>=0.8)

3.5 Discussion

Mutations within the FUS gene and the levels at which it is expressed have been associated with diseases of the central nervous system such as ALS and Frontotemporal lobar degeneration (FTLD) (Neumann et al. 2009; Mackenzie et al. 2010; Mitchell et al. 2013), however little is known about the parameters regulating FUS gene expression. The FUS gene locus was analysed for potential regulatory and polymorphic domains and a candidate was identified in a SVA retrotransposon 10kb upstream of the TSS of the gene (Figure 3.1). SVAs are highly CG-rich and are assumed to be silenced within the genome through mechanisms such as methylation in an attempt to stop their retrotransposition. However there is evidence for this methylation to be lost over retrotransposons in cancer, for somatic retrotransposition in the aging brain and for transposable elements to become active in replicatively senescent cells (Szpakowski et al. 2009; Baillie et al. 2011; De Cecco et al. 2013). The activity of the SVAs may be variable depending on their environment and corresponding epigenetic structure.

Retroviruses, exogenous and endogenous, have been linked with ALS (Alfahad and Nath 2013). There was an increased prevalence of reverse transcriptase, a key enzyme in the retrovirus life cycle converting RNA to DNA, observed in the serum of patients with SALS (Andrews et al. 2000; Steele et al. 2005). In the second study (Steele et al. 2005) the elevated reverse transcriptase levels were interpreted as indicative of involvement of an endogenous retrovirus rather than an exogenous retrovirus as blood relatives also had elevated levels whereas spouses were the same as controls. Another study also implicated the role of retrotransposons in ALS as HERV-K transcripts and reverse transcriptase protein in autopsy brain tissue of patients with ALS was detected along with the abberant expression of TDP-43 (Douville et al. 2011). The cellular environment that led to this increased expression of HERV-K transcripts and reverse transcriptase may be a global change that could influence the expression or activity of other retrotransposons in the genome including the SVAs. Alteration of epigenetic factors may modulate any transcriptional properties embedded within the SVA affecting the regulation of the FUS gene and this may be dependant on the genetic variation at this locus.

The SVA upstream of the FUS gene is polymorphic in the second of its central repetitive domains with two alleles identified in the population named long and short (for structure and sequence see figure 3.2). The frequency of these two alleles was analysed in a CEU HapMap cohort with the long allele most common in the population; LL-41.9%, LS-45.3% and SS-12.8% (Figure 3.3). The demonstration of polymorphic variation in this region in a VNTR domain, a class of domain that has extensively been shown to act as a risk factor in a number of diseases suggested that the region could support differential gene expression under appropriate environmental cues.

The potential for the SVA to act as a transcriptional regulator was tested in a reporter gene construct by other members of the research group (Thomas Wilm and Kejhal Kursheed) and is summarised in figure 3.4 from a submitted manuscript (Savage et al). The regulatory properties of both alleles, when part of the complete SVA and when alone as the TR/VNTR, were tested in the neuroblastoma cell line SK-N-AS (Figure 3.4A). It is interesting that while both the L & S TR/VNTR regions were enhancers the L & S SVA acted as a repressor in the SK-N-AS cell line. This would suggest that in addition to the activator region in the TR/VNTR the SVA contains a strong silencer element, flanking this central TR/VNTR region, which is functional in the SK-N-AS neuroblastoma cell line. There was no significant difference between the activities of the two alleles of the TR/VNTR when tested alone but there was a small but significant difference between the two alleles when tested as part of the complete SVA ($P<0.05$). To further validate the regulatory properties of this domain the long allele of the complete SVA and of the TR/VNTR domain were tested in the neural tube of the chick embryo. This region contains motor neurons which are the appropriate cell type to test a domain that might be involved in ALS; however FUS is a ubiquitously expressed protein. The expression of the FUS mRNA and protein was confirmed in the chick embryo at stage E5 (Kejhal Khursheed personal communication). As in the cell line model the TR/VNTR acted as an activator but in this model the SVA also demonstrated activator properties. This would be consistent with previous definitions of VNTRs from the both the serotonin and the dopamine transporters having cell line specific properties in reporter gene constructs (Michelhaugh et al. 2001; Haddley et al. 2008; Paredes et al. 2012) and that of the serotonin transporter having tissue specific properties in a transgenic mouse model (MacKenzie and Quinn 1999a). The data also validates again the fact that primate-specific domains such as the serotonin transporter VNTR and now the FUS SVA can function in other species. This has been previously proposed as a mechanism in evolution to generate different transcriptome profiles that would involve the promoters or regulatory domains of a gene altering to take advantage of the transcription factor complement in a cell to expand or lay extra complexity upon the gene's regulation.

Genotypic analysis of the TR/VNTR of the SVA was analysed in a SALS and control cohort provided by collaborators at King's College London. There were two alleles identified in this cohort (Figure 3.5A) that had also been observed in the

CEU HapMap cohort. The frequency of LL, LS and SS was found not to be significantly different in the sporadic cases compared to the matched controls, although minor differences could be seen between the frequency of individuals with a SS genotype in SALS and controls (15.4% vs 11.0%) when analysed using CLUMP (Sham and Curtis 1995). A much larger cohort will be required to validate such variation as an association in the SALS cohort.

A major problem performing genotyping analysis with large domains such as the SVA or the TR/VNTR domain is its time consuming and labour intensive involving PCR and its subsequent analysis through gel electrophoresis. Therefore the genotyping data for the CEU HapMap cohort from Figure 3.3 was used to identify a tagging SNP for the alleles of the FUS SVA. Several SNPs were shown to be in linkage disequilibrium with the FUS SVA (Figure 3.6), which may enable much larger cohorts to be analysed for variation of the FUS SVA. Large numbers of SNPs can be analysed at once and there are many genome wide association studies carried out determining the genotypes of SNP across the human genome. This data could be utilised along with the FUS SVA tagging SNP to enlarge the numbers for the SALS cohort already generated or within different cohorts.

In summary this chapter has demonstrated a novel primate tissue specific regulator that could play a role in FUS regulation in the form of a SVA located 10kb upstream of the FUS gene. This regulation could be modified by a number of environmental challenges including the changes correlated with the increased reverse transcriptase activity seen in the serum of ALS patients that could affect the epigenetic structure of the FUS locus in a manner analogous to epigenetic changes in cancer observed over retrotransposons. There was no association of the FUS SVA genotype with the SALS cohort analysed, however the identification of a tagging

SNP for this domain may make it easier to increase the numbers of this cohort and analyse the SVA variation within other disease cohorts.

Chapter 4

Global analysis of the distribution and structure of SVAs within the human genome

4.1 Introduction

The analysis of the SVA upstream of the FUS gene (chapter 3) has identified an element that has the ability to affect gene expression in a reporter gene model *in vitro* and *in vivo* and is polymorphic for the number of repeats in its central VNTR. Other SVAs throughout the human genome may also share these functional properties of the FUS SVA so the location of these elements is important for identifying the potential impact of SVAs on the genome on a global scale. There were 2762 SVAs identified in the human genome by Wang et al, so providing a manageable number in this retrotransposon family to analyse in a global approach to determine the distribution of the SVAs across the human genome.

The distribution of Alus has been associated with increased GC content whereas L1 elements are predominantly found in AT-rich regions (Medstrand et al. 2002). Analysis of the distribution of SVAs by Wang et al determined they were associated with regions of higher GC content but not with genes despite the SVAs preference for gene dense chromosomes (Wang et al. 2005). However the analysis of somatic retrotransposon insertions in the brain by Baillie et al revealed that these insertions were prevalent in actively transcribed genes and therefore more likely in regions of active chromatin (Baillie et al. 2011). This would correlate with the hypothesis that retrotransposons insert into active regions of chromatin due to its more open and accessible state for the insertions to occur. Therefore I expanded upon the analysis by Wang et al to determine if the SVA insertions do correlate with specific features of the genome, if the distribution varies between subtypes and how this could affect their influence on the human genome.

The nature of the sequence of SVAs also provides the potential to influence gene expression through multiple mechanisms such as differential methylation or G4 formation. Therefore a global *in silico* analysis of the potential of SVAs to form G4 DNA was undertaken to determine the extent to which SVAs may contribute to G4 within the genome and if amount of G4 potential was equal across the different subtypes. A handful of elements were also analysed for their potential to share characteristics of CpG islands.

This global analysis will hopefully provide greater insight into the pattern of insertions of the SVAs within the human genome and their potential functional impact.

4.2 Aims

- To determine if the site of insertion of SVA elements across the genome is random or correlated with features of the genome using the UCSC genome browser.
- To analyse the prevalence of SVAs key regions of the genome such as introns and promoters and their orientation in relation to the nearest gene.
- To analyse the potential of the SVA sequence using an *in silico* approach to form a type of secondary structure called G4.
- To calculate the potential of SVAs by subtype to act as CpG islands.

4.3 Methods

4.3.1 Manual annotation of SVA coordinates from UCSC SVA track and creation of a new SVA track for use in analysis

A list of coordinates was generated for all SVAs identified in the Hg19 from the repeat masker track of UCSC genome browser (http://genome.ucsc.edu/index.html). To download information from the tracks available in the UCSC genome browser the table browser needs to be accessed either through the homepage of the browser or the 'tools' menu. The table browser can be used to gather summary/statistics of each track, for example the number of LINE elements in the repeat masker track and the number of bases they contribute to the whole genome etc, or it can be used to extract data. The data from each track can be extracted as sequence data or chromosomal loci and can be viewed instantly or sent directly to the Galaxy software for further analysis. The following example of extracting the chromosomal loci of SVA elements from the repeat masker track will be used to illustrate the method employed and Figure 4.1 shows a screen shot of the table browser during this process. The correct species and version of the genome are selected along with the track of interest. In this instance the group 'variation and repeats' and the track 'repeat masker'. The filter button was then used to ask for the repeat of choice which was the SVAs. The SVAs are split into subtypes A-F so the data had to be extracted by subtype. The elements in the repeat masker track were filtered by the 'repName' option for each subtype in turn for example SVA_A. The output format of the data was chosen as 'BED-browser extensible data' and then the 'get output' button was pressed. There is then the option of creating one record per item or adding flanking regions to each item. The 'whole gene' option is chosen as this will create one record per SVA without additional flanking regions and the 'get

BED' button is used. The chromosomal loci of your chosen items, in this case SVA for each subtype, will be listed.

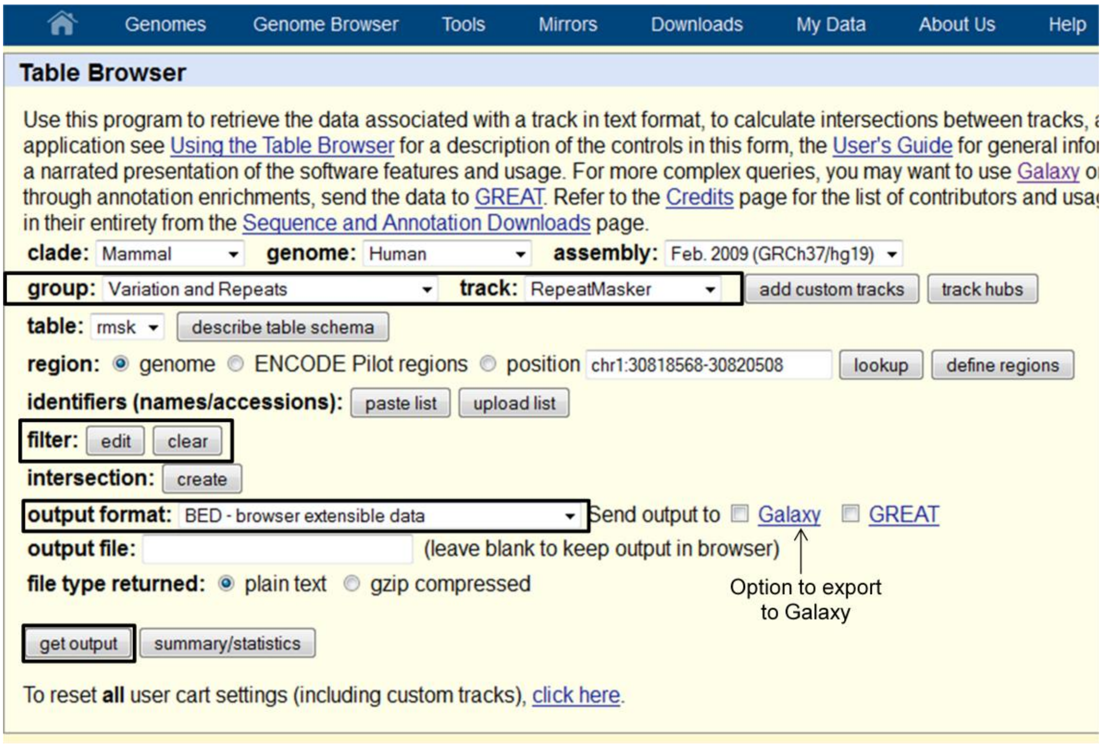


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, a
application see Using the Table Browser for a description of the controls in this form, the User's Guide for general info
a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy o
through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usa
in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)

group: Variation and Repeats track: RepeatMasker add custom tracks track hubs

table: rmsk describe table schema

region: ● genome ○ ENCODE Pilot regions ○ position chr1:30818568-30820508 lookup define regions

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection: create

output format: BED - browser extensible data Send output to □ Galaxy □ GREAT

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Figure 4.1: Screen shot of the table browser available on UCSC used to extract track data. The parameters were set to the required species and the repeat masker track selected. The filter was then used to select for SVAs by subtype. The data was collected in the BED format and retrieved by selecting the get output button. The data can also be exported to Galaxy which is freely available software on the internet that can be used to analyse the relationship between data sets or create new chromosomal coordinates surrounding the regions of interest. The summary/statistics button also allows for information such as the numbers of items in the selected track.

Many of these SVAs on the repeat masker track were fragmented and sometimes did not include the CCCTCT hexamer repeat at the 5'end (this component was often classified as a separate simple repeat). The seventh subtype F1 was also not present in the UCSC genome browser as the members of this group were included in the subtype F. Therefore a new list of SVA coordinates was to be created to be uploaded to UCSC as a custom track for use in analysis and this process is summarised in Figure 4.2 and the following paragraphs. Firstly with the aid of the interrupted repeat masker track in UCSC that recombined some of the SVAs that were fragmented into two or more pieces the coordinates were manually altered so there was one set of coordinates per SVA in Microsoft Excel. Secondly to include the CCCTCT repeat and poly A tails that are part of a canonical SVA but sometimes not contained within the coordinates of the SVA by the UCSC browser the 100bp upstream and downstream of the SVA coordinates (created in Galaxy http://galaxyproject.org/) were intersected with the simple repeat track from UCSC. This generated a list of all the simple repeats upstream and downstream of an SVA with their coordinates and their sequences. These coordinates of the simple repeats containing the CCCTCT like repeats or the poly A-tails of the SVA were merged with the coordinates of the SVAs. Finally to define which of the SVA Fs classified by UCSC were actually subtype F1 the Blat tool from UCSC was used to search for all sequences in the human genome sharing homology to the sequence of exon1 of the MAST2 gene (subtype F1 was generated when a 5' transduction integrated sequence from the MAST2 gene with an SVA F). Any results that were upstream of an SVA F were used to generate the loci of the subtype F1.

This list of SVA loci was uploaded to UCSC genome browser as a custom track called 'annotated SVAs' for use in the analysis and included 2676 elements. To upload the track the table browser was opened in UCSC and then the add custom track was chosen. The species and the version of the genome were set to human and version Hg19 that the track will be uploaded to. The chromosomal position (coordinates) and the subtype of the SVA of all the elements identified in the manual annotation process were pasted into the box on the page to create the new track, for example: chr1 30819215 30819861 SVA_A. The submit button is pressed to create the new track; this can then be accessed and viewed like any other UCSC track.

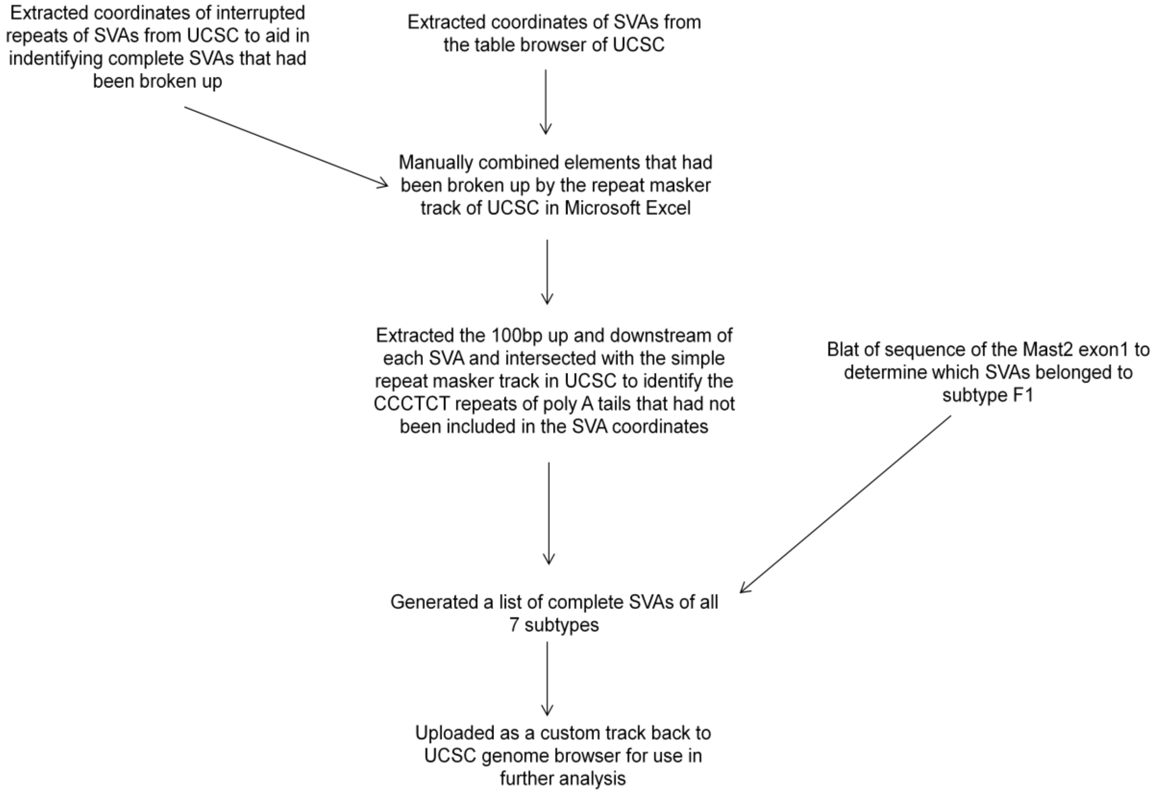

Extracted coordinates of interrupted repeats of SVAs from UCSC to aid in indentifying complete SVAs that had been broken up
Extracted coordinates of SVAs from the table browser of UCSC
Manually combined elements that had been broken up by the repeat masker track of UCSC in Microsoft Excel
Extracted the 100bp up and downstream of each SVA and intersected with the simple repeat masker track in UCSC to identify the CCCTCT repeats of poly A tails that had not been included in the SVA coordinates
Blat of sequence of the Mast2 exon1 to determine which SVAs belonged to subtype F1
Generated a list of complete SVAs of all 7 subtypes
Uploaded as a custom track back to UCSC genome browser for use in further analysis

Figure 4.2: Flow diagram describing the steps taken to generate the annotated SVA track for use in analysis of their distribution. To accurately analyse the distribution of the SVAs across the human genome a new track had to be generated as many of the SVAs were fragmented in the UCSC genome browser, some did not contain their CCCTCT repeat and poly A tails in the coordinates and the subtype F1 was not present.

4.3.2 Analysis of SVA distribution across human chromosomes and their correlation with gene density

Information on chromosome size and the number of genes present was collected from the NCBI website for HuRefChr37.3 (http://www.ncbi.nlm.nih.gov/).

The size of each chromosome was used to generate the number of SVAs expected for that particular chromosome and was plotted against the number of observed SVAs. The percentage of each SVA subtype present in the whole genome and the number of actual SVAs on each chromosome was used to calculate the expected number of each subtype for each chromosome. The Chi squared statistical test was used to determine if the distributions of SVAs and the subtypes across chromosomes were significantly different to that of the expected distribution.

The gene density (number of genes per million bases) and the SVA density (number of SVAs per million bases) were calculated for each chromosome and the two values plotted against each other on a scatter graph. The correlation coefficient for the relationship between the two sets of data was calculated and this was repeated for each subtype individually.

4.3.3 Analysis of SVA distribution within genes, intergenic regions and gene deserts

To analyse the distribution of SVAs further the genome was separated into three broad categories genes, intergenic regions and gene deserts. Gene deserts were defined as regions between genes which were over 250kb away from the start or end of a known gene, intergenic as regions between genes that are less than 250kb from the start or end of a known gene and genes were determined by the UCSC gene track from the UCSC genome browser. The regions were defined in this manner to assess if the retrotransposons had preferentially inserted into regions devoid of genes (gene deserts) or regions of the genome that could include active chromatin where genes and intergenic regions potentially containing regulatory domains (up to 250kb from

TSS) are located (Visel et al. 2009b; Shanley et al. 2010). The custom track 'annotated SVAs' was intersected with tracks for the three categories in the table browser of UCSC and the number of SVAs in these three regions was recorded. The three lists were analysed for any overlap (i.e. SVAs that spanned a border between two categories) and a decision was made as to which category the SVA should be assigned to. The number of observed SVAs in genes, intergenic regions and gene deserts was compared to the expected for the size of those regions. This analysis was also completed for three other retrotransposons: SINEs, LINEs and LTRs. The SVAs' distribution was also broken down into subtypes and compared to their proportional distribution across the whole genome.

4.3.4 Analysis of SVA distribution in defined regions up to 20kb up and downstream of genes

To assess the prevalence of SVA insertions within close proximity of genes and if they had been selected against in these regions the number of SVAs within set distances upstream of genes was determined. Galaxy software was used to generate coordinates of 1kb, 10kb and 20kb upstream of genes. These were uploaded back into UCSC and the table browser was used to analyse the number of SVAs in these regions. These numbers were compared to the expected number of SVAs determined by the size of these regions if the insertion was random. The same analysis was carried out for three other classes of retrotransposons (LTRs, LINEs and SINEs) and for each class the fold difference of the observed number of elements in these set regions of the genome to the expected was calculated. The distribution of the SVAs was broken down into subtypes for each of these set regions of the genome and compared to their proportional distribution across the whole genome. The same method was followed to analyse the presence of retrotransposons in the regions 1kb, 10kb and 20kb downstream of genes.

4.3.5 Analysis of orientation of SVAs in relation to their nearest gene

The strand of DNA the genes that had been previously indentified to contain a SVA within an exon or intron, 10kb upstream and 10kb downstream (4.3.3 and 4.3.4) was compared to the strand of DNA the SVA was located on. The coordinates of the SVAs and the strand of DNA they are located on (- or +) were downloaded from the table browser in UCSC as were the coordinates and the strand of DNA for the genes containing a SVA within their sequence or their 10kb flank. These were aligned manually in Microsoft Excel and whether the SVA and gene were on the same or opposite strand was recorded. The percentage of genes and SVAs on the same strand of DNA was calculated.

4.3.6 Prediction of sequences of SVAs that could form G-quadruplex DNA

The ‘annotated SVAs’ track (see 3.3.1) was used along with a track provided by a colleague (Mr Paul Myers) that contained all the sequences in the human genome that would potentially form G4 as predicted by Quadparser software (http://www.quadruplex.org/). This software uses a predictive algorithm that identifies potential G4 sequences on either strand of DNA and uses the following formula $G_{3+}N_{1–7}G_{3+}N_{1–7}G_{3+}N_{1–7}G_{3+}$ (Wong et al. 2010). The process of quantifying the amount of G4 potential in each SVA domain and other repetitive elements of the genome is summarised in Figure 4.3. The SVA and the G4 prediction track were intersected to provide the number of bases within the SVAs that have the potential to form G4 DNA. This number was calculated as a proportion of the total amount of G4 DNA predicted sequences in the genome. This was also done for several other elements which were either repetitive or mobile structures with the human genome; including simple repeats, microsatellites, LTRs, LINEs, SINEs and DNA transposons for comparison to the data for the SVAs. The fold difference of the proportion of these elements that are contributing to potential G4 DNA formation to their proportion of the whole genome was calculated and plotted on a bar graph.

To determine the amount of G4 DNA that could be formed separately by the CCCTCT hexamer repeat and the VNTRs of each SVA subtype two new tracks in UCSC were created; one containing the loci of the CCCTCT hexamer repeat alone and the other containing the loci of the VNTRs alone. This was achieved by intersecting the ‘annotated SVAs’ track with the Simple Repeat track of UCSC. This generated a list of all simple repeats within the SVAs, which was then merged in Galaxy so any overlapping repeats became one. This list contained the CCCTCT hexamer repeats, the VNTRs and the poly A-tails of the SVAs so the genomic sequence of all the loci were extracted. The sequence of all the repeats was used to delete the loci containing the poly A tails and then the remaining coordinates could be separated into two separate lists; one for the hexamer repeats and one for the VNTRs. The two new lists of coordinates were uploaded back to UCSC to create new tracks for the hexamer repeats and the VNTRs of the SVAs. This was intersected with the track containing all the sequences in the genome that could form G4 DNA and the number of bases of the hexamer repeat and the VNTRs that overlapped for each subtype was recorded.

The number of bases that could form G4 from the CCCTCT and the VNTRs for each subtype was converted to a proportion of the total number bases of each subtype and was plotted on a bar graph. Any G4 DNA of the SVAs that could not be attributed to the CCCTCT element of the VNTR was called ‘other’.

The tracks of the hexamer repeats and the VNTRs of the SVAs were used to calculate the average size of each subtypes’ hexamer repeat and VNTR. To determine if the differences in the amount of G4 formed by the hexamer repeats and the VNTRs of the subtypes was due to size or sequence a comparison across the subtypes of the their hexamer repeat and VNTR size and the percentage that they could form G4 was carried out.

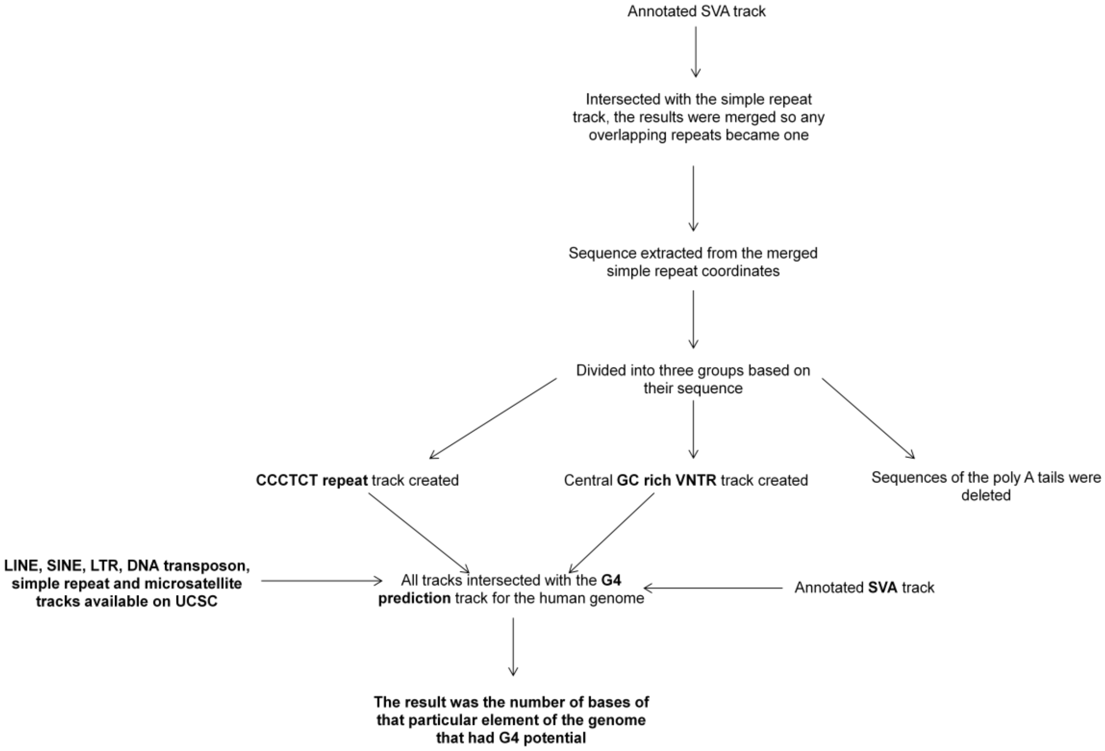

Annotated SVA track
Intersected with the simple repeat track, the results were merged so any overlapping repeats became one
Sequence extracted from the merged simple repeat coordinates
Divided into three groups based on their sequence
CCCTCT repeat track created
Central GC rich VNTR track created
Sequences of the poly A tails were deleted
LINE, SINE, LTR, DNA transposon, simple repeat and microsatellite tracks available on UCSC
All tracks intersected with the G4 prediction track for the human genome
Annotated SVA track
The result was the number of bases of that particular element of the genome that had G4 potential

Figure 4.3: A flow diagram demonstrating the process of identifying the amount of G4 potential with specific elements within the human genome. The amount of G4 in certain elements of the genome was determined using tracks within UCSC browser and a track containing the coordinates of the regions of the genome with G4 potential based on the equation from Quadparser. Two new tracks were generated of the coordinates of the CCCTCT repeat and the GC-rich VNTR of the SVAs to determine which domains of the SVA contribute to their G4 potential.

4.3.7 Potential of the sequences of the SVA subtypes to provide CpG islands across the human genome

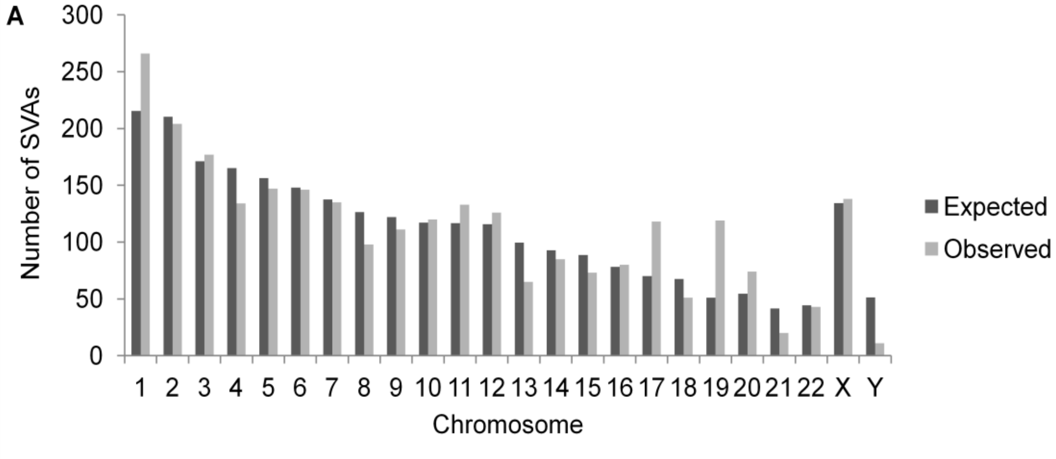
The UCSC genome browser contains a track predicting CpG islands across the genome using the following criteria for a region to be considered a CpG island: a GC content of 50% or more, length greater than 200bp and a ratio greater than 0.6 of observed number of CG dinucleotides to the expected number on the basis of the number of Gs and Cs. This ratio is calculated using the following formula: Obs/Exp CpG = Number of CpG * length of sequence / (Number of C * Number of G) (Gardiner-Garden and Frommer 1987). The above criteria were used to analyse SVAs and their central VNTR domain for their potential as CpG islands. The sequence of the SVAs was extracted from the 'annotated SVAs' track on UCSC genome browser for each subtype. This was a preliminary analysis of a handful of SVAs therefore 3 SVAs from each subtype were chosen at random, but it was confirmed they contained all the domains of a canonical SVA, and analysed for their ability to meet the criteria listed.

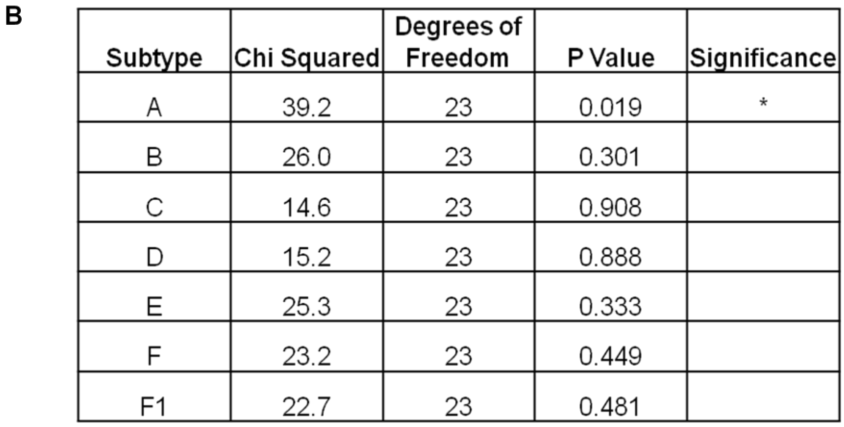
4.4 Results

4.4.1 Distribution of SVAs is not correlated to chromosome size but to their gene density

A previous study carried out by Wang et al had shown that the distribution of SVAs across chromosomes was significantly different to the predicted distribution, which would assume a random insertion model so the number of insertions was directly proportional to chromosome size (Wang et al. 2005). However I decided to replicate that study as to see if the data generated in my study was comparable to the published data to validate the methods in use (4.3.1). This analysis also showed that the expected and the actual distributions were significantly different if the expected number of SVA insertions for each chromosome was based on its size (X^2=220.0, df=23, P<0.001) (Figure 4.1A). Chromosome 1, 17, 19 and 20 showed a much higher number of SVAs than predicted for their size and this was noted that they are some of the most gene dense chromosomes (in particular 17 and 19). Conversely the chromosomes with lower than expected SVAs (13, 21 and Y) have much lower gene densities.

In addition to this the distribution of the SVAs by subtype across chromosomes was analysed to determine if the subtypes followed the general trend of all SVAs being located in greater numbers on gene dense chromosomes or if the subtypes had inserted differently to each other. The predicted number of each SVA subtype was calculated for each chromosome based on the actual number of SVAs on each chromosome and the percentage of each subtype across the whole genome. This demonstrated that subtype A's distribution was significantly different to the predicted (P<0.05) (Figure 4.1B). All the other subtypes were distributed as predicted by the actual number of SVAs identified on each chromosome. SVA As appear to have the opposite chromosomal distribution to the general distribution of all SVAs; being found in general in larger numbers on the chromosomes with lower gene densities.

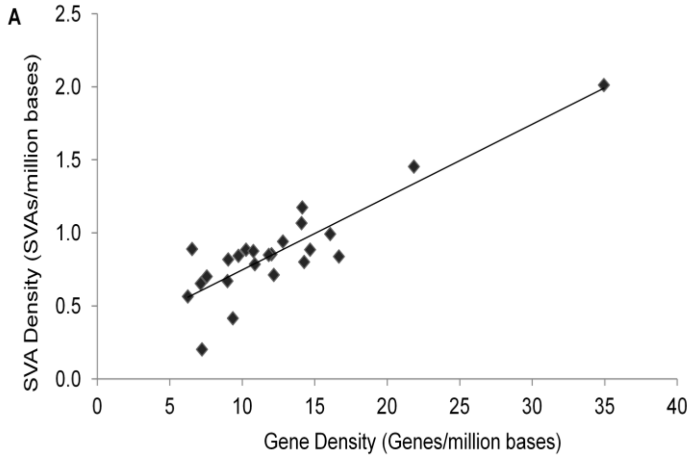




Subtype	Chi Squared	Degrees of Freedom	P Value	Significance
A	39.2	23	0.019	*
B	26.0	23	0.301	
C	14.6	23	0.908	
D	15.2	23	0.888	
E	25.3	23	0.333	
F	23.2	23	0.449	
F1	22.7	23	0.481	

Figure 4.1: Chromosomal distribution of SVAs is significantly different to the predicted. A - The number of SVAs expected on each chromosome based on their size compared to the observed. Chi squared test showed that the actual chromosomal distribution of the SVAs was significantly different to the predicted. (X^2=220.0, df=23, P<0.001) B – The table shows the results of the chi squared test for each SVA subtypes' predicted distribution compared to the actual with subtype A only showing a significant difference.

The hypothesis would therefore be that the SVA distribution (except subtype A) was related to the distribution of genes, which was investigated by comparing the SVA density to the gene density of each chromosome (Figure 4.2A). This showed a positive correlation (r=0.74) for the relationship of chromosomal SVA density and gene density calculated using bootstrap confidence of 95% to remove outliers. This correlation was tested for each individual SVA subtype (A-F1) and the positive correlation with gene density was maintained for all subtypes except A, which showed a negative correlation with gene density (Figure 4.2C).



C

SVA Subtype	Correlation Coefficient
All Subtypes	0.74
A	-0.55
B	0.32
C	0.59
D	0.72
E	0.36
F	0.70
F1	0.31

B

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Gene Density	14.1	9.7	6.5	7.6	9.0	12.0	11.8	9.0	10.9	10.3	16.1	12.8	6.3	14.3	12.2	14.7	21.8	7.1	34.9	14.1	9.3	16.7	10.8	7.2
SVA Density	1.1	0.8	0.9	0.7	0.8	0.9	0.8	0.7	0.8	0.9	1.0	0.9	0.6	0.8	0.7	0.9	1.5	0.7	2.0	1.2	0.4	0.8	0.9	0.2

Figure 4.2: SVA density across chromosomes is correlated with their gene density. A – The SVA density is positively correlated with the gene density across human chromosomes. The correlation coefficient (0.74) was calculated using bootstrap confidence interval of 95% to remove outliers. B – This table shows the raw data used to create figure 4.2A to highlight which chromosomes are contributing to each point on the graph. C – The table shows the correlation coefficient of each SVA subtype density with the gene density of each chromosome. All subtypes show a positive correlation to varying degrees except SVA As, which have a negative correlation with gene density across chromosomes.

4.4.2 SVAs show a reduced presence in gene deserts with deviation in the subtype distribution from the expected in these regions compared to genic regions

Figures 4.1 and 4.2 showed a correlation of SVA insertions with gene dense regions of the genome. To dissect the distribution of SVAs further, the genome was divided into the three following regions: genes, intergenic and gene deserts defined in 4.3.1 and the difference in the expected and the observed distributions was calculated. The distribution of three other classes of retrotransposon (LTRs, LINEs and SINEs) were also analysed for comparison to that of the SVAs. The distribution of the different classes of retrotransposons shared some similarities, in particular a lower number than expected were found in gene deserts and all classes showed a significant difference in their actual distribution to the expected across the three regions analysed, Figure 4.3A (SVAs X^2=339.5, df=2, P<0.001, SINEs X^2=170647, df=2, P<0.001, LINEs X^2=44320, df=2, P<0.001, LTRs X^2=77018, df=2, P<0.001).

The distribution of SVAs was further analysed by subtype within the previously defined regions: genes, intergenic and gene deserts, Figure 4.3B. The SVA subtypes showed a significant difference in their distribution within gene deserts compared to the whole genome (Gene deserts X^2=13.91, df=6, P<0.05) but not within genes and intergenic regions (Genes X^2=0.71, df=6, P=0.99, Intergenic X^2=0.47, df=6, P=0.99). Subtypes D, E and F1 were underrepresented in gene deserts whereas subtype B in particular was found in higher numbers.

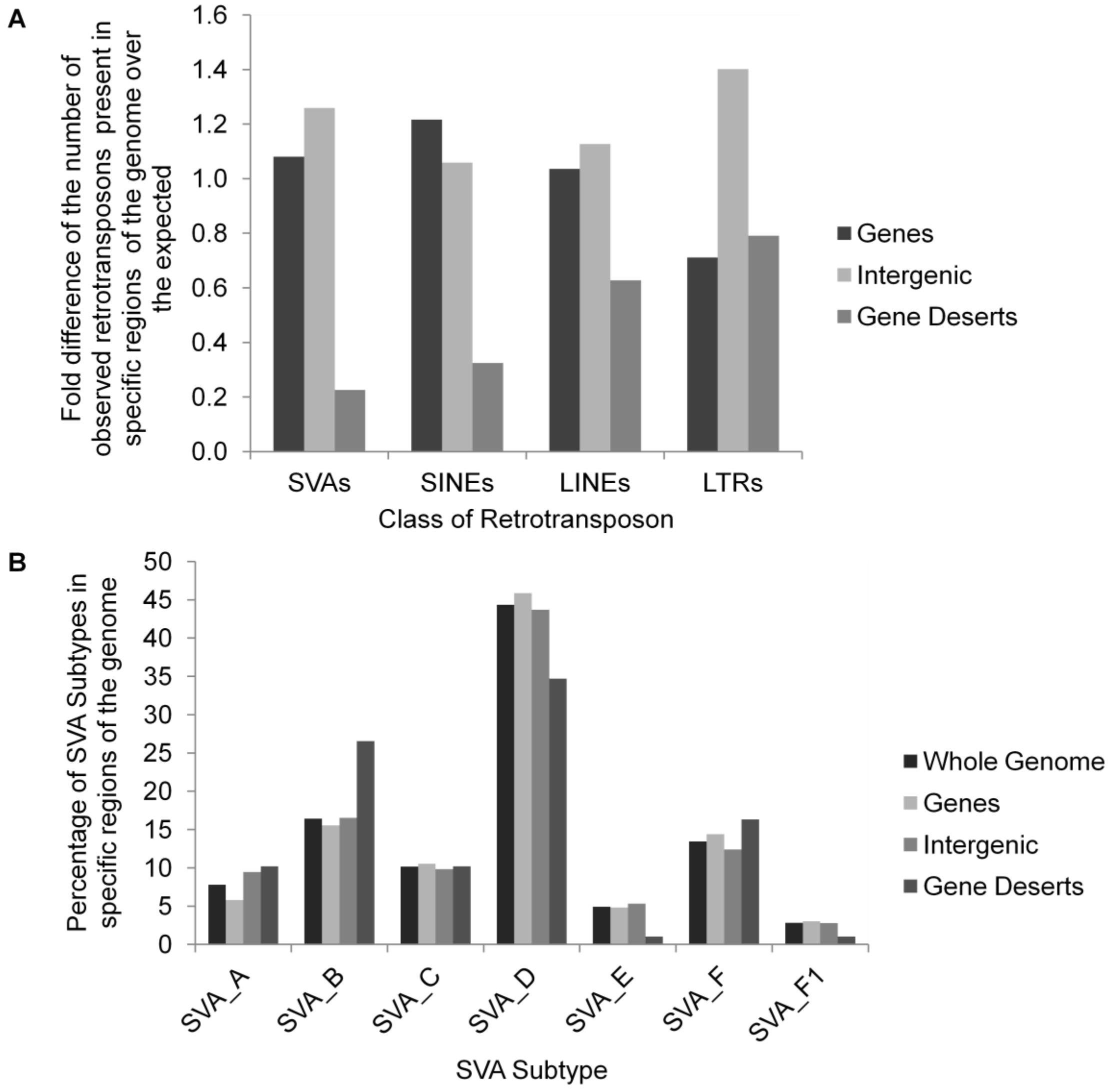


Figure 4.3: There are fewer than expected SVAs within gene deserts compared to genic regions. A – The number of observed retrotransposons in defined regions of the genome compared to the expected due to the size of the regions. Each class of retrotransposons are found in lower numbers in gene deserts most notably SVAs. For each class of retrotransposon the actual distribution was significantly different to the expected when analysed with the Chi squared test (Chi squared test carried out using raw data). (SVAs X^2=339.5, df=2, $P<0.001$), (SINEs X^2=170647, df=2, $P<0.001$), (LINEs X^2=44320, df=2, $P<0.001$), (LTRs X^2=77018, df=2, $P<0.001$). B – The distribution of SVAs within genes, intergenic regions and gene deserts was broken down by subtype and compared to their distribution across the whole genome. The subtype distribution in gene deserts showed a significant difference with B and F being more prevalent and D, E and F1s being underrepresented. (Genes X^2=0.71, df=6, P=0.99), (Intergenic X^2=0.47, df=6, P=0.99), (Gene deserts X^2=13.91, df=6, $P<0.05$).

4.4.3 The retrotransposons, SVAs and SINEs, show an increased presence in regions directly upstream and downstream of genes

The analysis of the presence of retrotransposons within regions directly upstream of the TSS of genes (within 20kb) was undertaken to assess the potential of these elements to influence gene expression. Retrotransposons contain characteristics such as internal promoters, CG rich sequences therefore potential sites of methylation and repetitive regions providing multiple sequence specific transcription binding sites. These are properties that could influence the loci of the site of their insertion and therefore may be selected against in areas such as promoters where they could interfere with the regulation of gene expression. The number of expected elements of the retrotransposon classes (SVAs, SINEs, LINEs and LTRs) in defined regions upstream of the start of the 5'UTR/TSS of genes (1kb, 10kb and 20kb) were compared to the observed (Figure 4.4A). SVAs and SINEs were significantly overrepresented throughout the 20 kilobases upstream of TSS of genes (SVAs X^2=221.9, df=2, P<0.001 and SINEs X^2=138825, df=2, P<0.001). LINEs and LTRs also show a significant difference in their distribution with fewer present in the first kilobase upstream of the TSS than expected but with less deviation in the 10 and 20 kilobase regions (LINEs X^2=2343, df=2, P<0.001 and LTRs X^2=2057, df=2, P<0.001). The same analysis was completed for defined regions (1kb, 10kb and 20kb) downstream of the 3'UTR of genes for comparison to the upstream data (Figure 4.4B). The SVAs and SINEs were also overrepresented throughout the 20kb downstream of genes and to a greater extent in the first kilobase of that region (SVAs X^2=227.6, df=2, P<0.001 and SINEs X^2=89097, df=2, P<0.001). The LINEs and LTRs showed a more similar distribution compared to the expected across the 20 kilobases downstream of genes, however it was still significantly different (LINEs X^2=695.2, df=2, P<0.001 and LTRs X^2=741.0, df=2, P<0.001).

The SVA distribution in these two regions was analysed further to determine if the percentage of each subtype present differed to that across the whole genome (Figure 4.5). The subtype distribution was not significantly different except within the first kilobase upstream of the start of transcription; subtypes A, B and E were found in lower numbers than expected and there were a greater number of subtypes C and D.

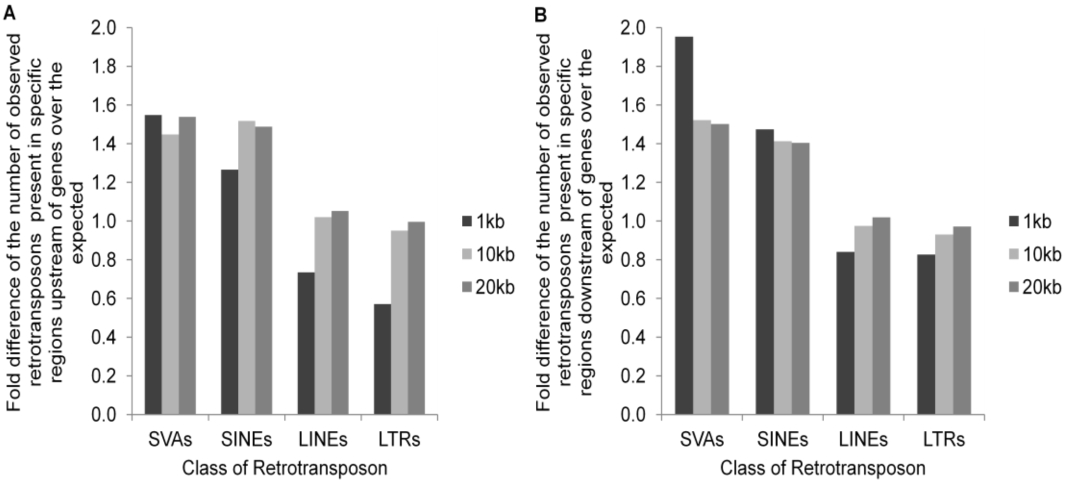


Figure 4.4: The number of SVAs and SINEs is greater than expected in regions upstream and downstream of genes. A- The fold difference of the number of retrotransposons within 1kb, 10kb and 20kb of the 5’UTR of genes compared to the expected. The distribution of the expected of the four classes of retrotransposon were all significantly different from the observed using the Chi squared test (Chi squared test carried out using raw data). (SVAs X^2=221.9, df=2, P<0.001), (SINEs X^2=138825, df=2, P<0.001), (LINEs X^2=2343, df=2, P<0.001), (LTRs X^2=2057, df=2, P<0.001). B- The fold difference of the number of retrotransposons within 1kb, 10kb and 20kb of the 3’UTR of genes compared to the expected. The distribution of the expected of the four classes of retrotransposon were all significantly different from the observed using the Chi squared test (Chi squared test carried out using raw data). (SVAs X^2=227.6, df=2, P<0.001), (SINEs X^2=89097, df=2, P<0.001), (LINEs X^2=695.2, df=2, P<0.001), (LTRs X^2=741.0, df=2, P<0.001)

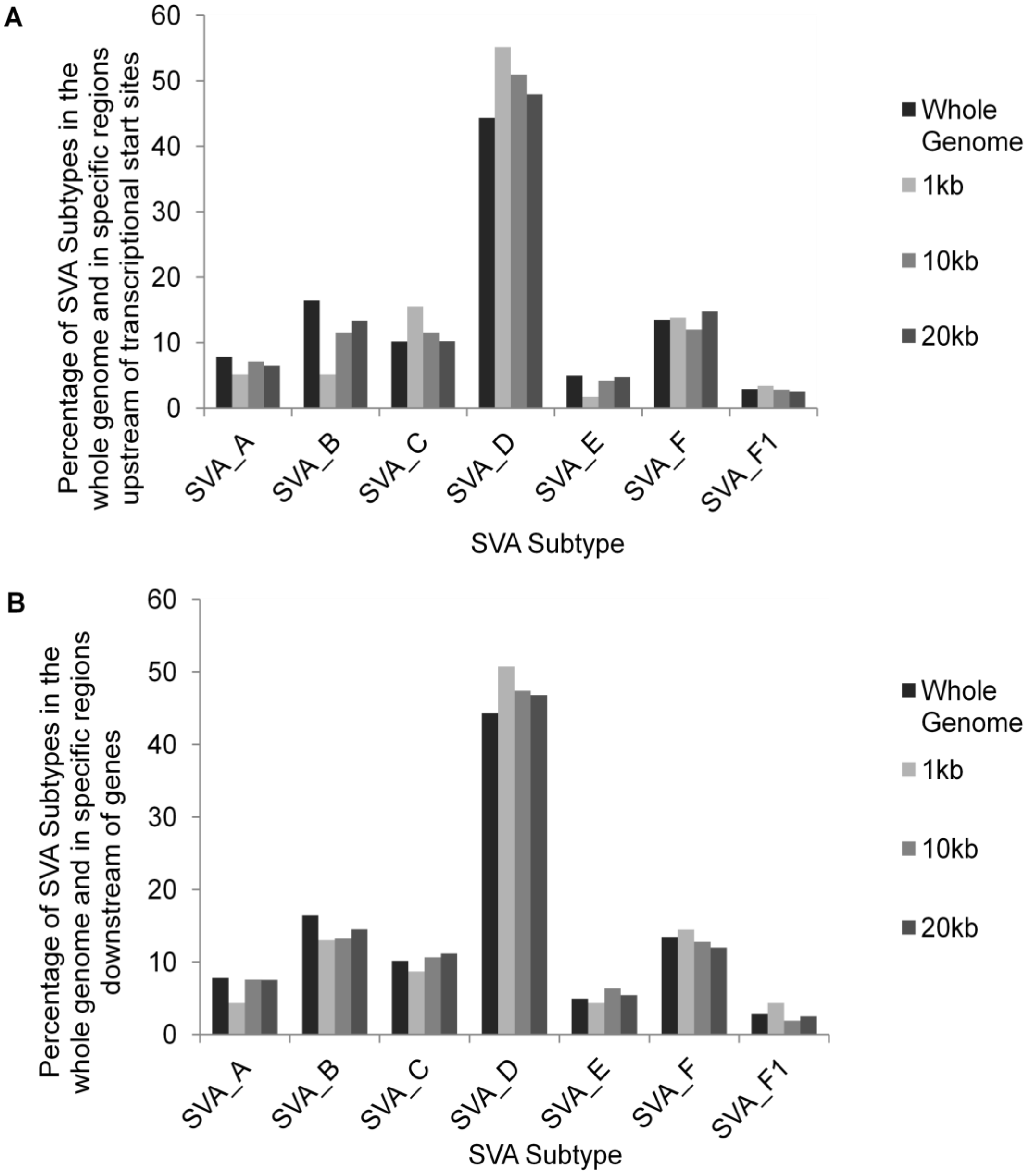


Figure 4.5: The distribution of SVA subtypes within the regions upstream and downstream of the whole gene. A- The distribution of the SVA subtypes across the whole genome compared to the distribution in 1kb, 10kb and 20kb directly upstream of the 5'UTR of genes. There is a significant difference in the distribution of the subtypes within the first kilobase compared to the whole genome. (1kb X^2=16.30, df=6, p<0.05), (10kb X^2=2.97, df=6, p=0.81), (20kb X^2=1.30, df=6, p=0.97). B- The distribution of the SVA subtypes across the whole genome compared to the distribution in 1kb, 10kb and 20kb directly downstream of the 3'UTR of genes. There no significant difference in the subtype's distribution within regions downstream of genes (1kb X^2=4.31, df=6, p=0.63), (10kb X^2=1.63, df=6, p=0.95), (20kb X^2=0.71, df=6, p=0.99).

4.4.4 SVA insertions are found on the opposite strand more frequently when exonic or intronic but there is not a preference for relative orientation when inserted upstream or downstream of the gene

When analysing the distribution of SVAs across the human genome a list of genes that had a SVA within an intron, exon or their 10kb flank was generated and this was used to compare the orientation of the SVA to the gene to determine if SVAs preferentially inserted in a particular orientation relative to the nearest gene. The SVAs within the genes' 10kb flank showed close to a 50/50 relationship to whether they were on the same strand or opposite as each other; showing there was not a preference for the orientation for the SVA insertion within these regions (Table 4.1). Of the SVAs that had inserted into introns or exons only 26.1% of them were on the same strand as the gene indicating that SVAs are preferentially found on the opposite strand to the gene.

	Percentage of SVA subtypes on same strand of DNA as the nearest gene within defined parameters (%)		
Subtype	**Within 10kb upstream of a gene**	**Within 10kb downstream of a gene**	**Within an exon or intron of a gene**
A	55.9	47.2	37.5
B	43.1	42.9	18.0
C	48.3	44.1	16.8
D	39.9	43.8	27.3
E	36.8	32.4	29.7
F	52.7	46.9	24.9
F1	68.8	60.0	28.2
All	49.3	45.3	26.1

Table 4.1: Intronic and exonic SVA insertions occur more frequently on the opposite strand to the gene they have inserted into. Using information from UCSC genome browser the orientation of SVAs inserted within a gene or within its 10kb flank was compared the orientation of the gene itself and the percentage of SVAs on the same strand as the gene was calculated.

4.4.5 SVAs have the potential to form G-quadruplex DNA to varying degrees between subtypes

The nature of the sequence contained within SVAs shows the potential for formation of secondary structures such as cruciforms and G4 DNA (Hancks and Kazazian 2010). G4 DNA is a secondary structure predicted to form in guanine-rich sequences and its function was discussed in section 1.3.2. To determine the exact nature of the relationship of the SVAs' sequence and G4 formation software was used to analyse the SVAs potential to form this type of secondary structure.

Of the total genomic DNA that can form G4 (predicted by Quadparser software (Wong et al. 2010) 1.88% is due to SVAs which only constitute 0.13% of the human genome. When repetitive or mobile DNA elements, which include simple repeats, microsatellites, LTRs, LINEs, SINEs and DNA transposons (as defined by UCSC genome browser Hg19) are compared; SVAs have the greatest potential contribution to G4 DNA for their size for any specific element (Figure 4.6A).

It was found that the percentage of sequence in each SVA subtype with the potential to form G4 increased as the age of the subtype decreased, thus subtypes E, F and F1 have the greatest potential for G4 formation (Figure 4.6B). This can be explained by the increase in the potential of the central VNTR region to form G4 DNA from subtype D through to F1. The possible amount of G4 formed by the CCCTCT repeat was found to increase through subtypes A to E; however the proportion it contributed to the total G4 potential of each subtype decreased. Subtype F1 does not contain a CCCTCT repeat therefore all of its G4 potential is within the central VNTR.

The average number of repeats in the CCCTCT domain and therefore its length varied between subtypes (Figure 4.6C) which accounts for the difference in G4 potential between the SVA subtypes in that particular domain; the longer the CCCTCT domain the greater the G4 potential. The average length of the GC-rich VNTRs also varied between subtypes but length did not show the same direct correlation with G4 potential as in the CCCTCT domain. For example the VNTRs of subtype A are just under half the length of those of subtype F1, however they have only a hundredth of the potential to form G4 DNA when compared to the VNTR sequences of subtype F1 (Figure 4.6D). It appears that the subtypes fall into two main groups when analysing their G4 potential in the VNTRs. Subtypes A, B and C have very low G4 potential in their VNTRs compared to subtypes E, F and F1 with subtype D bridging the difference between the older hominid specific and younger human specific subtypes. This would be explained by the development of the additional second VNTR of the younger subtypes with differences in the primary nucleotide content to the first VNTR containing sequences that have the potential for G4 DNA.

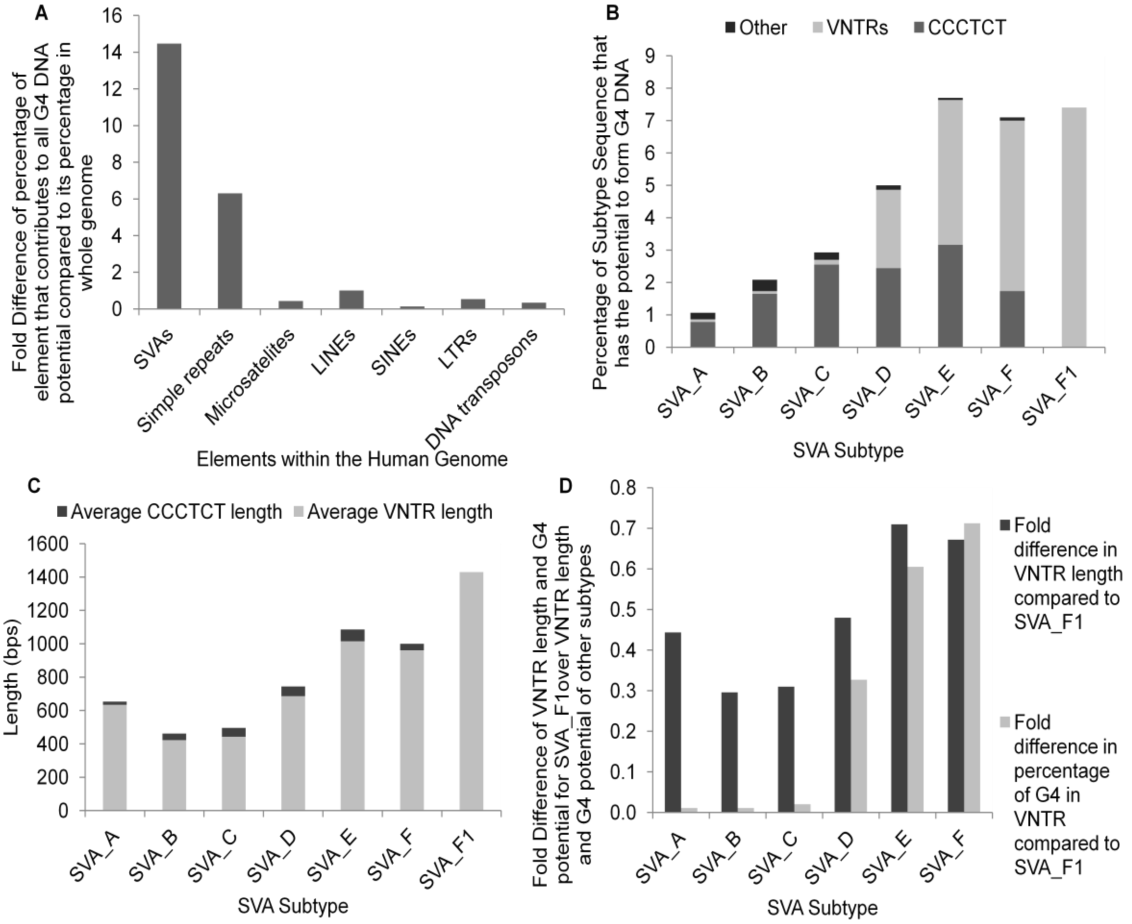


Figure 4.6: The sequences of SVAs have potential to form G-quadruplex DNA.

A – Potential G4 DNA formation was analysed *in silico.* The fold difference in the relative contribution of each element to their proportion in the whole human genome was calculated and is displayed.

B - The percentage of sequence from each SVA subtype that could potentially form G4 DNA in the human genome according to Quadparser software is shown; it was further sub-divided into the following elements: CCCTCT hexamer repeat, VNTRs and the remainder of the sequence (other).

C – Illustrates the relationship between VNTR and hexamer repeat length during evolution of the SVA subtypes. The average lengths are shown in base pairs.

D – The fold difference in size of each of the central VNTRs from the SVA subtypes in the human genome, and their percentage contribution to form G4 compared to the value for SVA subtype F1 which has the highest value for both central VNTR length and G4 potential of the central VNTR.

4.4.6 SVAs and their VNTRs show some characteristics of CpG islands

The definition of a CpG island was taken from the UCSC genome browser and the literature as the following: greater than 200bp in length, more than 50% GC content and a ratio of observed to expected CG as greater than 0.6. The potential of the GC-rich sequences of the SVAs and in particular their central VNTR to be CpG islands were calculated for 3 of each SVA subtype. The SVAs and their VNTRs of all the subtypes met two of the criteria as were larger than 200bp in length and had a GC content greater than 50%. SVAs have a high GC content, in this analysis an average of 64.4% for the whole SVA and 74.0% for the central VNTR (Table 4.2A). The GC content of the SVAs increased as the as class of the subtype decreased. The SVAs analysed did not have a ratio of observed to expected CG dinucleotides greater than 0.6 with an average ratio of 0.53 for the whole SVA and 0.54 for the VNTR alone (Table 4.2A). The CpG ratio and the GC content of three CpG islands located at genes of interest are shown in table 4.2B for comparison to the values for the SVAs and their VNTRs. The GC content of the CpG islands is similar to that of the SVAs and lower compared to that of the VNTRs however the CpG ratio of the SVAs and their VNTRs is lower than that of the CpG islands.

A

	SVA		VNTRS	
Subtype	**CpG Ratio**	**GC Content (%)**	**CpG Ratio**	**GC Content (%)**
A	0.48	60.2	0.51	69.9
B	0.54	60.6	0.54	71.9
C	0.50	61.3	0.54	73.5
D	0.54	65.1	0.55	74.4
E	0.54	67.3	0.54	75.8
F	0.54	66.6	0.54	75.6
F1	0.58	69.4	0.55	77.0
Average	0.53	64.4	0.54	74.0

B

CpG Island	CpG Ratio	GC Content (%)
FUS gene	0.94	66.0
Major TSS of PARK7 gene	0.92	65.9
Minor TSS of PAR7 gene	0.87	65.8

Table 4.2: CpG island ratio and the GC content of a small number of the different SVA subtypes and their central VNTRs. A - The CpG ratio of the SVAs and their VNTRs using the following formula Obs/Exp CpG = Number of CpG * length of sequence / (Number of C * Number of G) was calculated for 3 of each subtype and then averaged. For a region to be considered a CpG island the ratio of observed CG nucleotides to the expected needs to be greater than 0.6. The sequences of the SVAs and their VNTRs have ratios close to but lower than the required ratio of 0.6. This ratio is similar between the SVA and their VNTRs alone and across the subtypes with subtypes A and F1 showing the greatest difference to the rest. The GC content of the VNTRs was greater than that of the whole SVA and for both the whole SVA and its central VNTR the GC content showed a general increase from oldest to youngest across the subtypes. B – The CpG ratios and the GC content of three CpG islands located at the TSS of two genes of interest in this project (see chapter 3 section 3.4.1 for FUS and see chapter 5 section 5.4.1 for PARK7) for comparison to the values for the SVA values.

4.5 Discussion

SVA density was correlated with gene density across chromosomes (Figures 4.1 and 4.2) and displayed a preferential site of insertion into genic regions as opposed to gene deserts (Figure 4.3), which may reflect the more accessible and open nature of the chromatin to allow for transcription and therefore more amenable to retrotransposon insertions than inactive chromatin. SVAs showed a greater than expected presence in the 20kb flanking genes similar to that of the SINE elements (Figure 5.4) again supporting the hypothesis these elements insert more frequently into active chromatin. Further evidence for the insertion of TEs into active chromatin was demonstrated when a preference of Alu elements and to a lesser extent SVAs inserting into the region upstream of genes active in the germ line (Warnefors et al. 2010). Similar to the data generated in this chapter Alus (majority of SINEs in humans) have been shown to be located at higher than expected numbers within genes and their flanking regions whereas LINEs and LTRs are reduced in the flanking regions of genes (Medstrand et al. 2002). A recent publication of the distribution of TEs in the mouse genome has also shown that SINEs are enriched within genes and their flanking regions with LINEs and ERVs depleted in functional regions (Nellaker et al. 2012). SVAs show a similar distribution to that of SINEs reported in the literature.

The presence of SVAs in regions 1-20kb upstream of TSSs (Figure 5.4A) demonstrates that these elements are located in promoter regions and have the potential to influence gene expression. SVAs share similar primary sequences even across subtypes; which provides the potential for binding similar sequence specific binding factors that could affect aspects of transcription. Depending on the chromatin structure or access to these elements the end result could be subsets of SVAs which respond to similar cellular signalling pathways. SVAs were also found to be overrepresented in the region up to 20kb downstream of genes. This was noted in particular in the first kilobase downstream which showed greater numbers of SVAs present than in the first kilobase upstream of genes. This may due to fewer selective pressures in the region directly downstream of genes compared to upstream, allowing for the greater accumulation of SVA elements.

The SVAs that had inserted into the 10kb flank of genes did not show a preference for the orientation of their insertion relative to the gene. However the orientation of the SVAs when inserted into an intron or exon showed a strong preference to inserting on the opposite strand to the gene. Only 26.1% of SVAs have inserted into the same strand as the gene (Table 4.1) indicating selective pressures against this occurring. This may be due to the poly A tail that is part of a canonical SVAs structure as if on the same strand as the gene may cause pausing or even termination of transcription resulting in truncated mRNA. Also several splice sites have been identified in the sense strand of the SVA sequence but not the antisense with 6 out of the 7 intronic or exonic disease causing SVA insertions on the sense strand further indicating the potential negative effects intronic/exonic sense strand insertions (Kaer and Speek 2013). Alus also contain consensus splice sites within their sequence on both their sense and antisense strands and may introduce new exons into existing genes and are an important source of new exons in the primate genome (Sorek 2009). A study into the expression of transcripts containing 330 Alu-derived exons across 11 adult tissues revealed that most of these new exons are expressed in transcripts at low levels, however in a small number of cases the Alu-derived exons were constitutively spliced and in others the inclusion of the Alu-derived exon was tissue specific (Lin et al. 2008). The splice sites present in the SVA sequences could potentially be involved in evolutionary process similar to Alus in generating alternative transcripts.

Primary DNA sequence which contains stretches of guanine nucleotides can fold into four-stranded structures called G4 DNA, which are implicated in gene expression, replication and telomere maintenance (Huppert and Balasubramanian 2005; Oganesian and Bryan 2007; Gonzalez and Hurley 2010), chapter 1 Figure 1.5. Also the presence of G4 sequences along with abnormal hypomethylation was shown to be enriched in breakpoints mapped in cancer genomes, leading to the hypothesis that loss of methylation in regions with G4 sequences is part of the mutagenic processes in cancer (De and Michor 2011). SVAs contain sequences with G4 potential, specifically in their CCCTCT hexamer and central VNTR (Figure 5.8), therefore they could show similar properties to already characterised functions of G4 DNA discussed previously (chapter 1 section 1.3.2). Along with the regulatory properties of G4 DNA the hypothesised mutagenic properties of G4 sequences in demethylated regions in cancer (De and Michor 2011) is of interest as it has been demonstrated that SVAs experience a loss of methylation in cancer (Szpakowski et al. 2009). Therefore the loss of methylation of these elements in cancer may allow formation of G4 DNA which could play a role in gene regulation or could add to the mutagenic process in the tumour. The occurrence of somatic retrotransposition in the aging brain indicates a reduction in the silencing of these elements and could also allow G4 formation along with activating other transcriptional properties within the SVAs. The amount of G4 potential and the domain of the SVA it was predominantly located in varied across the different subtypes. The older subtypes (A, B and C) had the lowest potential; which was mostly located within the 5′ CCCTCT repeat, whereas the younger human specific (E, F and F1) demonstrated the greatest potential for G4 with an increase in the amount located in the central VNTR. Subtype D showed itself to be an intermediate of the two groups (Figure 4.6).

The CG-rich nature of the primary sequence of the SVAs provides potential regions for methylation and with many SVAs located near the transcriptional start site of genes therefore the methylation status of these elements could influence the expression of genes. A small number of SVAs from each subtype were analysed for characteristics of CpG islands. The SVAs reached two out of three of the thresholds in the definition of a CpG island and were close to the third requirement. The SVAs analysed may not meet all the requirements to be defined as a CpG island but SVAs still contain many CG dinucleotides that if methylated could potentially reduce the expression of transcripts originating near them.

The analysis in this chapter of the distribution of SVAs within the human genome showed a preference for genic regions over gene deserts and an overrepresentation of these elements in regions directly upstream of TSSs placing them in location where they may influence gene expression. This may be through the binding of transcription factors, their methylation status or even the formation of secondary structures such as G4 DNA.

Chapter 5

Characterisation of the PARK7/DJ-1 gene loci and the structure and function of a SVA upstream

5.1 Introduction

A SVA within 10kb upstream of the Parkinson protein 7 (PARK7) gene, a similar upstream locus to that found for the FUS SVA, was identified during the global analysis of SVAs in the human genome. This SVA was of particular interest as it was upstream of another gene involved in a neurodegenerative disease and was both a human specific element and contained all the domains of a canonical SVA. This SVA shared similar features to the FUS SVA, for example they are both of subtype D, but also differences that may impact on the function of these elements, for example the PARK7 SVA contains a CCCTCT hexamer repeat at its 5' end whereas the FUS SVA does not.

The PARK7 gene is located on chromosome 1p36.23 and encodes for a 189 amino acid multifunctional protein. PARK7 is also known as DJ-1 but will be referred to as PARK7 within this thesis. PARK7 was first identified as a protein that in conjugation with ras transformed the mouse NIH3T3 cell line, suggesting its oncogenic potential (Nagakubo et al. 1997). The PARK7 protein was found to be ubiquitously expressed across human tissues and found both in the nuclei and cytoplasm within the HeLa cell line (Nagakubo et al. 1997). The rat homologue of PARK7 (CAP1/sP22) is highly conserved (91% homology) and was identified as playing a role in male fertility (Wagenfeld et al. 1998a; Wagenfeld et al. 1998b). PARK7 acts as a positive regulator of the androgen receptor (AR) by binding to the PIASxα (protein inhibitor of activated STAT) preventing it from binding to the AR (Takahashi et al. 2001). Further regulatory properties of PARK7 involve the human specific regulation of the tyrosine hydroxylase (TH) gene, an enzyme involved in dopamine synthesis, by binding to a repressor of the TH preventing it from binding to the TH gene promoter and inhibiting expression (Ishikawa et al. 2010). PARK7 has also been identified as a negative regulator of the tumour suppressor PTEN's function (Kim et al. 2005). Another of PARK7's key functions is to protect against oxidative stress preventing cell death (Taira et al. 2004). PARK7 protects against oxidative stress and mitochondrial damage by undergoing oxidisation at a highly conserved cysteine residue (C106) causing a shift of the protein to a more acidic form (Canet-Aviles et al. 2004). Mitochondrial dysfunction and oxidative stress are mechanisms proposed to be involved in neuronal damage and ultimately cell death associated with Parkinson's disease (PD) and the reduction or loss of function of the PARK7 protein is a trigger for PD (Ariga et al. 2013). Mutations in the PARK7 gene including a 14kb deletion of a large coding portion of the gene and a missense mutation (L166P) affecting the function of the protein are a cause of autosomal recessive early onset PD (Bonifati et al. 2003). PARK7 has also been shown to be associated with cancer including breast and non-small cell lung carcinoma (Le Naour et al. 2001; MacKeigan et al. 2003). PARK7 and its role in disease processes are discussed in further detail in chapter 6.

Detailed analysis of the PARK7 gene locus and the SVA upstream was undertaken and was used to attempt to determine the impact of the SVA on the site of insertion and how this may affect gene expression.

5.2 Aims

- Analyse the PARK7 gene loci and the expression of PARK7 transcripts in two different cell lines; a human neuroblastoma cell line (SK-N-AS) and human adenocarcinoma cell line (MCF-7) to reflect the range of PARK7 activities.
- Analyse the epigenetic modifications and chromatin structure of the PARK7 locus including the two promoters.
- Determine the polymorphic nature of the PARK7 SVA using the CEU HapMap cohort.
- Identify the potential regulatory domains of the PARK7 SVA and differential function of its four alleles using reporter gene constructs.

5.3 Methods

5.3.1 Bioinformatic Analysis of the PARK7 gene loci

The PARK7 loci was analysed using the genome browser UCSC Hg19 (http://genome.ucsc.edu/index.html) and Archive Ensembl 10:Jan 2013 (http://www.ensembl.org/info/website/archives/index.html) was used to characterise the transcripts of the PARK7 gene.

5.3.2 Analysis of expression of PARK7 transcripts

Section 2.2.3 outlines the methods for analysing endogenous gene expression in SK-N-AS and MCF-7 cell lines and Table A1 contains PCR conditions for the specific primer sets used to analyse the expression of the different PARK7 transcripts. The fragments corresponding to the PARK7 transcripts on the gel were extracted (2.2.4.4) and sent for sequencing (2.2.4.10) with the primers that were used in the amplification of the transcripts.

5.3.3 Analysing the chromatin structure at the PARK7 loci using chromatin immunoprecipitation

Chromatin immunoprecipitation was used to analyse the epigenetic parameters and transcription factor binding at specific loci of the PARK7 gene and SVA in the SK-N-AS and MCF-7 cell lines (2.2.9). For details of the antibodies used see Table A2 in the appendix. The regions of interest which were the major and minor promoters of the PARK7 gene and 5’ of the PARK7 SVA were amplified using Go Taq Flexi polymerase (2.2.3.3.2) under standard conditions and primers listed in Table A1 of the appendix. A negative control primer set for a gene desert was used to determine if the binding observed was specific (Table A1).

5.3.4 Characterisation of the PARK7 gene promoter loci in gDNA of cell lines and breast tumour samples

The two PARK7 gene promoter regions (major and minor) and several other loci 5' and 3' of the major PARK7 TSS were amplified using GoTaq Flexi polymerase (2.2.3.3.2) with the addition of 1M betaine final concentration. gDNA was extracted from the cell lines using QIAamp DNA mini kit (2.2.7) and 10ng was used as template in the reaction. The gDNA from breast cancer tumour samples were obtained from the Liverpool Cancer Tissue Bank and 5ng of gDNA was used in the reaction. The PCR conditions and primers used are listed in Table A1 of the appendix. The PCR products were analysed using gel electrophoresis on a 1.2% agarose gel (2.2.3.4).

5.3.5 Genotyping the PARK7 SVA

The PARK7 SVA was amplified using KOD Hot Start DNA Polymerase (Novagen) under standard conditions (see 2.2.4.2 for components of KOD Hot Start master mix) with the addition of betaine (Sigma) at 0.5M final concentration and the reaction volume was reduced and optimised to 20µl to save reagents. For primers and cycling condition see Table A1 in the appendix. 1ng of gDNA from the CEU HapMap cohort was used as template in the reaction. The PCR products were run on a 1% agarose gel (2.2.3.4) for up to four hours for adequate separation of the alleles due to the large size of the SVA. The allele frequency data was used to calculate the expected genotype frequencies which were compared to the observed to determine if the PARK7 SVA locus is in HWE.

5.3.6 Cloning four identified alleles from the HapMap into an intermediate vector and subsequent sequencing

For accurate sequencing and downstream applications the four alleles of the PARK7 SVA were cloned into the intermediate vector pCR-Blunt (2.2.5.1). The plasmids containing each allele were sent for sequencing (2.2.4.10) with two primers in the vector (M13 reverse and M13 forward) and two within the SVA itself (5'CTCAGTGCTCAATGGTGCC 3' and 5'CCGCCTTTCTATTCCACAAA 3') to ensure the whole of the insert was sequenced. The plasmids generated were used as template in amplification of the individual repetitive regions (hexamer VNTR and TR/VNTR) to demonstrate the differences in size using KOD Hot Start polymerase (2.2.4.2), for primers and conditions used see Table A1 in the appendix.

5.3.7 Generation of PARK7 SVA reporter gene constructs

The different sized fragments in forward and reverse orientation and the four alleles of the PARK7 SVA were cloned into the reporter gene vector pGL3P (2.2.5.2). The PARK7 SVA was also cloned in to pGL3B to test its ability to act as a promoter (2.2.5.2).

5.3.8 Cell culture of SK-N-AS and MCF-7 cell lines, transfection and dual luciferase assay of reporter gene constructs

SK-N-AS and MCF-7 cells were cultured as outlined in 2.2.2.1. For transfection methods used and analysis of reporter gene expression methods see 2.2.6.

5.4 Results

5.4.1 Chromosomal Loci of the PARK7 gene and the SVA upstream of its transcriptional start site

The PARK7 gene is located on chromosome 1p36.23 and the locus of this gene was analysed using genome browsers UCSC Hg19 and Archive ensembl 10:Jan2013. The TSS of the gene was identified in UCSC by the origin of the transcripts of the gene, a CpG island and ENCODE data such as active histones marks and DNase 1 hypersensitivity clusters which corresponded to the locus of the promoter characterised by Taira et al (Taira et al. 2001). A region 7kb upstream of the TSS of the PARK7 gene also shared many of the features of the well characterised TSS for the PARK7 gene such as a CpG island, DNase 1 hypersensitivity clusters and active histone marks however there was not a known PARK7 transcript originating at that locus in UCSC genome browser. Intragenic CpG islands have previously led to the identification of previously unknown TSSs and promoters (Gardiner-Garden and Frommer 1994; Macleod et al. 1998; Kleinjan et al. 2004) and are often associated with the 5' end of genes. The CpG alone therefore indicates the potential for transcription occurring at this locus. There were also expressed sequence tags (ESTs) expressed within the region of the second CpG island and the Affymetrix Human Exon Array extended probe sets predicted exons at this site. Data from the Archive ensembl database predicted a transcript originating from this second CpG island (Figure A1 of the appendix). The ENCODE data of this locus and the predicted transcript led me to this region being putatively termed the minor TSS of the PARK7 gene and the already characterised promoter termed the major TSS (Figure 5.1). A human specific SVA D (classified by the interrupted repeat masker track on UCSC genome browser) was identified 8kb upstream of the major TSS and 1kb upstream of the minor TSS in UCSC genome browser (Figure 5.1).

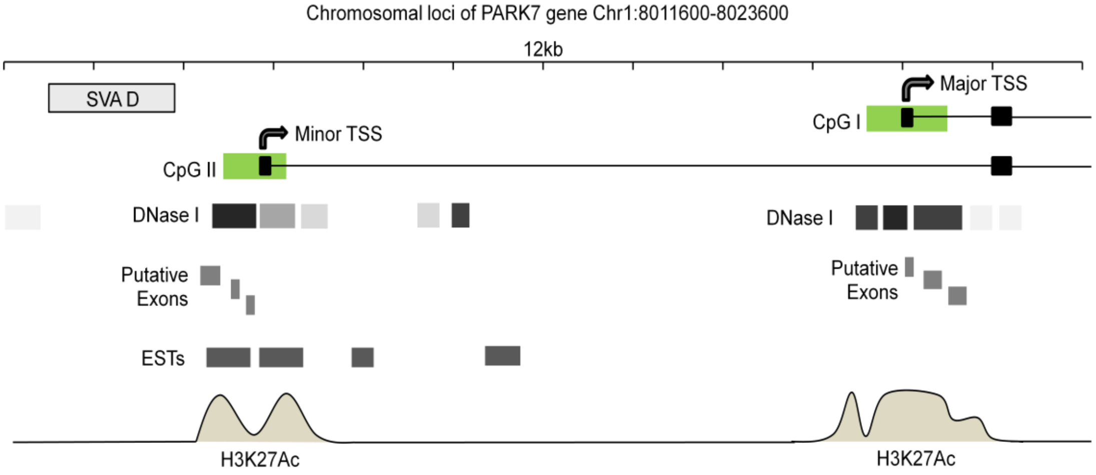


Figure 5.1: Schematic of chromosomal loci of PARK7 gene. Data was compiled from UCSC genome browser (Hg19) and Archive Ensembl (Ensembl 10:Jan 2013) (see Figure A1 of the appendix for data screenshots from UCSC). The SVA D is located 8kb upstream of the major transcriptional start site (TSS) of the PARK7 gene and 1kb upstream of the minor TSS. There are two CpG islands present with UCSC browser showing ENCODE data for H3K27Ac histone mark often found near active regulatory elements (7 cell lines), DNase hypersensitivity clusters and transcription factor binding across both of these regions indicating transcriptionally active areas. The majority of PARK7 transcripts begin at the major TSS but Archive Ensembl browser indicates there is a transcript at the minor TSS adjacent to the SVA D. UCSC Affymetrix Human Exon Array extended probe sets show putative exons at this second TSS and there are also expressed sequenced tags (ESTs) at the site.

5.4.2 PARK7 transcripts are differentially expressed in SK-N-AS and MCF-7 cell lines

The expression of the PARK7 gene mRNA was analysed under basal growth conditions in a SK-N-AS cell line and two MCF-7 cell lines supplied by different sources and of different passage number. Primers sets designed to amplify the transcripts originating at the major TSS and the predicted transcript originating at the minor TSS were used in this analysis (Figure 5.2A). The primers of PARK7 primer set 1 were located in exon 2 and 6 all of the known PARK7 transcripts and for primer set 2 the primers were within exon 1 and 2 of the transcripts originating at the major TSS with two specific products due to two exon 1s of different lengths. The third primer set was located within exon 1 and 2 of the predicted transcript in the Archive Ensembl database originating at the minor TSS.

PCR products from all three of these primers sets were identified in the SK-N-AS cell line (Figure 5.2B). These products numbered 1 to 5 were gel extracted and sent for sequencing. Products 1 to 3 corresponded correctly to the sequence of the PARK7 gene in the UCSC genome browser. From primer set 3 there was one product of 108bps expected from the predicted transcript however there were two products from the PCR. The sequence of PCR product 4 corresponded to the known exon 2 of the PARK7 gene and to sequence of exon 1 of the predicted transcript in Archive Ensemble originating at the minor TSS. However there was 8bps less within the sequenced exon 1 compared to the predicted. The sequencing of PCR product 5 showed some similarities with the PARK7 exon 2 sequence but the origin of this DNA could not be identified.

The PCR analysis of the lower passage MCF-7 cells (P.13) detected the expression of the PARK7 transcripts originating at the major promoter of the gene but not the transcript originating at the minor (Figure 5.2C) indicating the expression of the latter transcript is tissue specific. The PCR analysis of the expression of the PARK7 gene in the higher passage MCF-7 cells (P.32) could not detect the expression any of the transcripts of the gene (Figure 5.2B). The housekeeping gene β-tubulin was consistently expressed across several cDNA samples from the high passage MCF-7 cells indicating the RNA extraction and cDNA conversion were successful, however the known oestrogen responsive gene trefoil factor 1 (TFF1) in MCF-7 cells (see section 6.4.3) could not be detected. The extraction of total RNA from these cells and subsequent conversion to cDNA was carried out multiple times to ensure this was a not a problem associated with the processing of the sample but the results remained the same. The inconsistencies of this particular passage of cells led to its discontinuation for further experiments. See section 5.4.5 for the analysis of the gDNA from the high passage MCF-7 cell line in an attempt to determine the mechanism underlying the lack of PARK7 expression in these cells.

Our analysis in these two cell lines is consistent with ENCODE data for this locus available on the UCSC genome browser showing that the chromatin structure in 9 different human derived cells lines revealed a greater variation at the minor promoter of PARK7 compared to the major across the different cell lines (Figure 5.3). This information is based in ChIP data using nine factors targeting a variety of histone modifications that are associated with different chromatin states. The chromatin structure indicates the major promoter is active in all nine of the cell lines analysed whereas the minor is active in some of the cell lines (GM12878, K562, HepG2, NHEK and NHLF), weakly active in two (HMEC and HSMM) and inactive in another cell line (H1-hESC). There are also ESTs located in the region of the minor promoter of the PARK7 gene and this indicated transcriptional activity may also be in part related to the expression of these ESTs not just the PARK7 transcript. The region of the minor PARK7 promoter is showing tissue specificity across the 9 cell lines tested whereas the major is not.

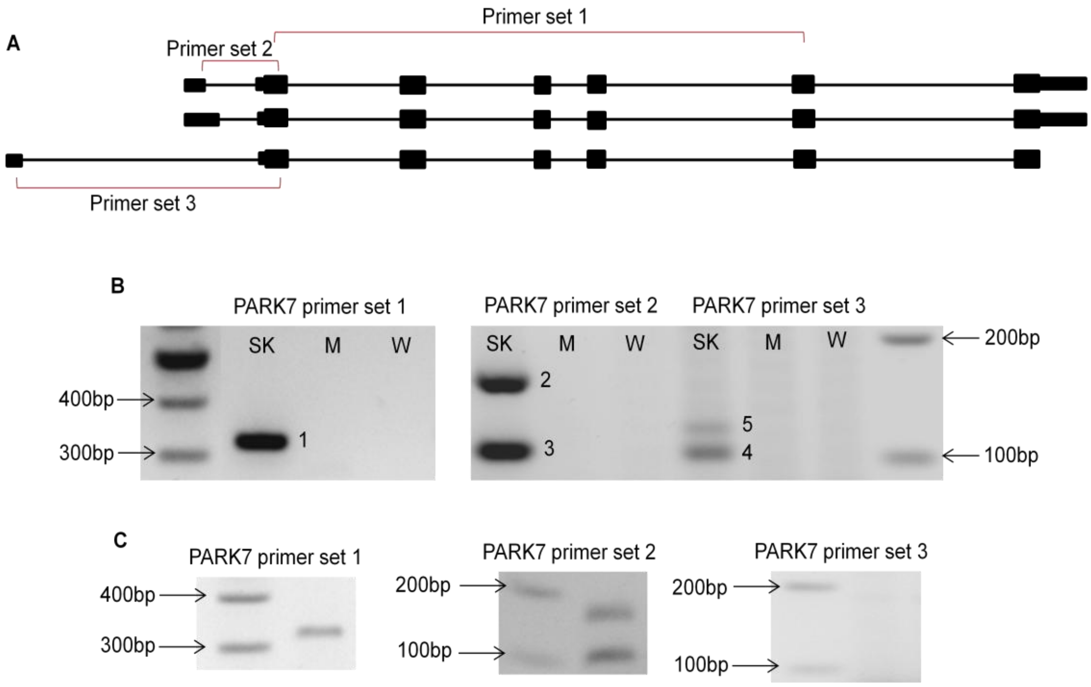


Figure 5.2: Expression of PARK7 transcripts in a SK-N-AS cell line and two MCF-7 cell lines of different passage.

A – Schematic of the PARK7 transcripts and the primers used in their amplification. PARK7 primer set 1 targets exon 2 and 6 of all identified PARK7 transcripts with a product size of 325bp. PARK7 primer set 2 targets exon 1 and 2 of the transcripts originating at the major TSS and due to transcripts with different lengths of exon 1 would produce two specific products at 102bp and 159bp. PARK7 primer set 3 targets exon 1 and 2 of the transcript predicted by Archive Ensembl with a product size of 108bp.

B – In the SK-N-AS cell line PARK7 was shown to be expressed when amplified with primer set 1 however was not present in the high passage MCF-7 cell line. The PCR product indicated with a number 1 was gel extracted and sequenced. The transcripts originating at the major TSS amplified by primer set 2 and the predicted transcript originating at the minor TSS were present in the SK-N-AS cell line but not the high passage MCF-7. The PCR products labelled 2-5 were gel extracted and sequenced, however the origin of the unexpected band 5 could not be determined. S=SK-N-AS cell line, M=MCF-7 cell line (high passage), W=water control.

C– Expression of mRNA targets in lower passage number MCF-7 cell line. The PARK7 transcripts amplified by primer set 1 and 2 are expressed in a MCF-7 cell line of a lower passage, however the transcript amplified by primer set 3 is not expressed.

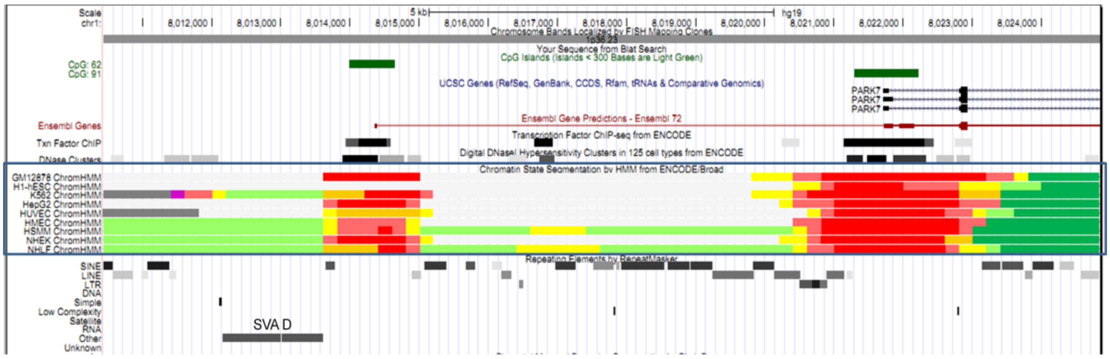

Scale
chr1:
5 kb
hg19
8,012,000
8,013,000
8,014,000
8,015,000
8,016,000
8,017,000
8,018,000
8,019,000
8,020,000
8,021,000
8,022,000
8,023,000
8,024,000
Chromosome Bands Localized by FISH Mapping Clones
1p36.23
Your Sequence from Blat Search
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 62
CpG: 91
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PARK7
PARK7
PARK7
Ensembl Genes
Ensembl Gene Predictions - Ensembl 72
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Chromatin State Segmentation by HMM from ENCODE/Broad
GM12878 ChromHMM
H1-hESC ChromHMM
K562 ChromHMM
HepG2 ChromHMM
HUVEC ChromHMM
HMEC ChromHMM
HSMM ChromHMM
NHEK ChromHMM
NHLF ChromHMM
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
SVA D

Figure 5.3: ENCODE data from UCSC browser shows greater variation in chromatin state at the minor promoter than the major promoter across 9 different cell lines. This screen shot from UCSC genome browser shows the PARK7 gene locus and upstream SVA. Boxed in blue is ENCODE data on the state of the chromatin in the region in 9 cell lines. The colours represent different chromatin states: bright red – active promoter, light red – weak promoter, orange – strong enhancer, yellow – weak/poised enhancer, dark green – transcriptional elongation, light green – weakly transcribed, dark gray - repressed and light gray - heterochromatin. The major promoter is shown as active across the 9 cell lines; however the chromatin state of the minor promoter is variable across the 9 cells lines showing tissue specificity. GM12878 - B-lymphocyte, lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasian, H1-hESC – human embryonic stem cell line, K562 – cell line derived from pleural effusion of a 53-year-old female with chronic myelogenous leukaemia in terminal blast crises, HepG2 - hepatocellular carcinoma cell line, HUVEC - umbilical vein endothelial cells, HMEC - mammary epithelial cells, HSMM - skeletal muscle myoblasts, NHEK – normal human epidermal keratinocytes, NHLF - lung fibroblasts.

5.4.3 Affymetrix probe sets show expression of regions corresponding to exons of the PARK7 gene at the major TSS and predicted exons at the minor TSS in different regions of the human brain

Transcriptional activity was hypothesised in the region 3’ of the PARK7 SVA approximately 7kb upstream of the PARK7 gene based on the ENCODE data available on the UCSC genome browser (Figure 5.1 and Figure A1). This in part was confirmed by the detection of a transcript in the SK-N-AS cell line originating in that region termed the minor promoter/TSS of the PARK7 gene (Figure 5.2B). Affymetrix probes were present in the region of the minor TSS and a collaboration was arranged with Dr Mina Ryten at University College London to address expression from this locus in the human CNS using global Affymetrix data she had generated on normal brain expression. Probes from the Affymetrix human exon array core probe set corresponding to the known exons of the PARK7 gene using the major TSS were used to detect the expression of these exons in different regions of the human brain (Table 5.1). The levels detected were significant enough to confirm the expression of the PARK7 exons in all of the brain regions tested. Probes were also used from the Affymetrix human exon array extended probe sets to look for expression of other exons in the region of the minor TSS of the PARK7 gene. The location of these three probes used are shown in Figure 5.1 called putative exons and are downstream of the PARK7 SVA and overlapping the second CpG island of the gene locus. Probes 2318734 and 2318735 demonstrated significant levels of expression in the regions of the brain tested (Table 5.2). This shows that there is transcriptional activity close to the PARK7 SVA within the human brain and that the area is not transcriptionally silenced.

This successful use of the Affymetrix human exon array extended probe sets to detect transcription activity in close proximity to a SVA, which would most likely based on dogma for the action of SVAs be considered a region of inactivity due to epigenetic modifications to suppress the retrotransposition of the SVA, led to the analysis of other probe sets located in close proximity to SVAs. The list of probes from the extended array was taken from UCSC genome browser and intersected with regions 1kb up and downstream of all the SVAs identified in the Hg19 in the Galaxy program. There were 1248 extended probe sets located within 1kb up or downstream of 729 different SVAs. These probe sets could be potentially used to assess the transcriptional activity occurring close to SVAs.

	Affymetrix human exon array core probes						
	2318743	**2318744**	**2318746**	**2318747**	**2318751**	**2318754**	**2318755**
CRBL	8.4399	10.6363	10.6403	9.62593	7.37805	7.31186	6.22524
FCTX	8.58727	10.7618	10.7633	9.81867	7.66961	7.65795	6.31523
HIPP	8.48885	10.6152	10.6728	9.82037	7.8162	7.57378	6.28323
MEDU	8.37343	10.588	10.6495	9.88626	7.81835	7.65326	6.27821
OCTX	8.46021	10.6742	10.6983	9.88322	7.69187	7.59721	6.29395
PUTM	8.45257	10.6337	10.675	9.80688	7.7563	7.55543	6.28135
SNIG	8.3719	10.5689	10.6267	9.79588	7.7587	7.58947	6.25403
TCTX	8.59431	10.7368	10.7672	9.84449	7.75681	7.6704	6.37456
THAL	8.3771	10.5694	10.6311	9.85412	7.84935	7.62727	6.32961
WHMT	8.08756	10.2715	10.3104	9.43152	7.31883	7.3547	6.0241

Table 5.1: Expression analysis using Affymetrix human exon array core probes corresponding to exons of the PARK7 gene. The probes used corresponded to the exons of the PARK7 gene to detect expression in several regions of the human brain. All the probes showed high enough levels to confirm expression of these exons. CRBL-cerebellum, FCTX-frontal cortex, HIPP-hippocampus, MEDU-medulla, OCTX-occipital cortex, PUTM-putamen, SNIG-substantia nigra, TCTX-temporal cortex, THAL-thalamus, WHMT-white

	Affymetrix human exon array extended probes		
Region of the Brain	2318733	2318734	2318735
CRBL	3.31776	4.93371	5.84348
FCTX	3.37886	5.10389	6.09532
HIPP	3.41042	4.997	6.14168
MEDU	3.4595	5.14324	6.00946
OCTX	3.44164	5.09881	6.06441
PUTM	3.44475	5.06815	6.1385
SNIG	3.45463	5.14249	6.10432
TCTX	3.28828	5.01947	6.02525
THAL	3.43769	5.194	6.01592
WHMT	3.40956	5.20551	6.08322

Table 5.2: Expression analysis using Affymetrix human exon array extended probes shows expression occurring in the region 3'of the SVA. Probe sets were used from the Affymetrix extended exon array corresponding to the region 3'of the PARK7 SVA and over the minor TSS of the PARK7 gene with their location shown in Figure 5.1. Probe sets 2318734 and 2318735 demonstrated significant levels of expression within all the brain regions analysed. CRBL-cerebellum, FCTX-frontal cortex, HIPP-hippocampus, MEDU-medulla, OCTX-occipital cortex, PUTM-putamen, SNIG-substantia nigra, TCTX-temporal cortex, THAL-thalamus, WHMT-white matter.

5.4.4 Epigenetic modifications and transcription factor binding at the PARK7 locus in SK-N-AS and MCF-7 cell lines

The histone marks and transcription factor binding across the major and minor promoters of the PARK7 gene and 5' of the PARK7 SVA were analysed using ChIP. ChIP allows for the identification of proteins bound to DNA and histone modifications to interrogate the potential transcriptional activity that is present at specific loci, in this case in the two cell lines, SK-N-AS and MCF-7. The histone modifications of interest in this assay were active (H3K4m2) and inactive (H3K9m3) to determine chromatin state. The presence of histone 3 (H3) was used as a positive control for the ChIP protocol. The regions were also analysed for the presence of the following proteins; RNA polymerase II, nucleolin, Sp1, CTCF and hnRNPK. Sp1 and hnRNPK in part were chosen due the model of regulation at the c-MYC gene at the NHE III$_1$ involving G4 DNA (section 1.3.2 Figure 1.5) and Sp1 is known to regulate the expression of the PARK7 gene at the major promoter (Taira et al. 2001). Nucleolin has been identified as binding to G4 DNA and stabilising its formation, therefore this was included to determine potential G4 DNA structures in the region of the SVA (Gonzalez et al. 2009). Finally the binding of CTCF was determined due the sequence of the hexamer repeat of the SVA that would provide multiple binding sites for this transcription factor. A primer set for a region located in a gene desert where little or no transcription is likely to be occurring was used a negative control to ensure the binding seen was specific. This negative primer set was chosen as had been used previously by a group (Richard Young's) in the analysis of the c-MYC promoter using ChIP. In both ChIP assays this negative primer set only showed the presence of inactive histone marks and H3 indicating this region is not an active region for transcription and provides a good negative control for ensuring the binding of the antibodies is specific (Figure 5.4 and 5.5).

The ChIP data for the SK-N-AS cell line indicates that the major and minor PARK7 promoters are transcriptionally active in this cell line due to the presence of RNA Pol II, active histone modifications (H3K4m2) and the transcription factor Sp1. These factors are also shown to be present 5’ of the SVA and therefore the SVA is flanked by active histone marks indicating it is an active genomic locus and not silenced as might be hypothesised to prevent the retrotransposition of the element. The lack of nucleolin at this locus, binding of Sp1 and weak binding of hnRNPK would suggest there is no G4 formed at this time in this cell line. The presence of G4 is associated with the repression of transcription therefore its absence at this locus is consistent with the active state of the region. The ChIP data for the MCF-7 cell line is incomplete due to problems occurring during the PCR analysis of the immunoprecipitated material. The amplification of immunoprecipitated material and sheared chromatin used as a positive control for the PCR over the major PARK7 promoter was unsuccessful despite the successful amplification of the same target in genomic DNA from MCF-7 cells carried out at the same time. This PCR was completed multiple times with the same outcome. The reason for this occurrence is unclear but could indicate the major PARK7 promoter region is particularly susceptible to sonication and there were not an adequate number of fragments in the sonicated material of this region for amplification. However the ChIP was successful over the minor promoter and 5’of the SVA. The data showed active transcription at the minor PARK7 promoter region with the presence of RNA pol II, active histone marks, Sp1 and hnRNPK binding. The 5’ of the SVA did not show the same active histone marks as in the SK-N-AS cell line with the lack of RNA pol II and Sp1 binding and the inconclusive equal intensity of inactive and active histone marks.

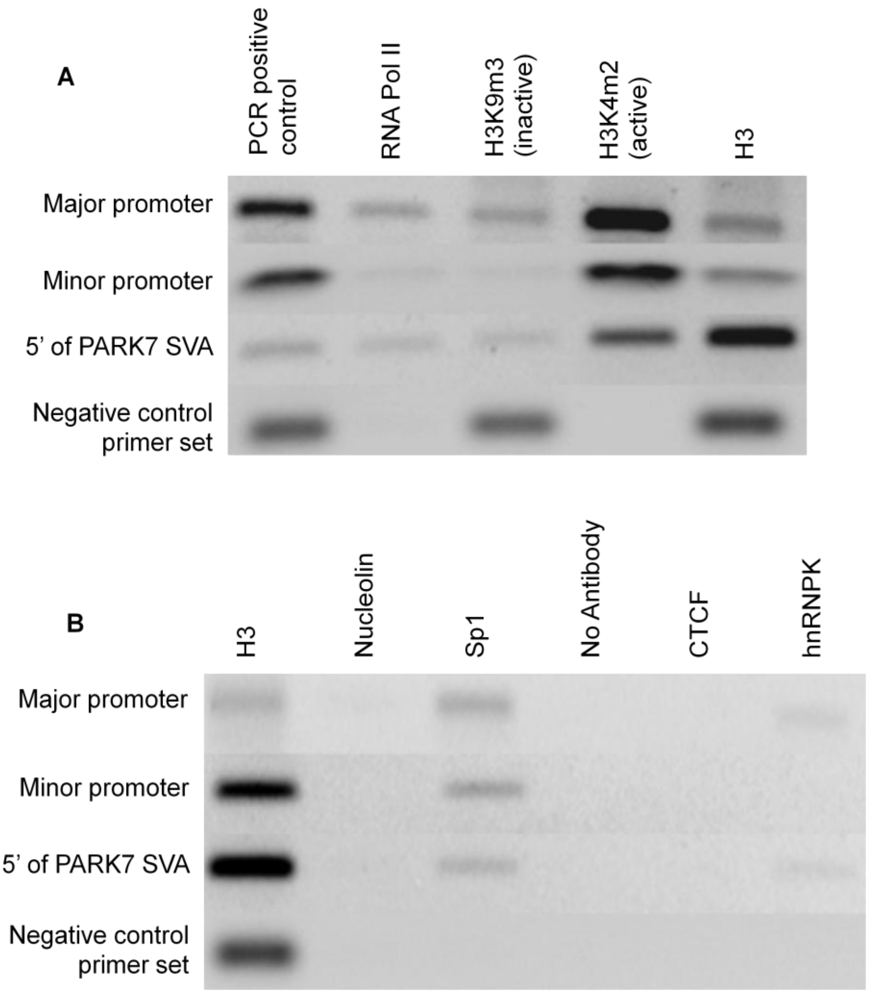


Figure 5.4: Chromatin modifications and transcription factor binding indicates activity at both PARK7 TSSs and the 5' of the PARK7 SVA in the SK-N-AS cell line. The presence of specific histone modifications, RNA polymerase II and the binding of transcription factors across the two PARK7 TSSs and 5' of the SVA in the SK-N-AS cell line was assessed using ChIP.

A – RNA pol II was shown to be present across the regions of the major and minor PARK7 TSS and upstream of the PARK7 SVA. Both active and inactive histone modifications were identified at all three loci but the intensity of the active marks was much greater indicating these regions contain active chromatin. There was no RNA polymerase II and active histone modifications detected at the negative primer set locus however the presence of histone 3 (H3) and inactive histone modifications were detected.

B - There was binding of the transcription factor Sp1 at the two PARK7 TSSs and 5' of the SVA but not at the negative primer set locus. There was weak binding of hnRNPK at the major PARK7 promoter and 5' of the SVA. There was no binding of CTCF and nucleolin at all three loci.

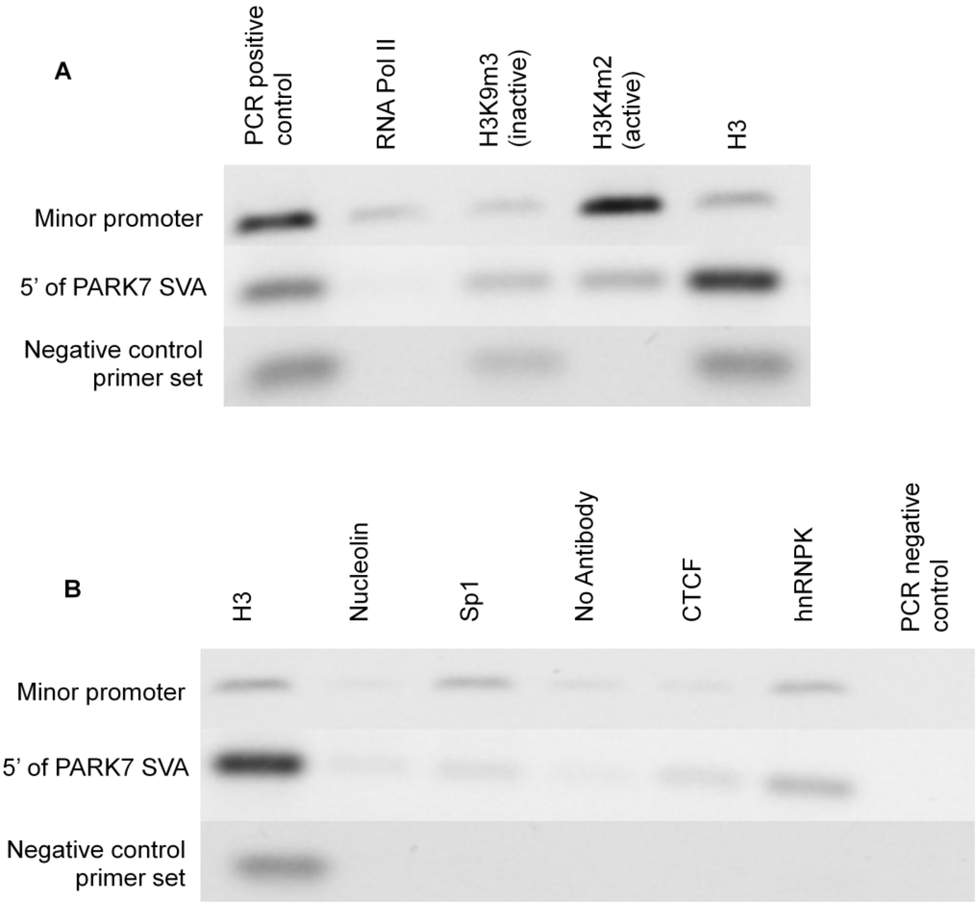


Figure 5.5: Chromatin modifications and transcription factor binding at the minor TSS of the PARK7 gene and 5' of the PARK7 SVA in MCF-7 cell line. The presence of specific histone modifications, RNA polymerase II and the binding of transcription factors across the minor PARK7 TSS and 5' of the SVA in the MCF-7 cell line was assessed using ChIP.

A – The presence of RNA pol II and active histone modifications is identified across the minor promoter indicating this region is transcriptionally active. RNA pol II is absent from the 5'of the SVA indicating this region is not transcriptionally active. However the active and inactive histone marks show equal intensity and is therefore inconclusive. The negative primer set showed the presence of inactive histone marks and H3.

B – There was binding of Sp1 and hnRNPK at the minor promoter which would indicate active transcription. There was also binding of hnRNPK in the 5' region of the SVA which may indicate the occurrence of single stranded DNA at this locus. The faint bands seen for nucleolin and CTCF for the minor PARK7 promoter and 5' of the SVA are at the same intensity at the no antibody control suggesting non-specific binding of DNA to the beads during the ChIP protocol and should not therefore been associated with the binding of the factors CTCF and nucleolin. The negative control for the PCR shows this is not due to contamination in the PCR.

5.4.5 Analysis of gDNA from MCF-7 cell lines and breast cancer tumours may suggest a deletion at the locus of the PARK7 gene

The lower passage MCF-7 cell line were chosen to be used over the high passage cell line based on the expression of the PARK7 transcript, however the reason for the differences between the two lines of MCF-7s was unclear. In a preliminary experiment it was addressed whether this could be due to a deletion or rearrangement at the PARK7 gene locus. The two PARK7 promoter regions were analysed in the genomic DNA from both these MCF-7 cell lines, the SK-N-AS cell line and from ten breast cancer tumour samples. Both promoters were present in the SK-N-AS and the low passage MCF-7 cell line but this region could not be amplified in the high passage MCF-7 cell line (Figure 5.6A). The gDNA extraction of the high passage MCF-7 cell line was carried out three times on separate occasions and the same result was seen. The PARK7 gene locus was also amplified in two other cell lines available in the lab (SH-SY5Y and JAr) and was present (data not shown). The minor promoter was present in all ten of the tumour samples tested and the major promoter was identified in eight of the samples (Figure 5.6A). The major promoter PCR was carried out in total three times and the breast cancer tumour samples F1 and B2 did not contain a PCR product for that region.

To analyse the PARK7 gene locus further primers sets were designed within the 7kb that separated the two promoters and up to 6kb downstream of the major TSS of the PARK7 gene. These regions were amplified in the SK-N-AS and the two MCF-7 cell lines and three of the breast cancer tumour samples (F1, G1 and B2). All four loci amplified were present in the SK-N-AS and low passage MCF-7 cell line but not in the high passage MCF-7 cell line (Figure 5.3B). This indicates that there may have been a deletion on chromosome 1 where the PARK7 gene is located in the high passage MCF-7 cell line and is not present in another line of MCF-7s. The three breast cancer tumour samples (F1, G1 and B2) tested for the presence of these four regions all contained PCR products of the correct size (Figure 5.6B), which included the two samples where the major promoter of the PARK7 gene was unable to be amplified. This suggested that much of the gene locus was intact apart from the region of the major promoter in the breast tumour samples F1 and B2. There may have been a deletion of that region or a mutation in the region corresponding to one of the primers used. Therefore another primer set for a larger fragment over the major PARK7 promoter (PCR product of 673bp) was used to determine which scenario seemed most likely. The region was present in the low passage MCF-7 cell line, SK-N-AS cell line and tumour sample G1 but was not present in tumour samples F1 and B2 and the low passage MCF-7 cell line (Figure 5.6C). This indicated there has been a small deletion of the major PARK7 promoter in the tumour samples F1 and B2 and the nature of this to be the subject of further analysis by the lab.

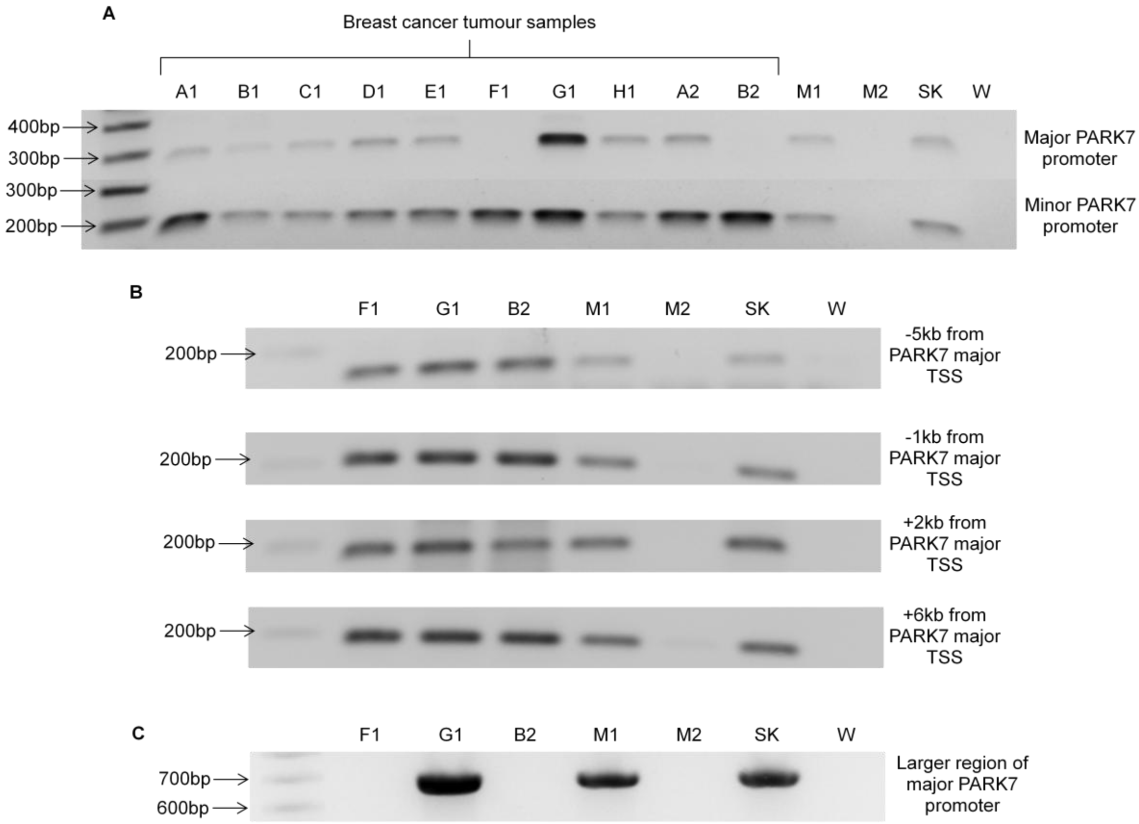

A
Breast cancer tumour samples
A1
B1
C1
D1
E1
F1
G1
H1
A2
B2
M1
M2
SK
W
400bp
300bp
300bp
200bp
Major PARK7 promoter
Minor PARK7 promoter
B
F1
G1
B2
M1
M2
SK
W
200bp
-5kb from PARK7 major TSS
200bp
-1kb from PARK7 major TSS
200bp
+2kb from PARK7 major TSS
200bp
+6kb from PARK7 major TSS
C
F1
G1
B2
M1
M2
SK
W
700bp
600bp
Larger region of major PARK7 promoter

Figure 5.6: PARK7 gene locus is altered in a high passage MCF-7 cell line.

A - The major PARK7 promoter is present within 8/10 breast cancer tumour samples, the SK-N-AS and low passage MCF-7 cell lines (expected product 270bp). For samples F1, G1 and B2 this was repeated in total 3 times to confirm the result. The minor PARK7 promoter was present in all 10 breast cancer tumour samples, the SK-N-AS and low passage MCF-7 cell lines (expected product 158bp).

B - The PARK7 locus was analysed further due to the unexpected lack of PCR products for the major promoter in samples F1 and B2. Regions upstream and downstream of the major TSS were amplified to determine if a deletion of this locus had occurred. In the high passage MCF-7 cell line all the regions of interest could not be amplified indicating a deletion which would support the lack of PARK7 mRNA in these cells. These regions up and downstream of the major PARK7 TSS were amplified in the breast tumour samples F1 and B2 that did not have a PCR product for the major promoter for PARK7.

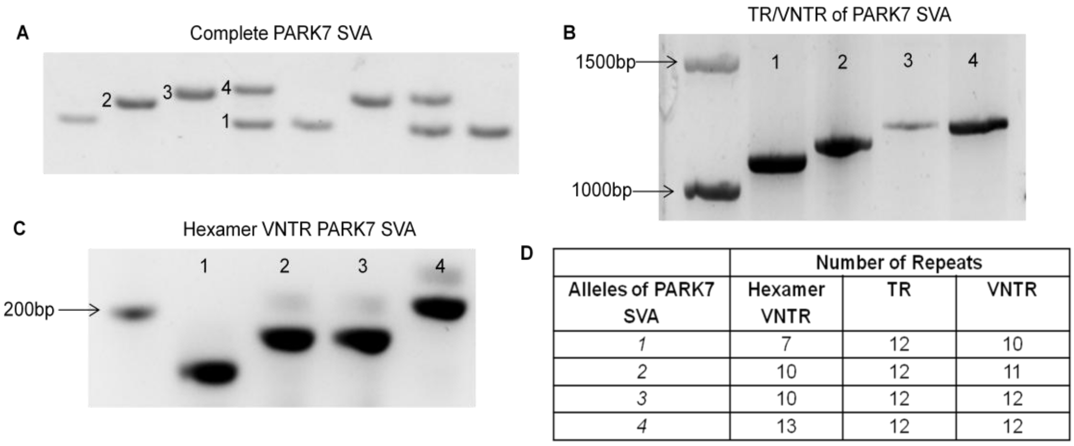
C – Amplification of a larger fragment surrounding the major promoter of the PARK7 gene in three tumour samples (F1, G1 and B2) and the three cell lines (2xMCF-7 and SK-N-AS). The expected PCR product was 673bp which is present in tumour sample G1 and the low passage MCF-7 and SK-N-AS cell lines. This would suggest that much of the gene locus is intact but there has been a deletion in the region of the major promoter. M1=low passage MCF-7 cell line, M2=high passage MCF-7 cell line, S=SK-N-AS cell line. W=water control.

5.4.6 Genotype profile of CEU HapMap cohort for the SVA upstream of the PARK7 gene

Genotypic analysis of the PARK7 SVA identified four distinct alleles which were polymorphic in length, in 87 individuals from the CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) HapMap cohort. An example image of the different alleles of the PARK7 SVA in the cohort is shown in figure 5.7A. The alleles are numbered from 1 to 4 from the shortest to the longest and the PCR products labelled with these numbers were gel extracted and cloned into an intermediate vector. This enabled the fragments to be sequenced to determine where the variation in this SVA was located and the plasmids generated were used in amplification of specific domains of the SVA. The intermediate vectors containing four alleles of the SVA were used as template to amplify the two repetitive regions of the SVA; the 5’ hexamer repeat and the central GC-rich domain. The central repetitive GC-rich domain was shown to be variable in length between alleles 1, 2 and 3 with this region in allele 4 running at the same size as in allele 3 (Figure 5.7B). When the hexamer repeat was amplified from the alleles within the intermediate vector this was also shown to be variable between alleles 1, 2 and 4 with allele 3 running at the same size as allele 2 (Figure 5.7C). This indicates that the central GC-rich repetitive region is the same size in alleles 3 and 4 and that the CCCTCT hexamer repeat is the same size in alleles 2 and 3.

The sequence analysis of the four alleles confirmed the allelic variation was found to be generated by differences in the number of repeat units present in the CCCTCT hexamer repeat and in central GC-rich VNTR region. The hexamer repeat was either a 7, 10 or 13 repeat domain and was termed a hexamer VNTR. The central VNTR domain of the younger SVA subtypes can be split into two domains due to differences in the sequence of the repeat units. The sequence analysis revealed there was no variation observed in the number of repeats in the 5’ VNTR of the central dual VNTR region, which was a stable 12 copy variant of 37-40bp repeat length and therefore termed a tandem repeat (TR). Variation was observed in the second 3’ VNTR domain and consisted of either 10, 11 or 12 repeats with a repeat length of 37-52bp in this cohort. This is summarised in figure 5.7D.

The genotype and allele frequencies of the CEU HapMap cohort are shown in Table 5.3A and B respectively. Alleles 1 and 3 were the most common within this cohort with 94% of the individuals having at least one of these alleles. HWE analysis was carried out for the PARK7 SVA locus in the CEU HapMap cohort showed it is not in HWE as the expected genotype frequencies were significantly different to the observed ($p<0.05$).



	Number of Repeats		
Alleles of PARK7 SVA	Hexamer VNTR	TR	VNTR
1	7	12	10
2	10	12	11
3	10	12	12
4	13	12	12

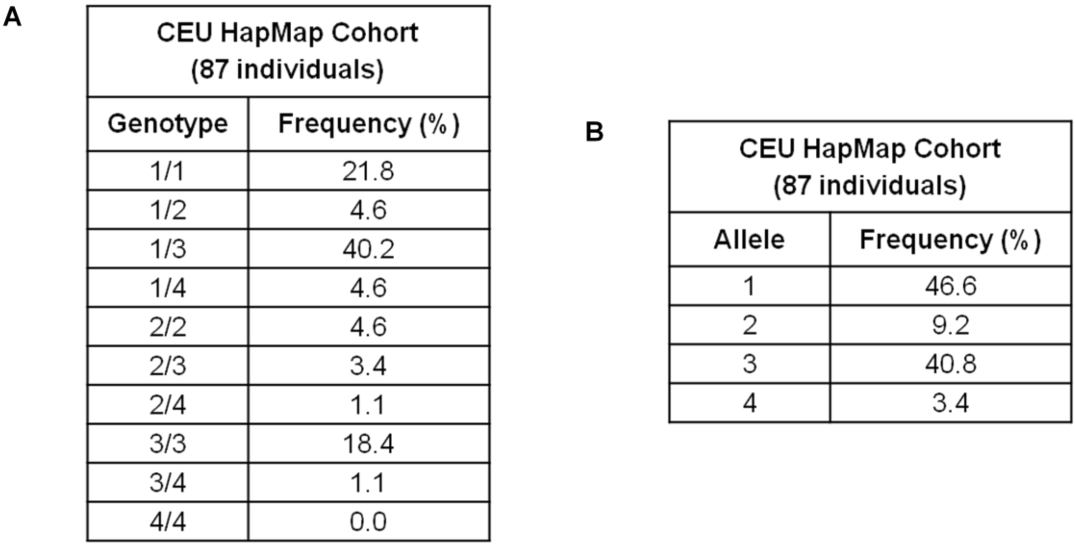
Figure 5.7: The PARK7 SVA has four alleles identified in a CEU HapMap cohort.

A- Image of PARK7 SVA amplified in individuals from a CEU HapMap cohort and run on a 1% agarose gel. The four alleles are numbered 1-4 from smallest to longest. The products labelled 1-4 were extracted from the gel and cloned into an intermediate vector (pCR-Blunt) and sequenced.

B – An image of the amplification of the central repetitive region of the intermediate vectors containing each of the four alleles. This central repetitive region has 3 alleles of different sizes with allele 3 and 4 containing a central repetitive region of the same length.

C - An image of the amplification of the hexamer repeat of the intermediate vectors containing each of the four alleles. This hexamer repeat has 3 alleles of different sizes with allele 2 and 3 containing a hexamer repeat of the same length.

D - Table displaying the number of repeats present within the three repetitive regions of the SVA. The 5’CCCTCT hexamer repeat termed hexamer VNTR was shown to have between 7 and 13 repeats across the four alleles. The next repetitive region was not variable between the four alleles and contained a stable 12 copies of its repeat. The second variable region was a GC-rich VNTR with 10-12 repeats 3’ of the TR region (see figure 5.9A for schematic of PARK7 SVA structure).



A

CEU HapMap Cohort (87 individuals)	
Genotype	Frequency (%)
1/1	21.8
1/2	4.6
1/3	40.2
1/4	4.6
2/2	4.6
2/3	3.4
2/4	1.1
3/3	18.4
3/4	1.1
4/4	0.0

B

CEU HapMap Cohort (87 individuals)	
Allele	Frequency (%)
1	46.6
2	9.2
3	40.8
4	3.4

Table 5.3: Genotype and allele frequencies of individuals in the CEU HapMap cohort for the PARK7 SVA. A - Table showing the frequency of each genotype within the 87 individuals genotyped from the CEU HapMap cohort. B - Table showing the frequency of each allele within the CEU HapMap cohort. The PARK7 SVA locus is not in HWE; $p<0.05$

5.4.7 Sequence of the SVA upstream of the PARK7 gene

The four alleles of the PARK7 SVA identified in individuals from the CEU HapMap were sequenced once cloned into the intermediate vector pCR-Blunt (2.2.5.1). The complete sequence of all four alleles of the SVA was obtained except for a portion of the Alu-like sequence of allele 4 as the sequencing proved difficult over the area after the CCCTCT hexamer repeat. The sequence of the SVA is shown in Figure 5.8. The additional repeats found in alleles 2-4 are underlined. The sequences that have G4 potential on either strand are highlighted in bold and these are located in the CCCTCT hexamer VNTR and the GC-rich VNTR. The sequences of SVAs contain many CG sites for methylation therefore the CG dinucleotides of the PARK7 SVA are in green to illustrate this.

TCCTCT***CCCTCTCCCTCTCCCTCTCCCTCTCCCTCCTCTCCCTCTCCCTCTCCCTCTCCCTC***
TCCCTCTCTCTCTCC

ACGGTCTCCTTCCACGGTCTCCCTCTGATGCCGAGCCAAAGCTGGACGGTACTGCTGCCATC
TCGGCTCACTGCAACCTCCCTGCCTGATTCTCCTGCCTCAGCCTGCCGAGTGCCTGCGCACG
CCGCCACGCCTGACTGGTTTTCGTTTTTTTTTTTTTGTGGAGACGGGGTTTTGCTGTGTTGGC
CGGGCTGGTCTCCAGCTCCTAACCACGAGTGATCCGCCAGCCTCGGCCTCCCGAGGTGCCGG
GATTGCAGACGGAGTCTCGTTCACTCAGTGCTCAATGGTGCCCAGGCTGGAGTGCAGTGGCG
TGATCTCGGCTCGCTACAACCTCCACCTCCCAGCCGCCTGCCTTGGCCCCCCAA

AGTGCCGAGATTGCAGCCTCTGCCCAGCCGCCACCCCGTC
TGGGAAGTGAGGAGCGTCTCTGCCTGGCCCCCCATCGTC
TGGGATACGAGGAGCCTCTCTGCCTGGCTGCCCAGTC
TGGAAAGTGAGGAGCGTCCCTGCCCGGCCGCCATCCCATC
TAGGAAGCGAGGAGCGCCTCTTCCCCGCCGCCATCCCATC
TAGGAAGTGAGGAGCGTCTCTGCCCGGCCACCCATCGTC
TGAGATGTGGGGAGCACCTCTGCCCCGCCGCCCTGTC
TGGGATGTGAGGAGCGCCTCTGCTGGGCCGCAACCCTGTC
TGGGAGGTGAGGAGCGTCTCTGCCCGGCCGCCCCGTC
TGAGAAGTGAGAAAACCCTCTGCCTGGCAACCGCCCCGTC
TGAGAAGTGAGGAGCCCCTCCGTCCGGCAGCCACCCCGTC
TGGGAAGTGAGGAGCGTCTCCGCCCGGCAGCCACCCCGTC

TR

T***GGGAGGGAGGTGGGGGGGGG**G*TCAGCCCCTGCCCGGCCAGCTGCCCTGTC
CGGGAGGTGAGGGGCTCCTCTGCCCGGCCAGCCGCCCCGTC
CG***GGAGGGAGGTGGGGGGG***TCAGCCCCCGCCCGGCCAGCCGCCCCGTC
CGGGAGGGAGGTGGGGGGATCAGCCCCCGCCCGGCCAGCCGCCCCGTC
CGGGAGGGAGGTGGGGGGTCAGCCCCCGCCCGGCCAGCCGCCCCGTC
CGGGAGGGAGGTGGGGGGATCAGCCCCCGCCCGGCCAGCCGCCCCGTC
CG***GGAGGGAGGTGGGGGGG***TCAGCCCCCCGCCCGGCCAGCCGCCCTATC
CAGGAGGTGAGGGGCGCCTCTGCCCGGCCGCCCCTAC
TGGGAAGTGAGGAGCCCCTCTGCCTGGCCAGCCGCCCCGTC
CG***GGAGGGTGGTGGGGGGG***TCAGCCCCCGCCCGGCCAGCCGCCCCATC
CGGGAGGTGAGGGGCGCTTCTGCCCGGCCGCCCCTAC
TGGGAAGTGAGGAGCCCCTCTGCCCGGCCAGGACCCCGTC

VNTR

TGGGAGGTGTGCCCAGCGGCTCATTGGGGATGGGCCATGATGACAATGGCGGTTTTGTGGAA
TAGAAAGGCGGGAAGGGTGGGGAAAAAATTGAGAAATCGGATGGTTGCCGGGTCTGTGTGGA
TAGAAGTAGACATGGGAGACTTTTCATTTTGTTTTGTACTAAGAAAAATTTTTTTGCCTTGG
AAAAAAAAAAAAAAAAAAAAAAA

Figure 5.8: Sequence of the PARK7 SVA. The complete sequence of the SVA upstream of the PARK7 gene is shown with the additional repeats found in the CCCTCT hexamer VNTR and the GC-rich VNTR of the different alleles underlined. The sequences that have the potential to form G4 DNA predicted by Quadparser software is shown in italics. SVAs are GC rich and the CGs in the sequence are coloured green to show the many potential sites for methylation.

5.4.8 Distinct components of the SVA upstream of the PARK7 gene shows differential activity in a reporter gene construct which is orientation dependant in the SK-N-AS and MCF-7 cell lines

The ability of the intact PARK7 SVA and its distinct individual domains to act as transcriptional regulators was tested in a cell line model using the dual luciferase assay. SVA insertions can be found in the same, or opposite orientation to a gene locus. When analysed, 49.3% of the SVAs found within 10kb upstream of TSSs were on the same strand as the gene (Table 4.1), for this reason we also tested whether their function was orientation dependant. Eight reporter gene constructs were generated (Figure 5.9A) containing the following fragments in both forward and reverse orientations:

- the whole SVA (SVA)
- SVA with the SINE region deleted (SVAΔSINE)
- central TR and VNTR (TR/VNTR) as a dual component due to the difficulties associated with designing a specific primer to clone the TR and VNTR individually
- a 5’ truncation with only the CCCTCT hexamer, Alu-like sequence and 10 of the 12 repeats of the TR of allele 1 of the PARK7 SVA (truncated SVA) present

SVAs are described as having a CCCTCT domain at their 5’ end and a poly A-tail at their 3’ end and this was used to define the forward orientation. The ability of the eight fragments to support reporter gene expression (luciferase) directed by a heterologous minimal promoter was tested in two cell lines, SK-N-AS and MCF-7.

In the SK-N-AS cell line (Figure 5.9B) the intact PARK7 SVA in forward orientation did not alter the levels of reporter gene expression, when compared to the minimal promoter alone (pGL3P) however when the SINE domain was deleted reporter gene activity was significantly enhanced (P<0.05). The TR/VNTR and the truncated SVA in the forward orientation acted to significantly repress luciferase activity when compared to the minimal promoter alone (pGL3P) (P<0.001, P<0.01 respectively). When the domains were tested in the reverse orientation the reporter gene levels were all significantly different when compared to the levels seen in the forward orientation (SVA P<0.001, SVAΔSINE P<0.05, TR/VNTR P<0.05, truncated SVA P<0.01). The activity of the SVA and SVAΔSINE in reverse orientation were reduced compared to when in the forward orientation whereas the activity of the TR/VNTR and truncated SVA showed the opposite trend.

The reporter gene constructs supported distinct reporter gene activity levels in the MCF-7 cell line when compared to that observed in the SK-N-AS cell line (Figure 5.9C). In the forward orientation the complete SVA had a significant increase in reporter gene activity in MCF-7 cells (P<0.01), distinct from its function in SK-N-AS, however similarly to SK-N-AS cells the SVAΔSINE showed the greatest ability to enhance reporter gene activity. In contrast the TR/VNTR showed similar activity to that of the minimal promoter alone. The truncated SVA acted as a repressor as it did in the SK-N-AS cell line (P<0.05). The domains in the reverse orientation all showed a significant difference to the activity of the domains in the forward orientation (SVA P<0.001, SVAΔSINE P<0.001, TR/VNTR P<0.001, truncated SVA P<0.01). The SVA, SVAΔSINE and TR/VNTR all showed decreased activity in the reverse orientation when compared to the domains in the forward orientation. The truncated SVA showed greater activity in the reverse orientation than when in the forward orientation.

Classical enhancers show similar functional properties in both orientations however the PARK7 SVA does not. Therefore the orientation of the SVA relative to the gene it has inserted near to could alter the affect of the SVA on transcription of the gene.

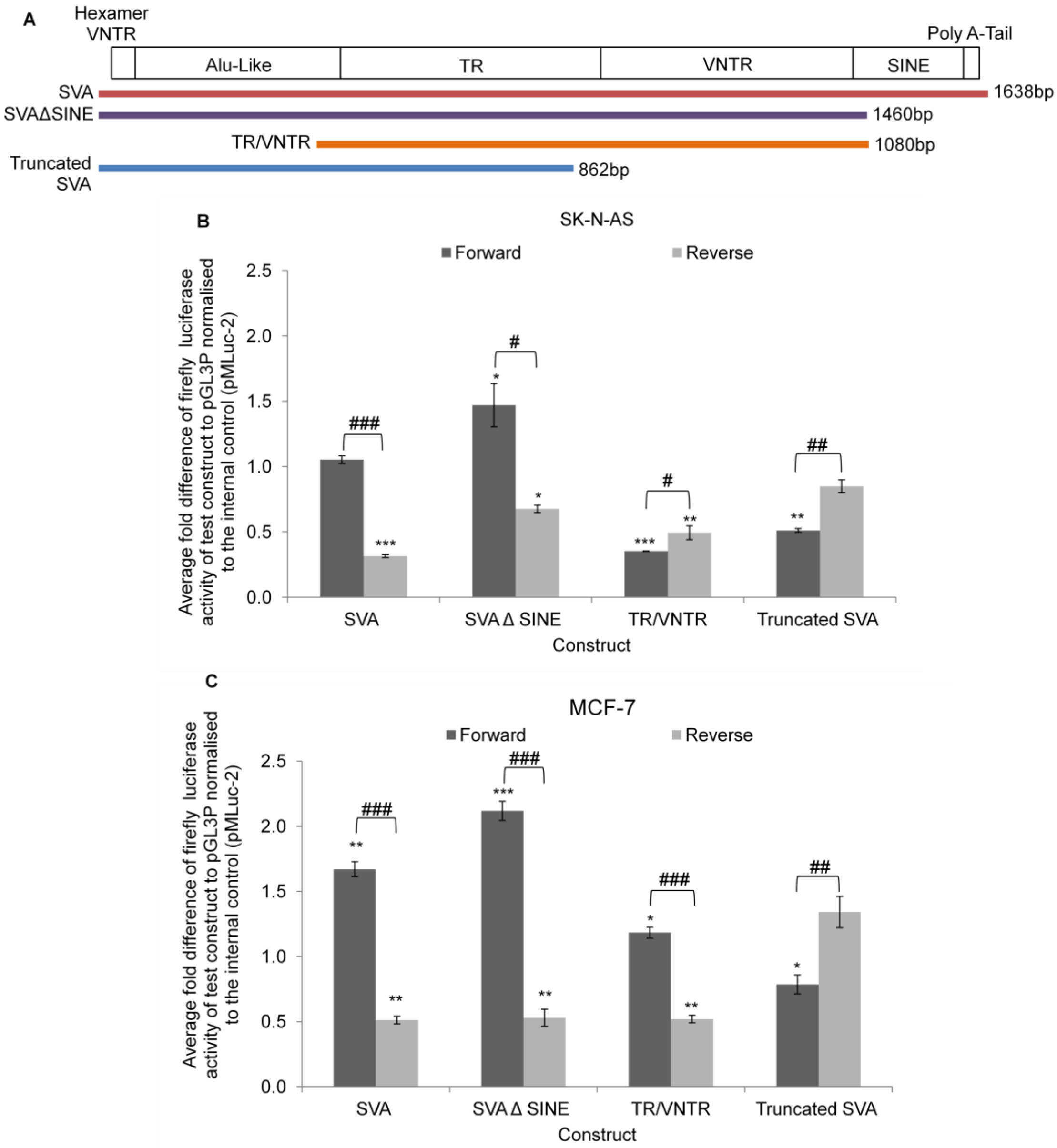


Figure 5.9: The PARK7 SVA showed the ability to affect expression in a reporter gene construct. A – Schematic showing the genomic structure of the PARK7 SVA and the relationship to the fragments tested in the reporter gene constructs. B - The average fold activity of the different fragments from the SVA tested in both forward and reverse orientation over the minimal SV40 promoter alone (pGL3P) in the SK-N-AS cell line. Data was normalised to compensate for transfection efficiency (N=4). C - The average fold activity in the MCF-7 cell line of the different fragments of the SVA in forward and reverse orientation over the minimal SV40 promoter alone (pGL3P) normalised to the internal control to account for transfection efficiency. N=4. One tailed t-test was used to measure significance of fold activity of PARK7 SVA fragments over SV40 minimal promoter alone (pGL3P) *P<0.05, **P<0.01, ***P<0.001 and to compare fold activity of forward and reverse orientations #P<0.05, ##P<0.01, ###P<0.001.

5.4.9 The PARK7 SVA did not display the ability to initiate transcription in a promoter less reporter gene vector in SK-N-AS and MCF-7 cell lines

The ability of the whole PARK7 SVA in forward and reverse orientations to initiate transcription was tested in the promoter less vector pGL3B in both the SK-N-AS and MCF-7 cell lines (Figure 5.10). The activity of the SVA in the forward orientation was significantly increased over that of the pGL3B vector in both the SK-N-AS and MCF-7 cell lines (P<0.01). Despite this increased activity I did not think to a great enough degree to determine that the SVA can act as promoter, compared to pGL3P. At the very low levels of expression observed in pGL3B the experiment is very sensitive to small changes which could occur from background. Even though the pGL3B vector alone does not contain a promoter the bioluminescence from the luciferase assay was above that of background indicating minimal amount of transcription of the firefly luciferase reporter gene from spurious binding of RNA polymerases, which is 2-3% of that of the pGL3P vector containing the minimal SV40 promoter. Therefore the PARK7 SVA maybe enhancing this spurious activity when in the forward orientation. The SVA in the reverse orientation showed a significant decrease in activity over that of the pGL3B vector alone in both the SK-N-AS and MCF-7 cell lines (P<0.01 and P<0.001 respectively). However again due to the low levels of activity of the pGL3B vector this difference is most likely due to the SVA in the reverse orientation repressing this minor amount of transcription. In both cell lines there was a significant difference between the activity of the SVA in the forward and the reverse orientations (P<0.001). This data has been added for completeness but due to the low levels of expression in this assay the resulting fold changes of the constructs containing the SVA cannot determine that the SVA is acting as a promoter.

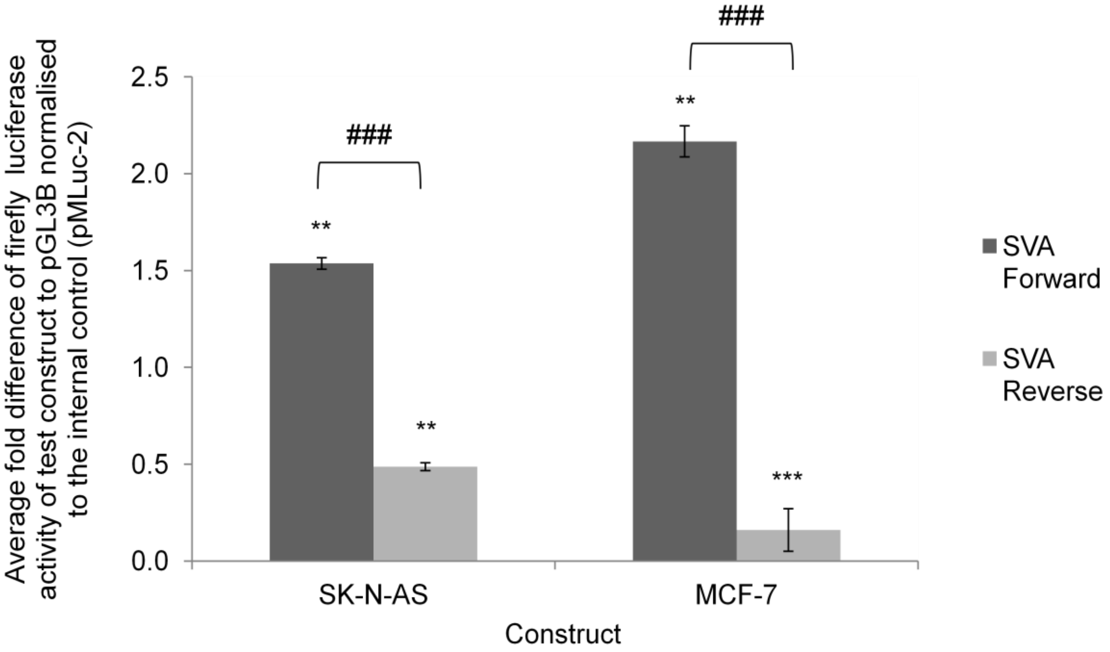
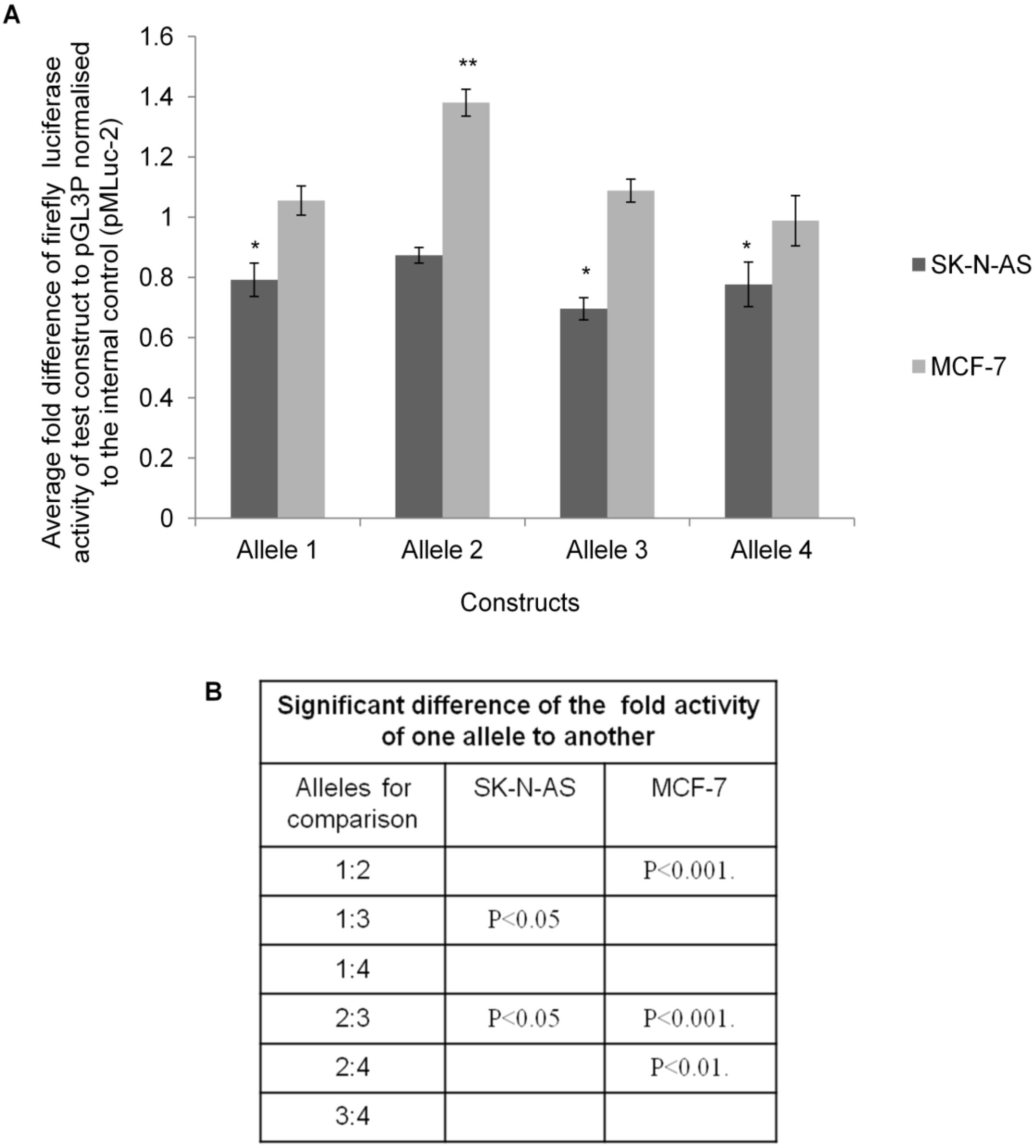


Figure 5.10: The PARK7 SVA did not display the ability to initiate transcription in a promoter less reporter gene vector. The average fold activity of the whole PARK7 SVA tested in both forward and reverse orientation over the promoter less vector pGL3B in the SK-N-AS and MCF-7 cell lines. Despite showing significant differences over the empty vector the SVA could not be defined as having promoter function as the fold activity change was low over the very low levels of background transcription occurring from the pGL3B vector. Data was normalised using an internal control to compensate for transfection efficiency. One tailed t-test was used to measure the significance of fold activity of the PARK7 SVA in each orientation over the promoter less vector pGL3B. **P<0.01, ***P<0.001 and to compare the activity of the SVA in the forward and reverse orientation to each other ###P<0.001 N=4

5.4.10 Alleles of SVA upstream of the PARK7 gene show differential activity in a reporter gene construct in SK-N-AS and MCF-7 cell lines

The regulatory activity of the four alleles of the PARK7 SVA was tested in the pGL3P vector, which contains a minimal SV40 promoter, in the SK-N-AS and MCF-7 cell lines (Figure 5.11A). In the SK-N-AS cell line all four alleles showed a minor repressive function with allele 1, 3 and 4 with activity significantly below that of the pGL3P vector alone ($P<0.05$). The four alleles showed very similar levels of activity with a small but significant difference between alleles 1 and 3 and alleles 2 and 3 ($P<0.05$) (Figure 5.11B).

All four alleles showed a differential level of activity in the MCF-7 cell line compared to in the SK-N-AS cell line (Figure 5.11A). Alleles 1, 3 and 4 showed similar activity to that of pGL3P alone and were not significantly different. Allele 2 in the MCF-7 cell line showed the greatest activity of all the alleles and was significantly above that of pGL3P alone ($P<0.01$). There was a significant difference between the activity of allele 2 when compared to the activity of allele 1, 3 and 4, $P<0.001$, $P<0.001$, $P<0.01$ respectively (Figure 5.11B).



Significant difference of the fold activity of one allele to another		
Alleles for comparison	SK-N-AS	MCF-7
1:2		P<0.001.
1:3	P<0.05	
1:4		
2:3	P<0.05	P<0.001.
2:4		P<0.01.
3:4		

Figure 5.11: Reporter gene constructs containing alleles of the SVA upstream of PARK7 show differential activity in MCF-7 and SK-N-AS cell lines. A – The average fold activity of the reporter gene constructs containing the four alleles of the SVA over control vector pGL3P normalised to the internal control TK renilla in SK-N-AS and MCF-7 cell lines. *P>0.05, **P>0.01 B – Table showing the significant difference in fold activity between the alleles of the SVA within SK-N-AS and MCF-7 cell lines. N=4

5.5 Discussion

The bioinformatic analysis of the PARK7 gene locus identified a human specific SVA D 8kb upstream of the start of transcription. In the UCSC genome browser the TSS of the PARK7 gene showed characteristics such as a CpG island, active histone marks and DNase 1 hypersensitivity clusters which were also present in the region less than 1kb downstream of the SVA and the browser Archive Ensembl predicted a transcript originating at this locus (Figure 5.1). This indicated a second putative TSS of the PARK7 gene adjacent to the SVA. The expression of the well characterised PARK7 transcripts originating from the major TSS were confirmed in the SK-N-AS human neuroblastoma cell line and the human adenocarcinoma MCF-7 cell line whereas as the predicted transcript originating near the SVA was only identified in the SK-N-AS cell line (Figure 6.2). The two TSSs of the PARK7 gene were named major (the already characterised promoter) and minor (the putative promoter). ENCODE data showing the state of the chromatin across 9 different cell lines available in UCSC also indicated a tissue specific expression at the minor promoter but not at the major (Figure 6.3). The new transcript differs from the known transcripts by differential exon usage for exon 1 which is non-coding therefore would not result in differences at the protein level. However it may be the regulation of the expression of these transcripts which is more important with the two promoters differentially responding to environmental cues as noted by the tissue specific expression pattern of the new transcript.

ChIP data in the SK-N-AS cell line supports the finding that both of these promoters are active with the identification of active histone marks (H3K4m2), RNA polymerase II and Sp1 binding at both these sites (Figure 5.4). There is also evidence of active chromatin at the 5’ of the PARK7 SVA indicating this region is not silenced as may have been assumed to prevent this element from undergoing retrotransposition. The ChIP data from the MCF-7 cell line is inconclusive due to problems amplifying the major promoter in the sheared chromatin. Although there is evidence of transcriptional activity at the minor promoter the expression of the new PARK7 transcript was not identified in this particular cell line, this may be due to the presence of ESTs in this region. Affymetrix probe array data, carried out by a collaborator Mina Ryten, also confirmed expression occurring at this locus in several human brain regions (Table 5.1 and 5.2). This may not directly correspond to the new transcript but does confirm there is active transcription at this locus.

Retrotransposons have been thought to be silenced through mechanisms such as methylation to prevent their retrotransposition; however the data shown for the PARK7 SVA locus in the UCSC genome browser and the analysis of the transcripts of the PARK7 gene (5.4.1 and 5.4.2) indicate that the region near to the SVA is transcriptionally active and not silent at least in cell lines. This preliminary analysis identified another transcript of the PARK7 gene that originates close to the PARK7 SVA (within 750bp) that was predicted by the Archive Ensembl browser. Further bioinformatic analysis determined there are 172 SVAs in the human genome within 1kb upstream of Archive Ensembl predicted transcripts compared to the 58 SVAs within 1kb of known genes. The confirmation of the existence of this new PARK7 transcript indicates there could be other mRNAs originating close to SVAs that are yet to be identified. There are also 1248 Affymetrix extended probe sets located within 1kb up or downstream of 729 different SVAs which could be used to characterise expression patterns in those regions as seen for the PARK7 SVA.

Four alleles of the PARK7 SVA were identified in the CEU HapMap cohort (Figure 5.7). The PARK7 SVA shared similar characteristics as the FUS SVA in terms of the central repetitive region in which there was a TR and a VNTR. However the PARK7 SVA not only showed variation in the number of repeats in this central VNTR as was observed for FUS SVA but also in the CCCTCT hexamer repeat therefore this was termed a hexamer VNTR. The genetic variation at this locus provides potential for differential function and association with disease.

The ability of the PARK7 SVA to act as a transcriptional regulator in a classical reporter gene model was analysed similar to the assay carried out for the FUS SVA in chapter 3 section 3.4.3. Different sized fragments of the PARK7 SVA were tested in a reporter gene vector with a minimal promoter to determine the domains of the SVA may influence gene expression as were the four alleles of the SVA. The activity of these constructs was tested in both the SK-N-AS and MCF-7 cell lines. The TR/VNTR showed differential properties between the two cell lines; in the SK-N-AS the TR/VNTR showed repressive qualities whereas in the MCF-7 this construct showed activity just above that of the minimal promoter. The complete SVA showed no activity in the SK-N-AS cell line but enhanced reporter gene expression in MCF-7 cells. Interestingly the deletion of the SINE element from the SVA fragment resulted in significantly higher levels of reporter gene expression than the SVA alone in both cell lines which was consistent with the loss of the SINE from the FUS SVA resulting in increased reporter gene activity in the SK-N-AS cell line. There are therefore probably a minimum of three distinct functional elements in the PARK7 SVA that adjust its ability to modulate expression, the central TR/VNTR, SINE and the CCCTCT and Alu-like sequences. The activity of the constructs showed an orientation dependency and therefore the orientation of SVAs relative to genes may affect how they could impact on gene expression. The data on the central TR/VNTR indicated they support distinct transcriptional properties dependent on cell type. This is consistent with the action of VNTRs we have previously observed in the human serotonin and dopamine transporter genes (Guindalini et al. 2006; Haddley et al. 2008; Ali et al. 2010; Vasiliou et al. 2012) with different complements of transcription factors present in both these cell lines responsible for the activity of the reporter gene directed by the TR/VNTR. The four alleles of the PARK7 SVA showed very similar activity to each other in the SK-N-AS cell line. In the MCF-7 cell line allele 2 showed the greatest activity and was significantly different to all the other 3 alleles. The alleles may show greater functional differences in response to challenge than under basal conditions. In an attempt to identify if the PARK7 SVA can initiate transcription itself the whole SVA was tested in a promoter less reporter gene construct in which the SVA did not display the ability to act as a promoter.

The differences in the expression of the PARK7 gene between two different passages of MCF-7 cell lines led to the identification of a large deletion of the PARK7 locus in the higher passage cells. A deletion across the major promoter of the PARK7 gene was also indicated in two out of ten breast tumour samples obtained from the Liverpool Cancer Tissue Bank (Figure 5.6). The exact nature of this deletion is yet to confirmed but further analysis is to be completed by the research group, for example if this variant is present in the population or is the consequence of a deletion during the tumourigenic process. In the literature there have been examples of variants in the promoter region of the PARK7 gene but these could not explain the finding observed in the tumour samples. There is a 18bp insertion/deletion polymorphism is located in the promoter region of PARK7 and a

16bp deletion across the major TSS of the PARK7 gene was identified in a South African PD patient (Eerola et al. 2003; Keyser et al. 2009).

This chapter analysed the PARK7 SVA in detail, demonstrated its ability to differentially affect transcription within a reporter gene construct in two different cell lines with multiple regulatory domains. A new transcript of the PARK7 gene was identified originating within 1kb of the SVA with tissue specific expression. Active histone marks, RNA polymerase II and Sp1 binding was determined in the region of both PARK7 promoters and 5’ of the SVA in the SK-N-AS cell line supporting the hypothesis that the region of the PARK7 SVA is active and is not transcriptionally silenced. The genetic variation of the PARK7 SVA also provides potential functional differences between individuals and this variation would be important to be assessed in terms of association with disease.

Chapter 6

Can SVAs provide novel biomarkers for disease?

6.1 Introduction

There were 2676 SVAs identified in the Hg19 with a preference for inserting into genic regions (chapter 4) with two of these elements demonstrating the ability to affect transcription in a reporter gene model *in vitro* and one of these element tested demonstrated this ability *in vivo* as well (chapter 3 and 5). Two approaches were undertaken in an attempt to functionally group sets of these elements and identify potential genetic markers linked to disease through the analysis of the PARK7's primary sequence and the presence of SVAs in genes linked to the same disease.

The primary sequence of an element can define potential function such as methylation state and secondary structure formation (discussed for SVAs in chapter 4) but also the transcription factors that could bind. Of particular interest to me in that regard were the repetitive domains of the SVA, such as the central VNTR, as this would provide multiple transcription factor (TF) binding sites. It is hypothesised that weak binding sites that are repeated several times can maintain the presence of a TF better than a single site as when a TF dissociates from its binding site it scans the neighbouring region of DNA for additional sites before completely leaving that section of DNA (Breen et al. 2008). For example the VNTR located within intron 2 of the serotonin transporter gene has been shown to bind the TF YB-1 and CTCF and the transcriptional properties of this VNTR and its alleles are differentially regulated by the action of theses TF *in vitro* (Klenova et al. 2004; Roberts et al. 2007). SVAs are located throughout the human genome and share similar primary sequence therefore could provide multiple loci that could bind and respond to similar TFs. The sequences of the repeat units of the central VNTR of the SVAs are imperfect and vary, most notably the presence of two VNTR domains, especially in the younger SVAs, that have distinct sequences. Therefore three of the repeat units of the PARK7

SVA were taken and used to search the genome for other SVAs with high homology for these specific sequences to identify other SVAs and the genes they are near. The genes generated from this list would be analysed for their potential to be linked in networks and response to specific TFs and if these same TFs may modulate the SVA to identify TF pathways that link the SVAs with a subset of target genes. This model will use the nearest gene to the SVA as the target gene however the SVAs do not have only modulate the nearest but provides a good starting point.

The second method of creating a subset of SVAs involved the analyses of a list of genes related in a disease process. Mutations in the PARK7 gene have been linked to Parkinson's disease (PD) along with many other genes, including α synuclein (SNCA), parkin (PARK2) and leucine repeat rich kinase 2 (LRRK2) (Corti et al. 2011). The mutations within these genes cause autosomal dominant (SNCA, LRRK2) or recessive PD (PARK2 and PARK7) with their own particular disease progression and onset (Houlden and Singleton 2012). PD is a neurodegenerative disease associated with the loss of dopamine producing neurons in the substantia nigra pars compacta with symptoms only becoming apparent when 70-80% of nigrostriatal terminals have been lost (Bernheimer et al. 1973). PD affects 0.5-1% of people aged 65-69 and 1-3% of those aged over 80 (Nussbaum and Ellis 2003) with 5-10% of these patients carrying a mutation in one of the known disease causing genes (Corti et al. 2011). The vast majority of PD cases are sporadic and are due to a combination of genetic and environmental factors (Migliore and Coppede 2009). Since the identification of the mutations in the SNCA causing PD (Polymeropoulos et al. 1997) there have been hundreds of studies to identify mutation in genes linked to familial PD and to identify genetic variants that are risks for sporadic PD (Coppede 2012). A list of known PD causing genes were taken from a review on the genetics of PD (Corti et al. 2011) and UCSC genome browser was used to determine if a SVA was present within those genes or their flanking region. This could provide novel genetic variation within these genes and it may be important to consider genetic variants across several genes in disease association studies. It could be a combination of genetic variants in a pathway that confer a risk for a specific disease; for example specific alleles of the VNTR within the 3’UTR of the SLC6A3 gene and a VNTR within the dopamine 4 receptor (DRD4) in combination are considered a risk for attention deficit hyperactivity disorder but these genotypes do not show the same association when analysed on their own (Carrasco et al. 2006). The presence of SVAs in PD genes may affect the transcriptional regulation of the gene through binding of TFs, secondary structure formation, introduction of new splice sites, transcription pausing at their poly A tail and differential methylation patterns. An SVA insertion in the TAF1 gene has been linked to a disease with parkinsonism (X-linked dystonia-parkinsonism) in a Philippine cohort through differential methylation of this element in the caudate nucleus of sufferers resulting in reduced mRNA reduction of TAF1 (Makino et al. 2007). Differential methylation state of PD genes has also been shown in specific brain regions; decreased methylation of intron 1 of the SNCA was identified in the substantia nigra, putamen and cortex of patients with sporadic PD indicating a role for epigenetic regulation of gene expression in PD (Jowaed et al. 2010). Genetic variation of the elements, including both variants in copy number within the VNTR and the absence or presence of these elements, could result in differential response to their environment in a tissue specific manner.

6.2 Aims

- Use two different approaches to analyse the potential role of SVAs across networks of genes; firstly use the primary sequence of the PARK7 SVA in search of other SVAs within the genome and secondly take a known group of genes associated with PD and determine how many contain a SVA.

Section A

- Analyse the primary sequence of the PARK7 SVA to identify potential transcription factors that may act on the domain and how this may be linked in networks across genes with SVAs near their locus.
- Expand on the bioinformatic data generated to test the hypothesis how transcription factors may regulate endogenous gene expression and potentially influence regulatory properties of the SVA as well, using the oestrogen receptor alpha as an example.
- Determine if the genetic variation of the PARK7 SVA is associated with a disease cohort of breast cancer patients with wild-type BRCA genes.

Section B

- To use a list of PD associated genes from Corti et al 2011 to determine how many contain a SVA.
- Analyse the genetic variation of any SVAs identified within the gene loci.

6.3 Methods

6.3.1 Identification of other SVAs containing similar repeat unit sequence to the PARK7 SVA

The primary sequence of the central repetitive region of the PARK7 SVA was used to identify potentially similar SVAs based on the TR and VNTR repeat unit sequences. The Blat search tool on the UCSC genome browser was used to search the whole genome for the sequences of interest. The Blat tool results provided the locus of the sequence identified within the genome, the percentage homology of the sequence compared to the original query sequence and the number of bases that the sequence spans. The sequences from the TR and VNTR shown in Figure 6.1 were chosen as they contained features consistent throughout many of the repeat units of those repetitive regions.

The results from each Blat search were removed if they didn't reach the following criteria: at least 90% homology with the original sequence and to have a score and span of +/- 15% of the length of the original sequence. After the exclusions 178 SVAs containing the repeat sequence from the TR, 168 SVAs containing the repeat sequence from the VNTR and 94 SVAs containing the sequence of the junction of the TR and VNTR remained. The results from all three Blat searches were then aligned to generate a list of SVAs containing the repeat units from the TR and VNTR of the PARK7 SVA, however there were no SVAs containing all the threes repeat sequences when using the criteria previously described.

The loci of these SVAs were analysed for the genes within 100kb of them. The loci of the gene, the strand the gene was located on and the distance of the SVA from gene was recorded.

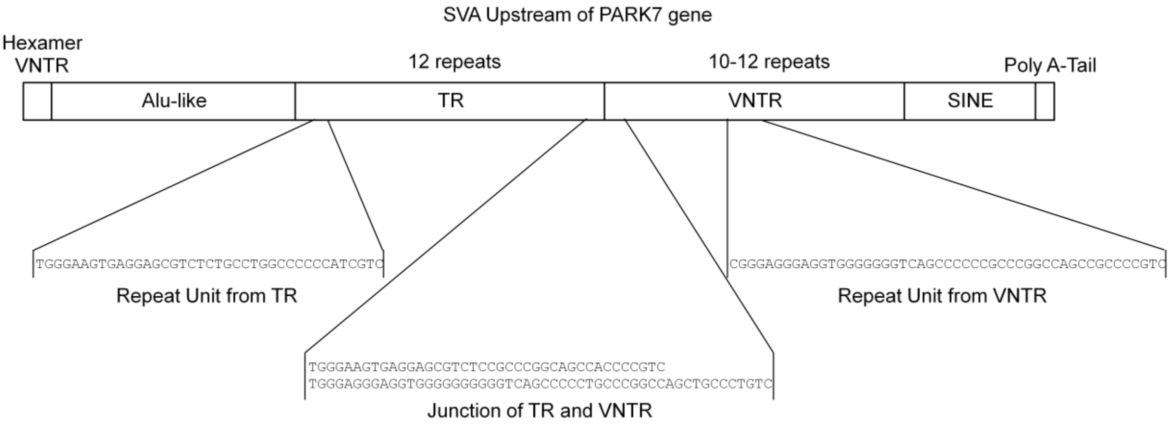

SVA Upstream of PARK7 gene
Hexamer VNTR
12 repeats
10-12 repeats
Poly A-Tail
Alu-like
TR
VNTR
SINE
TGGGAAGTGAGGAGCGTCTCTGCCTGGCCCCCCATCGTC
Repeat Unit from TR
CGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
Repeat Unit from VNTR
TGGGAAGTGAGGAGCGTCTCCGCCCGGCAGCCACCCCGTC
TGGGAGGGAGGTGGGGGGGGGTCAGCCCCCTGCCCGGCCAGCTGCCCTGTC
Junction of TR and VNTR

Figure 6.1: Schematic of the PARK7 SVA and sequences used in the Blat search analysis. The structure of the PARK7 SVA and the sequence of the repeat units used in the Blat search analysis.

6.3.2 Pathway analysis of genes containing SVAs with similar repeat unit sequence to the PARK7 SVA

Using the analytical program MetaCore from Thomson Reuters (https://portal.genego.com/) the list of genes generated from the search in 6.3.1 was analysed for potential regulatory factors and interactions within pathways. The MetaCore software can be used to analyse experimental data or gene lists to look for protein-protein, protein-DNA and protein-compound interactions. It highlights pathways that may be enriched for your genes of interest how they are interacting with one another.

The gene list was uploaded and analysed using the transcription regulation tool which generated a list of transcription factors that regulated or were regulated by the list of genes generated. The list was also analysed to determine the pathway these genes are linked in using the build network tool. This tool allows the user to analyse direct interactions of their chosen genes or indirect interactions that may involve one or more factors separating your genes of interest. For this particular list of genes the pathway analysis was completed for interactions within two steps, therefore there would be at most one factor between the genes in the network. The resulting network would not be too complex for analysis but would allow identification of common pathways involving these genes.

To complement the transcription factor regulation data for the gene lists the sequence of the PARK7 SVA was analysed for predicted transcription factor binding sites using PROMO software to determine if any of these factors may overlap. The PROMO program available on the internet (http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3) and is a virtual laboratory for the study of transcription factor binding sites in DNA sequences and uses the TRANSFAC 8.3 database (Messeguer et al. 2002; Farre et al. 2003).

6.3.3 Analysis of endogenous gene response to 17-β-estradiol in the MCF-7 cell line

MCF-7 cells (low passage, see chapter 5) were maintained in normal media in T175 flasks (2.1.4.2). For treatment with 17-β-estradiol the growth media of the cells was replaced with phenol red free and charcoal stripped media (2.1.4.3) for two days. The cells were then counted (2.2.2.2) and plated into 6-well plates (400000 cells per well) to complete the experiment in triplicate and left for 24hrs. The media was then changed for each well and replaced with either fresh media, media containing 1µl/ml of molecular grade ethanol or media with 10nM of 17-β-estradiol (dissolved in molecular grade ethanol). 1µl/ml of molecular grade ethanol was used as this was the same volume that was added for the 17-β-estradiol to test the effect of the vehicle of the stimulus. The cells were then incubated for 18hrs and RNA extracted for analysis of endogenous gene expression (2.2.3). Several genes were analysed including the housekeeping gene β-actin, a known oestrogen responsive gene trefoil factor 1(TFF1), and test genes of interest (PARK7, MAN1C1 and TFPI). GoTaq flexi polymerase was used in the amplification of these targets (2.2.3.3.2) and the conditions for each primer set are detailed in Table A1 of the appendix. The PCR products were run on 1.2-1.5% agarose gels for analysis (2.2.3.4).

6.3.4 Analysis of the response of reporter gene plasmids to 17-β-estradiol in the MCF-7 cell line

The MCF-7 cells were grown in normal media (2.1.4.2) in T175 flasks until two days prior to being plated into 24 well plates the media was changed for phenol red free media with charcoal stripped serum, to remove steroid hormones (2.1.4.3) which the cells were maintained in for the remainder of the experiment. The cells were transfected and the dual luciferase assay performed as outlined in 2.2.6. However for the 17-β-estradiol treatment when the media was to be changed 4hrs after transfection it was replaced with either fresh media for the basal condition, media containing 1µl/ml of molecular grade ethanol for the vehicle control or media with 10nM of 17-β-estradiol (dissolved in molecular grade ethanol) and incubated with the cells for 18hrs. The media was then changed again after 18hrs with fresh media and the dual luciferase assay was performed 48hrs after the cells had been transfected.

6.3.5 Genotyping the PARK7 SVA in a breast cancer cohort without BRCA mutations

The PARK7 SVA was genotyped in the breast cancer BRCA wild-type and matched female control cohort from the National Genetics References Laboratory, St Mary's Hospital, Manchester with 5ng of gDNA as template. The controls were sex matched to the breast cancer samples and screened for BRCA mutations. The final concentration of betaine for this PCR was optimised at 1M, which was higher than previously used. The data set was also analysed to determine if the PARK7 SVA locus was in HWE.

6.3.6 Identification of tagging SNPs for the PARK7 SVA alleles in the HapMap cohort

In the identification of tagging SNPs for the PARK7 SVA alleles the same method was used as for the FUS loci for SNPs that would tag the FUS SVA (section 3.3.4). The genotype of the individuals from the CEU HapMap cohort for the PARK7 SVA was used in conjunction with the SNP data available for the PARK7 locus. Four 'SNPs' were created for the four alleles of the PARK7 SVA and the genotype of each individual for the PARK7 SVA was converted into the genotypes of these 'SNPs'. The genotype data for the individuals in the CEU HapMap cohort for the SNPs 200kb either side of the PARK7 SVA were downloaded from the International HapMap database (http://hapmap.ncbi.nlm.nih.gov/) using release 28. The 'SNP' genotypes for the PARK7 SVA were inserted into this data at the locus corresponding to the SVA. This was then uploaded into Haploview for linkage analysis. All the genotyped individuals' data for all four alleles of the SVA were uploaded and then separately the data for just the individuals that had allele 1 and/or 3 of the SVA (the most common alleles).

6.3.7 Identification of Parkinson's disease causing genes containing a SVA within 10kb

A list of Parkinson's disease associated genes was taken from table 1 of Corti et al 2011 and UCSC genome browser was used to identify if a SVA was present within the gene or up to 10kb in its flanking region. The SVA subtype and relative position and orientation to the gene were recorded.

6.3.8 Amplification of the PARK2 SVA

The SVA F within an intron of the PARK2 gene was amplified using GoTaq polymerase (master mix outlined in 2.2.3.3.2) with the addition of betaine at a final concentration of 1M. 10ng of gDNA from the CEU HapMap cohort was used as template in the reaction. The primer set and cycling conditions can be found in Table A1 of the appendix. The PCR products were run on a 1% agarose gel (2.2.3.4).

6.3.9 Amplification of the LRRK2 SVA

KOD Hot Start polymerase (2.2.4.1) was used to amplify the SVA C located within an intron of the LRRK2 gene with the addition betaine to a final concentration of 1M. 1ng of gDNA from individuals in the CEU HapMap cohort was used as template. The PCR products were run on a 1% agarose gel (2.2.3.4). The primers and cycling conditions are shown in Table A1 of the appendix.

6.3.10 Amplification of the TAF1 SVA

The SVA located within intron 32 of the TAF1 gene that is present in the UCSC genome browser was amplified using KOD Hot Start polymerase as outlined in 2.2.4.1 however the $MgSO_4$ was optimised to a final concentration of 2mM and betaine was added to a final concentration of 1M. 1ng of gDNA from the CEU HapMap cohort was used as template for the reaction. The PCR products were run on a 1% agarose gel (2.2.3.4). The primers and cycling conditions are shown in Table A1 of the appendix.

6.4 Results

Section A

6.4.1 List of genes within 100kb of a SVA sharing similar primary sequence to specific repeat units of the TR/VNTR of the PARK7 SVA

As outlined in section 6.3.1 a list of 19 SVAs with repetitive central domains that shared homology to specific repeats of the TR/VNTR of the PARK7 SVA were identified and the genes within 100kb of these SVAs were recorded and are shown in Table 6.1. Regulatory domains for genes can be located considerable distances from the genes they regulate (Shanley et al. 2010; Shanley et al. 2011), however the nearest gene within 100kb of the SVA up and downstream was chosen as this would provide a manageable number of genes for this preliminary analysis and would include the genes closest to the SVA. The majority of these SVAs were from the subtype D with three from each subtype E and F. All except one of these SVAs were human specific; the SVA within ZNF670-ZNF695 was also found in the chimpanzee genome according to the UCSC genome browser. The length of the repetitive domains between each of these individual SVAs varied, this is not a polymorphic range but the differences in size of SVA insertions at different loci. The hexamer repeat at the 5' end of each of the SVAs ranged from 4-20 repeat units in length with three SVAs without a hexamer domain. The domain termed a TR in the PARK7 SVA ranged from 10-13 repeat units and the domain termed a VNTR ranged from 5-25 repeat units in length across the 19 SVAs. There were 32 genes identified that were within 100kb of the 19 SVAs and the SVAs were found in a range of distances up or downstream of the genes and intronically. For simplicity this list of genes will be referred to as the SVA gene list.

Chr	Start	Stop	Size	Subtype	Gene	Upstream	Downstream	Intronic
1	**8012111**	**8013640**	**1529**	**SVA D (+)**	**PARK7 (+)** **TNFRSF9 (-)**	**8kb** **9kb**	**-** **-**	**-** **-**
1	21141242	21143749	2507	SVA E (+)	EIF4G3 (-) HP1BP3 (-)	- 28kb	- -	Y -
1	25929721	25931433	1712	SVA D (+)	MAN1C1 (+) LDLRAP1 (+)	12kb -	- 34kb	- -
1	26921936	26924007	2071	SVA D (+)	RPS6KA1 (+) ARID1A (+)	- 99kb	20kb -	- -
1	37545339	37547744	2405	SVA D (+)	GRIK3 (-)	45kb	-	-
1	39699459	39701100	1641	SVA F (-)	MACF1 (+)	-	-	Y
1	46189629	46191758	2129	SVA E (+)	IPP (-) TMEM69 (+) GPBP1L1 (-)	- - 37kb	- 30kb -	Y - -
1	63938105	63940580	2475	SVA D (+)	ITGB3BP (-) ALG6 (+) EFCAB7 (+)	- - 48kb	- 34kb -	Y - -
1	112149661	112151498	1837	SVA D (-)	RAP1A (+) ADORA3 (-)	11kb of transcripts 1 and 3 43kb	- -	Y transcript 2 -
1	170387365	170389034	1669	SVA D (+)	-	-	-	-
1	176364624	176366749	2125	SVA D (-)	PAPPA2 (+)	-	66kb	-
1	206276044	206278006	1962	SVA D (-)	C1orf186 (-) CTSE (+) AVPR1B (+)	- 39kb -	- - 45kb	Y - -
1	227897703	227899534	1831	SVA D (-)	ZNF678 (+) SNAP47 (+) JMJD4 (-)	- 17-23kb -	48kb - 19kb	Y - -
1	247194849	247196269	1420	SVA D (+)	ZNF670-ZNF695 (-) ZNF670 (-) ZNF695 (-)	- - 23kb	- 2kb -	Y - -
2	11042073	11044034	1961	SVA D (-)	KCNF1 (+) PDIA6 (-)	8kb 64-89kb	- -	- -
2	42831720	42834257	2537	SVA F (-)	MTA3 (+)	-	-	Y
2	63948844	63950437	1593	SVA F (+)	-	-	-	-
2	188420653	188422627	1974	SVA D (+)	TFPI (-)	1kb	-	-
4	109329937	109331742	1805	SVA E (+)	-	-	-	-

Table 6.1: Table showing the SVAs that share the repeat sequence from the TR and VNTR of the PARK7 SVA and the genes that are within 100kb of these SVAs. This includes the chromosomal location, size, subtype of the SVA and the relative loci of the genes within 100kb of these genes. The SVAs without a gene listed are not within a 100kb of a gene. (+) – located on sense strand, (-) – located on antisense strand. See list below for full name of each gene.

Gene ID	**Gene Name**
PARK7 (DJ-1)	Parkinson protein 7
TNFRSF9	tumour necrosis factor receptor superfamily, member 9
EIF4G3	eukaryotic translation initiation factor 4 gamma 3
HP1BP1	heterochromatin protein 1 binding protein 3
MAN1C1	mannosidase, alpha, class 1C, member 1
LDLRAP1 (ARH)	low density lipoprotein receptor adaptor protein
RPS6KA1 (p90RSK1)	ribosomal protein S6 kinase, 90kDa, polypeptide 1
ARID1A (BAF250)	AT rich interactive domain 1A (SWI-like)
GRIK3 (GluR7)	glutamate receptor, ionotropic, kainate 3
MACF1	microtubule-actin crosslinking factor 1
IPP	intracisternal A particle-promoted polypeptide
TMEM69	transmembrane protein 69
GPBP1L1	GC rich promoter binding protein 1 like 1
ITGB3BP (NRIF3)	integrin beta 3 binding protein (beta3-endonexin)
ALG6	alpha 1,3 glucosyltransferase,
EFCAB7	EF hand calcium binding domain 7
RAP1A	member of RAS oncogene family
ADORA3	adenosine A3 receptor
PAPPA2	pappalysin 2
C1orf186	chromosome 1 open reading frame 186
CTSE	cathepsin E
AVPR1B	arginine vasopressin receptor 1B
ZNF678	zinc finger protein 678
SNAP47	synaptosomal-associated protein, 47kDa
JMJD4	jumonji domain containing 4
ZNF670-ZNF695	ZNF670-ZNF695 readthrough
KCNF1 (Kv5.1)	potassium voltage-gated channel, subfamily F, member 1
PDIA6 (ERP5)	protein disulfide isomerase family A, member 6
MTA3	metastasis associated 1 family, member 3
TFPI	tissue factor pathway inhibitor

6.4.2 Transcriptional regulation and pathway analysis of the genes from the SVA gene list

The MetaCore program from Thomson Reuters was used to determine if the genes from the SVA gene list were linked by similar transcriptional regulation or within specific pathways. The analysis of the transcriptional regulation of these genes generated a list of 31 networks involving transcription factors and the genes uploaded. The top five of these networks were listed in Table 6.2 and include the transcription factors CREB1, ESR1, Sp1, GCR alpha and PR. It is interesting to note that three of these factors (ESR1, PR and GCR alpha) are regulated through hormones and that CREB1 can induce transcription of genes in response to hormonal stimulation of the cAMP pathway. 15 genes from the SVA gene list are regulated by these factors or regulate the factors themselves and are listed in the final two columns of Table 6.2.

The SVA gene list was then analysed to determine pathways linking these genes within two steps of each other. 14 of these genes were identified to be linked through pathways using this analysis and are shown circled in blue in Figure 6.2. Two of the transcription factors listed in table 6.2 (CREB1 and ESR1) are also shown in this pathway.

The repetitive nature of the sequence of SVAs could provide multiple sequence specific transcription factor binding sites within the genome if the region was not ‘silenced’ and available for interaction with such transcription factors. Therefore these SVAs could potentially be involved in gene regulation through the binding of these factors. The sequence of the PARK7 SVA was analysed using the PROMO software to predict transcription factor binding sites (Figure 6.3A). 62

different transcription factors had predicted binding sites within the PARK7 SVA sequence; many of these had multiple predicted binding sites. Four of the top five transcription factors regulating the genes from the SVA gene list (Table 6.2) had predicted binding sites within the SVA sequence and are boxed in red (Figure 6.3A) (ESR1, Sp1, GCR alpha and PR). The specific sequence of the three repeat units used in the Blat search (section 6.3.1) were also analysed for predicted transcription factor binding (Figure 6.3B) and 19 factors had predicted binding sites. Sp1, ESR1 and GCR alpha (boxed in red in Figure 6.3B) that regulate the genes from the SVA gene list were found to potentially bind to the SVA sequence.

This bioinformatic analysis of factors that regulate a number of genes from the SVA gene list and predicted transcription factor binding sites within the SVA itself provided a list of factors that may regulate these genes and have the potential to alter transcription through binding to the SVA. The potential for ESR1 to regulate both the endogenous PARK7 gene and the ability of the PARK7 SVA to alter transcription in a reporter gene model was chosen out of the list of transcription factors (Table 6.2) to be investigated in a cell line model. ESR1 was chosen for several reasons: it was shown to regulate several genes on the SVA gene list and there are predicted half sites of the palindromic oestrogen response elements (ERE) in the sequence of the SVA, the response of MCF-7 cells to oestrogen is widely used in the literature as they are oestrogen responsive and PARK7 is reported in the literature to be associated with breast cancer therefore the oestrogen pathway could potentially be a important in the modulation of this gene.

	Network	Total nodes	Seed nodes	p-Value	Genes regulated by TF	Genes regulating the TF
1	CREB1	9	8	1.770E-27	DJ-1, MTA3, p90RSK1, GluR7, ADORA3, JMJD4, Kv5.1, PAPPA2	p90RSK1
2	ESR1 (nuclear)	7	6	1.550E-20	ERP5, MTA3, GluR7, ADORA3	p90RSK1, NRIF3
3	Sp1	5	4	1.020E-13	DJ-1, MTA3, ADORA3, CTSE	-
4	GCR-alpha	4	3	2.300E-10	ADORA3, TFPI, AVPR1B	-
5	PR (nuclear)	4	3	2.300E-10	MAN1C1, ADORA3, MACF-1	-

Table 6.2: Data generated on the transcriptional regulation of the gene list from table 6.1 using MetaCore. The table lists the top five transcription factors that regulate or are regulated by the genes generated by sequence analysis of the PARK7 SVA TR/VNTR when analysed using MetaCore. CREB1 - cAMP responsive element binding protein 1, Sp1 – specificity protein 1, ESR1 – oestrogen receptor 1, GCR alpha – glucocorticoid receptor alpha, PR – progesterone receptor, TF – transcription factor. PARK7 is referred to as DJ-1 in the MetaCore software.

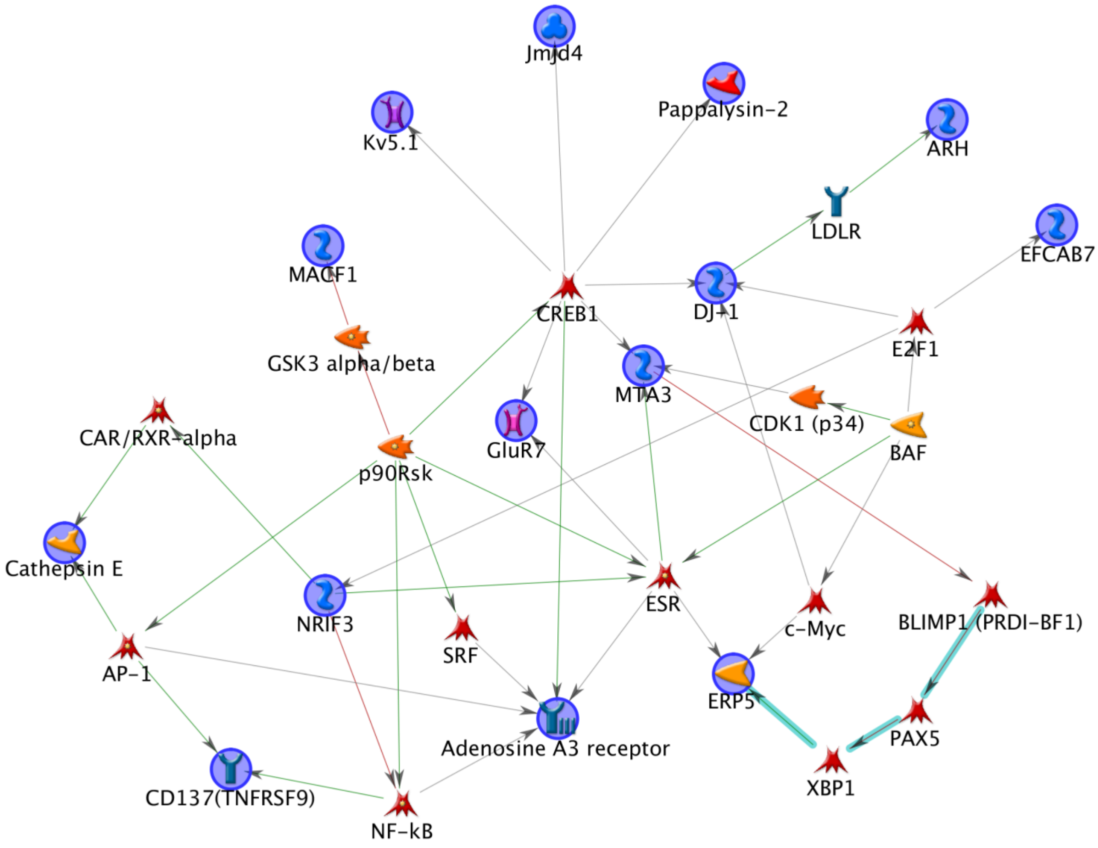
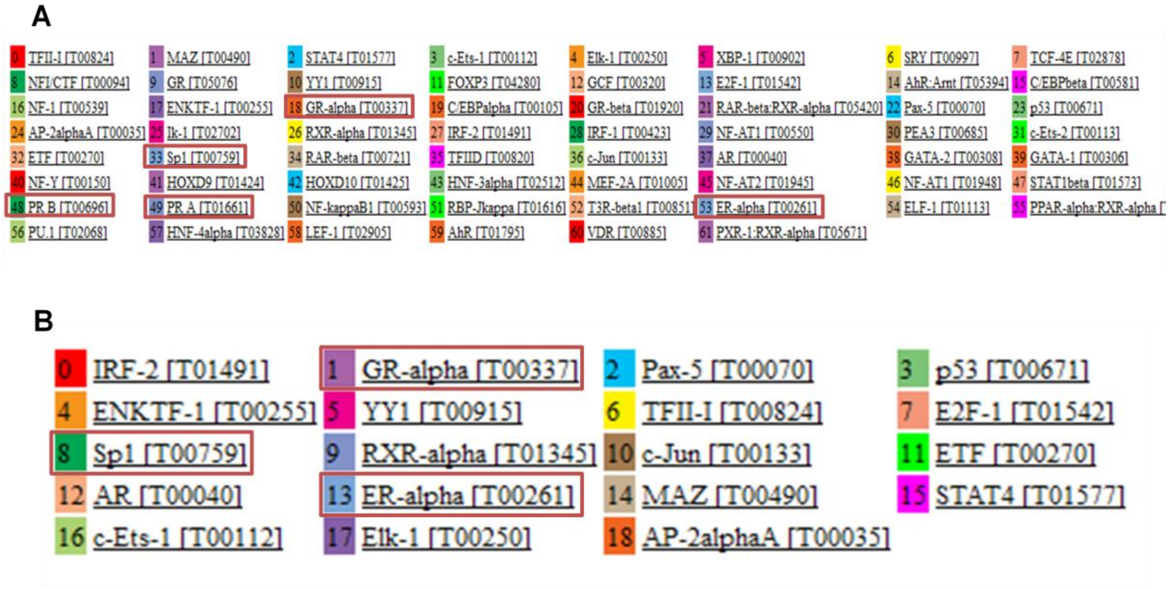


Figure 6.2: Pathway analysis for the list of genes generated from the SVAs sharing similar repeat units with the PARK7 SVA TR/VNTR. The list of genes was analysed to determine the shortest pathway within two steps of each other. The genes from Table 6.1 are circled in blue with 14 out of 32 linked within this pathway. Two of the top five transcription factors from table 6.2 (CREB1 and ESR) are found within this network.



0 TFII-I [T00824]	1 MAZ [T00490]	2 STAT4 [T01577]	3 c-Ets-1 [T00112]	4 Elk-1 [T00250]	5 XBP-1 [T00902]	6 SRY [T00997]	7 TCF-4E [T02878]
8 NFI/CTF [T00094]	9 GR [T05076]	10 YY1 [T00915]	11 FOXP3 [T04280]	12 GCF [T00320]	13 E2F-1 [T01542]	14 AhR:Arnt [T05394]	15 C/EBPbeta [T00581]
16 NF-1 [T00539]	17 ENKTF-1 [T00255]	18 GR-alpha [T00337]	19 C/EBPalpha [T00105]	20 GR-beta [T01920]	21 RAR-beta:RXR-alpha [T05420]	22 Pax-5 [T00070]	23 p53 [T00671]
24 AP-2alphaA [T00035]	25 Ik-1 [T02702]	26 RXR-alpha [T01345]	27 IRF-2 [T01491]	28 IRF-1 [T00423]	29 NF-AT1 [T00550]	30 PEA3 [T00685]	31 c-Ets-2 [T00113]
32 ETF [T00270]	33 Sp1 [T00759]	34 RAR-beta [T00721]	35 TFIID [T00820]	36 c-Jun [T00133]	37 AR [T00040]	38 GATA-2 [T00308]	39 GATA-1 [T00306]
40 NF-Y [T00150]	41 HOXD9 [T01424]	42 HOXD10 [T01425]	43 HNF-3alpha [T02512]	44 MEF-2A [T01005]	45 NF-AT2 [T01945]	46 NF-AT1 [T01948]	47 STAT1beta [T01573]
48 PR B [T00696]	49 PR A [T01661]	50 NF-kappaB1 [T00593]	51 RBP-Jkappa [T01616]	52 T3R-beta1 [T00851]	53 ER-alpha [T00261]	54 ELF-1 [T01113]	55 PPAR-alpha:RXR-alpha [
56 PU.1 [T02068]	57 HNF-4alpha [T03828]	58 LEF-1 [T02905]	59 AhR [T01795]	60 VDR [T00885]	61 PXR-1:RXR-alpha [T05671]		

0 IRF-2 [T01491]	1 GR-alpha [T00337]	2 Pax-5 [T00070]	3 p53 [T00671]
4 ENKTF-1 [T00255]	5 YY1 [T00915]	6 TFII-I [T00824]	7 E2F-1 [T01542]
8 Sp1 [T00759]	9 RXR-alpha [T01345]	10 c-Jun [T00133]	11 ETF [T00270]
12 AR [T00040]	13 ER-alpha [T00261]	14 MAZ [T00490]	15 STAT4 [T01577]
16 c-Ets-1 [T00112]	17 Elk-1 [T00250]	18 AP-2alphaA [T00035]	

Figure 6.3: Predicted transcription factor binding sites within the sequence of the PARK7 SVA using PROMO software. A – These are the transcription factor binding sites that were predicted within the sequence of the complete PARK7 SVA (allele 1). B – The predicted transcription factor binding sites of the sequence of the three repeat units of the TR/VNTR used in the Blat search. In both A and B the transcription factors that were identified in Table 6.2 are highlighted with a red box.

6.4.3 Endogenous PARK7 mRNA expression is modulated by 17-β-estradiol in a MCF-7 cell line

The expression levels of several of our target genes were analysed in response to 17-β-estradiol in a MCF-7 cell line. The genes of interest were amplified with specific primers sets and the PCR products run on an agarose gel (Figure 6.4). The housekeeping gene, β-actin, expression remained stable across the three conditions; basal, vehicle control (ethanol) and 10nM 17-β-estradiol (bottom panel of Figure 6.4). The gene TFF1 (trefoil factor-1) is a known oestrogen responsive gene in the MCF-7 cell line in the literature (Amiry et al. 2009) and was therefore used as a positive control for this experiment. The increase in expression of TFF1 in response to 10nM 17-β-estradiol can be seen in the top panel of Figure 6.4. The expression of PARK7 and MAN1C1 was also increased in response to exposure to 10nM 17-β-estradiol but not that of TFPI (see middle three panels of figure 6.4). The expression of the genes MAN1C1 and TFPI from the SVA gene list were chosen to be analysed due to location of their SVA relative to the gene, 12kb and 1kb upstream respectively, that was similar to the position of the PARK7 SVA relative to the PARK7 gene (1-8kb upstream).

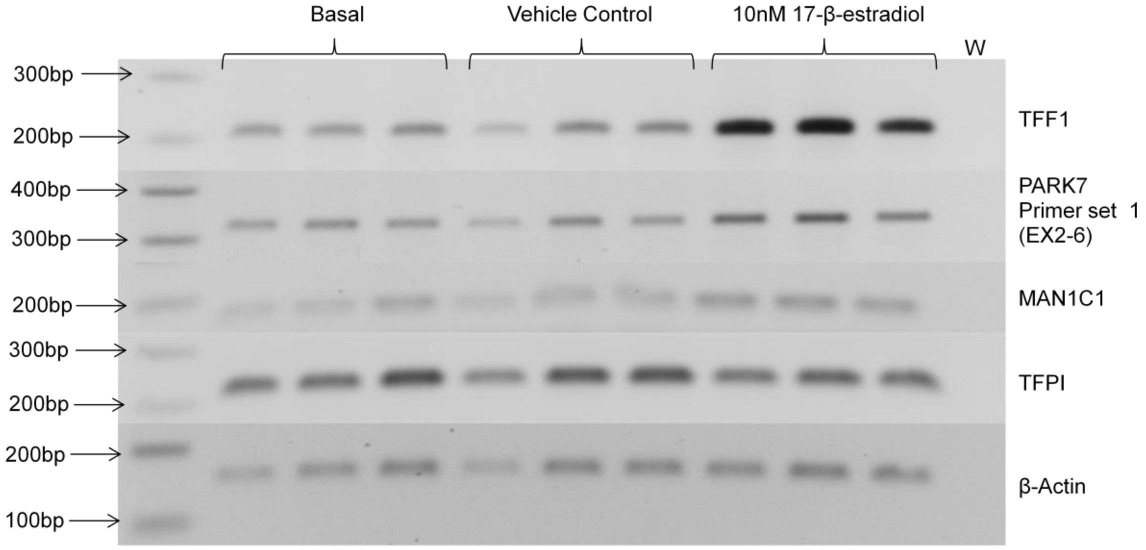


Figure 6.4: The endogenous levels of TFF1, PARK7 and MAN1C1 mRNA are upregulated in the MCF-7 cell line in response to oestrogen. Agarose gel electrophoresis images of the PCR products from the amplification of the following genes TFF1 (209bp), PARK7 primer set 1 (325bp), MAN1C1 (159bp), TFPI (218bp) and β-actin (158bp) under different conditions; basal, vehicle control (ethanol) and 10nM 17-β-estradiol. N=3

6.4.4 PARK7 SVA does not respond to oestrogen in a transient transfection reporter gene assay

An ERE-TK Luc vector and the control vector (TK Luc) were provided by Karen Chapman (Martin et al. 2004) and were used in the validation of the oestrogen response of a transiently transfected vector in the MCF-7 cell line. The ERE-TK Luc vector had significantly increased activity over the minimal promoter alone under basal conditions ($P<0.01$). When exposed to the vehicle control (ethanol) the activity of the ERE-TK Luc vector over the minimal promoter alone did not significantly increase when compared to the vector under basal conditions. When the MCF-7 cells were exposed to 10nM 17-β-estradiol the activity of the ERE-TK Luc vector significantly increased by 8.3 fold when compared to the activity of the vector exposed to the vehicle control ($P<0.001$).

The different sized fragments of the PARK7 SVA in the forward orientation used in 5.4.8 were tested in this validated MCF-7 cell model response to 17-β-estradiol. There was no significant difference in the activity of these reporter gene constructs containing the different domains of the PARK7 SVA in response to stimulation of the MCF-7 cells with 10nM 17-β-estradiol.

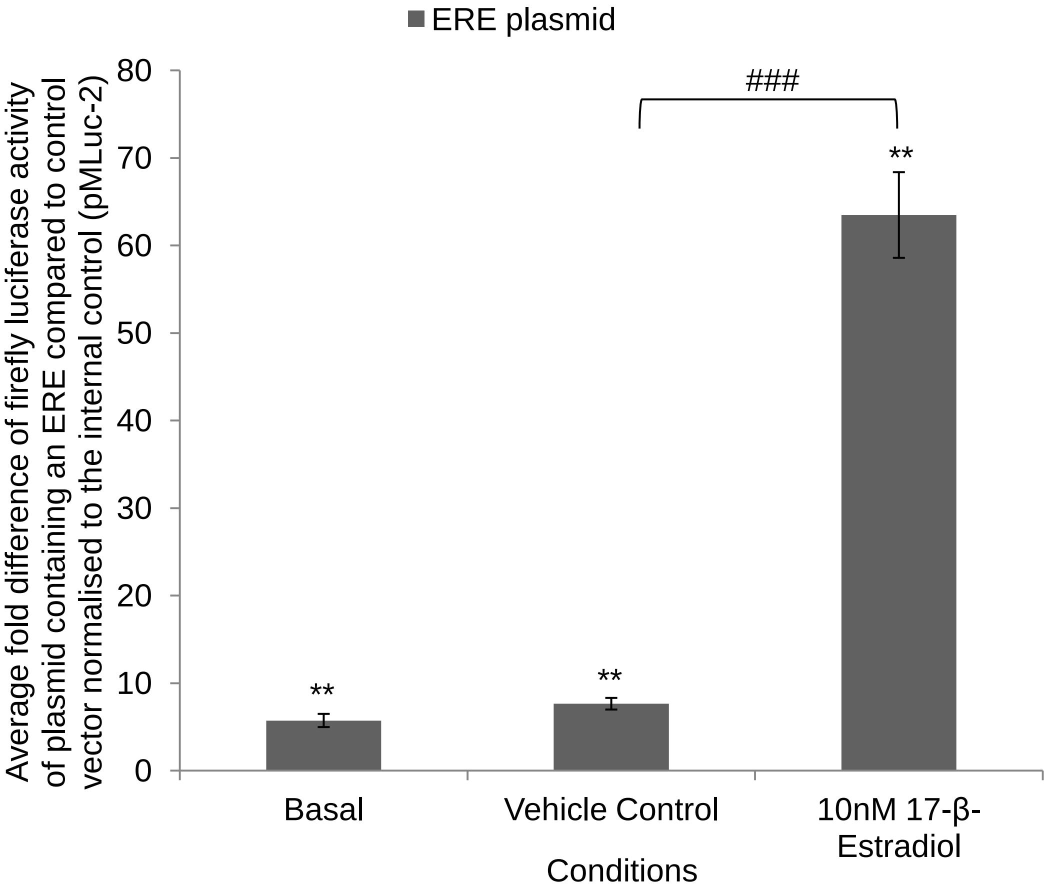


Figure 6.5: The activity of the ERE containing plasmid increases after MCF-7 cells are exposed to 10nM 17-β-estradiol. The average fold activity of the ERE containing vector (ERE-TK Luc) over the TK promoter alone (TK Luc) under different conditions: basal, vehicle control (ethanol) and 10nM 17-β-estradiol. Data was normalised to the internal control to compensate for transfection efficiency. One tailed t-test was used to measure significance of fold activity of ERE over TK promoter alone (TK Luc) **P<0.01, ***P<0.001 and to compare the fold activity of ERE-TK Luc across different conditions ### P<0.001. N=4

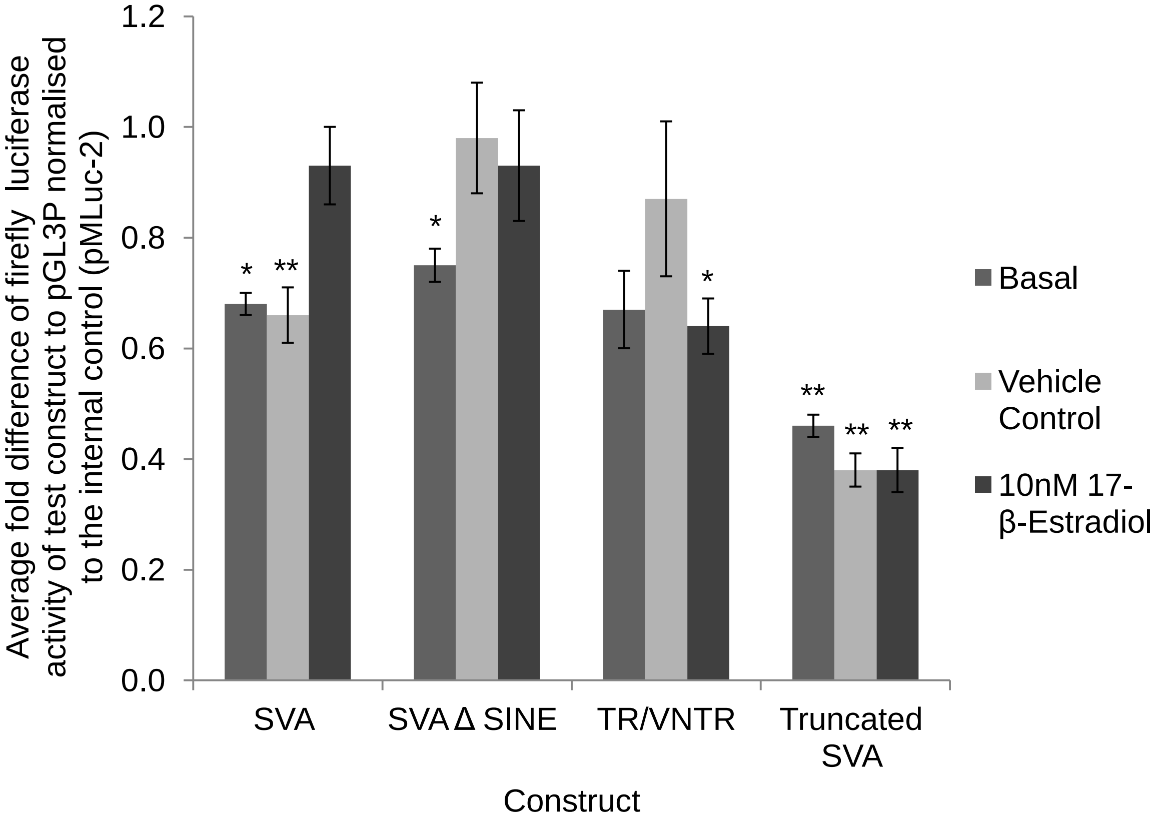


Figure 6.6: There was no significant difference in the activity of the reporter gene constructs containing the different domains of the PARK7 SVA in response to 10nM 17-β-estradiol. The average fold activity of the different fragments of the PARK7 SVA in forward orientation over the minimal SV40 promoter alone (pGL3P) under different conditions: basal, vehicle control (ethanol) and 10nM 17-β-estradiol. The activity of the test constructs were normalised to the internal control to account for transfection efficiency. N=4. One tailed t-test was used to measure significance of fold activity of PARK7 SVA fragments over SV40 minimal promoter alone (pGL3P) and to compare fold activity across the different conditions. * $P<0.05$, **$P<0.01$, N=4

6.4.5 In a preliminary analysis there was no association of the PARK7 SVA genotype with a BRCA wild-type breast cancer cohort

The PARK7 SVA was genotyped in a cohort of BRCA wild-type breast cancer patients and matched controls as PARK7 has been linked to breast cancer (Le Naour et al. 2001; Oda et al. 2012). The frequencies of the different genotypes between the breast cancer patients and controls were not significantly different when analysed using CLUMP (T1 P=0.65 and T4 P=0.65) and are shown in Table 6.3. The most notable difference between the two groups was a 7.2% reduction in the 1/1 genotype from the control group compared to the breast cancer patients.

The PARK7 SVA had also been genotyped in the control samples from the King's cohort used in chapter 3 for studies of the FUS SVA and its association to ALS and it was noted that there were differences between the all female controls from the breast cancer cohort and the King's control cohort of mixed gender. Therefore the sexes of the individuals from the King's control cohort were obtained from collaborators at King's College London who had provided the samples and this cohort was broken down into females and males (Table 6.3). There were differences between the frequencies of the genotypes of the PARK7 SVA across the four groups but most notably were the differences between the female groups and the male group. In particular was the underrepresentation of genotype 1/3 in the King's control male group of 28.7% compared to 41.1%, 46.5% and 40.4% in the three female groups. The frequency of genotype 3/3 was higher in the male group (25.3%) compared to the three female groups (19.9%, 19.1% and 15.4%). Also an allele 5, larger than allele 4, had been identified in this King's control cohort and all three individuals with this previously unidentified allele were male. The frequency of the alleles between the four groups were then compared (Table 6.4). Despite the differences in genotype frequencies the frequency of the alleles was much more similar between males and females. There was a slightly lower frequency of allele 2 (3.4%) in the males compared to the three female groups, (6%, 7.2% and 5.8%). It was also noted than the frequency of allele 1 (46.5%) in the BRCA wild-type breast cancer group was lower than that of the other three groups (50.7%, 51.9% and 51.1%).

These unexpected differences between males and females for the genotypes of the PARK7 SVA were not significant when the frequencies for the males and females of the King's control cohort were analysed using CLUMP (T1 P=0.11 and T4 P=0.16). These are still relatively small cohorts and if this difference between males and females was to be validated, larger cohorts would be needed and is not clear why this difference may occur. The genotyping of such large pieces of DNA is time consuming and labour intensive so if a tagging SNP could be identified for the different alleles of the SVA larger cohorts could be analysing in a more efficient way. This had already been achieved for the two alleles of the FUS SVA in chapter 3, however the larger number of alleles of the PARK7 SVA would make it more difficult to attempt to identify tagging SNPs.

HWE was analysed for all four cohorts mentioned in this section and all were found to be in HWE with the following p values: female breast cancer control p=0.30, BRCA wildtype p=0.72, females in ALS control p=0.88 and males in ALS control p=0.37.

	Frequency of Genotype (%)			
Genotype	**Female Controls (151)**	**BRCA Wild-type Breast Cancer (188)**	**Females from ALS control cohort (104)**	**Males from ALS control cohort (87)**
1/1	25.8	18.6	28.8	33.3
1/2	6.0	7.4	2.9	4.6
1/3	*41.1*	*46.3*	*40.4*	*28.7*
1/4	2.6	2.1	2.9	2.3
2/2	1.3	1.1	1.0	0
2/3	2.6	4.8	6.7	2.3
2/4	0.7	0	0	0
3/3	*19.9*	*19.1*	*15.4*	*25.3*
3/4	0	0.5	1.9	0
3/5	0	0	0	*3.4*

Table 6.3: Frequencies of the PARK7 genotypes in a BRCA wild-type breast cancer and control cohort. The frequencies are shown of the different genotypes of the PARK7 SVA in the breast cancer BRCA wild-type and control cohort that are all female and the frequencies in the control samples from the King’s control cohort from chapter 3 broken down by sex to compare the differences between the two sexes.

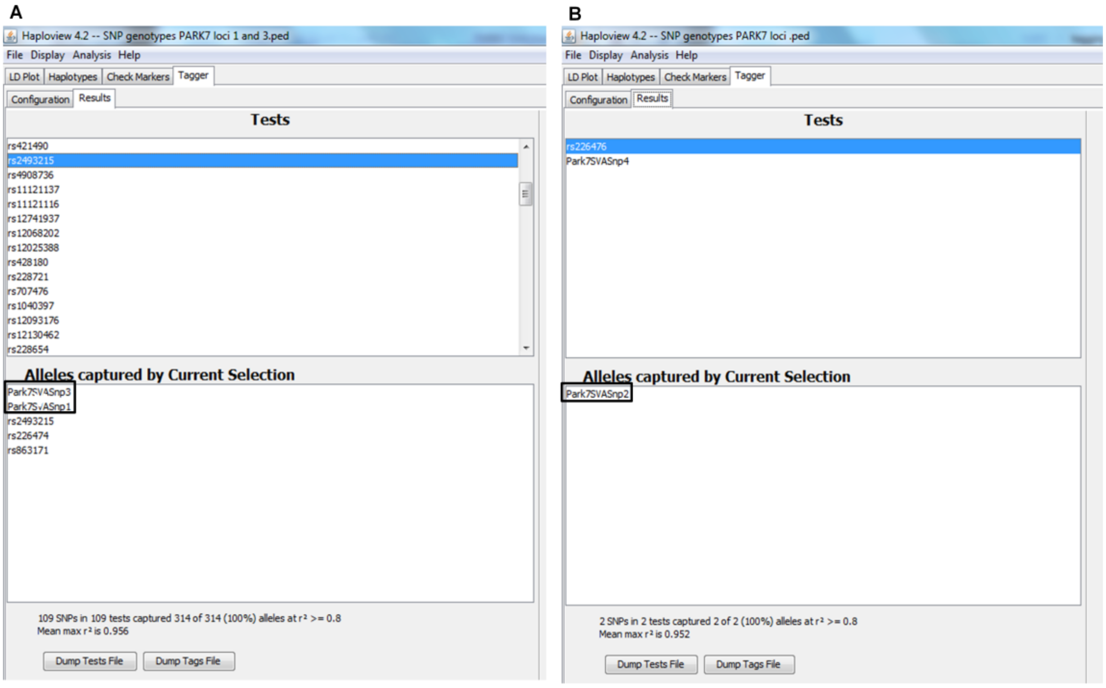
	Frequency of Alleles (%)			
Allele	**Female Controls (302)**	**BRCA Wild-type Breast Cancer (376)**	**Females from ALS control cohort (208)**	**Males from ALS control cohort (174)**
1	50.7	*46.5*	51.9	51.1
2	6.0	7.2	5.8	*3.4*
3	41.7	44.9	39.9	42.5
4	1.7	1.3	2.4	1.1
5	0	0	0	*1.7*

Table 6.4: Frequencies of the PARK7 alleles in the BRCA wild-type breast cancer and control cohort. The frequencies are shown of the different alleles of the PARK7 SVA in the breast cancer BRCA wild-type and control cohort that are all female and the frequencies in the control samples from the ALS cohort from chapter 4 broken down by sex to compare the differences between the two sexes. All four cohorts were shown to be in HWE with the following p values corresponding to the cohorts in the table from left to right: p=0.30, p=0.72, p=0.88 and p=0.37.

6.4.6 Two tagging SNPs can be used to identify alleles 1, 2 and 3.

The PARK7 SVA has more than two alleles so more than one SNP would be required to tag the alleles. To identify tagging SNPs the genotyping of the CEU HapMap cohort (section 5.4.6) was used and the SNP data for that region of those individuals which was freely available from the International HapMap database. This however would limit the potential identification of tagging SNPs for alleles 1-4 as allele 5 was not present in this cohort (for methods see section 3.3.4 and 6.3.6).

Analysis of linkage in the PARK7 region revealed two SNPs that in conjunction can be used to identify the genotype of the SVA for alleles 1, 2 and 3, however a SNP to tag allele 4, the rarest of the alleles, could not be identified. The analysis of the individuals in the CEU HapMap cohort with alleles 1 and/or 3, the most common of the alleles, determined at the SNP rs2493215 a genotype of G corresponded to allele 1 and a genotype of A corresponded to allele 3 with a r^2 of 0.909 (Figure 6.7A). When the analysis was expanded to include individuals with allele 2 and 4 it was shown that the A genotype corresponded to these alleles as well. The genotype data was analysed further in Haploview software to include all four alleles of the PARK7 SVA, which identified a SNP (rs226476) that would tag allele 2 with a r^2 of 0.903 (Figure 6.7B); a genotype of T corresponds to allele 2 and a genotype of G corresponds to alleles 1, 3 and 4. Therefore using these two SNPs in combination alleles 1, 2 and 3 can be tagged by their specific genotypes. This work was ongoing at the end of my thesis by others in the group as a method to more rapidly link SVA polymorphisms to the vast SNP data available and therefore to disease pathways.



A

Tests

```
rs421490
rs2493215
rs4908736
rs11121137
rs11121116
rs12741937
rs12068202
rs12025388
rs428180
rs228721
rs707476
rs1040397
rs12093176
rs12130462
rs228654
```

Alleles captured by Current Selection

```
Park7SVASnp3
Park7SVASnp1
rs2493215
rs226474
rs863171
```

```
109 SNPs in 109 tests captured 314 of 314 (100%) alleles at r² >= 0.8
Mean max r² is 0.956
```

B

Tests

```
rs226476
Park7SVASnp4
```

Alleles captured by Current Selection

```
Park7SVASnp2
```

```
2 SNPs in 2 tests captured 2 of 2 (100%) alleles at r² >= 0.8
Mean max r² is 0.952
```

Figure 6.7: The genotype of the SNPs rs2493215 and rs226476 can be used to determine the genotype of the PARK7 SVA alleles 1, 2 and 3. A – A screen shot from Haploview software showing that the SNP rs2493215 (highlighted in blue) tags the ‘SNPs’ for alleles 1 and 3 of the PARK7 SVA (outlined in the black box) with a r^2>0.8. B – A screen shot from Haploview software showing that the SNP rs226476 (highlighted in blue) tags the ‘SNP’ for allele 2 of the PARK7 SVA (outlined in the black box) with a r^2>0.8.

Section B

6.4.7 Five Parkinson's disease associated genes contain a SVA within 10kb of their gene loci

In the first section of this chapter a list of SVAs were identified by their primary sequence and transcription factors that might regulate them and the genes they are near. In the second section another list of SVAs was generated by identifying the genes that are known to be associated with the same disease, determining if there is an SVA within their locus and if the numbers of SVAs in the gene subset are overrepresented. Mutations in the coding region the PARK7 gene have been associated with autosomal recessive Parkinson's disease (PD) (Bonifati et al. 2003). There are several other PARK genes including SNCA, LRRK2 and PINK1 that are causative genes for PD. Therefore a list of known associated genes for PD was taken from Corti et al 2011. Five out of the thirteen genes analysed using UCSC genome browser had a SVA that was within the gene or its 10kb flanking region (PARK2, PARK7, LRRK2, PLA2G6 and ATXN2). The gene locus, association with PD and whether a SVA was present was summarised in Table 6.5. The SVAs identified belonged to either subtype C, D or F and their conservation across primate species varied. Four of the SVAs were located within the intron of one of the genes (PARK2, LRRK2, PLA2G6 and ATXN2) whereas the previously characterised PARK7 SVA was 1-8kb upstream of the major TSS. Previous analysis in chapter 4 analysing the relative orientation of the SVAs to genes they have inserted into or within the 10kb flank had generated the number of genes with a SVA in these regions. 6.3% of genes had a SVA within an intron or their 10kb flank whereas 38% (5/13) of the PD associated genes had a SVA in these regions. This preliminary data suggests the subset of PD associated genes have a greater number of SVA insertions than genes on average. It has already been shown that the PARK7 SVA is polymorphic (5.4.6 and 6.4.5) so the genetic variation of two more SVAs from this list was analysed (6.4.8).

PARK Loci	Gene	Chr Position	Involvement in PD	SVA	Chr Loci of SVA	SVA loci to Gene	Alleles of SVA
PARK1/4	SNCA	4q21	Early onset, dominant	N	-	-	-
PARK2	Parkin (-)	6q25-q27	Juvenile and early onset, recessive and sporadic.	SVA F (+)	Chr6:162759277-162761189	Intron	2
PARK5	UCHL 1	4p14	Late onset, dominant	N	-	-	-
PARK6	PINK 1	1p35-p36	Early onset, recessive	N	-	-	-
PARK7	PARK7/DJ-1 (+)	1p36	Early onset, recessive	SVA D (+)	Chr1:8012111-8013640	1-8kb upstream	5
PARK8	LRRK2 (+)	12q12	Late onset, dominant and sporadic	SVA C (-)	Chr12:40746271-40747834	Intron	2
PARK9	ATP13A2	1p36	Early onset recessive	N	-	-	-
PARK11	GIGYF2	2q36-q37	Late onset, dominant	N	-	-	-
PARK13	Omi/HTRA2	2p13	Unclear	N	-	-	-
PARK14	PLA2G6	22q12-q13	Atypical PD, recessive.	Fragment of SVA F	Chr22:38549308-38549389	Intron	ND
PARK15	FBXO7	22q12-q13	Atypical PD, recessive.	N	-	-	-
-	ATXN2/SCA2 (-)	12q24.1	Unclear	SVA D (+)	Chr12:111944423-111945974	Intron	ND
-	GBA	1q21	Unclear	N	-	-	-

Table 6.5: Five of these Parkinson's Disease associated genes contain a SVA. This table lists the genes that are known to be associated with PD taken from table 1 of Corti et al (2011). The UCSC genome browser (Hg19) was used to identify the presence of a SVA within the gene or within 10kb. Three out of the four SVAs identified within this gene set were analysed for their polymorphic nature. N= no SVA present (+) – on sense strand, (-) – on antisense strand. SNCA – alpha synuclien, UCHL1 – ubiquitin COOH-terminal hydrolase 1, PINK 1 – PTEN-induced kinase 1, LRRK2 – leucine rich repeat kinase 2, ATP12A2 – ATPase type 13A2, GIGYF2 – GRB10-interacting GYF protein 2, HTRA2 - HtrA serine peptidase 2, PLA2GB – group VI phopholipase A2, FBXO7 – F-box protein 7, ATXN2/SCA2 – Ataxin 2, GBA – beta glucocerebrosidase. ND – not done.

6.4.8 The SVAs identified within PARK2 and LRRK2 genes are polymorphic

Two of the SVAs identified to be present with genes associated with Parkinson's disease were amplified in 11 of the CEU HapMap samples to determine if these SVAs were polymorphic as seen in the PARK7 and FUS SVAs. Both of the SVAs were located within the introns of their respective genes but were members of different subtypes. The SVA within the PARK2 gene is a member of the subtype F and is human specific. Two alleles were identified for this SVA in the small group of HapMap samples tested (Figure 6.8A). The SVA within the LRRK2 gene was a subtype C and is present within humans, chimpanzees and gorillas according to UCSC. There also appears to be two alleles of this SVA so far identified in the 11 HapMap samples tested however there were no heterozygous individuals (Figure 6.8B).

Previous analyses of the PARK7 and FUS SVAs in the HapMap cohort led to the identification of tagging SNPs for the different alleles which would enable much larger cohorts of individuals to be genotyped for these elements in a less labour intensive manner. Genetic variation implicated in disease processes may not be limited to alleles located at one specific gene but across several linked in similar pathways and it would be important to assess these genetic variants in combination as well as in their own right.

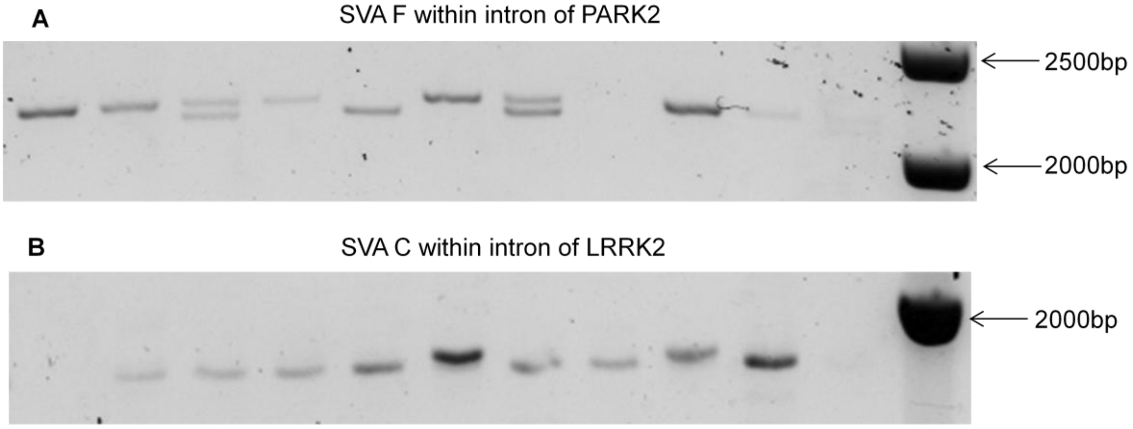


Figure 6.8: SVAs present within the PARK2 and LRRK2 genes are polymorphic. A - The SVA F within an intron of the PARK2 gene has two alleles in 11 individuals analysed form the CEU HapMap cohort. Expected product - 2090bp. The bands are running slightly higher than expected but are specific therefore the alleles of the SVA here may be larger than the sequence available on UCSC. B - The SVA C within an intron of the LRRK2 gene has two alleles in 11 individuals from the CEU HapMap cohort but no heterozygotes. Expected product – 1812bp.

6.4.9 The SVA associated with X-linked dystonia-parkinsonism inserted in to intron 32 of the TAF1 gene where a SVA was already present

An insertion of a SVA F has been linked to X-linked dystonia-parkinsonism (XDP) in a Philippine cohort (Makino et al. 2007; Hancks and Kazazian 2010) where males with this insertion into intron 32 of the TATA-binding protein-associated factor 1 gene (TAF1) located at Xq13.1 suffered from the disease. The SVA insertion was absent from other ethnic groups such as Japanese and European and African Americans tested and female carriers with this SVA insertion present on one of their X chromosomes were observed in the Philippine population. Therefore this SVA insertion is polymorphic in terms of its presence or absence and its presence is associated with disease. The sequence of the DYT3 (dystonia-parkinsonism syndrome locus) region provided by the authors (http://www.ddbj.nig.ac.jp, accession number AB191243) was analysed. The disease associated SVA (2627bp) within intron 32 was located and during this analysis I noted that there was the sequence of another SVA less than 11kb downstream and still within intron 32. This second SVA was also located on the negative strand of DNA but was shorter at 1606bp. The TAF1 gene was analysed on the UCSC genome browser and there was a SVA D found in humans and chimpanzees within intron 32 of the gene that was 1665bp long. The schematic in Figure 6.9 represents the structure of intron 32 from the sequence of the DYT3 region (AB191243) and from the UCSC genome browser Hg19.

The SVA F insertion was not present in the UCSC genome sequence; however the site of the insertion could be identified using the flanking sequence of SVA F in the published sequence from Makino et al (underlined in Figure 6.9).

There is evidence of a target site duplication (TSD) caused by the insertion of the SVA F which is highlighted in red, Figure 6.9. The flanking sequences of the SVA D from both sequences were compared and these matched indicating they were the same SVA insertion. I compared the sequences of the two SVA Ds and they were identical apart from within the central VNTR region. The VNTR of the SVA in the UCSC genome browser (Figure 6.10A) was larger and contained two additional repeats compared to the VNTR of the SVA in the sequence published by Makino et al (Figure 6.10B). There were also additional bases present in repeats 9 and 15 of the Makino et al sequence compared the UCSC sequence and are highlighted in red (Figure 6.10).

The difference in the VNTR region of the two SVA D sequences indicate that it is polymorphic in terms of the number of repeats as shown previously for the FUS and PARK7 SVAs (chapter 3 and 5/6 respectively). The potential variation of this SVA was analysed in a small number of individuals from the CEU HapMap cohort and the controls from the SALS cohort used in chapter 3. The SVA D was present in all 11 individuals from the CEU HapMap (Figure 6.11) and 21 individuals tested from the control cohort for the SALS group used in chapter 4 (data not shown). There was no variation that could be identified in the analysis in the size of the SVA D observed across these individuals. This could indicate the differences in VNTR size are unique to specific ethnic groups or is extremely rare and could not be identified in the small number of samples tested.

In chapter 4 the insertion of SVAs was determined to be more frequent in genic regions as opposed to gene deserts was therefore not randomly spread throughout the genome. This occurrence of a SVA inserting in close proximity to another, seen in the Makino et al sequence, raised the question of how many SVAs had inserted near to other SVAs. The coordinates of all the SVAs from the Hg19 from UCSC used in the analysis in chapter 4 was used to determine how many SVAs had inserted within 100kb and 10kb of another SVA. 863 SVAs were found to have inserted within 100kb of another SVA and there were 110 SVAs that had inserted within 10kb of another SVA similar to the SVA F and SVA D insertions. Many of these SVA insertions within 10kb of another SVA have occurred at evolutionary distinct periods due to differences in their presence across primate species. If the number of SVA insertions into these regions, within 100kb and 10kb of another SVA, were predicted based purely on the size of the region the number of SVA are overrepresented. The 100kb flanking regions of SVAs encompasses 14.3% of the human genome but contains 32.3% of the SVAs (a 2.3 fold increase) and the 10kb flanking regions of the SVAs constitutes 1.7% of the human genome but contains 4.1% of the SVAs (a 2.4 fold increase). It was also noted in UCSC there were two more SVAs within the TAF1 locus; a SVA A 14kb downstream of the SVA D and another SVA D 109kb upstream of the SVA D. There have been other ‘clusters’ of SVA insertions observed while using UCSC where several SVAs were located in relatively small regions. Examples of these ‘clusters’ include chr19p12 (genes in this region include ZNF90 and ZNF486) where 7 SVAs have inserted within a 235kb region and at chrXp11.22 (genes in this region include HUWE1, Mir98 and PHF8) where there are four SVA insertions within 270kb. If SVAs had inserted randomly across the human genome there on average would be less than one SVA per million bases so to have several loci of SVAs in such close proximity would further supports that the site insertion of these elements is not random and that certain regions of the genome are more susceptible to these events than others.

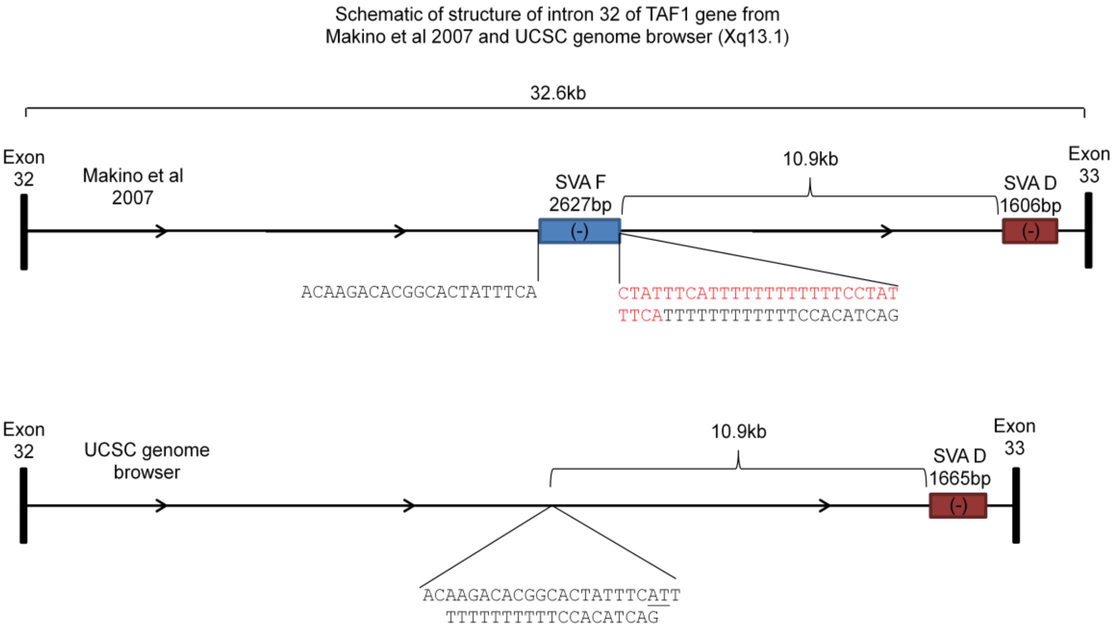
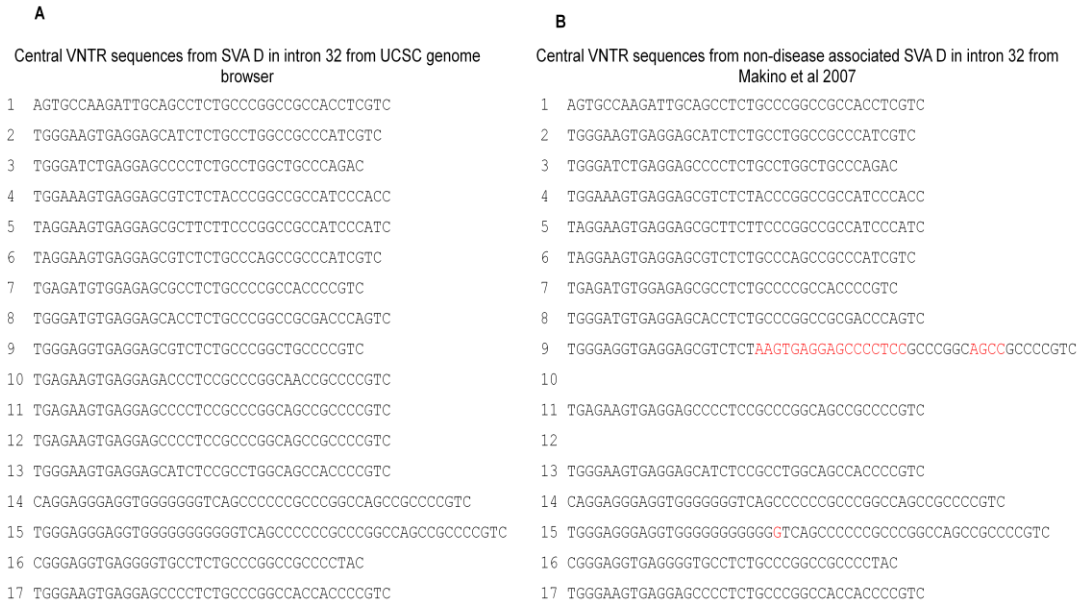


Figure 6.9: Schematic representing intron 32 of the TAF1 gene from the Makino at al 2007 and UCSC Hg19 sequences. The blue box in the top image represents the disease causing human specific SVA F insertion reported by Makino at al which is 2627bp in length and located on the negative strand (opposite to the orientation of the gene). This SVA is not present in the UCSC genome sequence but the site at which it inserted can be identified and shown underlined. The sequences flanking the SVA insertion identified in Makino et al correspond to this locus except there is a TSD highlighted in red caused by the insertion of the SVA. There is also a SVA D present within intron 32 (indicated by red box) 10.9kb downstream of the SVA F insertion or site of insertion in both the Makino et al and UCSC sequence. This SVA D is present with the chimpanzee genome as well as the human according UCSC and is a located on the negative strand. The flanking sequences of the SVA Ds match between the Makino et al and the UCSC sequences indicating it is the same insertion. The length of the SVA D differs between the two sequences by 59bp.



A

Central VNTR sequences from SVA D in intron 32 from UCSC genome browser

```
1  AGTGCCAAGATTGCAGCCTCTGCCCGGCCGCCACCTCGTC
2  TGGGAAGTGAGGAGCATCTCTGCCTGGCCGCCCATCGTC
3  TGGGATCTGAGGAGCCCCTCTGCCTGGCTGCCCAGAC
4  TGGAAAGTGAGGAGCGTCTCTACCCGGCCGCCATCCCACC
5  TAGGAAGTGAGGAGCGCTTCTTCCCGGCCGCCATCCCATC
6  TAGGAAGTGAGGAGCGTCTCTGCCCAGCCGCCCATCGTC
7  TGAGATGTGGAGAGCGCCTCTGCCCCGCCACCCCGTC
8  TGGGATGTGAGGAGCACCTCTGCCCGGCCGCGACCCAGTC
9  TGGGAGGTGAGGAGCGTCTCTGCCCGGCTGCCCCGTC
10 TGAGAAGTGAGGAGACCCTCCGCCCGGCAACCGCCCCGTC
11 TGAGAAGTGAGGAGCCCCTCCGCCCGGCAGCCGCCCCGTC
12 TGAGAAGTGAGGAGCCCCTCCGCCCGGCAGCCGCCCCGTC
13 TGGGAAGTGAGGAGCATCTCCGCCTGGCAGCCACCCCGTC
14 CAGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
15 TGGGAGGGAGGTGGGGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
16 CGGGAGGTGAGGGGTGCCTCTGCCCGGCCGCCCCTAC
17 TGGGAAGTGAGGAGCCCCTCTGCCCGGCCACCACCCCGTC
```

B

Central VNTR sequences from non-disease associated SVA D in intron 32 from Makino et al 2007

```
1  AGTGCCAAGATTGCAGCCTCTGCCCGGCCGCCACCTCGTC
2  TGGGAAGTGAGGAGCATCTCTGCCTGGCCGCCCATCGTC
3  TGGGATCTGAGGAGCCCCTCTGCCTGGCTGCCCAGAC
4  TGGAAAGTGAGGAGCGTCTCTACCCGGCCGCCATCCCACC
5  TAGGAAGTGAGGAGCGCTTCTTCCCGGCCGCCATCCCATC
6  TAGGAAGTGAGGAGCGTCTCTGCCCAGCCGCCCATCGTC
7  TGAGATGTGGAGAGCGCCTCTGCCCCGCCACCCCGTC
8  TGGGATGTGAGGAGCACCTCTGCCCGGCCGCGACCCAGTC
9  TGGGAGGTGAGGAGCGTCTCTAAGTGAGGAGCCCCTCCGCCCGGCAGCCGCCCCGTC
10
11 TGAGAAGTGAGGAGCCCCTCCGCCCGGCAGCCGCCCCGTC
12
13 TGGGAAGTGAGGAGCATCTCCGCCTGGCAGCCACCCCGTC
14 CAGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
15 TGGGAGGGAGGTGGGGGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTC
16 CGGGAGGTGAGGGGTGCCTCTGCCCGGCCGCCCCTAC
17 TGGGAAGTGAGGAGCCCCTCTGCCCGGCCACCACCCCGTC
```

Figure 6.10: Comparison of the central VNTR sequences of a SVA inserted at the same loci from UCSC genome browser and Makino et al 2007. The VNTR from the SVA sequence within the Makino et al publication is smaller with two fewer repeats (10 and 12) compared to the VNTR from the sequence available in UCSC but with additional bases present in repeat 9 and 15 (highlighted in red). A – The sequence of the VNTR of a SVA D from the UCSC genome browser located in intron 32 of the TAF1 gene. This VNTR if 688bp long with 17 repeat units. B - The sequence of the VNTR of an additional SVA to the disease associated SVA in Makino et al located in intron 32 of the TAF1 gene. This VNTR if 629bp long with 15 repeat units.

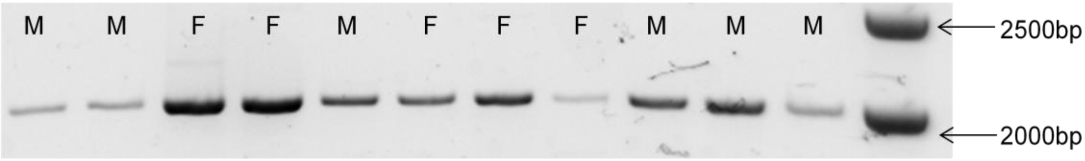


Figure 6.11: Amplification of the SVA D located within intron 32 of the TAF1 gene in UCSC genome browser in CEU HapMap samples. The presence of this SVA was confirmed in a Caucasian population however there was no variation noted in the 11 individuals tested from the CEU HapMap cohort. Expected product size 1956bp. This SVA is located on the X chromosome therefore the sex of the individual was noted (F=female M=male). This SVA was also tested in an additional 21 samples from the King's controls of the ALS cohort used in chapter 3 and no genetic variation was seen.

6.5 Discussion

A search for SVAs that shared sequence homology to repeats of the TR and VNTR of the PARK7 SVA identified 19 SVAs (including the PARK7 SVA) which were located within 100kb of 32 different genes (Table 6.1). The MetaCore programme was used to determine TFs known to regulate these genes with CREB1, ESR1, Sp1, GCRα and PR the top five of these factors (Table 6.2). Of these GCRα, Sp1 and ESR1 were shown to have binding sites present in the sequences of the repeats of the PARK7 SVA to identify this list of SVAs and subsequent genes (Figure 6.3B). It was hypothesised that these factors may bind to the SVAs and play a role in regulating the expression of the gene they are near. ESR1 was chosen to test this hypothesis as the MCF-7 cell line is oestrogen responsive so provided a good cell line model for this experiment and PARK7 has been linked to breast cancer so modulation of PARK7 by oestrogen could be an important mechanism in the disease process. The endogenous PARK7 mRNA levels increased when exposed to 10nM 17-β-estradiol for 18hrs as did the mRNA of the MAN1C1 gene, however the mRNA of TFPI gene did not increase (Figure 6.4). The expression levels of the reporter gene constructs containing different sized fragments of the PARK7 SVA (previously used in section 5.4.8) transfected into the MCF-7 cell line were not significantly altered by the exposure of the cells to 10nM 17-β-estradiol for 18hrs (Figure 6.6). A plasmid containing an oestrogen response element (ERE) was used as a positive control for this assay which showed a significant and robust increase in response to 17-β-estradiol (Figure 6.5). This would suggest that there is not modulation of these genes occurring through the binding of the ESR1 to the SVAs at multiple loci as hypothesised due to the lack of response of the expression of the TFPI gene and the PARK SVA modulation of the reporter gene unaffected by the cell exposed to oestrogen. The modulation of the PARK7 and MAN1C1 genes is likely to be through other oestrogen responsive pathways. Analysis using the MetaCore programme did not show that PARK7 is regulated directly by the ESR1 however it is by Sp1. It has been shown in the literature that the ESR1 and Sp1 can associate through protein-protein interactions and mediate a response to oestrogen that does not require the ESR1-DNA binding and may only require Sp1 binding sites (Porter et al. 1997). This may be the mechanism that is involved in the up regulation of the PARK7 gene in response to oestrogen demonstrated in this cell line model as there are Sp1 binding sites in the major PARK7 promoter and it is known to be modulated by Sp1 (Taira et al. 2001). MAN1C1 although not shown to be directly regulated by the ESR1 (Table 6.2) it is regulated by the PR that is an oestrogen responsive gene and therefore expression could be indirectly modulated by stimulation with oestrogen. The TFPI gene was not predicted to be modulated by the ESR1, Sp1 or the PR therefore its lack of response to oestrogen is consistent with TFPI not being part of an oestrogen responsive pathway.

The response of PARK7 expression to oestrogen could be important in disease process as PARK7 is known to be involved in breast cancer. PARK7 was first linked to breast cancer when the PARK7 protein was identified at higher levels in the sera of breast cancer patients than controls (Le Naour et al. 2001). PARK7 protein is secreted by breast cancer cells with PARK7 protein detected in the nipple fluid of patients with breast cancer at higher levels than those with benign lesions and this high level of PARK7 protein in the nipple fluid was correlated with low protein expression but high mRNA expression in the tumour (Oda et al. 2012). This low expression of PARK7 protein but high mRNA expression in the tumours of patients with invasive ductal carcinoma (IDC) was a predictor of poor outcome

(Tsuchiya et al. 2012). Although in another study low PARK7 protein levels was an indicator of pathological complete remission after neoadjuvent chemotherapy in patients with IDC indicating chemotherapy is an important component of improving the survival of patients with IDC and low levels of PARK7 protein (Kawate et al. 2013). Increased levels of PARK7 protein offer resistance to cellular apoptosis most likely through multiple mechanisms including its protection against oxidative stress and PARK7's negative regulation of PTEN an inducer of apoptosis (Canet-Aviles et al. 2004; Kim et al. 2005; Clements et al. 2006). Therefore the levels of PARK7 are important for the potential treatment of the disease with a reduction in PARK7 increasing the response to chemotherapy. The up regulation of PARK7 in response to oestrogen could be an important mechanism in oestrogen receptor positive cancer cells as increased levels of PARK7 protein can increase the cells resistance to apoptosis.

The genetic variation of the PARK7 SVA was analysed in a cohort of breast cancer patients without mutations in their BRCA genes and a female matched control group from the National Genetics References Laboratory, St Mary's Hospital, Manchester. There were a reduced number of 1/1 genotypes (7.2%) for the PARK7 SVA in the breast cancer patients compared to the control group (Table 6.3). However there was not a significant difference between the breast cancer patients and the control group for the genotype of the PARK7 SVA when analysed using CLUMP (Sham and Curtis 1995). A greater in depth analysis of the PARK7 SVA variants and their association with particular characteristics of the tumour of the patients would be valuable especially in light of the up regulation of PARK7 gene in response to cellular stimulation with oestrogen. The frequency of the genotypes of the PARK7 SVA in the breast cancer and control group were noted to be different to the frequency of the genotypes in the King's control group from the ALS cohort used in chapter 3 for the FUS SVA analysis. The King's control group consisted of male and females whereas the breast cancer cohort was all female, therefore the King's control group were broken down by sex. The frequency of the genotypes in the females of King's control group was more similar to those of the breast cancer control group than the males (Table 6.3) with fewer 1/3 genotypes in the males and allele 5 only present in the males. These cohorts are still small so this may be a factor of not a high enough N number and the reason for these sex differences is unclear.

In a similar process as for the FUS SVA tagging SNPs were identified for the PARK7 SVA. This was a more complicated process due the increased number alleles of the PARK7 SVA. Two tagging SNPs were identified that could be used to determine the most common alleles 1, 2 and 3: rs2493215 a G corresponds to allele 1 and a A corresponds to allele 3 and at rs226476 a T corresponds to allele 2 (section 6.4.6). Unfortunately a tagging SNP for the rarer allele 4 and allele 5 as not present in the HapMap cohort could not be identified. These tagging SNPs could be used, similar for the FUS SVA, to analyse the genetic variation of the PARK7 SVA in larger cohorts as analysis of the SVA variants is time consuming and labour intensive. This model is being continued by others in our group.

The analysis of the presence of SVAs in causative genes for PD identified 5 of these genes, including PARK7, with a SVA within the gene or its 10kb flank (Table 6.5) which was an overrepresentation when compared to the percentage of genes across the whole genomes with SVAs in those relative regions. The potential of these SVAs to be polymorphic was tested in two elements in addition to the already characterised genetic variation of the PARK7 SVA. In a small sample of individuals from the CEU HapMap cohort the SVAs within introns of the PARK2 and LRRK2 genes were identified to contain two alleles each (Figure 6.8). This initial analysis shows that there are at least two other polymorphic SVAs in known PD causing genes providing further genetic variation to be considered in disease association studies. If the numbers of HapMap samples genotyped were to be expanded a tagging SNP may be identified for the alleles of the SVAs within LRRK2 and PARK2, as had been completed for the FUS and PARK7 SVAs. This could make the studies of potential genetic association with disease simpler.

The intron 32 of the TAF1 gene was analysed further due to the disease causing insertion of an SVA associated with X-linked dystonia-parkinsonism. The disease causing insertion was identified to have inserted less than 11kb upstream of a SVA already present in the intron 32 when I analysed the sequence online provided by the authors of the manuscript. There was also a SVA identified in the UCSC genome browser at the same locus (Figure 6.9) and the presence of this SVA in Caucasian individuals was confirmed (Figure 6.11). The size of the VNTR of this SVA varied between the sequence from the Philippine individual and the sequence from UCSC (Figure 6.10) but did not show variation across the Caucasian individuals tested. This difference in VNTR size may be due to the ethnic differences. It is the second SVA insertion in this intron that is associated with disease but the presence of one SVA is not sufficient to be pathogenic. This could mean that the impact of SVAs on the genomic locus is amplified when more than one element in present in a relatively small region. Therefore the frequency of these close proximity insertions was analysed using UCSC and identified 110 SVAs within 10kb of another SVA and 863 had inserted within 100kb of another SVA. The preference of SVA insertions had been shown to be in genic regions over gene deserts (chapter 4) but this data indicates there may be regions that are particularly susceptible to SVA insertions over others. The hypothesis that there are regions of the genome that are particularly vulnerable to retrotransposon insertions has been suggested previously due to the insertion of a SVA and an AluY in two different individuals at the exact same nucleotide of exon 9 of the BTK gene resulting in X-linked agammaglobulinemia (Conley et al. 2005). There have been other documented independent retrotransposon insertions occurring at the same genomic loci of Factor IX and APC of different individuals (Miki et al. 1992; Vidaud et al. 1993; Halling et al. 1999; Wulff et al. 2000) and Conley et al suggest due to the lack of similarities of the DNA sequence across these insertion sites it is the surrounding DNA and chromatin structure that make them vulnerable to retrotransposon insertions. Although a full analysis of whether these regions acted as hot spots for insertions for all classes of retrotransposons was outside the scope of this thesis. Alternatively the close proximity of multiple SVA insertions described in this chapter such as at chr19p12 (ZNF90 and ZNF486) and at chrXp11.22 (HUWE1, Mir98 and PHF8) may be due to the initial SVA insertion increasing the susceptibility of the genomic locus to further SVA insertions perhaps through altering the structure of the locus.

The chapter has highlighted important preliminary data including the regulation of the PARK7 gene, identification of additional polymorphic SVAs in disease associated genes and indicated the potential greater complexity to the site of SVA insertions and that their influence on the genome may be amplified when two elements are in close proximity to one another.

Chapter 7

Discussion

Over 98% of the human genome is non-coding containing domains of various regulatory functions, including controlling the specific genes expressed and the levels of their expression that is crucial to direct and maintain the function and phenotype of the cell. The regulation of gene expression can occur in many forms from the proximal promoter, enhancers and repressors to the packaging of the DNA into chromatin and its state of methylation, the latter parameters representing epigenetic regulation. To identify potential regulatory domains within the non-coding DNA, genome browsers have proved a useful tool for bioinformatic analysis of the plethora of data now available. This has dramatically increased in recent years with the publication of data from the ENCODE consortium (Kavanagh et al. 2013b). For example genome browsers have provided an easy method to identify regions that are highly conserved between species indicating potential function which from further analysis proved to be regulators of gene expression *in vitro* and *in vivo* (Davidson et al. 2006; Shanley et al. 2010; Paredes et al. 2011). There is also a wealth of data available within these browsers, in particular UCSC, not just on the sequence of the genome but the chromatin state (inactive and active histone marks), transcription factor (TF) binding and DNase clusters across cell lines that can be used to predict transcriptionally active regions. One of the areas of expertise of the laboratory in which I did my study lay in the analysis of VNTR domains and their ability to affect gene expression in a tissue specific and stimulus inducible manner and their association to disease which was correlated with specific copy number of the VNTR (Guindalini et al. 2006; Breen et al. 2008; Haddley et al. 2008; Ali et al. 2010; Vasiliou et al. 2012). Such variation in an individual's genetics to another can influence the levels of gene expression and the response to their environment not only in the short term but in the long term by altering epigenetic factors. This gene-environment interaction can play an important role in phenotypic differences and disease susceptibility. Therefore the analysis of specific genomic regions encompassing genes of interest was undertaken to identify potential novel regulatory and polymorphic domains.

The analysis of the 10kb upstream of the FUS gene for regions of DNA that may be functional and polymorphic identified a large repetitive region which was part of a composite element called a SVA (Figure 4.1). SVAs are a non-autonomous hominid specific member of the non-LTR retrotransposons. 2676 SVAs were identified across the human genome using the UCSC genome browser and therefore have the potential to affect gene expression at multiple loci. The repetitive nature of the domains within their structure, the central VNTR and 5' CCCTCT repeat, provide sources of genetic variation that may result in differential regulation between individuals. There is also growing interest in the ability of retrotransposon insertions to create intra-individual genetic variation through somatic retrotransposition events, generating genetically distinct cells and providing the cell with additional mechanisms of transcriptional control (Faulkner 2011). Somatic retrotransposition has been shown to occur in the human brain altering the genetic landscape of the individual (Baillie et al. 2011). The functional properties of the SVAs therefore may differentially affect populations of cells within an individual due to somatic insertions.

The sequence of the SVAs indicated their role in the regulation of gene expression through several mechanisms including the provision of multiple TF binding sites, G4 structural formation and sites of methylation (Hancks and Kazazian 2010). In chapter 4 a global analysis of the distribution of SVAs across the human reference genome 19 was undertaken to determine if the sites of their insertions correlated with any particular features of the genome. The insertion of the SVAs showed a preference for genic regions as opposed to gene deserts and their density was correlated with gene density across chromosomes (Figure 4.2 and 4.3). This is an indicator that their site of insertion is linked to active chromatin as the more open nature of the chromatin would make it more accessible to retrotransposon insertions. Once inserted into these transcriptionally active regions retrotransposons may be under negative selection if they are imposing detrimental affects to the host genome. LINE and LTR elements are depleted in introns and flanking regions of genes in particular when in the same orientation as the gene, however Alus are enriched in genes and their flanking regions suggesting LINEs and LTRs are selected against in these regions whereas Alus are not (Medstrand et al. 2002). SVAs show similar distribution to Alus due to their enriched presence in the flanking regions of genes (Figure 4.4) but show similarities to LINEs as appear to be selected against when inserting intronically in the same orientation as the gene (Table 4.1). Only 26.1% of intronic SVA insertions are in the same orientation as the gene whereas this is near to 50% when they have inserted in within the 10kb flank of the gene. The negative impact of SVAs inserted into an intron in the same orientation of the gene may be related to the introduction of alternative splice sites as exemplified by disease causing insertions (Kaer and Speek 2013) or the presence of polyadenylation signals within the SVA sequence affecting the transcriptional machinery.

The retrotransposition of the SVAs, along with other retrotransposons, is inhibited by the cell to prevent potential *de novo* insertions that may have a negative impact through heterochromatin formation and methylation. SVAs have a high GC content (~60%) and were hypothesised to act as mobile CpG islands (Wang et al. 2005). Therefore a handful of SVAs from each subtype were analysed for their ability to meet a set of three criteria used to predict CpG islands by the UCSC genome browser taken from the literature (Gardiner-Garden and Frommer 1987). The SVAs in this analysis met two of the criteria but not the third as they had an average ratio of the observed number of CG dinucleotides to the expected of 0.53 which was below the 0.6 threshold. However there are still many sites for methylation within the sequence of the SVAs and may interfere with regulatory domains in their proximity, for example hypermethylation of the disease causing SVA insertion into intron 32 of the TAF1 gene is associated with a reduction of TAF1 mRNA in the caudate nucleus (Makino et al. 2007).

There is growing evidence for the loss of silencing of retrotransposons, including SVAs, in specific conditions such as cancer, the aging brain and replicatively senescent cells (Szpakowski et al. 2009; Baillie et al. 2011; De Cecco et al. 2013), which could lead to the activation of regulatory properties within the SVAs sequence. Analysis within this project of a specific SVA, upstream of the PARK7 gene, demonstrated the locus of this element and the adjacent sequence is the site of active chromatin and transcription in the SK-N-AS (human neuroblastoma) cell line. The majority of the transcripts of the human PARK7 gene originate at the already characterised promoter (Taira et al. 2001) which is 8kb downstream of the PARK7 SVA and their expression was confirmed in the SK-N-AS and MCF-7 (human breast adenocarcinoma) cell line (Figure 5.2). There was also evidence of another TSS of the PARK7 gene which included a CpG island, ENCODE data of active histone marks and DNase hypersensitivity clusters and a transcript predicted by Archive Ensembl within less than 1kb of the PARK7 SVA (Figure 5.1). The expression of this transcript was confirmed in the SK-N-AS cell line but not in the MCF-7s (Figure 5.2). There are 243 Archive Ensembl predicted transcripts that contain a SVA within 1kb upstream compared to the 58 UCSC genes which indicates there could be more TSSs within 1kb of a SVA than indicated by the number of known transcripts if these were confirmed as in the PARK7 example. ChIP analysis of the minor and major TSSs of the PARK7 gene and 5' of the SVA in the SK-N-AS cell line identified active histone marks, RNA polymerase II and Sp1 present at these loci showing this region is transcriptionally active and the SVA is not epigenetically silenced (Figure 5.4). Both of the PARK7 TSSs were shown to be transcriptionally active in several regions of the human brain using Affymetrix human exon array probes which provides further evidence that the region adjacent to the SVA is active (Table 5.1 and 5.2). The Affymetrix human exon array probes could provide a tool for analysing the prevalence of transcription occurring close to SVAs on a global scale as there are 1248 extended probe sets located within 1kb up or downstream of 729 different SVAs.

The evidence suggesting that all SVAs are not always silenced suggest they could support regulatory properties within their sequence and influence the expression of gene near to their site of insertion (we are not ruling out their regulation of more distant genes but this is the most straightforward hypothesis to test currently). The ability of two SVAs, one located upstream of the PARK7 gene and the other upstream of the FUS gene, to affect gene expression were tested in reporter gene models. Both of these SVA are members of the subtype D, however the PARK7 is human specific and contains all the domains of a canonical SVA whereas the FUS SVA is found in both the human and chimpanzee genomes and is missing the CCCTCT repeat located at its 5' end. I demonstrated both FUS and PARK7 SVAs to be polymorphic in terms of the number of repeats in their VNTR domains, which is discussed in further detail later. VNTRs have previously been shown to affect gene expression in a tissue specific manner and their function dependant on the allele present *in vitro* and *in vivo* (MacKenzie and Quinn 1999b; Haddley et al. 2008; Ali et al. 2010; Haddley et al. 2012) and it was therefore hypothesised that the SVAs or at least their VNTR domains may function in a similar manner.

Both alleles (long and short) of the intact FUS SVA and its TR/VNTR domain (due to adjacent TR the VNTR could not be cloned alone) were tested in a reporter gene vector with a minimal promoter in the SK-N-AS cell line (Figure 3.4B) (completed by Thomas Wilm). The long and short complete SVAs showed repressive function whereas the long and short of the TR/VNTR increased transcription of the reporter gene. This suggested the addition of the SINE element in the complete SVA compared to the TR/VNTR was acting as a dominant repressor within the SK-N-AS cell line. There was a small but statistically significant difference between the long and short alleles when tested in the complete SVA however this difference was not seen when the alleles were tested in the TR/VNTR. The ability of the long allele of the FUS SVA and TR/VNTR to affect gene expression was tested in an *in vivo* chick embryo model (Figure 3.4A) (completed by Kejhal Khursheed). Both the complete SVA and TR/VNTR supported expression of the reporter gene in the neural tube of the chick embryo which was not seen when the minimal promoter alone was transfected. The intact FUS SVA in the SK-N-AS cell line showed repressive properties and in the chick embryo enhancer properties therefore it affects gene expression in a tissue specific manner.

The ability of the different domains of the PARK7 SVA and four of its alleles to affect reporter gene expression was tested in the SK-N-AS and MCF-7 cell lines. The domains of the PARK7 SVA demonstrated differential function and the

TR/VNTR domain demonstrated tissue specific regulatory properties (Figure 5.9). The construct containing the fragment of the PARK7 SVA without the SINE domain showed increased expression levels over the intact SVA suggesting it is a repressor which was also noted for the SINE of the FUS SVA. The PARK7 SVA fragments showed different regulatory functions depending on whether they were in the forwards or reverse orientation within the reporter gene therefore the relative orientation of a SVA to a gene could influence its affect on gene expression. The complete SVA was also tested in a promoter less reporter gene but did not display the ability to initiate transcription itself. The ability of the four alleles of the PARK7 SVA identified and cloned from individuals from the CEU HapMap cohort to differentially affect reporter gene expression were tested as part of the complete SVA sequence. All four alleles showed similar levels of reporter gene expression in the SK-N-AS cell line as did alleles 1, 3 and 4 in the MCF-7, however allele 2 showed increased levels of reporter gene expression that were significantly different to the other three alleles (Figure 5.11). The full potential of the alleles to differentially affect transcription may only be seen in response to a stimulus.

The primary sequence of DNA determines the TFs that may bind and in turn regulate gene expression. SVAs at multiple loci of the human genome may respond to similar factors and stimuli due to the similarities in sequence they share. The primary sequence of specific repeats from the PARK7 central TR and VNTR was used to define a subset of SVAs and the closest genes to their insertion. MetaCore software was used to determine the factors that regulate these genes and PROMO software was used to define potential factors that may bind to the sequence of the PARK7 SVA. ESR1 was shown to regulate a number of the genes containing a SVA and there were binding sites present in the sequence of the SVA (Table 6.2 and

Figure 6.3B) therefore the response of the endogenous expression of a small number of genes on the list and the PARK7 SVA to oestrogen in the MCF-7 cell line was tested. The expression of the endogenous PARK7 gene responded to signalling pathways stimulated by the exposure of the cell to oestrogen (Figure 6.4), however the regulatory properties of the PARK7 SVA were not altered by the same stimulus (Figure 6.6). It might be that the full complexity of SVA transcriptional regulation will be required to be addressed in reporter gene analysis from a chromosomal or nucleosomal construct to address the structural importance that might reside in such retrotransposons.

The sequence of several of the SVAs domains also provides the potential for formation of G-quadruplex (G4) DNA (runs of guanine nucleotides) which is involved in the repression of gene expression (Gonzalez and Hurley 2010). All of the SVA subtypes contained G4 potential however the amount of G4 potential in general increased as the age of the subtype decreased with the human specific subtypes E-F1 showing the greatest G4 potential in particularly in the emerging second central VNTR domain within their structure (Figure 4.6). There is evidence in the literature for the binding of nucleolin to G4 structures and has been shown through ChIP to be bound to the c-MYC NHE III_I domain in HeLa cells (Gonzalez et al. 2009). Therefore the binding of nucleolin in the region of the PARK7 SVA was analysed in SK-N-AS cell line to potentially validate the presence of G4 DNA at this locus. There was no binding of nucleolin under the conditions the ChIP was performed (Figure 5.4), however this PARK7 locus was associated with active transcription (presence of active histone marks and RNA Pol II) which would be consistent with the lack G4 formation as this would linked to repression of gene expression.

Both the FUS and PARK7 SVAs were shown to be polymorphic with two alleles and five alleles respectively. In both cases the central repetitive region consisted of a TR and a VNTR, however in addition the alleles of the PARK7 SVA showed differences in the number of repeats of the CCCTCT domain which was not present in the FUS SVA (Figure 3.2 and 5.7). The number of repeats of the long and short alleles of the FUS SVAs and alleles 1-4 of the PARK7 SVAs were confirmed by sequence analysis. The fifth allele of the PARK7 SVA was identified at a later stage in a separate control cohort and was not part of the initial analysis of the variants of this SVA carried out in the CEU HapMap cohort and therefore has not been sequenced at this time. Many multifactorial diseases are caused by both a genetic and an environmental component and a specific allele of a common variant within the population can provide a genetic predisposition to a specific disease. For example VNTRs of the SLC6A4 and SLC6A3 gene have demonstrated ‘risk’ alleles for a variety of disorders including depression, addiction and PD (reviewed in Haddley et al. 2008; Haddley et al. 2012). It was therefore hypothesised that the VNTRs within the FUS and PARK7 SVAs may also be involved in genetic predisposition to disease.

In 3-5% of FALS and ~1% of SALS cases mutations in the FUS gene have been identified as causative for the disease (Kwiatkowski et al. 2009; Vance et al. 2009; Corrado et al. 2010). Genes known to be involved in the familial form of a disease can indicate loci of common variants that may contribute a predisposition to the sporadic form which may be involved in differential gene expression in response to the environment. Therefore the FUS SVA, shown to be a regulatory of gene expression *in vitro* and *in vivo*, was genotyped in a SALS and matched control cohort provided by collaborators at King’s College London. There was a small difference of 4.4% between the SS genotype frequency of the SALS and control cohorts (15.4% vs 11%) (Figure 3.5) which was not significant when analysed using CLUMP (T1 P=0.36 and T4 P=0.33) (Sham and Curtis 1995).

The protein levels of PARK7 within breast tumours has been linked to the prognosis of the patient and the response to neoadjuvent chemotherapy and has been considered as a useful biomarker for breast cancer due to elevated levels identified in the sera and nipple fluid of breast cancer patients (Le Naour et al. 2001; Oda et al. 2012; Tsuchiya et al. 2012; Kawate et al. 2013). Expression of PARK7 and its presence within cancer cells could be an important factor in the cancer cells' survival. The frequency of the PARK7 SVA genotypes were analysed in a cohort of breast cancer patients without mutations in their BRCA genes and matched controls to determine if the variation of the SVA may contribute a risk factor for the disease. There was no significant difference between the patients and controls when analysed using CLUMP (T1 P=0.65 and T4 P=0.65).

The genotyping analysis of SVAs is time consuming and labour intensive and would be an impractical method, although feasible, for analysing large cohorts. Therefore using the genotype data for the CEU HapMap cohort generated for the FUS and PARK7 SVAs in conjunction with SNP data available for these individuals on the International HapMap website tagging SNPs could be identified for the alleles of the SVAs. A tagging SNP for both alleles of the FUS SVA and alleles 1, 2 and 3 (the most common) for the PARK7 SVA were successfully identified (section 3.4.5 and 6.4.6). This could provide a more practical method for analysing large disease cohorts for variants of the SVAs. This could also be employed for other polymorphic SVAs of interest. For example a list of known PD associated genes were analysed for the presence of a SVA and 5 out of 13 contained a SVA within their intron or their 10kb flank (Table 6.5). Two of these SVAs, within an intron of PARK2 and LRRK2, were analysed further and shown to be polymorphic (Figure 6.8) and a tagging SNP for these SVAs may prove a useful tool in disease association studies. The study of SVAs as regulatory and polymorphic domains using data generated in this thesis as a starting point is continuing within the research group.

The work carried out in this project has contributed to the field adding to the understanding of the functional properties of the SVAs along with identification of novel genetic variation within these elements. Data presented here in this thesis has been published (Savage et al. 2013) and further publications are within the submission process. The analysis of this family of hominid specific retrotransposons has demonstrated their ability to modulate gene expression *in vitro* and *in vivo* and in a tissue specific and allelic dependant manner. SVAs located in many functional regions are a source of genetic variation throughout the human genome contributing to the genetic differences between individuals and potentially to the phenotypic differences and disease susceptibility. The functional properties of the SVAs may be activated in specific conditions such as cancer or even the aging brain due to the loss of silencing across retrotransposons. The emerging field of somatic retrotransposition may provide further roles for the differential regulation of gene expression by SVAs within an individual and SVAs should be considered important sources of genetic variation and function when analysing genomic loci.

Appendix

A.1 The sequence, target region and cycling conditions used for each primer set

Primers	Application for Primer Use	Target Region	Expected Product Size	PCR Conditions
For 5’cagttttccctcagacccagcac 3’ Rev 5’gagctgttgggtacacctcccagac 3’	Genotyping	VNTR of FUS SVA chr16:31180471-31181135	615/665bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 60.3^{o}C for 30s and 72^{o}C for 1min, 1 cycle of 72^{o}C for 5mins
For 5’ ttcattttcagcctggtgtg 3’ Rev 5’ cgtctccatttcctctgctc 3’	mRNA expression levels	Exon 1 and 2 of transcripts of PARK7 gene originating from major TSS	102bp and 159bp (cDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 60^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ catcctggctaaaggagcag 3’ Rev 5’ ttcatgagccaacagagcag 3’	mRNA expression levels	Exon 2 and 6 of transcripts PARK7 gene	325bp (cDNA)	1 cycle of 95^{o}C for 2mins, 28 cycles of 95^{o}C for 30s, 62^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ agtggacctacgtcatgcag 3’ Rev 5’ cgtctccatttcctctgctc 3’	mRNA expression levels	Exon 1 and 2 of predicted transcript of PARK7 gene	108bp (cDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 57^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ caccttctacaatgagctgcgtgtg 3’ Rev 5’atagcacagcctggatagcaacgtac3’	mRNA expression levels	Exon 3 and 4 of the β-Actin gene	158bp (cDNA)	1 cycle of 95^{o}C for 2mins, 25 cycles of 95^{o}C for 30s, 60^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ ttgtggttttcctggtgtca 3’ Rev 5’ ccgagctctgggactaatca 3’	mRNA expression levels	Exon 2 and 3 of the Trefoil factor 1 (TFF1) gene	209bp (cDNA)	1 cycle of 95^{o}C for 2mins, 28 cycles of 95^{o}C for 30s, 62^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ ttccgaataaaggccatcag 3’ Rev 5’ caagtgcagggatccaaact 3’	mRNA expression levels	Exon 5 and 6 of the Mannosidase α class 1C member 1 (MAN1C1)	159bp (cDNA)	1 cycle of 95^{o}C for 2mins, 30 cycles of 95^{o}C for 30s, 60^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins

Primers	**Application for Primer Use**	**Target Region**	**Expected Product Size**	**PCR Conditions**
For 5’ gcctgggcaatatgaacaat 3’ Rev 5’ ctcattggcacgacacaatc 3’	mRNA expression levels	Exon 5 and 7 of the Tissue factor pathway inhibitor (TFPI) gene	218bp (cDNA)	1 cycle of 95^{o}C for 2mins, 30 cycles of 95^{o}C for 30s, 60^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ agttgtgcaggttggctagg 3’ Rev 5’ ccttgaaaagtctgccctga 3’	Analysis of genomic DNA	5kb upstream of major PARK7 promoter chr1:8016444-8016598	155bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 63.1^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ aggaagggtagccaggagaa 3’ Rev 5’ cttgctgaggctgcactctt 3’	Analysis of genomic DNA	1kb upstream of major PARK7 promoter chr1:8020452-8020653	202bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 63.1^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’tcgaactcctggcttcaagt 3’ Rev 5’ggaaggaagaaggggcatag 3’	Analysis of genomic DNA	2kb downstream of major PARK7 promoter chr1:8023653-8023829	177bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 63.1^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ ctgagcccaggagttttgag 3’ rev 5’ gagtgcagtgatgcaatcgt 3’	Analysis of genomic DNA	6kb downstream of major PARK7 promoter chr1:8027761-8027948	188bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 63.1^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ aggcctggaccagagtccta 3’ Rev 5’ cggtcagtcaaatccaacg 3’	Analysis of genomic DNA	Major PARK7 promoter chr1:8021368-8022040	673bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 60.6^{o}C for 30s and 72^{o}C for 1min, 1 cycle of 72^{o}C for 5mins
For 5’ggctttttgataacccctga 3’ Rev 5’tttcggatcacaggcatgagc 3’	Cloning	Whole PARK7 SVA chr1:8012044-8013681	1638bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 20s, 61^{o}C for 10s and 70^{o}C for 1min, 1 cycle of 70^{o}C for 2mins
For 5’ggctttttgataacccctga 3’ Rev 5’ ccgcctttctattccacaaa 3’	Cloning	SVA Δ SINE PARK7 SVA chr1:8012044-8013503	1460bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 20s, 59.6^{o}C for 10s and 70^{o}C for 1min, 1 cycle of 70^{o}C for 2mins
For 5’ctcagtgctcaatggtgcc 3’ Rev 5’ ccgcctttctattccacaaa 3’	Cloning	VNTRs of PARK7 SVA chr1:8012424-8013503	1080bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 20s, 60^{o}C for 10s and 70^{o}C for , 1 cycle of 70^{o}C for 2mins

Primers	**Application for Primer Use**	**Target Region**	**Expected Product Size**	**PCR Conditions**
For 5’tgtaggtaccggctttttgataaccc 3’ Rev 5’gtaactcgagtttcggatcacaggc 3’	Cloning	Alleles of PARK7 SVA with restriction sites chr1:8012034-8013691	1658bp (+VNTR variation) (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 20s, 60^{o}C for 10s and 70^{o}C for 1min, 1 cycle of 70^{o}C for 2mins
For 5’ggctttttgataacccctga 3’ Rev 5’gcaaggcttagcttggacag 3’	Genotyping	PARK7 SVA chr1:8012044-8013878	1835bp (+VNTR variation) (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 20s, 63.9^{o}C for 10s and 70^{o}C for 1min, 1 cycle of 70^{o}C for 2mins
For 5’ gcatttgctcctgacttcaa 3’ Rev 5’ tctgattttcttggttctcacg 3’	Genotyping	PARK2 SVA chr6:162759153-162761242	2090bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 59.3^{o}C for 30s and 72^{o}C for 3min, 1 cycle of 72^{o}C for 5mins
For 5’ tggcagacaagttttgccta 3’ Rev 5’gagatctggacatggctcct 3’	Genotyping	LRRK2 SVA chr12:40746123-40747934	1812bp (gDNA)	1 cycle of 95^{o}C for 2mins, 30 cycles of 95^{o}C for 20s, 64.9^{o}C for 10s and 70^{o}C for 1min, 1 cycle of 70^{o}C for 2mins
For 5’ccaactggtctggagtaagaca 3’ Rev 5’aactacaagccaccagtttgc 3’	Genotyping	TAF1 SVA D2 chrX:70671054-70673009	1956bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 20s, 64.9^{o}C for 10s and 70^{o}C for 1min, 1 cycle of 70^{o}C for 2mins
For 5’agccactgcttaccaacctc 3’ Rev 5’atttctgtcaatgggcaagg 3’	ChIP analysis	5’ of PARK7 SVA chr1:8011671-8011836	166bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 64^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’cggaagtggacctacgtcat 3’ Rev 5’ tgacaccgagttctgtgagg 3’	ChIP analysis	Minor Promoter of PARK7 gene chr1:8014352-8014509	158bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 64^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins
For 5’ agggtggcggtagagactgt 3’ Rev 5’ cacaccaggctgaaaatgaa 3’	ChIP analysis	Major Promoter of PARK7 gene chr1:8021524-8021793	270bp (gDNA)	1 cycle of 95^{o}C for 2mins, 35 cycles of 95^{o}C for 30s, 64^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins

Primers	**Application for Primer Use**	**Target Region**	**Expected Product Size**	**PCR Conditions**
For 5’ agggtggcggtagagactgt 3’ Rev 5’ cacaccaggctgaaaatgaa 3’	ChIP analysis	Region within a gene desert chr8:127755406-127755513	108bp (gDNA)	1 cycle of 95^{o}C for 2mins, 30 cycles of 95^{o}C for 30s, 60^{o}C for 30s and 72^{o}C for 30s, 1 cycle of 72^{o}C for 5mins

Table A1: Details of the primers used in amplification of target regions using PCR. The sequence of the forward and reverse primers, the application for their use, the target for amplification, product size and cycling conditions are listed for all primers used.

A.2 Information regarding the antibodies used in the ChIP protocol

Antibody	Host Species	Immunogen	Company and Cat No.	Amount used per IP with 10µg of sheared chromatin
Anti - H3	Rabbit (Polyclonal)	Synthetic peptide conjugated to KLH derived from within residues 100 to the C-terminus of Human Histone H3.	Abcam 1791	5µg
Anti - H3K9M3	Rabbit (Polyclonal)	Synthetic peptide conjugated to KLH derived from within residues 1-100 of Human H3, trimethylated at lysine 9.	Abcam, 8898	3.5µg
Anti - H3K4M2	Rabbit (Monoclonal)	A synthetic (Dimethyl K) peptide corresponding to residues surrounding Lys4 of Histone H3.	Abcam 32356	3µg
Anti - RNA PolII CTD phospho Ser5	Rat (Monoclonal)	This RNA pol II CTD phospho Ser5 antibody was raised against a peptide containing the RNA pol II CTD sequence phosphorylated at serine 5.	Active Motif 61085	5µg
Anti - Nucleolin	Mouse (Monoclonal)	Human nucleolin protein from Raji cell extract.	Abcam 13541	4µg
Anti - Sp1	Rabbit (Polyclonal)	Full length human Sp1 protein.	Millipore 17-601	4µg
Anti - hnRNPK	Rabbit (Polyclonal)	Synthetic peptide corresponding to a sequence from the C-terminus of isoform a of human hnRNP K.	Abcam 70492	2µg
Anti - CTCF	Mouse (Monoclonal)	His-tagged recombinant protein corresponding to human CTCF.	Millipore 17-10044	2µg

Table A2: Details of the antibodies used in chromatin immunoprecipitation protocol. The target of the antibody, host species, company the antibody was obtained from and the amount used in the ChIP protocol are listed in the table.

A.3 Screen shot of the PARK7 gene locus in UCSC genome browser

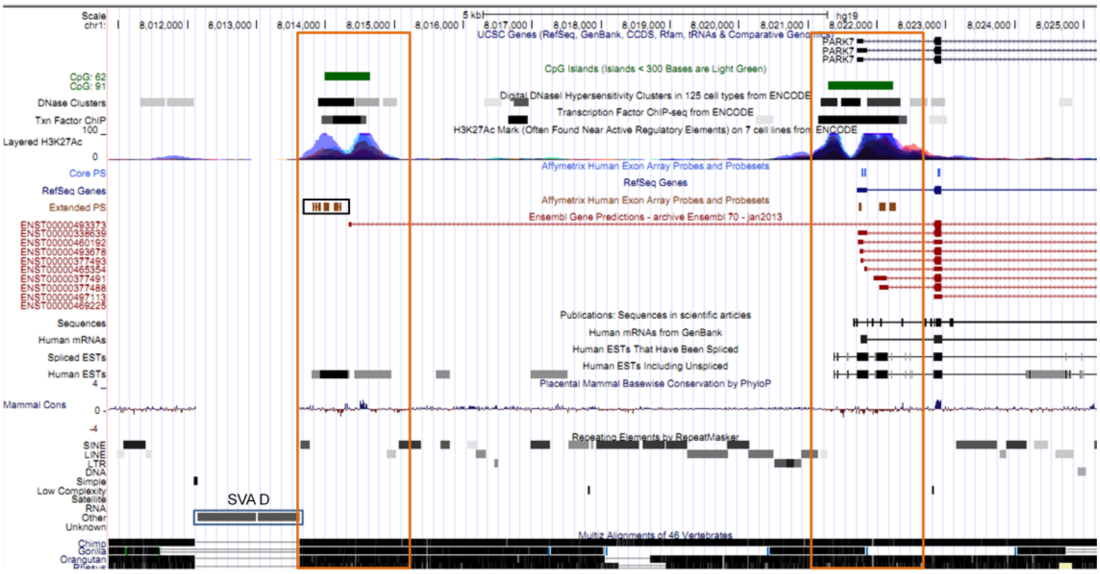

Scale
chr1:
5 kb
hg19
8,012,000
8,013,000
8,014,000
8,015,000
8,016,000
8,017,000
8,018,000
8,019,000
8,020,000
8,021,000
8,022,000
8,023,000
8,024,000
8,025,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PARK7
PARK7
PARK7
CpG: 62
CpG: 91
CpG Islands (Islands < 300 Bases are Light Green)
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
100 _
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
0
Core PS
Affymetrix Human Exon Array Probes and Probesets
RefSeq Genes
RefSeq Genes
Extended PS
Affymetrix Human Exon Array Probes and Probesets
Ensembl Gene Predictions - archive Ensembl 70 - jan2013
ENST00000493373
ENST00000338639
ENST00000460192
ENST00000493678
ENST00000377493
ENST00000465354
ENST00000377491
ENST00000377488
ENST00000497113
ENST00000469225
Sequences
Publications: Sequences in scientific articles
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Human ESTs
Human ESTs Including Unspliced
4 _
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0 -
-4
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
SVA D
Multiz Alignments of 46 Vertebrates
Chimp
Gorilla
Orangutan
Rhesus

Figure A1: UCSC genome browser displaying data suggesting there are two PARK7 TSSs. The two orange boxes highlight the two PARK7 TSSs which are both demonstrating features such as a CpG island, DNase clusters, transcription factor binding and histone marks associated with active regulatory elements. The region boxed in orange on the right is the characterised promoter with the PARK7 gene transcripts originating here and is termed the major promoter. The orange box on the left of the image was predicted to be another TSS and was termed the minor promoter. There was a transcript predicted to originate from this minor promoter by the Archive Ensembl database and the transcripts from this browser are shown in red. There are Affymetrix Human Exon Array extended probe sets in the region of the minor TSS boxed in black and human ESTs also present. There is a SVA D located less than 1kb upstream of the minor TSS and 8kb upstream of the major TSS which is boxed in blue.

Publications

Published

Savage AL, Bubb VJ, Breen G, Quinn JP. 2013. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC evolutionary biology* **13**(1): 101.

Submitted

An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS

Abigail L. Savage*, Thomas P. Wilm*, Kejhal Khursheed*, Aleksey Shatunov, Karen E. Morrison, Pamela J. Shaw, Christopher E. Shaw, Bradley Smith, Gerome Breen, Ammar Al-Chalabi, Diana Moss, Vivien J. Bubb and John P. Quinn.

*Authors contributed equally to production of data for this manuscript

The evolution of neuropeptide gene expression patterns driving changes in behaviour?

John P. Quinn, Alix Warburton, Paul Myers, Abigail L. Savage and Vivien J. Bubb.

Abbreviations

ChIP – chromatin immunoprecipitation

CNV – copy number variation

ECR – evolutionary conserved region

ESTs – expressed sequence tags

FALS – familial amyotrophic lateral sclerosis

GWAS – genome wide assocication studies

G4 – G-quadruplex

HERV – human endogenous retrovirus

LINE – long interspersed element

LTR – long terminal repeats

NMD – nonsense mediated mRNA decay

ORF – open reading frame

PBS – phosphate buffered saline

PD – Parkinson’s Disease

PIC – protease inhinitor cocktail

PMSF – phenylmethanesulfonyl fluoride

PP – processed pseudogene

RNP – ribonucleoprotein complex

SALS – sporadic amyotrophic lateral sclerosis

SINE – short interspersed element

SNP – single nucleotide polymorphism

SVA – SINE-VNTR-Alu

TE – transposable element

TF – transcription factor

TPRT – target primed reverse transcription

TR – tandem repeat

TSD – target site duplication

TSS – transcriptional start site

UTR – untranslated region

VNTR –varaiable number tandem repeat

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Akman HO, Davidzon G, Tanji K, Macdermott EJ, Larsen L, Davidson MM, Haller RG, Szczepaniak LS, Lehman TJ, Hirano M et al. 2010. Neutral lipid storage disease with subclinical myopathy due to a retrotransposal insertion in the PNPLA2 gene. *Neuromuscular disorders : NMD* **20**(6): 397-402.

Al-Chalabi A, Fang F, Hanby MF, Leigh PN, Shaw CE, Ye W, Rijsdijk F. 2010. An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of neurology, neurosurgery, and psychiatry* **81**(12): 1324-1326.

Alfahad T, Nath A. 2013. Retroviruses and amyotrophic lateral sclerosis. *Antiviral research* **99**(2): 180-187.

Ali FR, Vasiliou SA, Haddley K, Paredes UM, Roberts JC, Miyajima F, Klenova E, Bubb VJ, Quinn JP. 2010. Combinatorial interaction between two human serotonin transporter gene variable number tandem repeats and their regulation by CTCF. *Journal of neurochemistry* **112**(1): 296-306.

Amiry N, Kong X, Muniraj N, Kannan N, Grandison PM, Lin J, Yang Y, Vouyovitch CM, Borges S, Perry JK et al. 2009. Trefoil factor-1 (TFF1) enhances oncogenicity of mammary carcinoma cells. *Endocrinology* **150**(10): 4473-4483.

Andrews WD, Tuke PW, Al-Chalabi A, Gaudin P, Ijaz S, Parton MJ, Garson JA. 2000. Detection of reverse transcriptase activity in the serum of patients with motor neurone disease. *Journal of medical virology* **61**(4): 527-532.

Anguelova M, Benkelfat C, Turecki G. 2003. A systematic review of association studies investigating genes coding for serotonin receptors and the serotonin transporter: II. Suicidal behavior. *Molecular psychiatry* **8**(7): 646-653.

Ariga H, Takahashi-Niki K, Kato I, Maita H, Niki T, Iguchi-Ariga SM. 2013. Neuroprotective function of DJ-1 in Parkinson's disease. *Oxidative medicine and cellular longevity* **2013**: 683920.

Babushok DV, Kazazian HH, Jr. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Human mutation* **28**(6): 527-539.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**(7374): 534-537.

Balaj L, Lessard R, Dai L, Cho YJ, Pomeroy SL, Breakefield XO, Skog J. 2011. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nature communications* **2**: 180.

Bantysh OB, Buzdin AA. 2009. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochemistry Biokhimiia* **74**(12): 1393-1399.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nature reviews Genetics* **3**(5): 370-379.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics* **12**: 187-215.

Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M. 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *Journal of virology* **79**(19): 12507-12514.

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**(2): 933-951.

Bernheimer H, Birkmayer W, Hornykiewicz O, Jellinger K, Seitelberger F. 1973. Brain dopamine and the syndromes of Parkinson and Huntington. Clinical, morphological and neurochemical correlations. *Journal of the neurological sciences* **20**(4): 415-455.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes & development* **16**(1): 6-21.

Bonello GB, Pham MH, Begum K, Sigala J, Sataranatarajan K, Mummidi S. 2011. An evolutionarily conserved TNF-alpha-responsive enhancer in the far upstream region of human CCL2 locus influences its gene expression. *J Immunol* **186**(12): 7025-7038.

Bonifati V, Rizzu P, Squitieri F, Krieger E, Vanacore N, van Swieten JC, Brice A, van Duijn CM, Oostra B, Meco G et al. 2003. DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* **24**(3): 159-160.

Bratthauer GL, Cardiff RD, Fanning TG. 1994. Expression of LINE-1 retrotransposons in human breast cancer. *Cancer* **73**(9): 2333-2336.

Brazda V, Laister RC, Jagelska EB, Arrowsmith C. 2011. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC molecular biology* **12**: 33.

Breen G, Collier D, Craig I, Quinn J. 2008. Variable number tandem repeats as agents of functional regulation in the genome. *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society* **27**(2): 103-104, 108.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America* **100**(9): 5280-5285.

Byrne S, Walsh C, Lynch C, Bede P, Elamin M, Kenna K, McLaughlin R, Hardiman O. 2011. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of neurology, neurosurgery, and psychiatry* **82**(6): 623-627.

Canet-Aviles RM, Wilson MA, Miller DW, Ahmad R, McLendon C, Bandyopadhyay S, Baptista MJ, Ringe D, Petsko GA, Cookson MR. 2004. The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization. *Proc Natl Acad Sci U S A* **101**(24): 9103-9108.

Carrasco X, Rothhammer P, Moraga M, Henriquez H, Chakraborty R, Aboitiz F, Rothhammer F. 2006. Genotypic interaction between DRD4 and DAT1 loci is a high risk factor for attention-deficit/hyperactivity disorder in Chilean families. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **141B**(1): 51-54.

Chenais B. 2013. Transposable elements and human cancer: a causal relationship? *Biochimica et biophysica acta* **1835**(1): 28-35.

Chenais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509**(1): 7-15.

Chio A, Calvo A, Moglia C, Ossola I, Brunetti M, Sbaiz L, Lai SL, Abramzon Y, Traynor BJ, Restagno G. 2011. A de novo missense mutation of the FUS gene in a "true" sporadic ALS case. *Neurobiology of aging* **32**(3): 553 e523-556.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**(49): 19428-19433.

Clark DW, Phang T, Edwards MG, Geraci MW, Gillespie MN. 2012. Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free radical biology & medicine* **53**(1): 51-59.

Clements CM, McNally RS, Conti BJ, Mak TW, Ting JP. 2006. DJ-1, a cancer- and Parkinson's disease-associated protein, stabilizes the antioxidant transcriptional master regulator Nrf2. *Proceedings of the National Academy of Sciences of the United States of America* **103**(41): 15091-15096.

Cogoi S, Xodo LE. 2006. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic acids research* **34**(9): 2536-2549.

Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian HH, Jr. 2005. Two independent retrotransposon insertions at the same site within the coding region of BTK. *Human mutation* **25**(3): 324-325.

Coppede F. 2012. Genetics and epigenetics of Parkinson's disease. *TheScientificWorldJournal* **2012**: 489830.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nature reviews Genetics* **10**(10): 691-703.

Corrado L, Del Bo R, Castellotti B, Ratti A, Cereda C, Penco S, Soraru G, Carlomagno Y, Ghezzi S, Pensato V et al. 2010. Mutations of FUS gene in sporadic amyotrophic lateral sclerosis. *Journal of medical genetics* **47**(3): 190-194.

Corti O, Lesage S, Brice A. 2011. What genetics tells us about the causes and mechanisms of Parkinson's disease. *Physiological reviews* **91**(4): 1161-1218.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**(7259): 1127-1131.

Damert A, Raiz J, Horn AV, Lower J, Wang H, Xing J, Batzer MA, Lower R, Schumann GG. 2009. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome research* **19**(11): 1992-2008.

Davidson S, Miller KA, Dowell A, Gildea A, Mackenzie A. 2006. A remote and highly conserved enhancer supports amygdala specific expression of the gene encoding the anxiogenic neuropeptide substance-P. *Mol Psychiatry* **11**(4): 323, 410-321.

De Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA, Manivannan J, Peterson AL, Kreiling JA, Neretti N, Sedivy JM. 2013. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. *Aging cell* **12**(2): 247-256.

De S, Michor F. 2011. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nature structural & molecular biology* **18**(8): 950-955.

Deng HX, Zhai H, Bigio EH, Yan J, Fecto F, Ajroud K, Mishra M, Ajroud-Driss S, Heller S, Sufit R et al. 2010. FUS-immunoreactive inclusions are a common feature in sporadic and non-SOD1 familial amyotrophic lateral sclerosis. *Annals of neurology* **67**(6): 739-748.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics* **35**(1): 41-48.

Ding W, Lin L, Chen B, Dai J. 2006. L1 elements, processed pseudogenes and retrogenes in mammalian genomes. *IUBMB life* **58**(12): 677-685.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences of the United States of America*.

Douville R, Liu J, Rothstein J, Nath A. 2011. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. *Annals of neurology* **69**(1): 141-151.

Eerola J, Hernandez D, Launes J, Hellstrom O, Hague S, Gulick C, Johnson J, Peuralinna T, Hardy J, Tienari PJ et al. 2003. Assessment of a DJ-1 (PARK7) polymorphism in Finnish PD. *Neurology* **61**(7): 1000-1002.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A et al. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**(3): 483-496.

Fairbanks DJ, Fairbanks AD, Ogden TH, Parker GJ, Maughan PJ. 2012. NANOGP8: evolution of a human-specific retro-oncogene. *G3 (Bethesda)* **2**(11): 1447-1457.

Farre D, Roset R, Huerta M, Adsuara JE, Rosello L, Alba MM, Messeguer X. 2003. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic acids research* **31**(13): 3651-3653.

Faulkner GJ. 2011. Retrotransposons: mobile and mutagenic from conception to death. *FEBS letters* **585**(11): 1589-1594.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nature genetics* **41**(5): 563-571.

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**(5): 905-916.

Fletcher TM, Sun D, Salazar M, Hurley LH. 1998. Effect of DNA secondary structure on human telomerase activity. *Biochemistry* **37**(16): 5536-5541.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *Journal of molecular biology* **196**(2): 261-282.

-. 1994. Transcripts and CpG islands associated with the pro-opiomelanocortin gene and other neurally expressed genes. *Journal of molecular endocrinology* **12**(3): 365-382.

Gonzalez V, Guo K, Hurley L, Sun D. 2009. Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *The Journal of biological chemistry* **284**(35): 23622-23635.

Gonzalez V, Hurley LH. 2010. The c-MYC NHE III(1): function and regulation. *Annual review of pharmacology and toxicology* **50**: 111-129.

Goodier JL, Kazazian HH, Jr. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**(1): 23-35.

Griffiths DJ. 2001. Endogenous retroviruses in the human genome sequence. *Genome biology* **2**(6): REVIEWS1017.

Guindalini C, Howard M, Haddley K, Laranjeira R, Collier D, Ammar N, Craig I, O'Gara C, Bubb VJ, Greenwood T et al. 2006. A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample. *Proceedings of the National Academy of Sciences of the United States of America* **103**(12): 4552-4557.

Haddley K, Bubb VJ, Breen G, Parades-Esquivel UM, Quinn JP. 2012. Behavioural Genetics of the Serotonin Transporter. *Current topics in behavioral neurosciences*.

Haddley K, Vasiliou AS, Ali FR, Paredes UM, Bubb VJ, Quinn JP. 2008. Molecular genetics of monoamine transporters: relevance to brain disorders. *Neurochemical research* **33**(4): 652-667.

Halling KC, Lazzaro CR, Honchel R, Bufill JA, Powell SM, Arndt CA, Lindor NM. 1999. Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. *Human heredity* **49**(2): 97-102.

Han K, Konkel MK, Xing J, Wang H, Lee J, Meyer TJ, Huang CT, Sandifer E, Hebert K, Barnes EW et al. 2007. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. *Science* **316(**5822): 238-240.

Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HH, Jr. 2009. Exon-trapping mediated by the human retrotransposon SVA. *Genome research* **19**(11): 1983-1991.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH, Jr. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human molecular genetics* **20**(17): 3386-3400.

Hancks DC, Kazazian HH, Jr. 2010. SVA retrotransposons: Evolution and genetic instability. *Seminars in cancer biology* **20**(4): 234-245.

-. 2012. Active human retrotransposons: variation and disease. *Current opinion in genetics & development* **22**(3): 191-203.

Hancks DC, Mandal PK, Cheung LE, Kazazian HH, Jr. 2012. The minimal active human SVA retrotransposon requires only the 5'-hexamer and Alu-like domains. *Molecular and cellular biology* **32**(22): 4718-4726.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic acids research* **33**(8): 2374-2383.

Hassoun H, Coetzer TL, Vassiliadis JN, Sahr KE, Maalouf GJ, Saad ST, Catanzariti L, Palek J. 1994. A novel mobile element inserted in the alpha spectrin gene: spectrin dayton. A truncated alpha spectrin associated with hereditary elliptocytosis. *The Journal of clinical investigation* **94**(2): 643-648.

Herman AI, Kaiss KM, Ma R, Philbeck JW, Hasan A, Dasti H, DePetrillo PB. 2005. Serotonin transporter promoter polymorphism and monoamine oxidase type A VNTR allelic variants together influence alcohol binge drinking risk in young women. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **133B**(1): 74-78.

Hohjoh H, Singer MF. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *The EMBO journal* **16**(19): 6034-6043.

Houck CM, Rinehart FP, Schmid CW. 1979. A ubiquitous family of repeated DNA sequences in the human genome. *Journal of molecular biology* **132**(3): 289-306.

Houlden H, Singleton AB. 2012. The genetics and neuropathology of Parkinson's disease. *Acta neuropathologica* **124**(3): 325-338.

Huppert JL, Balasubramanian S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic acids research* **33**(9): 2908-2916.

-. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic acids research* **35**(2): 406-413.

Ishikawa S, Taira T, Takahashi-Niki K, Niki T, Ariga H, Iguchi-Ariga SM. 2010. Human DJ-1-specific transcriptional activation of tyrosine hydroxylase gene. *The Journal of biological chemistry* **285**(51): 39718-39731.

Jowaed A, Schmitt I, Kaut O, Wullner U. 2010. Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **30**(18): 6355-6359.

Kaer K, Speek M. 2013. Retroelements in human disease. *Gene* **518**(2): 231-241.

Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH, Jr. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes & development* **23**(11): 1303-1312.

Kavanagh DH, Dwyer S, O'Donovan MC, Owen MJ. 2013a. The ENCODE project: implications for psychiatric genetics. *Molecular psychiatry*.

-. 2013b. The ENCODE project: implications for psychiatric genetics. *Molecular psychiatry* **18**(5): 540-542.

Kawate T, Iwaya K, Kikuchi R, Kaise H, Oda M, Sato E, Hiroi S, Matsubara O, Kohno N. 2013. DJ-1 protein expression as a predictor of pathological complete remission after neoadjuvant chemotherapy in breast cancer patients. *Breast cancer research and treatment* **139**(1): 51-59.

Kazazian HH, Jr. 2004. Mobile elements: drivers of genome evolution. *Science* **303**(5664): 1626-1632.

Keyser RJ, van der Merwe L, Venter M, Kinnear C, Warnich L, Carr J, Bardien S. 2009. Identification of a novel functional deletion variant in the 5'-UTR of the DJ-1 gene. *BMC medical genetics* **10**: 105.

Kim HS. 2012. Genomic impact, chromosomal distribution and transcriptional regulation of HERV elements. *Molecules and cells* **33**(6): 539-544.

Kim RH, Peters M, Jang Y, Shi W, Pintilie M, Fletcher GC, DeLuca C, Liepa J, Zhou L, Snow B et al. 2005. DJ-1, a novel regulator of the tumor suppressor PTEN. *Cancer cell* **7**(3): 263-273.

Kim YJ, Lee J, Han K. 2012. Transposable Elements: No More 'Junk DNA'. *Genomics & informatics* **10**(4): 226-233.

Kleinjan DA, Seawright A, Childs AJ, van Heyningen V. 2004. Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. *Developmental biology* **265**(2): 462-477.

Klenova E, Scott AC, Roberts J, Shamsuddin S, Lovejoy EA, Bergmann S, Bubb VJ, Royer HD, Quinn JP. 2004. YB-1 and CTCF differentially regulate the 5-HTT polymorphic intron 2 enhancer which predisposes to a variety of neurological disorders. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **24**(26): 5966-5973.

Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M et al. 1998. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394**(6691): 388-392.

Kramerov DA, Vassetzky NS. 2011. SINEs. *Wiley interdisciplinary reviews RNA* **2**(6): 772-786.

Kubo S, Seleme MC, Soifer HS, Perez JL, Moran JV, Kazazian HH, Jr., Kasahara N. 2006. L1 retrotransposition in nondividing and primary human somatic cells. *Proceedings of the National Academy of Sciences of the United States of America* **103**(21): 8036-8041.

Kulski JK, Shigenari A, Inoko H. 2010. Polymorphic SVA retrotransposons at four loci and their association with classical HLA class I alleles in Japanese, Caucasians and African Americans. *Immunogenetics* **62**(4): 211-230.

Kwiatkowski TJ, Jr., Bosco DA, Leclerc AL, Tamrazian E, Vanderburg CR, Russ C, Davis A, Gilchrist J, Kasarskis EJ, Munsat T et al. 2009. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **323**(5918): 1205-1208.

Lai SL, Abramzon Y, Schymick JC, Stephan DA, Dunckley T, Dillman A, Cookson M, Calvo A, Battistini S, Giannini F et al. 2011. FUS mutations in sporadic amyotrophic lateral sclerosis. *Neurobiology of aging* **32**(3): 550 e551-554.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CpG islands as gene markers in the human genome. *Genomics* **13**(4): 1095-1107.

Lattante S, Conte A, Zollino M, Luigetti M, Del Grande A, Marangi G, Romano A, Marcaccio A, Meleo E, Bisogni G et al. 2012. Contribution of major amyotrophic lateral sclerosis genes to the etiology of sporadic disease. *Neurology* **79**(1): 66-72.

Le Naour F, Misek DE, Krause MC, Deneux L, Giordano TJ, Scholl S, Hanash SM. 2001. Proteomics-based identification of RS/DJ-1 as a novel circulating tumor antigen in breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **7**(11): 3328-3335.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, Lohr JG, Harris CC, Ding L, Wilson RK et al. 2012a. Landscape of somatic retrotransposition in human cancers. *Science* **337**(6097): 967-971.

Lee J, Ha J, Son SY, Han K. 2012b. Human Genomic Deletions Generated by SVA-Associated Events. *Comparative and functional genomics* **2012**: 807270.

Lin L, Shen S, Tye A, Cai JJ, Jiang P, Davidson BL, Xing Y. 2008. Diverse splicing patterns of exonized Alu elements in human tissues. *PLoS genetics* **4**(10): e1000225.

Logroscino G, Traynor BJ, Hardiman O, Chio A, Mitchell D, Swingler RJ, Millul A, Benn E, Beghi E. 2010. Incidence of amyotrophic lateral sclerosis in Europe. *Journal of neurology, neurosurgery, and psychiatry* **81**(4): 385-390.

MacKeigan JP, Clements CM, Lich JD, Pope RM, Hod Y, Ting JP. 2003. Proteomic profiling drug-induced apoptosis in non-small cell lung carcinoma: identification of RS/DJ-1 and RhoGDIalpha. *Cancer research* **63**(20): 6928-6934.

MacKenzie A, Quinn J. 1999a. A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer- like properties in the mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America* **96**(26): 15251-15255.

-. 1999b. A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer-like properties in the mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America* **96**(26): 15251-15255.

MacKenzie A, Quinn JP. 2004. Post-genomic approaches to exploring neuropeptide gene mis-expression in disease. *Neuropeptides* **38**(1): 1-15.

Mackenzie IR, Rademakers R, Neumann M. 2010. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *Lancet neurology* **9**(10): 995-1007.

Macleod D, Ali RR, Bird A. 1998. An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands. *Molecular and cellular biology* **18**(8): 4433-4443.

Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, Matsumoto S, Tabuena MD, Maranon E, Dantes M et al. 2007. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. *American journal of human genetics* **80**(3): 393-406.

Martin C, Ross M, Chapman KE, Andrew R, Bollina P, Seckl JR, Habib FK. 2004. CYP7B generates a selective estrogen receptor beta agonist in human prostate. *The Journal of clinical endocrinology and metabolism* **89**(6): 2928-2935.

Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254**(5039): 1808-1810.

Mc CB. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* **36**(6): 344-355.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**(7337): 216-219.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome research* **12**(10): 1483-1495.

Membrino A, Cogoi S, Pedersen EB, Xodo LE. 2011. G4-DNA formation in the HRAS promoter and rational design of decoy oligonucleotides for cancer therapy. *PloS one* **6**(9): e24421.

Menendez L, Benigno BB, McDonald JF. 2004. L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. *Molecular cancer* **3**: 12.

Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM. 2002. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* **18**(2): 333-334.

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**(6771): 785-789.

Michelhaugh SK, Fiskerstrand C, Lovejoy E, Bannon MJ, Quinn JP. 2001. The dopamine transporter gene (SLC6A3) variable number of tandem repeats domain enhances transcription in dopamine neurons. *Journal of neurochemistry* **79**(5): 1033-1038.

Migliore L, Coppede F. 2009. Genetics, environmental factors and the emerging role of epigenetics in neurodegenerative diseases. *Mutation research* **667**(1-2): 82-97.

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research* **52**(3): 643-645.

Mills RE, Bennett EA, Iskow RC, Luttig CT, Tsui C, Pittard WS, Devine SE. 2006. Recently mobilized transposons in the human and chimpanzee genomes. *American journal of human genetics* **78**(4): 671-679.

Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y. 1992. Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic acids research* **20**(12): 3139-3145.

Mine M, Chen JM, Brivet M, Desguerre I, Marchant D, de Lonlay P, Bernard A, Ferec C, Abitbol M, Ricquier D et al. 2007. A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. *Human mutation* **28**(2): 137-142.

Mitchell JC, McGoldrick P, Vance C, Hortobagyi T, Sreedharan J, Rogelj B, Tudor EL, Smith BN, Klasen C, Miller CC et al. 2013. Overexpression of human wild-type FUS causes progressive motor neuron degeneration in an age- and dose-dependent fashion. *Acta neuropathologica* **125**(2): 273-288.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**(5): 917-927.

Munafo MR, Johnstone EC. 2008. Smoking status moderates the association of the dopamine D4 receptor (DRD4) gene VNTR polymorphism with selective processing of smoking-related cues. *Addiction biology* **13**(3-4): 435-439.

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**(7044): 903-910.

Muotri AR, Marchetto MC, Coufal NG, Gage FH. 2007. The necessary junk: new functions for transposable elements. *Human molecular genetics* **16 Spec No. 2**: R159-167.

Nagakubo D, Taira T, Kitaura H, Ikeda M, Tamai K, Iguchi-Ariga SM, Ariga H. 1997. DJ-1, a novel oncogene which transforms mouse NIH3T3 cells in cooperation with ras. *Biochemical and biophysical research communications* **231**(2): 509-513.

Nellaker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome biology* **13**(6): R45.

Neumann M, Rademakers R, Roeber S, Baker M, Kretzschmar HA, Mackenzie IR. 2009. A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain : a journal of neurology* **132**(Pt 11): 2922-2931.

Nussbaum RL, Ellis CE. 2003. Alzheimer's disease and Parkinson's disease. *The New England journal of medicine* **348**(14): 1356-1364.

Oda M, Makita M, Iwaya K, Akiyama F, Kohno N, Tsuchiya B, Iwase T, Matsubara O. 2012. High levels of DJ-1 protein in nipple fluid of patients with breast cancer. *Cancer science* **103**(6): 1172-1176.

Oganesian L, Bryan TM. 2007. Physiological relevance of telomeric G-quadruplex formation: a potential drug target. *BioEssays : news and reviews in molecular, cellular and developmental biology* **29**(2): 155-165.

Ono M, Kawakami M, Takezawa T. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic acids research* **15**(21): 8725-8737.

Paredes UM, Bubb VJ, Haddley K, Macho GA, Quinn JP. 2011. An evolutionary conserved region (ECR) in the human dopamine receptor D4 gene supports reporter gene expression in primary cultures derived from the rat cortex. *BMC neuroscience* **12**: 46.

Paredes UM, Quinn JP, D'Souza UM. 2012. Allele-specific transcriptional activity of the variable number of tandem repeats in 5' region of the DRD4 gene is stimulus specific in human neuronal cells. *Genes, brain, and behavior*.

Pavlicek A, Gentles AJ, Paces J, Paces V, Jurka J. 2006. Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends in genetics : TIG* **22**(2): 69-73.

Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R et al. 1997. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**(5321): 2045-2047.

Porter W, Saville B, Hoivik D, Safe S. 1997. Functional synergy between the transcription factor Sp1 and the estrogen receptor. *Mol Endocrinol* **11**(11): 1569-1580.

Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome research* **16**(7): 855-863.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Lower J, Stratling WH, Lower R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic acids research* **40**(4): 1666-1683.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual review of genetics* **46**: 21-42.

Reik W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**(7143): 425-432.

Roberts J, Scott AC, Howard MR, Breen G, Bubb VJ, Klenova E, Quinn JP. 2007. Differential regulation of the serotonin transporter gene by lithium is mediated by transcription factors, CCCTC binding protein and Y-box binding protein 1, through the polymorphic intron 2 variable number tandem repeat. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **27**(11): 2793-2801.

Rohrer J, Minegishi Y, Richter D, Eguiguren J, Conley ME. 1999. Unusual mutations in Btk: an insertion, a duplication, an inversion, and four large deletions. *Clin Immunol* **90**(1): 28-37.

Savage AL, Bubb VJ, Breen G, Quinn JP. 2013. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC evolutionary biology* **13**(1): 101.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS one* **8**(2): e54710.

Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, Cooley T, Heath P, Smith KD, Margolet L. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**(2): 113-125.

Sham PC, Curtis D. 1995. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Annals of human genetics* **59**(Pt 1): 97-105.

Shanley L, Davidson S, Lear M, Thotakura AK, McEwan IJ, Ross RA, MacKenzie A. 2010. Long-range regulatory synergy is required to allow control of the TAC1 locus by MEK/ERK signalling in sensory neurones. *Neuro-Signals* **18**(3): 173-185.

Shanley L, Lear M, Davidson S, Ross R, MacKenzie A. 2011. Evidence for regulatory diversity and auto-regulation at the TAC1 locus in sensory neurones. *Journal of neuroinflammation* **8**: 10.

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *The Journal of biological chemistry* **269**(11): 8466-8476.

Shin W, Lee J, Son SY, Ahn K, Kim HS, Han K. 2013. Human-specific HERV-K insertion causes genomic variations in the human genome. *PloS one* **8**(4): e60605.

Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH. 2002. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences of the United States of America* **99**(18): 11593-11598.

Sorek R. 2009. When new exons are born. *Heredity* **103**(4): 279-280.

Sproviero W, La Bella V, Mazzei R, Valentino P, Rodolico C, Simone IL, Logroscino G, Ungaro C, Magariello A, Patitucci A et al. 2012. FUS mutations in sporadic amyotrophic lateral sclerosis: clinical and genetic analysis. *Neurobiology of aging* **33**(4): 837 e831-835.

Steele AJ, Al-Chalabi A, Ferrante K, Cudkowicz ME, Brown RH, Jr., Garson JA. 2005. Detection of serum reverse transcriptase activity in patients with ALS and unaffected blood relatives. *Neurology* **64**(3): 454-458.

Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, Sasaki C, Costa J, Lizardi PM. 2009. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene* **448**(2): 151-167.

Taira T, Saito Y, Niki T, Iguchi-Ariga SM, Takahashi K, Ariga H. 2004. DJ-1 has a role in antioxidative stress to prevent cell death. *EMBO reports* **5**(2): 213-218.

Taira T, Takahashi K, Kitagawa R, Iguchi-Ariga SM, Ariga H. 2001. Molecular cloning of human and mouse DJ-1 genes and identification of Sp1-dependent activation of the human DJ-1 promoter. *Gene* **263**(1-2): 285-292.

Takahashi K, Taira T, Niki T, Seino C, Iguchi-Ariga SM, Ariga H. 2001. DJ-1 positively regulates the androgen receptor by impairing the binding of PIASx alpha to the receptor. *The Journal of biological chemistry* **276**(40): 37556-37563.

Takasu M, Hayashi R, Maruya E, Ota M, Imura K, Kougo K, Kobayashi C, Saji H, Ishikawa Y, Asai T et al. 2007. Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon. *Tissue antigens* **70**(2): 144-150.

Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu CC, Mori K, Oda T, Kuga A, Kurahashi H, Akman HO, DiMauro S et al. 2011. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. *Nature* **478**(7367): 127-131.

Tsuchiya B, Iwaya K, Kohno N, Kawate T, Akahoshi T, Matsubara O, Mukai K. 2012. Clinical significance of DJ-1 as a secretory molecule: retrospective study of DJ-1 expression at mRNA and protein levels in ductal carcinoma of the breast. *Histopathology* **61**(1): 69-77.

Ullu E, Tschudi C. 1984. Alu sequences are processed 7SL RNA genes. *Nature* **312**(5990): 171-172.

van der Klift HM, Tops CM, Hes FJ, Devilee P, Wijnen JT. 2012. Insertion of an SVA element, a nonautonomous retrotransposon, in PMS2 intron 7 as a novel cause of Lynch syndrome. *Human mutation* **33**(7): 1051-1055.

Vance C, Rogelj B, Hortobagyi T, De Vos KJ, Nishimura AL, Sreedharan J, Hu X, Smith B, Ruddy D, Wright P et al. 2009. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* **323**(5918): 1208-1211.

Vasiliou SA, Ali FR, Haddley K, Cardoso MC, Bubb VJ, Quinn JP. 2012. The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro. *Addiction biology* **17**(1): 156-170.

Verma A. 2011. Altered RNA metabolism and amyotrophic lateral sclerosis. *Annals of Indian Academy of Neurology* **14**(4): 239-244.

Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. 1993. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *European journal of human genetics : EJHG* **1**(1): 30-36.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **103**(9): 3220-3225.

Visel A, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA. 2009a. Functional autonomy of distant-acting human enhancers. *Genomics* **93**(6): 509-513.

Visel A, Bristow J, Pennacchio LA. 2007. Enhancer identification through comparative genomics. *Seminars in cell & developmental biology* **18**(1): 140-152.

Visel A, Rubin EM, Pennacchio LA. 2009b. Genomic views of distant-acting enhancers. *Nature* **461**(7261): 199-205.

Wagenfeld A, Gromoll J, Cooper TG. 1998a. Molecular cloning and expression of rat contraception associated protein 1 (CAP1), a protein putatively involved in fertilization. *Biochemical and biophysical research communications* **251**(2): 545-549.

Wagenfeld A, Yeung CH, Strupat K, Cooper TG. 1998b. Shedding of a rat epididymal sperm protein associated with infertility induced by ornidazole and alpha-chlorohydrin. *Biology of reproduction* **58**(5): 1257-1265.

Wagner MJ. 2013. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics* **14**(4): 413-424.

Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV. 2001. Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **7**(6): 1553-1560.

Wang-Johanning F, Liu J, Rycaj K, Huang M, Tsai K, Rosen DG, Chen DT, Lu DW, Barnhart KF, Johanning GL. 2007. Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. *International journal of cancer Journal international du cancer* **120**(1): 81-90.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *Journal of molecular biology* **354**(4): 994-1007.

Warnefors M, Pereira V, Eyre-Walker A. 2010. Transposable elements: insertion pattern and impact on gene expression evolution in hominids. *Molecular biology and evolution* **27**(8): 1955-1962.

Watanabe M, Kobayashi K, Jin F, Park KS, Yamada T, Tokunaga K, Toda T. 2005. Founder SVA retrotransposal insertion in Fukuyama-type congenital muscular dystrophy and its origin in Japanese and Northeast Asian populations. *American journal of medical genetics Part A* **138**(4): 344-348.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Molecular and cellular biology* **21**(4): 1429-1439.

Wilund KR, Yi M, Campagna F, Arca M, Zuliani G, Fellin R, Ho YK, Garcia JV, Hobbs HH, Cohen JC. 2002. Molecular mechanisms of autosomal recessive hypercholesterolemia. *Human molecular genetics* **11**(24): 3019-3030.

Wingo TS, Cutler DJ, Yarab N, Kelly CM, Glass JD. 2011. The heritability of amyotrophic lateral sclerosis in a clinically ascertained United States research registry. *PloS one* **6**(11): e27985.

Wong HM, Stegle O, Rodgers S, Huppert JL. 2010. A toolbox for predicting g-quadruplex formation and stability. *Journal of nucleic acids* **2010**.

Wulff K, Gazda H, Schroder W, Robicka-Milewska R, Herrmann FH. 2000. Identification of a novel large F9 gene mutation-an insertion of an Alu repeated DNA element in exon e of the factor 9 gene. *Human mutation* **15**(3): 299.

Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences of the United States of America* **103**(47): 17608-17613.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res* **19**(9): 1516-1526.

Zabolotneva AA, Bantysh O, Suntsova MV, Efimova N, Malakhova GV, Schumann GG, Gayfullin NM, Buzdin AA. 2012. Transcriptional regulation of human-specific SVAF(1) retrotransposons by cis-regulatory MAST2 sequences. *Gene* **505**(1): 128-136.

Zhang J, Wang X, Li M, Han J, Chen B, Wang B, Dai J. 2006. NANOGP8 is a retrogene expressed in cancers. *The FEBS journal* **273**(8): 1723-1730.

Zhao Y, Du Z, Li N. 2007. Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS letters* **581**(10): 1951-1956.

Zhu ZB, Hsieh SL, Bentley DR, Campbell RD, Volanakis JE. 1992. A variable number of tandem repeats locus within the human complement C2 gene is associated with a retroposon derived from a human endogenous retrovirus. *The Journal of experimental medicine* **175**(6): 1783-1787.