Transposable elements and SVA insertion polymorphisms in gene regulation and Parkinson's Disease

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy (or other degree as appropriate) by

Ashley Hall BSc MSc MRes

September 2022

Acknowledgements

Firstly, I would like to express my immeasurable gratitude to my PhD supervisors Prof John Quinn, Dr Vivien Bubb and Dr Lakis Liloglou for their constant support and guidance throughout my studies. Even throughout the COVID-19 pandemic, their direction has allowed me to grow both a scientist and as a person. I am particularly grateful for the opportunities afforded to me by the lab, such as attendance of conferences and a placement at NIH in the US. Speaking of NIH, I would like to thank Dr Kimberley Billingsley for her guidance during my stay there and for facilitating my expeditious departure when the pandemic hit.

Secondly, I would like to thank all of the truly fantastic people that I have worked alongside in the lab for making every day a laugh, including Emma Price, Jack Marshall, Ana Illera López, Li Li, Sarah "High Key" Doran, Sarah Jones, Max Cadogan and Alexander Fröhlich. A special mention goes out to Ben Middlehurst, a man who truly knows his judo well, for not only matching my coffee addiction but for frequently lending his seemingly bottomless well of scientific knowledge. I will truly miss Burrito Fridays at Quinn lab.

I am truly grateful to my parents for their support on my academic path and for always providing me with the tools I needed to get to where I am now. I would like to also thank my incredible girlfriend, Josie, for her saintly patience and encouragement during the thesis writeup – I am not sure I would have made it without her provision of homemade dumplings.

I must also thank "The Bois" James, Ben and Dan for helping to keep me sane during the pandemic's lockdowns with our weekly gaming sessions. In a similar vein, I must acknowledge Michael, Harry, Charlie, Vin and Erika from my D&D group, "The Spicy Dice", for keeping me sane during the self-imposed lockdown of thesis writing.

I am also grateful to the Wellcome Trust for funding my PhD and providing me with this opportunity.

Publications

1. Grenn et al. The Parkinson's Disease Genome-Wide Association Study Locus Browser. Movement Disorders. 2020; 35(11): 2056-2067 https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mds.28197
2. Hall A, Bandres-Ciga S, Diez-Fairen M, Quinn JP, Billingsley KJ. Genetic Risk Profiling in Parkinson's Disease and Utilizing Genetics to Gain Insight into Disease-Related Biological Pathways. Int. J. Mol. Sci. 2020, 21(19), 7332. https://www.mdpi.com/1422-0067/21/19/7332
3. Hall A, Moore AK, Hernandez DG, Billingsley KJ, Bubb VJ, Quinn JP, NABEC (North American Brain Expression Consortium). A SINE-VNTR-*Alu* in the *LRIG2* Promoter Is Associated with Gene Expression at the Locus. Int. J. Mol. Sci. 2020, 21(22), 8486.
   https://www.mdpi.com/1422-0067/21/22/8486

Contents

List of figures

List of tables

Abbreviations

bp – Base pairs

cDNA – Complementary DNA

ChIP – Chromatin immunoprecipitation

CTCF – CCCTC-binding factor

C-terminus – Carboxyl-terminus

DMEM – Dulbecco's Modified Essential Media

DMSO – Dimethyl sulfoxide

EDTA – Ethylenediaminetetraacetic acid

eQTL – Expression quantitative trait locus/loci

ESC – Embryonic stem cell

FBS – Foetal Bovine Serum

FOUNDIN-PD – The Foundational Data Initiative for Parkinson's Disease

GALA – Gene-associated loop anchor

gDNA – Genomic DNA

GFP – Green fluorescent protein

GWAS – Genome-wide association study/studies

HERV – Human endogenous retrovirus

hg38 – Human genome build 38

iPSC – Induced pluripotent stem cell

KAP1 – KRAB associated protein 1

KRAB-ZFP – Krüppel-associated box zinc-finger protein

LD – Linkage Disequilibrium

LTR – Long terminal repeat

MCS – Multiple cloning site

Myo – Million years old

NIH – National Institutes of Health, USA

ORF – Open Reading Frame

PCR – Polymerase chain reaction

PD – Parkinson's Disease

piRNA – PIWI-interacting RNA

Poly-A – Polyadenine

PPMI – Parkinson's Progression Markers Initiative

RIP – Retrotransposon insertion polymorphism

SINE – Short interspersed elements

SNP – Single nucleotide polymorphism

SVA – SINE-VNTR-*Alu*

TE – Transposable element

TF – Transcription factor

TPM – Transcripts per kilobase million

TPRT – Target-primed reverse transcription

TSD – Target-site duplication

TSS – Transcriptional start site

UTR – Untranslated region

VNTR – Variable number tandem repeat

WGS – Whole genome sequencing

XDP – X-linked Dystonia Parkinsonism

Abstract

Transposable elements (TEs) have propagated throughout the genome over evolutionary history and now make up over half DNA of human DNA. Long-dismissed as parasitic selfish DNA elements, there is growing evidence that TEs can facilitate genomic evolution by introducing new regulatory factors to the loci into which they insert and modulating surrounding gene expression. Importantly, the recent and ongoing mobilisation of non-LTR retrotransposons L1, *Alu*, and SVA create loci in the genome where a particular insertion may or may not be present amongst the populace – known as retrotransposon insertion polymorphisms (RIPs). TEs are highly repetitive throughout the genome and as a result contemporary short-read genome sequencing technologies often struggle to map TEs back to a reference sequence. This means that in whole genome sequencing (WGS) projects TEs may not be captured, and sequence variation within TEs or RIPs become an underappreciated source of genomic variation. This is important not only in the study of normal physiology but for genetically complex diseases such as Parkinson's Disease (PD), where genome-wide association studies (GWAS) compare a great number of DNA sequences in search of genomic variants that are associated with the disease. The potential for functional regulatory elements such as TEs to be missed in such studies is particularly notable for PD since to date very few directly causative genetic variants have been identified. SVA retrotransposons are contemporarily active and are hominid-specific, and therefore represent prime candidates for drivers of human-specific and interpersonal differences in modulation of gene expression. Accordingly, candidate SVA elements were studied at the lab bench to functionally validate any regulatory effects associated with their sequence variation or presence vs absence

that may be missed by WGS approaches. Using a combination of wet lab techniques and bioinformatic interrogation a common SVA RIP upstream of the *LRIG2* gene promoter was examined, finding that allele dosage of this element was associated with decreased expression and increased methylation at the locus in both a cohort of human brain tissue data and in a CRISPR-genetically modified cell line model. Similarly, a novel SVA RIP within the *KANSL1* gene was studied at the *MAPT* locus, a PD-associated region for which functional variants have previously been difficult to identify due to the existence of large chromosomal inversions, termed the H1 and H2 haplotypes, with high levels of linkage disequilibrium that preclude convenient association of SNPs with local gene expression. By comparing functionally validated genotypes of this KANSL1 SVA to expression of genes at the *MAPT* locus it was found that the SVA was similar to a H2-associated SNP in prediction of expression of nearby genes, suggesting that the SVA might contribute to these H2-specific gene expression patterns. More interesting, however, was the finding that this SVA was a predictor of expression of the gene *WNT3*, which is potentially the most important gene in PD risk at the *MAPT* locus and is not associated with SNPs that predict H1/H2 haplotype. A strategy was then devised to validate these regulatory associations in a reporter gene construct model, the beginnings of which were described here. Finally, in light of these observations of gene regulatory potential associated with SVAs in normal and PD physiology it was examined whether SVAs and other TEs contribute to changes in chromatin architecture in development of the PD neuron, since TEs of all classes have been documented to provide binding sites for the architectural protein CTCF and facilitated chromatin looping. By overlapping coordinates of genes, TEs and chromatin loop anchors identified in induced pluripotent stem cells (iPSCs) before

and after dopaminergic differentiation, it was observed that the 3D architecture of the genome at genes was altered in cells derived from people with PD. Namely, the overall proportion of PD gene-associated loop anchors that colocalised with TEs was reduced in PD, and it was observed that differentiation of these iPSC lines was generally associated with an increase in loop anchor–TE–gene overlap in control lines but a decrease in overlap in PD lines.

Altogether, this thesis provides an assessment of the influences of TEs, particularly SVA retrotransposons, on the surrounding genome and their potential consequences for cellular physiology in health and disease. Furthermore, this thesis lays the groundwork for several avenues of investigation that may lead to greater understanding of how these elements shape the human genome.

Chapter 1 Introduction

1.1. Parkinson's Disease

1.1.1. Overview

By 2040 the global prevalence of Parkinson's Disease (PD) is expected to reach 12.9 million cases, doubling from 6.2 million in 2015 [1]. In addition to the heavy personal toll exacted by PD on patients and those around them, the economic costs are substantial; it has been estimated that the total economic burden of PD in the US in 2017 was $51.9 billion, which includes both direct medical costs and indirect costs such as lost earnings [2]. PD is therefore an increasingly important challenge for society, as current treatments only manage symptoms. It is imperative that a greater understanding of the aetiology of PD is achieved to lay the groundwork for treatments that may halt or reverse its progression, or even prevent its onset altogether.

PD is characterised by a loss of dopaminergic neurons in the brain, particularly in the substantia nigra of the midbrain, and widespread accumulation of intracellular α-synuclein protein aggregates such as Lewy bodies. Symptoms include motor features such as rigidity, resting tremor, loss of balance and bradykinesia (slowness of movement), amongst others [3]. A majority of PD patients also have non-motor symptoms, including cognitive impairment, mood disorders, constipation, REM (rapid eye movement) sleep perturbation, chronic pain and sensory symptoms such as hyposmia (reduced sense of smell) [4]. PD is a heterogeneous disorder in which these symptoms may manifest on wide spectrums and in varying combinations, highlighting

that clinical presentation and progression is likely influenced by a mixture of genetic and environmental factors.

1.1.2. The genetic component of PD

Perhaps surprisingly, until only two decades ago PD was considered to be wholly caused by environmental factors. Early epidemiology studies pointed to exposure to viruses and neurotoxins such as MPTP (a contaminant in the synthetic opioid MPPP) [5]; perhaps the most famous example was the strong association between the 1918 influenza pandemic and the increased rates of post-encephalitic parkinsonism that followed [6, 7]. Additionally, this non-genetic basis of PD was supported by the first cross-sectional twin studies of the disease [8]. It is now widely known, however, that PD is a complex disorder influenced by both genetic and environmental factors. Indeed, 5–10% of PD follows a classical Mendelian inheritance pattern, and around 15% of PD patients have family history of the disease [9]. The first direct evidence of a heritable component of PD came in 1997 with the identification of rare mutations in the *SNCA* gene (encoding α-synuclein) that were responsible for a monogenic (caused by a single gene) form of PD [10]. This was quickly followed by the discovery of additional rare recessive forms of PD caused by deleterious mutations in the genes *PINK1* (aka *PARK6*) [11], *PARK7* (encoding DJ-1) [12] and *PARK2 (*aka *PRKN*, encoding Parkin) [13], and the identification of autosomal dominant PD arising from mutations in *LRRK2* [14]. To date, mutations within several more genes have been associated with monogenic PD and increased disease risk (**Table 1.1**) [15].

Gene	Protein	Inheritance
ATP13A2	Lysosomal type 5 ATPase	Autosomal recessive
DNAJC16	DNAJ/HSP40 homolog subfamily C member 6	Autosomal recessive
EIF4G1	Eukaryotic translation initiation factor 4 gamma 1	Autosomal dominant
FBXO7	F-box only protein 7	Autosomal recessive
GBA	Glucocerebrocidase	Risk locus
GIGYF2	GRB interacting GYF protein 2	Autosomal dominant
HTRA2	HTRA serine peptidase 2	Autosomal dominant
LRRK2	Leucine-rich repeat kinase 2	Autosomal dominant
LRRK2	Leucine-rich repeat kinase 2	Risk locus
PARK2	Parkin	Autosomal recessive
PARK7	DJ-1	Autosomal recessive
PINK1	PTEN-induced putative kinase 1	Autosomal recessive
PLA2G6	Phospholipase A2	Autosomal recessive
SNCA	α-synuclein	Autosomal dominant
SNCA	α-synuclein	Risk locus
UCHL1	Ubiquitin c terminal hydrolase	Autosomal dominant
VPS35	Vacuolar protein sorting 35	Autosomal dominant

Table 1.1 – Loci associated with monogenic forms of PD and increased risk of disease. Adapted from Hernandez *et al.* 2016 [15].

However, most PD cases cannot be attributed to a single penetrant deleterious mutation. In light of this, many studies have pursued the 'common disease common variant' hypothesis, in which the genetic component of PD is instead considered to be the culmination of many common, low-risk alleles [16]. Genome-wide association studies (GWAS) have been invaluable in addressing this hypothesis. Briefly, the

premise of a GWAS is to determine which common genetic variants, usually single nucleotide polymorphisms (SNPs), are consistently associated with a trait by comparing the genotypes of many individuals, often in a case vs. control setup. The power of this analysis increases with the number of participants, as this allows the contributions of relatively low-effect variants to be detected. Accordingly, increasingly large PD GWAS have been performed across several populations [17-22]. The most recent and largest PD GWAS meta-analysis involved approximately 37,700 cases, 18,600 'proxy' cases (samples derived from individuals without PD but with an immediate relative with PD) and 1.4 million PD-free controls, and identified 90 independent genetic signals associated with genetically complex PD [23].

Recently, consideration of these low-effect alleles in WGS data has been shown to be useful for the generation of polygenic risk scores (PRS) – simple models that sum the weighted contributions of multiple risk variants. These have shown potential as a tool for disease prediction, with scores having been associated with risk of PD [24], age of onset [24, 25], and rate of motor and cognitive decline [26]. When the 90 risk loci identified in the most recent PD meta-analysis are incorporated into PRS for PD, those in the top PRS decile are nearly 6-fold more likely to develop PD than those in the bottom decile [23]. It has also been demonstrated that by factoring in covariates such as sex, age, family history and hyposmia, a combined risk score can be produced with considerable sensitivity and specificity [27]. In principle, the identification of individuals at greater risk of complex disease through PRS can allow the targeting of lifestyle interventions which may delay or slow disease progression.

Despite this progress in identifying genetic risk signals, it has proven difficult to directly ascribe function to the disease-associated SNPs identified through GWAS because most occur in 'non-coding' DNA. In other words, many of these genetic signals do not lie within genes or in elements known to regulate them, making their influence on protein function or expression unclear. Any regulatory effects upon genes may, therefore, be subtle and difficult to study, and SNPs may not even necessarily exert their effects on the nearest gene. Furthermore, a genetic signal may simply be in linkage disequilibrium (LD) with the true disease-causing variant – meaning that the SNP identified is typically inherited along with the causative DNA element that has, for whatever reason, escaped detection in a GWAS.

It is therefore clear that to better understand the genetic basis of PD and other genetically complex disease the context of disease-associated genetic signals must be considered, and a detailed understanding of the non-coding genome must be achieved. In doing so it is foreseeable that prediction of disease risk will be improved, for example through more refined PRS, which in turn may lead to advances in disease management.

1.2. Transposable elements

1.2.1. Overview

With the completion of the first sequence of the human genome it became apparent that as much as 98% of the genome did not encode proteins, the crucial macromolecules that participate in virtually every process in the cell [28]. It was observed that a large constituent of this non-coding DNA was sequence derived from transposable elements (TEs), which make up around 45% of the human genome (**Figure 1.1**). Simply put, TEs are sections of DNA capable of moving or copying from one site to another within the genome – although in humans the vast majority of TEs are now incapable of mobilisation due to inactivating mutations [28]. These were first described as 'mutable loci' in maize by Barbara McClintock in the 1950s [29], and are now often referred to as 'jumping genes'. Initially dismissed as parasitic junk DNA, it is now clear that TEs are important contributors to genetic diversity via their ability to introduce novel regulatory elements into a locus [30, 31], as will be discussed in this thesis introduction. Briefly, the composition of genomic transposon and retrotransposon families are outlined below, followed by detailed descriptions of the structures and transposition mechanisms of the contemporarily active LINE-1 (L1, **Section 1.2.2**), *Alu* (**Section 1.2.3**) and SINE-VNTR-*Alu* (SVA, **Section 1.2.4**) elements. Subsequently discussed are retrotransposon insertion polymorphisms (RIPs) and their roles in disease (**Section 1.2.5**), retrotransposition in somatic tissues (**Section 1.2.6**), TE control and co-option by host factors (**Section 1.2.7**), how TE insertions facilitate genome evolution by distributing regulatory elements (**Section 1.2.8**), and why SVA retrotransposons in particular represent prime candidates for driver of human-specific genome variation (**Section 1.2.9**). Finally, the challenges facing

identification of TEs using contemporary whole genome sequencing (WGS) approaches, and the implications this has for study of TEs in genome-wide association studies (GWAS) of genetically complex disease, are considered (**Section 1.2.10**). Indeed, it is immediately reasonable to speculate that such a large constituent of the genome may play a role in processes or diseases thought to have a strong basis in non-coding DNA, such as PD.

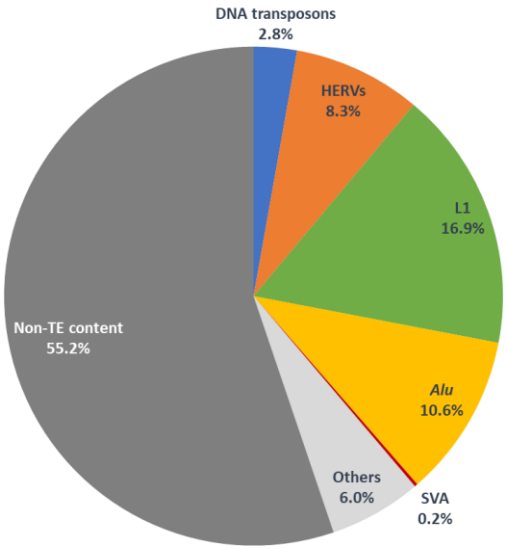


Figure 1.1 – Percentage contributions of TEs to the human genome. Adapted from Cordaux *et al.* 2009 [31].

TEs can be grouped into several families and subfamilies depending on their method of transposition and the presence of certain sequence motifs. Around 2.8% of the genome is derived from DNA transposons which can excise themselves and reinsert at distal genomic loci (**Figure 1.1**), however these elements have been inactive in

humans for around 37 million years [32]. The remaining ~42% of TE genomic content is classified as retrotransposons, which typically propagate throughout the genome via a 'copy-and-paste' mechanism first involving transcription into RNA followed by reverse transcription and insertion of a new copy at a distant genomic site, thereby accumulating in copy number over evolutionary timescales. Retrotransposon mobilisation is typically repressed in somatic cells through several mechanisms including DNA methylation and formation of heterochromatin [33], but transposition may occur during embryogenesis when euchromatin is pervasive [34, 35]. Should a novel insertion occur in a cell that later differentiates into part of the germline, it can be passed on to progeny. For TEs that have been active in recent evolutionary history there may be genomic sites where an insertion has not yet become fixed. In other words, the TE can be found to be present or absent at a given locus within the population – known as a retrotransposon insertion polymorphism (RIP).

Retrotransposons can be further subdivided into two groups based on the presence or absence of long terminal repeat (LTRs) sequences: termed LTR and non-LTR retrotransposons, respectively. LTR retrotransposons, also known as human endogenous retroviruses (HERVs), make up approximately 8.3% of the genome and the most ancient of these in humans are ~100 million years old (Myo) [28, 36]. Previous analyses have detected HERV RIPs in the human genome, identifying at least 120 HERV-K elements that are unique to humans when compared to chimpanzees and 15 that are polymorphic for presence among humans [37-40], suggesting transposition in the last ~6 million years since the chimpanzee-human divergence.

However, evidence of contemporary *de novo* HERV transposition activity was lacking until relatively recently when 4 apparently functionally intact proviruses were identified [41], with a lack of inactivating mutational drift implying recent insertion.

By contrast, the majority of TE content in the human genome is derived from the contemporarily active non-LTR retrotransposons. These include Long Interspersed Element 1 (LINE-1, abbreviated to L1), *Alu* and the composite element SINE-VNTR-*Alu* (SVA), which make up approximately 16.9%, 10.6% and 0.2% of the human genome respectively (**Figure 1.1**) [28]. Moving forward, it is important to note that much of the discussion in this thesis will refer to the 'reference' human genome – the publicly available haploid mosaic of human genomes. The most recent iteration, Genome Reference Consortium Human Build 38 (also referred to as Human Genome 38, Hg38), is an aggregation of over 60 individual genome assemblies [42]. This is a representation of an idealised human genome that is invaluable for contemporary genome analyses but does not comprehensively capture interpersonal genetic variation. In addition to SNPs, individuals may differ from the reference genome through structural variants such as RIPs.

1.2.2. L1 elements and mobilisation

The LINE retrotransposon family contains three subclasses (LINE-1, -2, and -3), but only the L1 group remains retrotransposition-competent in humans. There are over 500,000 L1 elements in the reference human genome, the vast majority of which are inactive due to 5' truncations, point mutations and rearrangements [28], with only

80-100 full-length and transposition-competent L1s in the germline of any given individual [43]. Of these, only six L1s account for more than 80% of all transposition activity, earning them the classification 'hot L1s' [43]. A full-length L1 is approximately 6 kb in length and contains a 5' untranslated region (UTR), two open reading frame (ORF) proteins ORF1p and ORF2p, a 3'-UTR and a long polyadenine (poly-A) tail [44] (**Figure 1.2**).

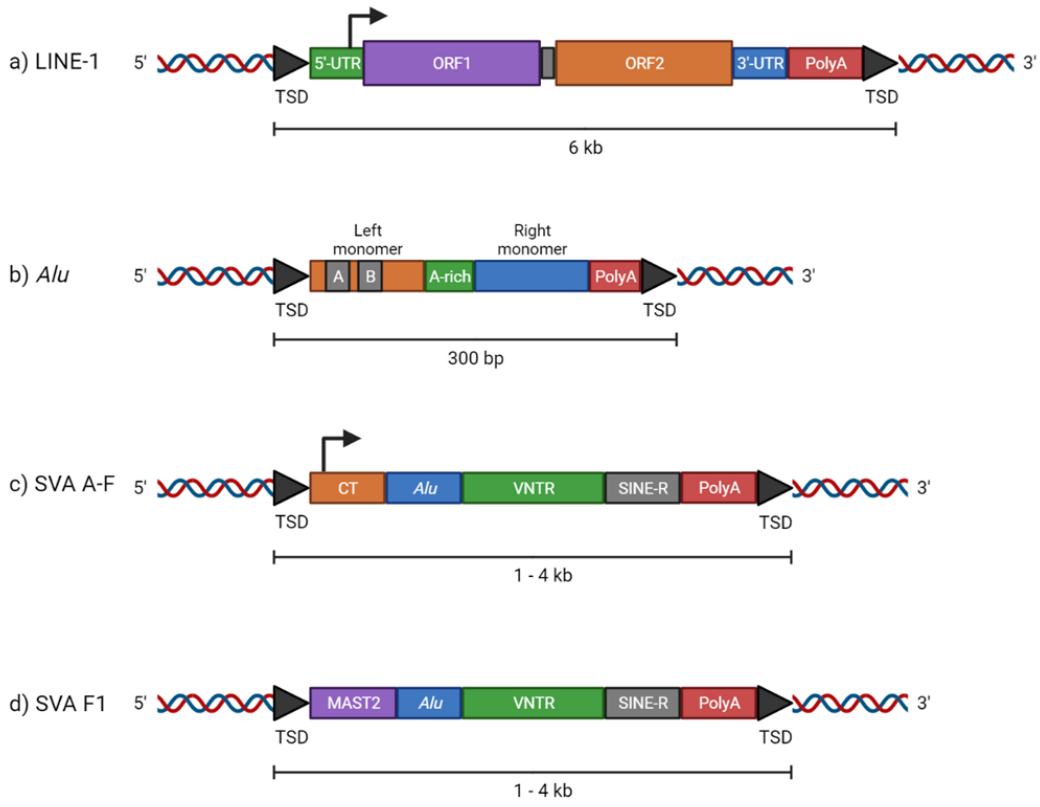


Figure 1.2 – Canonical structures of non-LTR retrotransposons. All TE components are described 5' to 3'. **a)** A full-length LINE-1 element is approximately 6 kb in length and possesses a 5'-UTR (green), encodes the ORF1 (purple) and ORF2 (orange) proteins, a 3'-UTR (blue) and a poly-A tail (red) which is thought to be important for retrotransposition by ORF1p and ORF2p. **b)** *Alu* SINE elements are around 300 bp long and consist of left (orange) and right (blue) monomers derived from 7SL RNA separated by an A-rich sequence (green), and a poly-A tail (red). *Alu* TEs possess their own transcriptional signals,

A-box and B-box elements (grey), within the left monomer. **c)** The SVA subclasses A – F feature a CT element which may be variable in copy number (orange), an *Alu*-like region (blue) composed of two antisense *Alu* sequences separated by a small intervening sequence, a VNTR region which may be one or two regions of 35–50 bp tandem repeats (green), a SINE region is derived from the 3' LTR of the retroviral HERV-K10 element (grey), and a canonical polyadenylated tail (red). **d)** The SVA F1 subclass features a 5' transduction of exon 1 of the *MAST2* gene in place of much or all of the CT element (purple). Intrinsic promoter regions are displayed as black right-angled arrows. Black arrowheads indicate target site duplication regions resulting from insertion.

ORF1p is a ~40 kDa RNA binding and chaperone protein while ORF2p is a ~150 kDa protein with endonuclease and reverse transcriptase (RT) activity [45, 46], with both being essential for *cis*-mobilisation of L1 transcripts [47]. Briefly, the lifecycle of an L1 element involves transcription from its own promoter within its 5'-UTR [48], transport to the cytoplasm for protein translation [49], formation of L1 ribonucleoproteins containing ORF1p, ORF2p and the L1 RNA [50, 51], and transport back to the nucleus followed by insertion into the genome via target-primed reverse transcription (TPRT) at a consensus sequence of 5'-YYYY/RR-3' (where Y = pyrimidine and R = purine) (**Figure 1.3a**) [46, 47, 52]. This insertion requires cleavage of the bottom DNA strand to release a 3'-OH group at the end of a T-rich sequence, which is thought to be base-paired by the L1 poly-A tail and used as a primer for ORF2p-mediated reverse transcription of the L1 complementary DNA (cDNA) strand (**Figure 1.3b**) [46, 50]. Cleavage of the top genomic DNA strand likely occurs after initiation of top strand L1 cDNA synthesis (**Figure 1.3b**) [53], with distance from the bottom strand nick determining the size of a target-site duplication (TSD) that is a hallmark for L1 retrotransposition and is typically 4 – 20 base pairs (bp). Subsequent steps in

TPRT require elucidation at molecular resolution, but it is presumed that the nascent bottom strand L1 cDNA is attached to the target site 3' overhang (**Figure 1.3c**) and then ORF2p-mediated top strand L1 cDNA synthesis proceeds using the 3' overhang as a primer and the bottom cDNA strand as a template (**Figure 1.3d**).

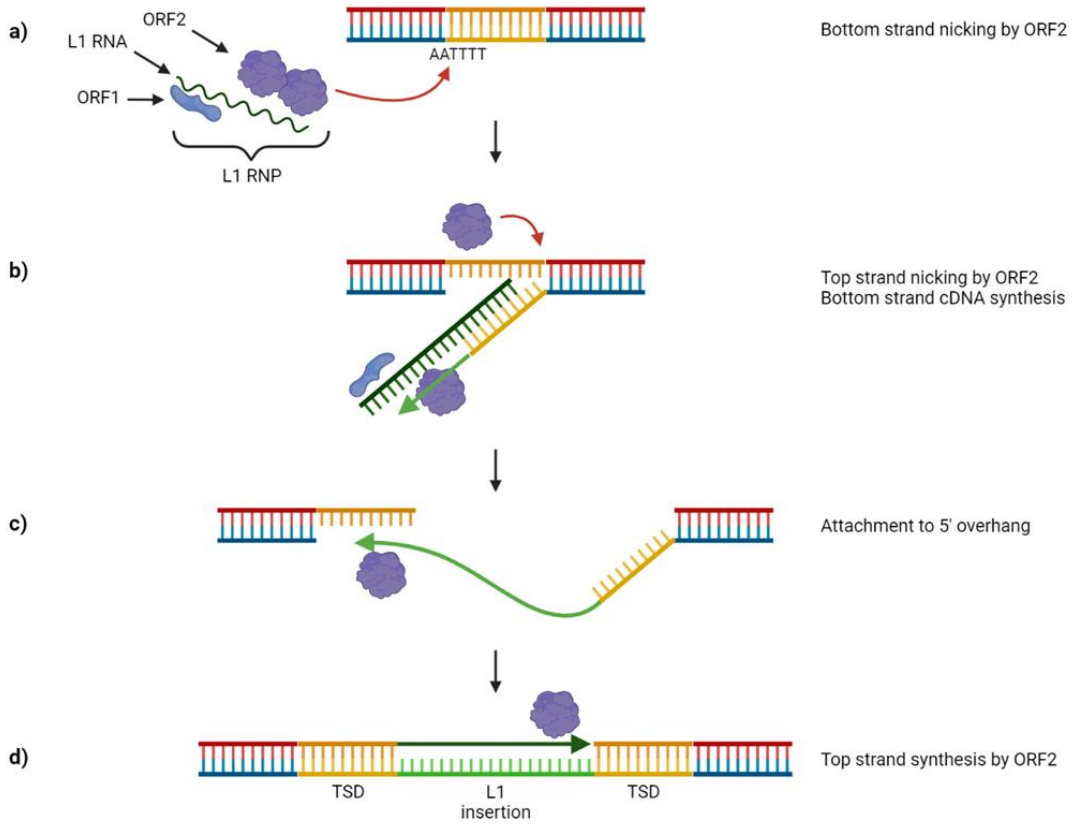


Figure 1.3 – L1 retrotransposition via target-primed reverse transcription. **a)** After the L1 is transcribed it associates with the L1 ORF1p chaperone and ORF2p endonuclease/reverse transcriptase in the cytoplasm and translocates to the nucleus. ORF2p endonuclease activity nicks the DNA bottom strand at the consensus sequence 5'-YYYY/RR-3', where Y = pyrimidine and R = purine. **b)** The L1 mRNA 3'-poly-A tail pairs with the liberated single-stranded T-rich DNA sequence and its 3'-OH is used by ORF2p to reverse transcribe the L1 bottom strand cDNA, templated by the L1 mRNA. After initiation of bottom strand synthesis the top genomic DNA strand is cleaved, presumably by ORF2p, with distance from the

bottom strand nick determining the size of cut site overhangs. **c)** The nascent bottom strand L1 cDNA engages with the upstream 3' overhang through an as-yet unknown mechanism. **d)** Through another currently unresolved mechanism, the top strand L1 cDNA is synthesised using the upstream target site 3'-OH as a primer and the L1 bottom strand cDNA as a template, presumably by ORF2p. Additionally, the bottom strand of the TSD is synthesised at the end of the L1 bottom strand.

This method of transposition presents multiple opportunities for structural polymorphisms to arise as L1s propagate throughout the genome. As mentioned previously the majority of genomic L1s (>99%) are 5'-truncated, suggesting that synthesis of the bottom strand cDNA is usually incomplete; it is speculated that the L1 RT competes with host proteins such as mRNA editing enzymes and DNA repair factors during extension (**Figure 1.4a**) [54, 55]. Additionally, the L1 poly-A constitutes a 'weak' transcription termination signal and frequently results in transcriptional read-through, causing a 3'-transduction when the sequence downstream of the L1 is copied to a new locus (**Figure 1.4b**) [56]. Similarly, it has been reported that 5' sequences may be transduced when transcription initiates upstream of the L1 (**Figure 1.4c**) [57], although this is considered rare for this element. L1-mediated insertion is also capable of causing target-site deletions: these range from several base pairs if top-strand nicking in TPRT produces a 5'-overhang that is processed by 5'-3' exonuclease activity (**Figure 1.4d**), to up to a megabase if a distant DNA nick or break is invaded by the nascent TE cDNA such that second-strand synthesis occurs from the distant 3'-OH group, excising the intervening sequence (**Figure 1.4e**) [58].

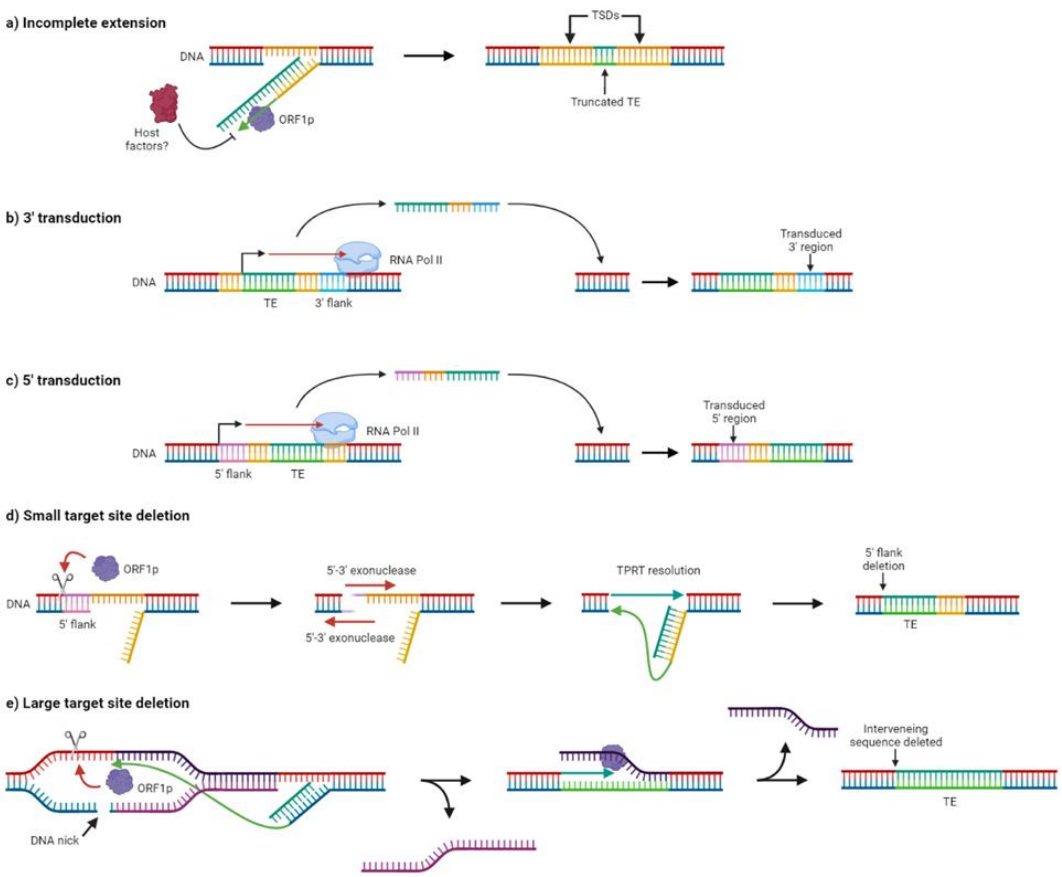


Figure 1.4 – Structural variation in L1-mediated retrotransposition. **a)** Since TPRT involves 5'-3' synthesis of the bottom strand first, events which perturb this extension – such as competition with host factors – can result in TPRT resolution of an incomplete TE cDNA that is 5' truncated relative to the top strand. **b)** RNA polymerase II can fail to terminate transcription at a TE poly-A tail, resulting in read-through and transduction of the 3' region to a new locus. **c)** Transcription initiation upstream of a TE can result in read-through and incorporation if the TE into the nascent transcript. This can be recognised by the L1 transcriptional machinery and inserted into the genome, thereby transducing the region 5' of the original TE locus. **d)** Should top strand cleavage by ORF2p endonuclease result in a 5'-overhang, it can be targeted by host 5'-3' exonuclease activity to produce a DNA blunt end upstream of the TE insertion. TPRT resolution repairs the DSB with deletion of the region targeted by the 5'-3' exonuclease. **e)** The 3' end of the nascent TE cDNA may invade a DNA nick or DSB that is distant from the insert site in the linear DNA sequence but proximal in 3D space. Ligation of the TE cDNA 3'-OH to

a free 5'-phosphate and cleavage of the DNA top strand with provide a template for cDNA top strand synthesis, excising the intervening sequence.

Notably, ORF1p and ORF2p preferentially bind their own RNA molecule for *cis* mobilisation but are also co-opted to mobilise Short Interspersed Elements (SINE), *Alu* and SVA transcripts in *trans* [59].

1.2.3. *Alu* retrotransposons

Alu TEs are members of the SINE family of genetic elements and are the most abundant non-LTR retrotransposons in the reference human genome, numbering around 1.1 million copies [60]. *Alu* elements are composed of two monomers – derived from an ancient duplication of a signal recognition particle RNA (commonly 7SL RNA) gene – which are separated by an A-rich sequence, with the *Alu* possessing its own A and B promotor 'boxes' and Pol III terminator signal [61, 62] (**Figure 1.2**). It is likely that *Alu* sequence similarity to 7SL RNA transcripts enhances their localisation at the ribosome by first binding to SRP9/14, which allows them to hijack the L1 protein machinery and has thereby enabled *Alu* TEs to become the most successful retrotransposons in the genome [63]. The *Alu* elements can be broadly divided into 3 subclasses based on evolutionary age, with oldest to youngest being *AluJ* at ~65 Myo, *AluS* at ~30 Myo and *AluY* at ~24 Myo [64, 65]. *AluJ* retrotransposons are considered to be largely extinct for mobilisation in humans due to the accumulation of inactivating mutations, whereas the younger *AluS* and *AluY* families exhibit low and high levels of mobilisation, respectively. While this is in part due to these TEs

possessing larger numbers of functionally intact elements, it has also been postulated that the *AluS* and *AluY* families have evolved lower affinity for SRP9/14 which enables them to disengage from the ribosome in order to co-opt the L1 ORF proteins for mobilisation [64].

1.2.4. SVA retrotransposons

As with *Alu* TEs, SVA transposons do not transpose autonomously but are mobilised *in trans* by the transposition machinery of LINE-1. In contrast to the highly numerous L1 and *Alu* families, there are only ~3000 full-length SVAs identified in the reference human genome [28]. Present only in hominids, SVAs are the evolutionarily youngest TEs in the genome. Considered 5′ to 3′, SVAs are composed of a $CCCTCT_n$ hexamer repeat (CT element), an antisense *Alu*-like sequence, one or two variable number tandem repeat (VNTR) regions, a SINE region derived from the retroviral HERV-K10 element, and a poly-A tail [66] (**Figure 1.2**). SVAs with all intact components vary from 1 to 4 kb in size and are classified A–F in order of evolutionary age based on their SINE region, with estimated ages ranging from 13.6 Myo for the SVA A subclass to 3.2 Myo for SVA Fs (**Figure 1.5**) [66]. Additionally, an F1 class of SVA contains a 5′-transduction of exon 1 of the *MAST2* gene [67], such that much or all of the CT element is replaced (**Figure 1.2**). SVA subfamilies E, F and F1 are human specific, and with the addition of the D subfamily, are active for retrotransposition within the human genome [66]. Although SVA D elements are also present in gorilla and chimpanzee genomes, their continued transposition has resulted in 67.5% of SVA D insertions in humans being species-specific [68]. SVAs may be polymorphic with regards to the length (repeat

copy number) of their CT element, VNTR and poly-A signal components [44], and may contain SNPs. Additionally, these elements are occasionally 3'-truncated owing to poly-A signals within the SINE region [69], and as many as 10% of genomic SVAs contain 5'-transduced sequence due to upstream transcription initiation followed by splicing to donor sites within the SVA [70]. Notably, the newest F1 SVA subclass was generated when the first exon of the *MAST2* gene was spliced to the 3'-end of the *Alu*-like domain of an SVA F [71]. It is perhaps unsurprising that SVAs have been speculated to be the most polymorphic structural variants in the genome [72].

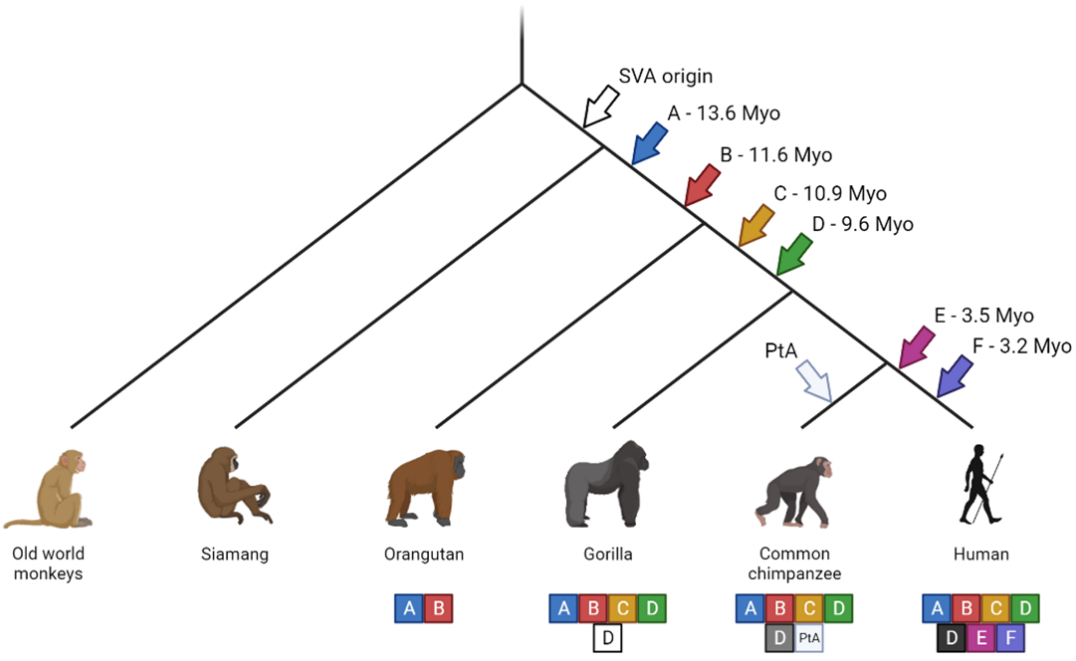


Figure 1.5 – Amplification dynamics of the SVA family of retrotransposons in primates. Estimated evolutionary age of SVA subclasses shown in million years old (Myo). The SVA D family remains contemporarily active, and so SVA D insertions that are unique to gorilla, chimpanzee and human

genomes are represented by white, grey and black 'D' boxes, respectively. Adapted from Wang *et al.* 2005 [66].

1.2.5. Non-LTR retrotransposon insertion polymorphisms and disease

It is well established that a burst of L1-mediated retrotransposition occurs in the early embryo when euchromatin is widespread [34, 35]. *In vitro* approaches have validated that human embryonic stem cells (ESCs) endogenously express L1 and support retrotransposition of engineered L1 expression constructs [73]. Additionally, generation of induced pluripotent stem cells (iPSCs) has been associated with activation of L1 during epigenetic reprogramming [74, 75], a process which creates a cellular environment broadly similar to that of ESCs [76]. Indeed, when 8 iPSC lines were reprogrammed from fibroblasts or endothelial cells a total of 7 L1, 2 *Alu* and 1 SVA *de novo* insertions were identified [77]. Novel insertions that occur in the primordial germline during embryogenesis can be passed on to progeny, resulting in new RIPs entering the human genome.

Amongst the general populace it has been estimated that one novel *Alu*, L1 or SVA germline insertion event occurs every 20, 100 – 200 and 900 live births, respectively [78, 79]. The global human population of ~7.9 billion (UN estimates as of April 2022) would therefore be expected to harbour approximately 462 million private RIPs, comprised of 4.0×10^8 *Alu*, 5.3×10^7 L1 and 8.8×10^6 SVA insertions. Early estimates have suggested that any two individuals would differ on average by 1283 *Alu,* 180 L1

and 56 SVA RIPs [72], highlighting the capacity for interpersonal genetic differences arising from retrotransposon activity.

An obvious potential consequence of retrotransposition is for insertions to disrupt gene function [44]. Indeed, the first identification of a *de novo* TE insertion was that of an L1 into exon 14 of the *Factor VIII* gene, which led to Haemophilia A [80]. Insertional mutagenesis results from TE insertion into exons, where the translated TE sequence directly causes dysfunction in the protein or where the inserted nucleotide sequence introduces a frameshift mutation that leads to nonsense-mediated decay (**Figure 1.6a**). Additionally, L1, *Alu* and SVA elements all contain multiple splice sites and have been reported to induce aberrant splicing when inserted intronically [70, 81, 82] (**Figure 1.6b**). This is exemplified by an SVA insertion into the 3'-UTR of the *fukutin* gene which can be alternatively spliced to exon 10 of the *fukutin* mRNA, truncating the protein's carboxyl-terminus (C-terminus) and causing mislocation from the Golgi to the endoplasmic reticulum, resulting in Fukuyama muscular dystrophy [83]. Intronic retrotransposon insertions may also cause premature termination of transcription through introduction of a polyadenylation signal [84, 85] (**Figure 1.6c**), resulting in a truncated protein.

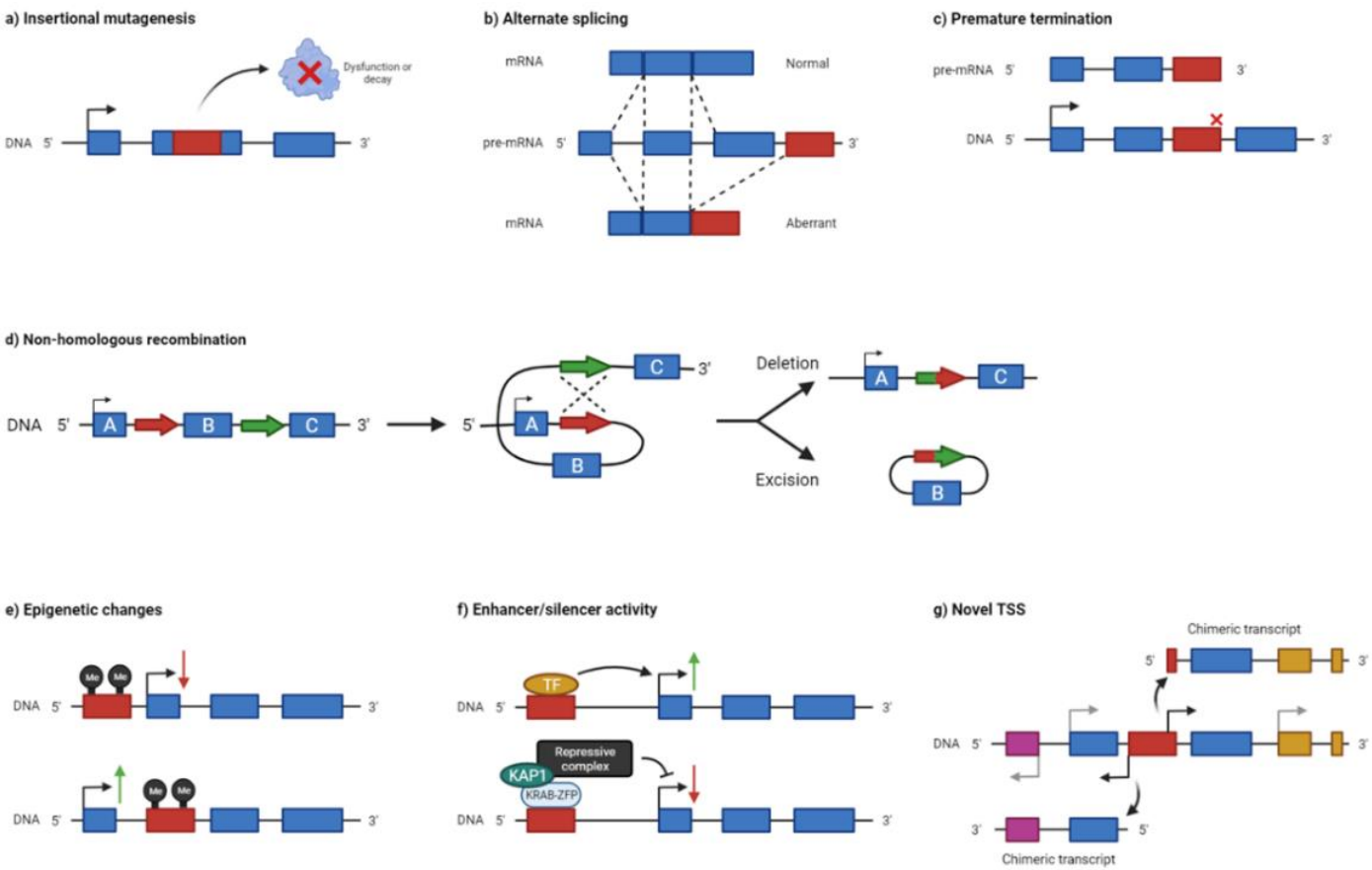


Figure 1.6 – Genomic impacts of TEs. Exons represented by blue blocks, TE insertion by red blocks. **a)** TE insertion into exons results in mutagenesis, creating a dysfunctional protein or inducing a frame-shift mutation that results in nonsense-mediated decay. **b)** Splice donor and acceptor sites in intronic TEs can override endogenous splice sites,

leading to aberrant splicing. **c)** Polyadenylation signals within intronic TEs can cause premature termination of gene transcription. **d)** TEs (red and green arrows) on the same chromosome may inappropriately strand invade in DSB repair, resulting in chromosomal rearrangements – a deletion is illustrated. **e)** Epigenetic modifications at TEs may induce changes in gene expression depending on the context of their insertion. **f)** TEs can act as enhancers or silencers depending on TF or repressor complex binding, respectively. **g)** Intrinsic promoter activity of some TEs can drive expression of novel chimeric transcripts (yellow and pink).

As noted previously, L1-associated retrotransposition can result in target-site deletions upon insertion (**Section 1.2.2**) with deletions ranging from a few base pairs to a megabase having have been associated with disease [86, 87]. Independent of TPRT and L1 ORF2p-mediated DNA breakage, the high copy numbers of L1 and *Alu* elements predisposes them to involvement in non-allelic homologous recombination which can generate duplications, inversions, deletions and chromosomal translocations [31, 88, 89] (**Figure 1.6d**). In further support for retrotransposon potential for inducing genome instability, analysis of a case of chromothripsis – a germline chromosome shattering event – identified an SVA insertion at a breakpoint associated with a 110 kb deletion flanked by four *Alu* elements [90]. Here, the authors propose that *Alu*-mediated chromosome looping brought distal regions of the chromosome together and poised them for recombination, with a concurrent SVA insertion event causing DNA cleavage followed by breakpoint resolution in which the 5'-end of the SVA transcript invaded the colocalised distal DNA, which led to deletion of the sequence in the crossover loop.

Importantly, non-LTR retrotransposon insertions are thought to alter the surrounding epigenome and are postulated to be capable of causing disease by inappropriately modulating gene expression. In somatic tissue L1s and SVAs are known to be densely methylated in their 5'-UTR and VNTR regions, respectively [91, 92]. Using the method of Gardiner-Garden and Frommer (1987), a CpG island is defined a sequence > 200 bp in length with a GC content >50% and a CpG_{obs}/CpG_{exp} (ratio of observed CpGs to expected number based on GC content) >0.6 [93]. The L1 5' UTR and SVA VNTR

regions have both been found to satisfy this definition [92, 94], so their transposition may therefore represent the introduction of a CpG island to a locus. Increased methylation within promoter or enhancer regions is associated with gene silencing and obstruction of transcription factor (TF) binding, while hypermethylation within the transcribed region can be associated with increased gene expression [95, 96] (**Figure 1.6e**). It has additionally been demonstrated that L1-mediated insertions are associated with local histone deacetylation through an as-yet unknown mechanism [97].

1.2.6. Somatic retrotransposition

L1-mediated retrotransposition may also occur in somatic cells and confer a 'somatic mosaicism' of cellular TE makeup in the adult, particularly if this occurs in embryonic development. Somatic insertions have been shown to be causative of at least one normally hereditary disease [98], and a number of insertions have been identified as drivers of cancer when they disrupt tumour suppressor genes [99-101]. In fact, it is likely that somatic L1-mediated insertions far outnumber those that affect the germline [102, 103]; in a study of a synthetic L1 expressed in transgenic mice >50 times more insertions were validated in the soma than in germ cells [104], and the authors estimate that the true total number of somatic insertions per animal to be several orders of magnitude higher. Furthermore, there is evidence that endogenous rates of L1 transposition are higher in neural progenitor cells and mature neurons than in other somatic tissues [105-107], with estimates ranging from 0.04 to 80 somatic L1 insertions per adult neuron made using a wide variety of approaches in

both bulk brain tissue and single neurons (methodologies reviewed by Faulkner *et al*. 2017) [108]. Retrotransposition-competent L1s have been shown to be hypomethylated in gDNA from adult motor cortex and cerebellum compared to matched blood [109], suggesting a level of tissue-specific de-repression that permits L1 activity. Additionally, single-cell genomic investigation of hippocampal neurons revealed that binding of the TF YY1 to the L1 5' region is important for L1 promoter methylation and repression, and that endogenous mutation or truncation of this region has allowed some young L1s to evade this repression [110] – providing further mechanism for L1 mosaicism in the brain. This raises questions regarding roles for L1-mediated transposition in neuronal plasticity and genetic diversity, and has implications for learning and cognition. However, to date this potential relationship remains almost entirely unexplored.

There is accumulating evidence that age-associated loss of repressive epigenetic marks is associated with de-repression of TEs in a wide range of cell types and species [111]. Considering the previously discussed role for retrotransposons in genomic instability this is potentially most damaging to the nervous system because, although small numbers of neural stem cells are retained in the adult brain [112], the vast majority of neurons are post-mitotic and cannot be replaced if deleterious insertions cause cell death. Furthermore, homology-directed DNA repair mechanisms are inactive in non-diving cells meaning that double-stranded breaks (DSBs) result in either NHEJ-directed repair, which may introduce mutations [113, 114], or cell death by triggering apoptosis [115]. In support of this, overexpression of L1 in HeLa cells

was associated with a mean occurrence of 68 DSBs per cell that was abolished when the L1 ORF2p endonuclease domain, essential for DNA nicking, was mutated [116]. The authors estimated that an L1 insertion event occurred once in every 10 – 100 ORF2p-mediated DSBs, and a subsequent study identified that this disparity between DNA breakage and L1 insertion results from retrotransposition-independent ORF2p DNA cleavage at genomic L1 loci, since the L1 sequence contains the ORF2p target sequence within its TSDs [117]. Also, in a mouse model of PD that featured loss of repressive chromatin marks RNA-seq detected 3 mouse L1 subfamilies that were upregulated, and it was shown that these were causative of DSBs that were similarly prevented by anti-L1 strategies such as RT inhibition and ORF2p knockdown [118]. In addition to genomic instability, expressed retrotransposon polynucleotides and proteins may induce cellular dysfunction directly. Detection of cytoplasmic retrotransposon double-stranded DNA by the cGAS-STING DNA sensing pathway leads to activation of interferon signalling, with L1 de-repression and cytoplasmic accumulation having been shown to contribute to chronic inflammation in mice, senescent cells and in human Aicardi-Goutières syndrome [119-121]. It is unclear how double stranded retrotransposon cDNA arrives in the cytoplasm since TPRT occurs in the nucleus, but it is theorised that these may be exported abortive products of reverse transcription or that cDNA synthesis can occur in the cytoplasm using as-yet unidentified primers [122]. There is also evidence that double-stranded character of L1 mRNA and inverted repeats in *Alu* transcripts can stimulate inflammation through detection by MDA5 or RIG-I proteins [123, 124]. In a similar vein, the neurotoxic HERV-K env protein has been observed to be elevated in the affected neural tissue of amyotrophic lateral sclerosis (ALS) patients [125]. Although endogenous RT activity

has been observed in the sera of ALS patients [126], it remains to be seen whether it is derived from HERV and/or L1 expression and what role, if any, this protein directly plays in disease pathology [127]. To summarise, while somatic retrotransposition seems to occur endogenously in the brain (and may even confer neuronal plasticity) it seems that de-repression of TEs as a result of ageing or disease may contribute to genome instability or inflammation, thereby providing a mechanistic link between disease onset and deterioration of cellular health.

1.2.7. Retrotransposon control and domestication

Unsurprisingly, this potentially harmful TE activity has driven the evolution of transcriptional and post-transcriptional mechanisms that repress TE activity. This places the retrotransposons themselves under considerable selective pressure to mutate and escape suppression, so host defence strategies need to be correspondingly adaptable.

One such TE silencing mechanism is the recognition of TEs by PIWI-interacting RNA (piRNA) in complex with PIWI-clade Argonaute proteins, a pathway which is restricted to the germline. piRNA clusters are genomic regions comprised of many TE remnants and nested fragments which are co-transcribed and then processed into individual RNA molecules. These then associate with Argonaute proteins and the RNA-protein complex is directed to nascent TE transcripts in the nucleus via complementary base pairing, leading to transcriptional repression at the genomic TE site via deposition of repressive histone modifications and DNA methylation (**Figure 1.7a**) [128]. It has

been proposed that piRNA clusters act as 'transposon traps' that passively acquire new TE sequences by chance insertion and then generate piRNAs targeting those TEs, with new piRNAs that confer a fitness benefit becoming fixed through positive selection [129]. Additionally, in a feedforward loop mechanism known as 'ping-pong amplification' TE transcripts are exonucleolytically cleaved by PIWI proteins guided by largely complementary antisense piRNAs derived from clusters (**Figure 1.7b**). Importantly, this allows a near-instant response to emerging TEs that are near-identical to sequences held within the 'genetic memory' of the piRNA cluster. The TE fragments are then taken up by PIWI proteins to target unprocessed piRNA transcripts, cleaving them into fragments usable by Argonaute and thereby amplifying the response to active retrotransposons (**Figure 1.7b**) [130]. There is evidence that human PIWI-like proteins are capable of the ping-pong cycle [131], and there are notable examples of PIWI-dependent control of TEs in humans – such as the regulation of active human-specific L1s in iPSCs [132]. However, TE control by piRNAs is not thought to occur at levels required to be a primary method of regulating TE activity in humans.

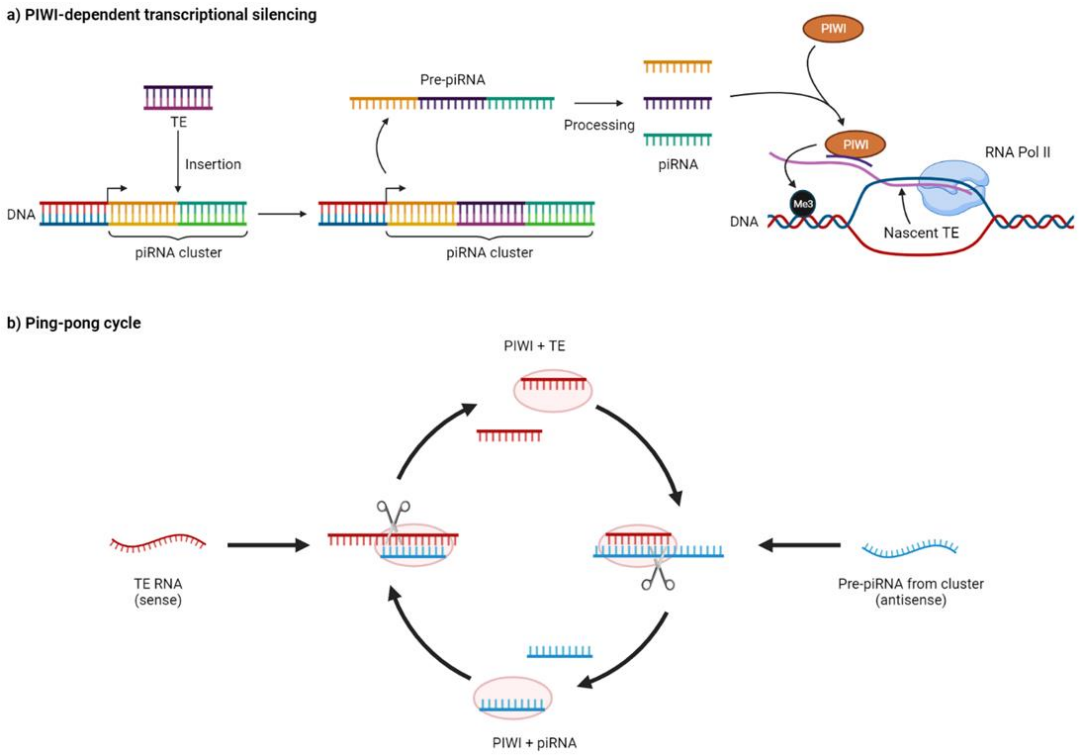


Figure 1.7 – Schematic of piRNA-mediated TE silencing. **a)** A genomic piRNA cluster acts as a 'transposon trap' when a TE randomly inserts into it. The new TE is incorporated into the co-transcribed pre-piRNA transcript and processed into a mature piRNA containing a fragment of the original TE. This piRNA associates with PIWI-clade Argonaute proteins and is guided to nascent transcripts of the same TE by complementary base pairing. Additional factors are recruited to deposit repressive epigenetic marks such as H3K9me3 (Black circle, Me3) and over time the TE is silenced genome-wide. **b)** In the ping-pong cycle piRNA-PIWI complexes are guided to TE RNAs that are largely complementary to the piRNA, leading to cleavage of the TE transcript. The resulting TE fragments are processed and taken up by PIWI proteins, and in turn guide them to pre-piRNA molecules containing the first piRNA by complementary base pairing. The pre-piRNA is cleaved and piRNAs released. In this way, TE degradation products are recycled to release more copies of the very piRNAs that target them, amplifying the PIWI-dependent response.

TEs of all classes may be recognised at the DNA level by Krüppel-associated box zinc-finger proteins (KRAB-ZFPs) which bind specific DNA sequences. Upon binding a genomic TE sequence KRAB-ZFPs then recruit KRAB associated protein 1 (KAP1; also known as TRIM28) which in turn acts as a scaffold for a silencing complex composed of the histone methyltransferase SETDB1, DNA methyltransferases, heterochromatin protein 1 (HP1), and the nucleosome remodelling and deacetylation (NuRD) complex [133]. Additionally, it has been demonstrated that the human silencing hub (HUSH) complex is recruited to the repressive histone mark H3K9me3 and cooperates with KAP1 in repression of TEs, particularly evolutionarily young L1s [134]. Altogether, these silencing pathways lead to transcriptional repression of TEs through formation of heterochromatin and deposition of DNA methylation (**Figure 1.6f, bottom**), and have been widely considered to irreversibly silence TEs and thereby protect genome integrity from rampant insertional mutagenesis [135, 136]. Importantly, while a given KRAB-ZFP might only be expressed during early development the repressive epigenetic state established at a TE is generally established permanently. KRAB-ZFPs are modular proteins that diversify the amino acid sequence of their DNA-binding zinc-finger domains through point mutations and larger scale substitutions of motifs with other KRAB-ZFP genes [137], allowing them to sample a wide range of DNA-binding sequences. Novel KRAB-ZFPs that target recently emerged retrotransposons will confer a fitness benefit by repressing deleterious widespread mobilisation and will become fixed by positive selection [138]. Highlighting their importance in TE control, it has been demonstrated that 159 out of 222 KRAB-ZFPs associate with at least one TE subfamily, and that many TE subfamilies are targeted by multiple KRAB-ZFPs [139]. While KRAB-ZFPs have been demonstrated to have roles besides TE

repression, it has been proposed that KRAB-ZFPs represent the primary method of retroelement control in higher vertebrates [133]. Specifically, they are best characterised for their role in induction of heterochromatin at TE loci during early embryogenesis to ensure a normal transcriptional environment for differentiation of embryonic stem cells [140]. Notable examples include zinc nuclease finger 91 (ZNF91) and ZNF93, which target SVA and L1 elements, respectively. ZNF91 acquired 7 new zinc fingers 8 – 12 million years ago that allowed it to bind to the VNTR region in the recently emerged SVA elements, and to this day continues to target all human SVA subclasses [141]. CRISPR KO of *ZNF91* leads to deposition of activating histone marks and transcription at SVAs [142], while overexpression of ZNF611, another SVA-targeting KRAB-ZFP, results in repression of SVAs [143]. Together, these studies demonstrate that functional redundancy appears to have been selected for in KRAB-ZFP-mediated control of SVAs. ZNF93 evolved earlier to repress L1 in primate genomes until ~12.5 million years ago when deletion of the ZNF93 binding site within the L1PA3 subfamily 5'-UTR enabled escape of repression and a burst of mobilisation [141]. In contrast to the piRNA-based TE silencing described previously, which is capable of rapid responses to emerging TE variants, KRAB-ZFP-based targeting is slower to respond to new TE threats as it requires gene duplication and tuning of DNA-binding sequences [144]. However, the element subfamily is then repressed globally. Furthermore, these pathways have been combined into a model of L1 regulation in which novel active integrants are initially repressed by piRNA-induced DNA methylation and then repressed by KRAB-ZFP until the TE loses its mobilisation potential through mutational drift [145].

Several lines of evidence suggest that retrotransposons and KRAB-ZFPs have co-evolved, including the observation that they have undergone parallel waves of genome expansion and that KRAB-ZFPs have undergone strong positive selection at DNA-binding residues [146, 147]. This has led to the proposition of an 'arms race' model in which repressed TEs mutate to escape KRAB-ZFPs recognition while the proteins in turn evolve novel DNA-binding capabilities, becoming fixed once they can repress the escaped TE and confer a fitness benefit [138]. However, this model appears to be an oversimplification; for instance, it should be nearly impossible for TEs to escape detection at observed rates when they are targeted by multiple distinct KRAB-ZFPs, and several TEs appear to have continued to mobilise – or even started mobilising – long after KRAB-ZFPs able to repress them had evolved [139]. Additionally, binding by certain KRAB-ZFPs is positively selected for by some TEs, such as ZNF382 and ZNF84 binding at human-specific L1s, and these particular KRAB-ZFPs do not recruit KAP1 [139]. It has been shown that some TE loci are active in adult tissues, not only being transcribed but providing alternate promoters [148], in contrast to the view that most KRAB-ZFP-targeted retroelements are irreversibly silenced through KAP1-mediated epigenetic modifications during early development. Indeed, one analysis found that only a fraction of genomic TEs were bound by KRAB-ZFPs (ranging from ~10% of L1s to ~50% of SVAs, for instance), although the authors did not determine whether the absence of binding was associated with age-associated mutational drift and TE inactivation [139]. Furthermore, in somatic cells gene expression near TEs can be tissue-specifically modulated by KRAB-ZFP regulation of target TEs or by TF binding to conserved sequences within TEs themselves (**Figure 1.6f, bottom**) [139, 149, 150]. For example, the primate-specific

LTR7 HERV-H element contains multiple LBP9, NANOG and OCT4 binding sites that drive expression of functional non-coding RNAs essential for the maintenance of human pluripotent stem cells in culture [151]. Importantly, the binding of KRAB-ZFPs is not necessarily mutually exclusive to the binding of TFs at TEs, as it has been shown that some KRAB-ZFPs are enriched in close proximity to TE-associated binding sites for proteins such as CCCTC-binding factor (CTCF), a master regulator of 3D chromatin structure [139]. Finally, KRAB-ZFPs demonstrate sophisticated patterns of expression, suggesting that their influence on TEs is a highly regulated process [133, 152]. Taken together these data have suggested a 'domestication' model in which KRAB-ZFPs, in addition to repressing deleterious TE mobilisation, allow the controlled release of some TEs and generally participate in their co-option by permitting their transcriptional *cis*-regulatory effects in a spatiotemporally restricted manner.

1.2.8. TEs in genome evolution

When the potential consequences of TE insertion are considered in light of their apparent domestication by host control factors, it is logical to conclude that TEs might be utilised by the host genome as a source of prefabricated genomic variation. Indeed, examples of retrotransposon-mediated genome evolution can be delineated – often through similar pathways to those that can result in disease (see **Section 1.2.5**). For example, although retrotransposon-mediated target-site deletions may cause disease several associated with L1, *Alu* or SVA insertions have been identified in the human genome versus primate genomes [153-155], suggesting that these deletions may have been selected for during genome evolution. Similarly, it was

recently demonstrated that a hominoid-specific *Alu* insertion into the *TBXT* gene causes a stem-loop structure to form in the pre-mRNA that promotes exon skipping, with the resulting protein isoform causing the absence of a tail in hominoids versus Old World monkeys [156]. It has been postulated that this phenotype was selected for by evolution as it contributed to upright bipedal locomotion, which arose around the same time [157].

TEs appear to contribute to coordinated genome regulation by dispersing sets of TF binding sites throughout the genome. Genome-wide chromatin immunoprecipitation (ChIP) maps indicate that diverse TFs bind within TE sequences [158-160], while ChIP-seq analysis of 26 TFs indicated that on average 20% of TF binding occurs within TE-derived sequences (ranging from 2% to 40%) [161]. Moreover, the timings of expansions of various TE subfamilies coincide with expansions in target bindings sites for certain TFs, resulting in the majority of TE-derived TF binding sites being species-specific [162]. Importantly, these novel binding sites can act as either enhancers or silencers (**Figure 1.6f**), leading to species-specific gene regulation patterns from binding of TFs. For example, in humans ~21% of binding sites for the pluripotency TF OCT4 are located within TEs, of which only 0.9% have a homologous OCT4-bound TE-derived region in the mouse genome [158]. The largest contributor of these TE OCT4 binding sites were LTR9B ERV sequences (8.3% of all binding sites), which demonstrated enhancer activity at examined loci. Similarly, binding sites for the tumour suppressor p53 have been distributed throughout the genome as part of HERV insertions, resulting in primate- and human-specific responses to DNA damage

[160], and propagation of MER20 DNA transposons introduced TF binding sites unique to placental mammals that contributed to the evolution of pregnancy [163]. It has also been observed that binding sites for CTCF have been propagated throughout the murine, canine and didelphine genomes by species-specific B2 SINE retrotransposon insertions, resulting in species-specific demarcation of chromatin boundaries [159]. Indeed, TEs contribute considerably to species-specific enhancers: for example, TEs overlap the majority of ape-specific and human-specific enhancers in the liver [164], while comparison of chimpanzee and human genomes reveals almost half of species-specific enhancers in cranial neural crest cells overlap TEs [165].

In contrast to enhancer regions, which can operate over very large genomic distances and can act independently of their orientation, promoter regions influence transcriptional start sites (TSS) in their immediate proximity via *cis*-regulatory signals that facilitate assembly of the Pol II pre-initiation complex. LTR regions in ERVs are well-documented as having provided ready-to-use promoters to loci in the human genome [166], such as the alternate transcription of *GSTO1* driven by a primate-specific MER4A ERV insertion [167]. Similarly, antisense promoter activity within L1 5'-UTRs is known to drive production of chimeric transcripts [168] (**Figure 1.6g**). These might then give rise to chimeric or truncated proteins, thereby facilitating protein evolution [169], and indeed over 1000 gene promoters in humans have been derived from co-opted TE sequences [170]. Furthermore, deletion of DNA methyltransferase 1 in human neuronal progenitor cells activated alternative

promoters of hominoid-specific L1s at neuron-specific genes, which was not due to upregulation of L1-targeting TFs upon global promoter demethylation but rather a result of L1 chromatin remodelling and accessibility for RNA polymerase II [171]. Interestingly, chimeric transcripts arising from L1 antisense promoter activity can exhibit altered patterns of tissue-specificity compared to the native transcript, as has been observed for *KIAA1797*, *CLCN5*, and *SLCO1A2* gene loci [172], thereby providing a mechanism for expanding gene expression programmes in evolution.

1.2.9. SVA retrotransposons in human-specific genomic variation

Taken together, the findings discussed thus far demonstrate that TEs have made important contributions to evolution of the human genome and its regulation over millions of years. A key question, therefore, is to what extent do the contemporarily active non-LTR retrotransposons (L1, *Alu* and SVA) continue to shape the human genome? Moreover, do RIPs arising from evolutionarily recent insertions result in meaningful, or even clinically-relevant, changes in gene expression?

In answering these questions SVA retrotransposons are of particular interest, as they represent prefabricated sources of hominid- or human-specific genome variation that might immediately contribute to interpersonal differences in gene regulation. Indeed, more than 60% of SVAs in the reference genome are located within gene bodies or less than 10 kb from a gene despite this amounting to only ~1% of the human genome [173], suggesting preferential insertion into actively transcribed regions [66]. SVAs possess a GC content of ~60% and this may exceed 70% within the

central VNTR region (**Figure 1.8**), which was noted to satisfy the >200 bp, >50% GC content and >0.6 CpG_{obs}/CpG_{exp} requirements for a CpG island described previously [92, 93]. Furthermore, it has been demonstrated that SVAs can be hypermethylated in somatic tissue [92], with evolutionarily older subclasses exhibiting greater methylation [174]. Methylation is capable of spreading laterally along DNA from CpG islands [175], although exact distances are difficult to predict as this is highly dependent on the presence of other heterochromatin-inducing factors such as histone modifying proteins and can be opposed by insulating TFs such as Sp1 [176]. Nevertheless, it is noteworthy that in mice B1 SINE retrotransposons have been demonstrated to cause transcriptional repression via spread of DNA methylation from ~1 kb away [177], and that in human colorectal cancer cell lines DNA methylation was observed to spread ~1.5 kb from a cluster of Alu elements into a promoter region [178]. Additionally, the 5' *MAST2* exon 1 transduction associated with SVA F1 retrotransposons is defined as a CpG island and has been shown to act as a promoter in human germline cells (**Figure 1.8b**) [179], which may have consequences for nearby gene expression. SVAs may also influence 3D chromatin structure – it has been demonstrated that CTCF can bind the SVA VNTR *in vitro*, while the germline-expressed paralog CTCF-like (CTCFL, also known as BORIS) can do so *in vivo* (**Figure 1.8**) [180]. Notably, this study also demonstrated that CTCFL may bind immediately upstream of the SVA F1 subfamily (**Figure 1.8b**). In reporter gene studies, gene expression was differently modulated by individual SVA components and the full-length SVA both *in vitro* and *in vivo* [173, 181]. Moreover, it was shown that stimulation of JAr cells with cocaine resulted in disparate changes in TF binding, histone acetylation and RNA Pol II binding at a GC-rich VNTR in the *SLC6A4* promoter

that were dependent on the element's repeat length [182], demonstrating the potential for VNTRs to act as length-dependent stimulus response elements. In X-linked Dystonia-Parkinsonism (XDP) the length of the variable CT hexamer repeat of a disease-specific SVA insertion within the *TAF1* gene is associated with repression of *TAF1* expression and is strongly inversely correlated with age of onset of disease [183-185], highlighting that sequence polymorphisms may be relevant in fine-tuning SVA effects at a locus. It has also been demonstrated that this intronic SVA insertion is associated with altered levels of acetylated histone H3 in the nearest TAF1 exon, although how this relates to XDP disease pathology is unclear and the authors do not compare this acetylation to SVA CT hexamer length [186]. Conversely, it has been observed that an intronic antisense SVA insertion within *CASP8*, a caspase involved in the extrinsic apoptotic pathway, is associated with protection against prostate cancer through a mechanism that is not yet fully defined but probably involves SVA-mediated intron retention [187]. Interestingly, it has recently been observed that in humans SVA insertions are overrepresented at zinc finger gene clusters on chromosomes 4, 7 and 19 [188]. Considering the potential for SVAs to influence the local genome, this raises the possibility of a type of feedback loop in genomic evolution involving hominid- or human-specific modulation of KRAB-ZFP expression by their own targets.

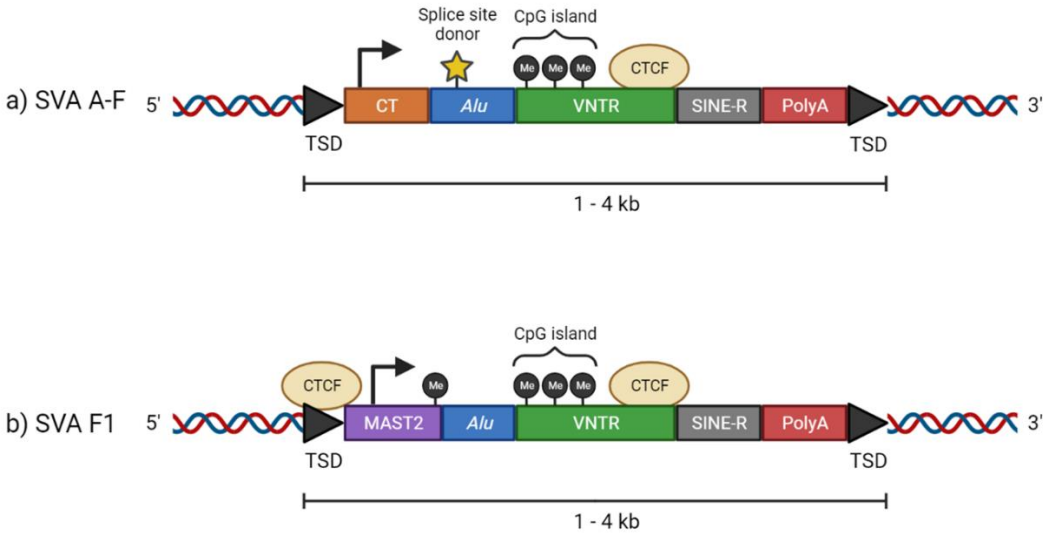


Figure 1.8 – Regulatory features of SVA retrotransposons. CpG islands with 5-methylcytosine residues are pictured. The noteworthy cryptic splice site within the SVA *Alu* region is highlighted at a yellow star. Potential CTCF/CTCFL binding sites are shown.

1.2.10. Non-LTR retrotransposons, whole genome sequencing, and complex disease

In summary, it is clear that TEs represent DNA elements that have the potential to exert strong influences on gene expression and genome integrity. The contemporary mobilisation of non-LTR retrotransposons gives rise to RIPs amongst the general populace on top of element length polymorphisms, which may contribute to interpersonal differences in gene expression. This is particularly true for SVA retrotransposons, which may be co-opted by the host genome as a source of human-specific genome variation. However, most work on non-LTR retrotransposons thus far has focussed on their properties *in vitro* or in notable disease cases; to date there is a scarcity of data on their functional influences in normal gene regulation *in situ* in humans.

This knowledge gap is especially prominent in high-throughput sequencing-based approaches such as those employed in GWAS, as mapping of reads from TEs has historically proven challenging [189]. The currently widely utilised short-read sequencing technologies have a read length upper limit of 300 bp, meaning that they will fail to sequence full-length TEs such as L1s and SVAs. Any reads that fall within the body of such elements cannot be readily mapped back to specific genomic coordinates, as the copy-and-paste nature of TE mobilisation means a given sequence likely occurs many times throughout the genome. Although paired-end sequencing technologies might improve mappability because the sequence of a TE may be 'paired' to non-repetitive sequence a known distance away, nesting of TEs within other TEs often precludes precise mapping [189]. Even when mapping can be accomplished it frequently only yields coarse-grain information on an insertion, such as its genomic coordinates and its transposon family. This would therefore miss any sequence or structural variants present within the TE, falling well short of the objective of WGS. Furthermore, as sequencing datasets have grown larger and more numerous manual annotation of repetitive sequences has quickly become impractical. Automated TE annotation approaches have therefore been developed that take a variety of approaches. 'General repeat finding' programs identify high copy number sequences in WGS data with high sensitivity but possesses poor capabilities in classifying TE superfamilies [190]. A 'sequence homology' approach, such as that used by the RepeatMasker program, that makes uses of pre-existing TE sequence databases affords relatively fast annotation but is limited by any gaps in the prior knowledge of TE variation [191, 192]. Furthermore, annotation tools can

exhibit differing levels of precision for TEs of varying size and complexity in a given genome [193].

As discussed previously, for diseases with complex genetic components such as sporadic PD great effort is being made to sequence the genomes of people with PD in greater numbers in order to pinpoint causal genetic variants at base pair resolution. The contemporary mobilisation of non-LTR retrotransposons is a potentially important source of genomic variation that is likely to be improperly annotated or even discarded in *de novo* short-read sequencing approaches. While it is impossible to validate and characterise every single RIP in the human genome at the lab bench, it is clear that entirely WGS-based approaches may not fully elucidate the relationship between retrotransposons and molecular phenotypes. However, a strategy that may bear fruit is to use GWAS to narrow the search for phenotype-associated genetic variance to a given locus and then the region may be manually inspected to nominate retrotransposons that are worthy of further investigation. More broadly, candidates may also be identified by simply overlaying *de novo* whole genome phenotypic datasets with transposon coordinates datasets such as the freely available RepeatMasker track on the UCSC genome browser (https://genome.ucsc.edu/).

As such, this thesis will attempt to combine targeted bench-side characterisation with broader bioinformatic approaches to investigate TEs, with particular focus on human-specific SVA RIPs and the context of PD. In this way, the *in situ* regulatory influences

of human TEs may be better elucidated, thereby improving understanding of both the normal intracellular environment and the complex genetic basis of PD.

1.3. General Aims

In this thesis retrotransposons will be studied both for their importance in PD and in normal gene regulation, with a particular focus on SVA elements. Firstly, a non-disease related SVA RIP will be studied as a model system of RIP influence in gene promoter regions. Secondly, PD-relevant SVA retrotransposons of interest identified in the reference genome and from automated annotation of PD genomes will be characterised. Finally, retrotransposon coordinates will be overlaid with chromatin structure datasets from PD-derived cell lines to gain an overview of TE involvement in any disease-specific phenomena, and potentially identify candidate retrotransposons for further study.

Chapter 2 Materials and Methods

2.1. Materials

2.1.1. Commonly used materials

Material	Components
TBE Buffer (5X) Diluted to 0.5X for use in running buffers and agarose-based gels.	108 g Tris base (Fisher, BP152) 55 g Boric acid (Sigma Aldrich, B0394) 5.84 g EDTA (Sigma Aldrich, E5134) Made up to 2 L with distilled water.
LB Broth, Miller (Sigma Aldrich, L3152)	25 g/L in distilled water, autoclaved.
LB Agar, Miller (Sigma Aldrich, L3027)	40 g/L in distilled water, autoclaved.

Table 2.1 – Constituents of commonly used lab materials.

2.1.2. DNA oligonucleotides

Oligonucleotides were obtained from Sigma Aldrich with desalt purification. Specific details of DNA oligonucleotide sequences used for PCR or pyrosequencing can be found in **Table 2.5**, while sequences for guide DNAs used in a CRISPR-Cas9 system can be found in table **Table 2.9**.

2.1.3. NABEC human frontal cortex DNA samples

Several retrotransposon targets of interest were genotyped in frontal cortex genomic DNA (gDNA) samples from the North American Brain Expression Consortium (NABEC) cohort, which were gifted by collaborators at the Laboratory of Neurogenetics, National Institutes of Health (NIH), USA. NABEC is a cohort of neurologically normal frontal cortex samples from the Database of Genotypes and Phenotypes (dbGaP) and is supported by publicly available datasets including Illumina genome-wide genotyping array data, whole genome sequencing (WGS), total RNA-Seq and Illumina

450K methylation data. Study details can be found at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001300.v2.p1&phv=495557&phd=&pha=&pht=6722&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1. Early NABEC research outputs are available as publications by Gibbs *et al*. 2010 [194], Hernandez *et al*. 2012 [195], and Kumar *et al*. 2013 [196].

2.1.4. The AMP-PD harmonised cohort dataset

To examine elements of interest in a PD-relevant dataset, access was gained to the Accelerating Medicines Partnership – Parkinson's Disease (AMP-PD, https://amp-pd.org/). AMP-PD is a large dataset made up of 8 previously separate study cohorts, including the Parkinson's Progression Markers Initiative (PPMI, http://www.ppmi-info.org/) and Parkinson's Disease Biomarkers Program (PDBP, https://pdbp.ninds.nih.gov/), which offers a wealth of clinical and genomic data that has been harmonised (made comparable). 'Tier 2' access enabled the download of genotyping and transcriptomic for selected loci and genes, which was then analysed locally using PLINK and the R software environment (**Section 2.2.1**).

2.1.5. Established human cell lines

Component	Supplier and catalogue number
Dimethyl sulphoxide (DMSO) (neat)	Fisher, D/4120/PB08
Foetal bovine serum (FBS), heat inactivated	Gibco, 10500-064
L-glutamine (200 mM)	Gibco, 25030-149
Phosphate-buffered saline (PBS) pH 7.2 (1X)	Gibco, 20012-019
Penicillin-Streptomycin (100X)	Sigma, P0781
Sodium pyruvate (100 nM)	Sigma, S8636
Trypsin-EDTA (0.25%)	Sigma, T4049

Table 2.2 – Commonly used tissue culture reagents.

All established cell lines were of human origin.

HEK293 (ATCC: CRL-1573): Immortalised embryonic kidney cell line of foetal origin.

Growth media: Dulbecco's Minimum Essential Media (DMEM; Sigma, D6429) supplied with 4500 mg/L D-glucose, 584 mg/L L-glutamine, and 110 mg/L sodium pyruvate, supplemented with 10 % (v/v) FBS and 1% (v/v) penicillin-streptomycin.

HeLa (ATCC: CCL-2): Isolated from cervical cancer of a 31-year-old.

Growth media: Growth media: DMEM (Sigma, D6429) supplied with 4500 mg/L D-glucose, 584 mg/L L-glutamine, and 110 mg/L sodium pyruvate, supplemented with 10 % (v/v) FBS, 1% (v/v) penicillin-streptomycin and 1% (v/v) MEM non-essential amino acids (Sigma, M7145).

JAR (ATCC: HTB-144): Trophoblastic tumour of the placenta of a male foetus.

Growth media: RPMI-1640 medium (Sigma, R0883) supplemented with 10% (v/v) FBS, 4500 mg/L D-glucose (Sigma, G5767), 1% (v/v) penicillin-streptomycin, 1% (v/v) L-glutamine, and 1% (v/v) sodium pyruvate.

MCF-7 (ATCC: HTB-22): Metastatic adenocarcinoma cells from breast tissue of a 69-year-old.

Growth media: DMEM (Sigma, D6429) supplied with 4500 mg/L D-glucose, 584 mg/L L-glutamine, and 110 mg/L sodium pyruvate, supplemented with 10 % (v/v) FBS and 1% (v/v) penicillin-streptomycin.

SH-SY5Y (ATCC: CRL-2266): Derived from cell line SK-N-SH, originally extracted from bone marrow metastasis of 4-year-old female with neuroblastoma.

Growth media: 1-to-1 mix of Minimal Essential Medium Eagle (Sigma, M2279) with Nutrient Mixture F-12 Ham (Sigma, N4888), supplemented with 10 % (v/v) FBS, 1% (v/v) penicillin-streptomycin, 1 % (v/v) L-glutamine, and 1 % (v/v) sodium pyruvate.

SK-N-AS (ATCC: CRL-2137): Bone marrow metastasis of poorly differentiated embryonal neuroblastoma from a 6-year-old female.

Growth media: DMEM (Sigma, D6429) supplied with 4500 mg/L D-glucose, 584 mg/L L-glutamine, and 110 mg/L sodium pyruvate, supplemented with 10 % (v/v) FBS and 1% (v/v) penicillin-streptomycin.

2.1.6. Plasmid vectors

Several commercially available plasmid vectors were used as part of the work presented here. pCR-Blunt from Invitrogen (46-0757, from kit K2700) was used for blunt-end ligation of PCR products (**Figure 2.1**). The pSpCas9(BB)-2A-GFP plasmid (also known as pX458) was used for CRISPR-Cas9-mediated deletions of SVA targets (**Figure 2.2**), and was gifted by Patrick Harrison, University College Cork, Ireland.

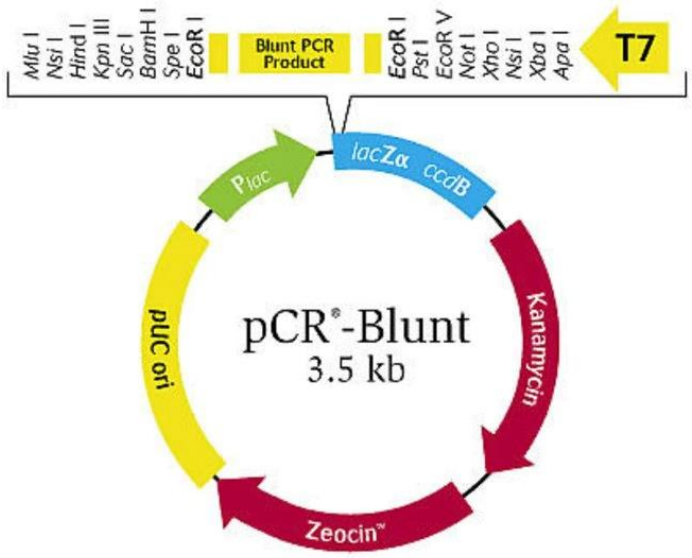


Figure 2.1 – Plasmid map of the pCR-Blunt vector from Invitrogen (image from manufacturer's website). Blunt-end ligation into this plasmid enables amplicons of interest to be multiplied with relative ease for downstream applications, while a multiple cloning site (MCS) flanking the ligation site affords several options for subsequent cloning into target expression vectors. Ligation of the plasmids own blunt ends together results in expression of the *E.coli*-lethal *ccdB* gene, permitting growth of only positive recombinants upon transformation.

Figure 2.2 – Plasmid map of the pSpCas9(BB)-2A-GFP vector (image from SnapGene, deposited by Zhang Lab – Addgene accession #48138). The Cas9 cassette expressing the Cas9 protein was derived from *S. pyogenes* (SpCas9). This enzyme is ideal for genome editing in human cell lines as it requires the protospacer adjacent motif (PAM) 5'-NGG-3', which is abundant in mammalian genomes. After insertion of a targeting oligonucleotide into the gRNA scaffold, the U6 promotor regulates synthesis of single guide RNA (sgRNA). The CBh promoter (chimeric CMV enhancer and chicken β-actin promoter) drives Cas9 expression followed at its C-terminus by a T2A self-cleaving peptide, releasing a co-translated green fluorescent protein (GFP) reporter.

2.2. Methods

2.2.1. Bioinformatic approaches

2.2.1.1. Software environments

All bioinformatic analyses were performed in the MobaXterm server and SSH client (https://mobaxterm.mobatek.net/), R software environment (https://www.rstudio.com/) or Windows Command Prompt.

2.2.1.2. Generation of proxy (tagging) SNPs

Genotypes for SVA presence/absence or specific length polymorphisms were determined via PCR (**Section 2.2.4**) in a subset of the NABEC cohort for which DNA was available in the lab. SVA genotypes were manually encoded in variant call file (vcf) format for each sample identifier, and then this subset was merged with SNP binary files for the entire cohort using the PLINK v1.90 whole genome association analysis toolset (https://www.cog-genomics.org/plink/) [197]. Proxy SNPs for each SVA genotype were then generated using the '--show-tags' function within PLINK while filtering for standard SNP and sample missingness. Proxy SNPs with LD values r^2>0.8 and D'>0.8 for each genotype were then determined using the PLINK --ld function. Since this yielded multiple proxy SNPs, those with the highest LD values were taken forward.

2.2.1.3. Linear regression of expression and methylation data versus SVA allele dosage

The chosen SVA-tagging SNP genotypes were imported into R Studio Version 1.2.1335 (Boston, MA, US) and merged with anonymised NABEC patient and sample information. These were: participant age (at time of death), gender, and ethnicity, along with sample Group (the institute that collected the sample) and RNA integrity number (RIN). These covariates were included in linear regression analyses to assess the relationship between SVA allele dosage and either expression or methylation data. Samples from individuals under 15 years of age (at time of death) were excluded to minimise developmental effects in the results. A linear regression model was generated and interpreted using the 'lm' and 'summary' functions, where test statistics follow a Student's t distribution, as follows:

Variable ~ SVA_genotype + Age + Gender + Ethnicity + Group + RIN_totalrna

Where 'Variable' is a set of expression or methylation values of interest.

In the linear models examining the LRIG2 SVA in RNA-seq and CpG methylation data for NABEC individuals, a total of 17 tests were performed: 2 RNA transcripts (*LRIG2* and *LRIG2-DT*) and 15 CpG probes (cg13503476, cg23932873, cg22598841, cg16709384, cg26091510, cg04139429, cg23175215, cg17310611, cg24448849, cg15031996, cg10983720, cg14912723, cg21504385, cg09332974, and cg23961141) were examined. The standard 0.05 alpha significance level for these analyses was therefore adjusted using Bonferroni correction for multiple tests: 0.05/17 = 2.94E-3.

2.2.1.4. Involvement of TEs in gene-associated chromatin loops

Hi-C data from 8 induced pluripotent stem cells (iPSCs) lines from The Foundational Data Initiative for Parkinson's Disease (FOUNDIN-PD, https://www.foundinpd.org/) were provided by collaborators at NIH, Maryland, USA. These iPSC lines were obtained from the Parkinson's Progression Markers Initiative (PPMI, http://www.ppmi-info.org/), an observational clinical study to identify progression markers in Parkinson's disease. The PPMI study was approved by the institutional review board at each sample collection site, and participants provided written informed consent (https://www.ppmi-info.org/study-design). Hi-C data for the 8 iPSC lines were provided for undifferentiated states (day 0) and at 65 days of a dopaminergic differentiation protocol, which had been pre-processed and quality controlled. There was no internal positive control for formation of chromatin loops detected by Hi-C, but successful differentiation of iPSCs was confirmed by immunocytochemistry of dopaminergic neuronal markers Tyrosine Hydroxylase and Microtubule-associated Protein 2. The iPSC lines were derived from 4 males and 4 females that were all of European ethnicity. 3 of these individuals were PD patients while the remaining 5 were neurologically healthy controls.

This Hi-C data was overlaid with retrotransposon coordinates – those of TEs in the reference genome were acquired from the RepeatMasker track on the University of California, Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu/). The RepeatMasker track in the latest human genome build, 38 (hg38), contains a large number of small fragments and simple repeats that are annotated as non-LTR

retrotransposons but are not informative in the study of full-length elements. TE annotations from the previous genome build, hg19, were therefore converted to hg38 coordinates using the UCSC browser's 'Liftover' function, as hg19 RepeatMasker contains many of the same full-length annotations as hg38 but far fewer small fragments. Many additional TEs have been identified that are not included in the reference genome - these 'non-reference' retrotransposon coordinates were obtained from The Genome Aggregation Database structural variant (gnomAD -SV) callset by filtering for 'insertions' (https://gnomad.broadinstitute.org/).

Reference and non-reference retrotransposon coordinates were separately overlaid with Hi-C coordinates using the 'intersect' function within the Bedtools computational toolset (https://Bedtools.readthedocs.io/en/latest/). Briefly, this function compares two lists of chromosome coordinates and reports and overlapping features. A given list of chromosome coordinates for a chromatin anchor was denoted 'A' and the list of TE coordinates was 'B'. Bedtools 'intersect' was then used with the flags: '-wa', which writes the coordinates of A for every overlap found with B; '-wb', which reports the coordinates of B found to overlap with A (without this Bedtools simply reports items from A that were a hit); and '-loj' ('left outer join') which individually reports every item from B that overlapped with a given item in A, and also reports NULL for B if no overlap found with A. Altogether, these flags produce a list of Hi-C coordinates in rows with features that they overlapped with – in this case TE coordinates. A simplified example of this code is shown below:

```
Bedtools intersect -wa -wb -loj \
-a A \
-b B \
> /destination_folder/overlap_file_AB
```

The resulting list of chromosome loop anchors supplemented with TEs were then intersected with coordinates of transcribed regions using the same methodology. The NCBI RefSeq Genes 'curated' subset (https://www.ncbi.nlm.nih.gov/refseq/about/, downloaded from USCS genome browser) was chosen as it did not contain unvalidated predicted transcripts.

2.2.2. PCR primer design

For standard PCR, primers were designed based on genomic DNA sequences obtained from UCSC genome browser hg38. Candidate primers with desirable thermodynamic properties were identified with Primer3 (http://primer3.ut.ee/) or NCBI Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/), the latter of which combines Primer3's thermodynamic analyses with a Basic Local Alignment Search Tool (BLAST) to simultaneously assess primer binding specificity. Desirable primers had a length of 18–20 bp in length, GC content of 40–60%, predicted melting temperature (T_m) of 55–65 °C, and minimal formation of stem-loop structures, heterodimers, or homodimers – defined as a ΔG value between –0 kcal/mole and –6 kcal/mole for these structures. These properties were double-checked against the OligoAnalyzer tool (https://eu.idtdna.com/pages/tools/oligoanalyzer).

2.2.3. Nucleic acid purification

2.2.3.1. DNA extraction

Prior to gDNA extraction, up to 5×10^6 cells from established lines were harvested by trypsinisation and pelleted by centrifugation at 300 x g for 5 min. All gDNA purifications were performed using the GenElute Mammalian Genomic DNA miniprep kit (Sigma, G1N350) using manufacturer's guidelines. DNA was eluted in Tris-EDTA (commonly TE) buffer (provided in kit) and for long term storage was stored at -20°C.

2.2.3.2. RNA extraction

As above, established cell lines were harvested by trypsinisation and centrifugation. RNA was purified using the Monarch Total RNA Miniprep Kit (NEB, T2010S) according to manufacturer's instructions. Extraction was performed on the same day as harvesting wherever possible to minimise degradation of RNA, otherwise cell pellets were stored at -80°C for no more than a week. On-column DNase I treatment (M0303S, provided in kit) was always performed in RNA preparations to remove any gDNA contaminants. RNA was eluted in 20 – 35 µl nuclease-free water (NFW) and stored at -20°C for up to 1 week, or -80°C for long-term storage.

2.2.3.3. Quantification of nucleic acid extract concentration and purity

The Nanodrop 8000 spectrophotometer (ThermoFisher Scientific, ND-8000-GL) was used for all nucleic acid quantification and quality control. Nucleic acid concentrations are based on absorbance at a 260 nm wavelength, whereas

absorbance at 280 nm or 230 nm is associated with unwanted protein or organic compounds, respectively. In line with generally accepted standards, DNA with a 260/280 ratio >1.8 and a 260/230 ratio >1.5 was considered pure while RNA with a 260/280 ratio >2.0 and 260/230 ratio >1.8 was considered pure.

2.2.4. Standard PCR reaction

Standard 'endpoint' PCR made use of GoTaq G2 Hot Start Taq Polymerase (Promega, M7408) for short amplicons, KOD Hot Start DNA Polymerase (Sigma Aldrich, 71086) for repetitive sequences or when proofreading activity was required, or KOD Xtreme™ Hot Start DNA Polymerase (Sigma Aldrich, 71975) for sequences that were otherwise difficult to amplify. Each polymerase was typically used with standard manufacturer-recommended reaction mixtures (**Table 2.3**) and cycling conditions (**Table 2.4**).

Polymerase	Reagent	Volume (µL)	Final Concentration
GoTaq G2 Hot Start Taq	5X Green Flexi Buffer	4	1X
	$MgCl_2$	3.2	4 mM
	dNTPs (10 mM)	0.4	0.2 mM
	Fwd primer (10 µM)	0.8	0.4 µM
	Rev primer (10 µM)	0.8	0.4 µM
	Polymerase (5U/µL)	0.1	1.25 U
	DNA (5-10 ng/µL)	1-2	0.25-0.5 ng/µL
	Nuclease free water	8.7-9.7	-
KOD Hot Start DNA Polymerase	10X KOD Buffer	2	1X
	dNTPs (2 mM)	2	0.2 mM
	$MgSO_4$ (25 mM)	1.2	1.5 mM
	Betaine (5 M)	4	1 M
	Fwd primer (10 µM)	0.6	0.3 µM
	Rev primer (10 µM)	0.6	0.3 µM
	Polymerase (1U/µl)	0.4	0.02 U
	DNA (5-10 ng/µL)	1-2	0.25-0.5 ng/µL
	Nuclease free water	7.2-8.2	-
KOD Xtreme™ Hot Start DNA Polymerase	2X Xtreme Buffer	10	1X
	dNTPs (2mM)	4	0.4 mM
	Fwd primer (10µM)	0.6	0.3 µM
	Rev primer (10µM)	0.6	0.3 µM
	Polymerase	0.2	0.01 U/µl
	DNA (5-10 ng/µL)	1-2	0.25-0.5 ng/µL
	Nuclease free water	2.6-3.6	-

Table 2.3 – Typical reaction mixtures used in PCR. DNA template input is specified where relevant.

	GoTaq G2 Hot Start		KOD Hot Start		KOD Xtreme™ Hot Start	
Polymerase activation	95 °C	2 min	95 °C	2 min	94 °C	2 min
Denaturation	95 °C	30 sec	95 °C	20 sec	98 °C	10 sec
Primer Annealing	X °C	30 sec	X °C	10 sec	X °C	30 sec
Extension	72 °C	1 min/kb	70 °C	<0.5 kb: 10 sec/kb 0.5 – 1 kb: 15 sec/kb 1 – 3 kb: 20 sec/kb >3 kb: 25 sec/kb	68 °C	30 sec – 1 min/kb
Final Extension	72 °C	2 min	NA	NA	NA	NA

Table 2.4 – Typical cycling conditions used in PCR where X is a previously experimentally determined primer annealing temperature.

Name	Sequences	Anneal Temp. (°C)	Product size (bp)	Application
LRIG2 SVA + Flanks	F: 5'-AGGAAGAGATGGAAGGAGACAA-3' R: 5'-GCCAAGACAGCGGAATGAAA-3'	56	+ SVA: 4302 - SVA: 1889	PCR
LRIG2 SVA Proximal	F: 5'-GCCCAGGTACTTTAGCACCC-3' R: 5'-CACACCCAGCCGCAAATATG-3'	63	+ SVA: 2586 - SVA: 131	PCR
LRIG2 SVA VNTR	F: 5'-CCTCCCAAAGTGCCGAGATT-3' R: 5'-CAAAGCCGCCATTGTCATCC-3'	60	~1847	Nested PCR
KANSL1 SVA + Flanks	F: 5'- CCCTCCAGCACTCCCATTTT-3' R: 5'- CGCCTAACATCACACTACTTGG-3'	56	+ SVA: 3814 - SVA: 1491	PCR
KANSL1 SVA Proximal	F: 5'- AGTGACAGGGAGAGACTTCATC-3' R: 5'-CATAAGTGAACGGAGATGTATGC-3'	60	+ SVA: 2558 - SVA: 235	PCR
KANSL1 SVA Combined VNTR	F: 5'-CCTCCCAAAGTGCCGAGA-3' R: 5'-TCCCGCCTTTCTATTCCACA-3'	58	1621	Nested PCR
KANSL1 SVA VNTR 1	F: 5'-CGACTCACTACAACCTACACCT-3' R: 5'-CGACTCACTACAACCTACACCT-3'	NA	614	Nested PCR
KANSL1 SVA VNTR 2	F: 5'-CCCGTCCGGGAGGGAGATGG-3' R: 5'-CTATTCCACAAAGCCGCCAT-3'	NA	953	Nested PCR
KANSL1 SVA CT	F: 5'-TGTTTTGGCTCTTGAAAATACT-3' R: 5'-GCAGCAGTACAGTCCAGC-3'	62	167	PCR
KANSL1 SVA Poly-A	F: 5'-CCCTCCACTATTGTCCCATGA-3' R: 5'-CATAAGTGAACGGAGATGTATGC-3'	62	181	PCR
***LRIG2* #1**	F: 5'-GGAACACAACAACCTTACAC-3' R: 5'-CAAGTTCGGATAGTCTTTGG-3'	60	140	qPCR
***LRIG2* #2**	F: 5'-TAGAAACTGGAACACAACAAC-3' R: 5'-GATAGTCTTTGGCAGAACTC-3'	60	140	qPCR
***ACTB* #1**	F: 5'-GACGACATGGAGAAAATCTG-3' R: 5'-ATGATCTGGGTCATCTTCTC-3'	60	131	qPCR
***ACTB* #2**	F: 5'-GATCAAGATCATTGCTCCTC-3' R: 5'-TTGTCAAGAAAGGGTGTAAC-3'	60	191	qPCR
cg23932873 Amplification	F: 5'- (Btn)GGAGGGATGTTGTTAAGG-3' R: 5'-TCCTCACATCCAATCTTTACT-3'	55	88	Pyrosequencing
cg23932873 Sequencing	5'-TACTCAACACCCTCTTATCTC-3'	NA	NA	Pyrosequencing

Table 2.5 – Details of PCR, qPCR and pyrosequencing primers. Sequences are displayed 5' to 3'. Btn = Biotin label.

2.2.5. Nested PCR

When amplifying SVA components using primer pairs in which both oligonucleotides anneal within the retrotransposon it is very likely that a large number of non-specific

products will be produced as the primers anneal to homologous sequences throughout the genome. To circumvent this it was necessary to carry out a 'nested' PCR in which the whole SVA was first amplified for around half the typical number of PCR cycles, and an aliquot of this reaction was then used as input for a PCR reaction with the SVA-internal primers for the remaining half of PCR cycles. The first part of this approach amplifies the whole SVA without producing a product that is visible on an agarose gel, and the second half then amplifies the internal component in a template mixture that has been enriched for the particular SVA of interest – resulting in only the targeted SVA-internal components yielding a visible product.

Region targeted	**First PCR**		**Second PCR**		
	Primers	Cycles	Input from 1st:	Primers	Cycles
LRIG2 SVA central VNTR	LRIG2 SVA Proximal	20	2 µl	LRIG2 SVA VNTR	20
KANSL1 SVA central VNTR	KANSL1 SVA Proximal	20	2 µl	KANSL1 SVA Combined VNTR	20-25
KANSL1 SVA specific VNTRs	KANSL1 SVA Proximal	25	1	KANSL1 SVA VNTR 1/2-specific	20

Table 2.6 – Nested PCR primer and cycle number combinations.

2.2.6. Agarose gel electrophoresis

Agarose gel electrophoresis was the typical mode analysis for DNA amplicons produced in a standard 'endpoint' PCR. PCR products were loaded onto agarose gels (Invitrogen, 16500) made up in 0.5X TBE containing 33.3 ng/ml ethidium bromide

(Sigma, E1385). PCR products that did not already contain loading dye as part of the reaction mixture were mixed 5:1 with blue/orange loading dye (Promega, G1881) prior to loading. Agarose gels were 0.8 – 3% (w/v), with density dependent on the required resolution; smaller fragments (<500 bp) were run on 2 – 3% agarose to achieve greater resolution while larger fragments (>1 kb) were run on 0.8 – 1% gels. 100 bp and 1 kb size markers (Promega, G2101 and G5711) were run alongside samples for sizing, and 0.5X TBE was used as running buffer. Volume of PCR products loaded, agarose gel percentages, voltages and running times are specified where relevant. Gels were visualised at 302 nm and imaged with the BioDoc-It imaging system UV transilluminator (UVP, WZ-97701).

2.2.6.1. DNA extraction from agarose gel

If a specific PCR product needed to be isolated from other products or unused reaction components, it was separated on an agarose gel, viewed under 302 nm UV light and excised with a scalpel. DNA was purified using the Wizard SV Gel and PCR Clean-Up System (Promega, A9282) and eluted in NFW.

2.2.7. Quantitative PCR (qPCR)

2.2.7.1. qPCR reaction setup

The GoScript Reverse Transcription System (Promega, A5000) was used for first-strand complementary DNA (cDNA) synthesis from total RNA (**Section 2.2.3.2**), according to manufacturer's instructions. The input RNA was first normalised to the 15 – 100 ng/µl range, depending on the lowest concentration for a given set of

samples. The 3 µl input RNA therefore corresponded to 45 – 300 ng for cDNA synthesis – exact quantities are specified where relevant. Resulting cDNA was then diluted 1:10 in NFW, and GoTaq qPCR Master Mix from Promega (A6002) was used for qPCR with CXR reference dye included in all reactions. Each reaction was performed in triplicate to assess technical precision and reproducibility – any quantification cycle (Cq) replicate values that varied by more than 0.2 standard deviations (SDs) had an outlier value discarded. If the SD was still greater than 0.2, replicates were discarded altogether. qPCR amplification and detection were performed in an Aria MX Real-time PCR System (Agilent), with analysis performed using Agilent Aria 1.8 Software. The oligonucleotides used here were from the predesigned KiCqStart range by Sigma (**Table 2.5**). For the mRNA targets that underwent qPCR amplification in this thesis, one reaction mixture and set of cycling conditions was found to work reliably – see **Table 2.7**:

Reaction Mixture			**Cycling conditions**		
Reagent	Volume (µL)	Final Concentration	Temp. (°C)	Time	Cycles
GoTaq® qPCR Master Mix (2X)	10	1X	95	2 min	1
Fwd primer (10 µM)	2	1 µM	95 60	15 sec 1 min	40
Rev primer (10 µM)	2	1 µM			
CXR dye	0.2	300 nM	95 55 95	1 min 30 sec 30 sec	1 (Melt curve)
cDNA	5	1.125 – 7.5 ng/µl			
NFW	0.8	-			

Table 2.7 – qPCR reaction mixture and cycling conditions.

2.2.7.2. Determining qPCR primer efficiency

The ΔΔCT method assumes a perfect doubling of the target sequence with each amplification cycle in PCR. This would be referred to as a primer efficiency of 100%, but in reality this efficiency is not often the case. For example, if the primers do not bind the target sequence with sufficiently high affinity then some target templates may not be annealed by primers during the extension step of the PCR, resulting in a less than 100% efficient doubling each cycle. MIQE guidelines state that primers should be 90 – 100% efficient for publication of qPCR data [198]. To determine primer efficiencies, primers were used to amplify a 5-fold or 10-fold dilution series of cDNA and the mean Cq was plotted across this range on a $\log_{10}$ scale. The slope of the resulting line can be used to calculate the primer efficiency as follows:

$$Efficiency\ (\%) = \left(10^{\left(\frac{-1}{Slope}\right)} - 1\right) \times 100$$

It is often stated in guidelines for qPCR that primer efficiency should be in the range of 90 – 110%. However, an efficiency over 100% suggests that components within the template are inhibiting qPCR and causing the most concentrated sample or two in the dilution series to produce Cq values closer together than they should, skewing the curve and inflating the efficiency value. For this reason, the Cq of the most concentrated sample was typically discarded when calculating primer efficiency.

2.2.7.3. Relative quantification of gene expression using the ΔΔCT method

In qPCR experiments, the cycle at which the fluorescence associated with amplification exceeds the background fluorescence has been given a variety of terms

including threshold cycle (Ct) and quantification cycle (Cq). Although Cq is used throughout this thesis, in accordance with MIQE guidelines [198], one of the most widespread methods for assessing gene expression changes in qPCR data is known as the ΔΔCT method. Cq is directly associated with target mRNA abundance in a sample, and so expression of a gene of interest can be determined by comparing its Cq to that of a housekeeping gene that is expected to be stable across conditions. This normalised target gene Cq value can then be compared between conditions, typically treatment vs control. This is summarised as:

$$\Delta CT_{Sample} = CT_{Target\ gene} - CT_{Housekeeping\ gene}$$

$$\Delta\Delta CT = \ \Delta CT_{Treated\ sample} - \ \Delta CT_{Control\ sample}$$

In summary, ΔΔCT is the difference in normalised Cq values (ΔCT) between a control sample and a given treated sample for a particular gene. ΔΔCT is a value for gene expression change in logarithm base 2, but the same change can be expressed as fold change using the equation below:

$$Fold\ change\ = \ 2^{-\Delta\Delta CT}$$

2.2.8. Pyrosequencing

Bisulphite conversion of 500 ng gDNA in 20 µl NFW was performed using an EZ DNA Methylation-Gold Kit from Zymo Research (D5005) according to manufacturer's instructions. The converted DNA was eluted in 10 µl, and concentration was therefore estimated to be 50 ng/µl. PCR primers capable of amplifying bisulphite converted DNA were designed using PyroMark Assay Design Software 2.0.2 (QIAGEN) to include cg23932873 at position chr1:113072514 (hg38), which is a CpG dinucleotide of

interest identified in methylation datasets from the NABEC cohort (**Section 3.2.5**). It is generally held that amplicons of bisulphite converted DNA should be no more 300 bp as the conversion process tends to fragment DNA, and amplification is more efficient for smaller amplicons – thus, the targeted amplicon was 88 bp in length (see **Table 2.5** for primer details). A fragment of the converted DNA including the CpG dinucleotide of interest was amplified using the Pyromark PCR Kit from QIAGEN (978703). The forward primer used in this PCR had a 5'-biotin tag for use in downstream pulldown purification. An excess of the reverse primer was therefore used in this PCR to ensure that the biotinylated forward primer was exhausted, as leftover biotin can interfere with downstream streptavidin pulldown steps. Pyromark PCR reaction conditions were as follows:

Reaction Mixture			**Cycling conditions**		
Reagent	Volume (µL)	Final Conc./Quantity	Temp. (°C)	Time	Cycles
PyroMark PCR Master Mix (2X)	15	1X	95	15 min	1
Fwd primer, biotin (10 µM)	0.45	0.15 µM	94 55 72	30 sec 30 sec 30 sec	42
Rev primer (10 µM)	0.9	0.3 µM			
DNA (~50 ng/µl)	2	100 ng	72	10 min	1
NFW to 30 µl	11.65	-			

Table 2.8 – Reaction conditions for Pyromark PCR amplification of bisulphite converted DNA.

The resulting PCR product was prepared for pyrosequencing on a QIAGEN Pyromark Q96 ID system according to manufacturer's guidelines. Briefly, the biotinylated PCR products were immobilised on streptavidin-coated Sepharose beads, the non-

biotinylated strand was removed with a proprietary 'Denaturation Solution', and a sequencing primer was annealed to the biotinylated strand (see **Table 2.5**). Pyrosequencing was then performed on the aforementioned Pyromark system. In short, pyrophosphate-labelled deoxynucleotides are incorporated into an extending sequencing oligonucleotide using the biotinylated strand of the Pyromark PCR amplicon as a template, with light emitted when a nucleotide is incorporated. If a cytosine residue of interest was originally 5-methylated, and therefore unable to be bisulphite converted to uracil, when the sequencing reaction progresses to this residue a dCTP is incorporated into the sequencing oligonucleotide (dGTP is incorporated if the opposite strand was targeted for sequencing). Alternatively, if the cytosine had been unmethylated it would have been converted to uracil during the bisulphite conversion and then replaced by thymine during the subsequent Pyromark PCR reaction, and consequently dTTP nucleotide will be incorporated at this position during the sequencing step (similarly, dATP is incorporated if opposite stand targeted). During pyrosequencing the system sequentially attempts to incorporate a C/G and then a T/A nucleotide (or vice versa), measuring the resulting light emitted by each reaction and thereby calculating a percentage of CpG dinucleotides that were methylated in the sample.

2.2.9. Molecular cloning

The LRIG2 and KANL1 were cloned and inserted into plasmid vectors using a similar methodology: briefly, the DNA of interest – containing the SVA – was amplified by

PCR, ligated into a vector, transformed into competent *E. coli*, positive clones were selected, and then the construct was purified.

2.2.9.1. Target amplification and ligation into pCR-Blunt

Target SVA sequences were amplified from gDNA using KOD Hot Start Polymerase (**Section 2.1.6**) as this enzyme possesses proof-reading activity and can efficiently amplify the GC-rich sequences (>60& GC content) often found within SVAs. This polymerase produces blunt-ended amplicons which can be ligated directly into the MCS of pCR-Blunt, a vector which confers kanamycin resistance and only permits bacterial growth when it is circularised by insert DNA (**Figure 2.1**). Therefore, when grown in kanamycin only cells with an insert-containing pCR-Blunt plasmid should survive.

The target sequence was amplified with KOD Hot Start DNA polymerase and the resulting amplicon of interest was isolated via gel agarose electrophoresis, excision from the gel and purification (**Sections 2.2.4. and 2.2.6.1**). The concentration of the purified DNA fragment was determined (**Section 2.2.3.3**) and the required quantity of insert was determined with the following equation:

$$Insert\ (ng) = \frac{x\ \times insert\ size\ (kb) \times vector\ quantity\ (ng)}{vector\ size\ (kb)}$$

Where x is the ratio of insert to vector – a ratio of 10:1 of insert:vector was typically used, and is specified where relevant. It was found that the best ligation efficiencies

were achieved when the pCR-Blunt vector was used in conjunction with T4 DNA Ligase from NEB (M0202) in a 10 µl reaction, as below:

Reagent	Volume (µl)
pCR-Blunt vector (25 ng/µl)	2 (50 ng)
Insert DNA	1 - 5
10X Ligase Buffer (NEB)	1
T4 DNA Ligase (NEB)	1
NFW to 10 µl	1 - 5

The ligation mixture was incubated a minimum of 2 hours at room temperature, but was typically incubated overnight.

2.2.9.2. Restriction digests

Following cloning into pCR-Blunt it was typical for the resulting construct to be digested with type II restriction endonucleases, either as part of restriction mapping or for subcloning into an expression vector. For subcloning, the insert was excised from the pCR-Blunt vector by digesting a large quantity of the construct (>1 µg) with an endonuclease that cleaved within the vector MCS. A typical reaction used enzymes and buffers from NEB, and is outlined below:

Component	Quantity
DNA (pCR-Blunt with insert)	1 – 5 µg
rCutsmart buffer (10X)	2 µl
Restriction endonuclease (20 U/µl)	0.05 – 0.25 µl (1 U/ng)
NFW	To 20 µl total

All digests were incubated at 37 °C for 1 hour, followed by a 20 min 65 °C heat inactivation if the RE was heat sensitive.

The insert was then separated from the backbone by agarose gel electrophoresis (**Section 2.2.6**), purified from the gel (**Section 2.2.6.1**), and ligated into a target vector (setup similar to **Section 2.2.9.1**) previously restricted to have complementary sticky-end overhangs to the insert.

2.2.9.3. Dephosphorylation of linearised DNA fragments

In some instances it was necessary to dephosphorylate DNA fragments prior to ligation, for example to prevent recircularization of a vector. Antarctic Phosphatase (NEB, M0289) was used as follows:

Component	**Quantity**
DNA fragment	500 ng
Antarctic Phosphatase (5 U/µl)	2 µl (20 U/µg)
Buffer (10X)	2 µl (1X)
NFW	To 20 µl total
Incubation Temperature	Time
37 °C	30 min
80 °C (Inactivation)	2 min

2.2.9.4. Transformation of chemically competent *E. coli*

All molecular cloning and subcloning propagation steps were carried out in 'subcloning efficiency' DH5α *E. coli* (Invitrogen, 18265017). These cells are sensitive to the ccdβ-mediated selection described in **Figure 2.1** and so are suitable for use

with pCR-Blunt. Transformation of chemically competent DH5α cells was carried out according to manufacturer's guidelines. Briefly, 2 µl of ligation mixture (quantity of DNA was variable, and specified where relevant) (**Section 2.2.9.1**) was added to a 50 µl aliquot of cells, mixed by flicking the tube and incubated on ice for 30 min. The cells were then heat shocked at 42 °C for 20 seconds and returned to ice for 2 min. 950 µl of pre-warmed LB broth was added to the tube and the cells were placed in a shaking incubator for 1 hour at 37 °C and 225 rpm. After incubation, 200 µl of transformant was spread on pre-warmed LB agar plates containing an appropriate antibiotic: 50 µg/ml kanamycin was used for pCR-Blunt transformants, and 100 µg/ml ampicillin was used for pSpCas9(BB)-2A-GFP. Plates were incubated overnight at 37 °C.

2.2.9.5. Extraction of plasmid DNA from transformed *E. coli*

Miniprep purification was used when relatively low quantities of plasmid DNA (~100 ng/µl) would suffice. From the plates on which transformed *E. coli* had been plated, individual colonies were transferred using a pipette tip to 5 ml LB broth cultures containing an appropriate antibiotic and incubated at 37 °C and 225 rpm overnight. The following day 1 ml of this culture was used for miniprep of plasmid DNA using the QIAprep Spin Miniprep kit (QIAGEN, 27106) according to manufacturer's instructions. Plasmid DNA was eluted in 25 – 40 µl NFW, quantified (**Section 2.2.3.3**) and stored at -20 °C. After miniprep purification it was typical to validate insert presence and orientation in the vector, either by restriction mapping (**Section 2.2.9.2)** or sequencing (**Section 2.2.10**).

Maxiprep purification of DNA was used when larger quantities of higher purity plasmid were required (~ 1 µg/µl), such as for transfection of cell lines. 50 µl of the 5 ml cultures used for minipreps was used to inoculate 100 ml LB broth containing appropriate antibiotic, which was incubated overnight at 37 °C and 225 rpm. DNA was purified using the QIAGEN Plasmid Maxi Kit (12163) following manufacturer's instructions. Following ethanol precipitation the DNA pellets were resuspended in 200 µl NFW, quantified (**Section 2.2.3.3**) and stored at -20 °C.

2.2.10. Sequencing

Sequencing of PCR products and plasmid constructs was performed externally by Source Bioscience via Sanger sequencing. SVAs are considered high GC content sequences (>60%) that are prone to secondary structure formation, so additional dGTP chemistry was included in sequencing of these elements.

2.2.11. Human cell line tissue culture

2.2.11.1. Culture of established cell lines

Several established human cell lines were used in the work presented here (details and media formulations in **Section 2.1.4**). All cell lines were typically cultured in T25 or T75 flasks (Corning) at 37 °C and 5% CO_2 in a humidified incubator, and passaged when at 80 – 90% confluency. Passaging involved aspiration of media, washing cells with pre-warmed PBS, addition of 0.25% trypsin-EDTA and incubation at 37 °C for ~5

min to achieve cell dissociation. Trypsin was quenched by the addition of media containing 10% FBS, and the cell suspension was transferred to a 15 ml falcon tube and centrifuged at 130 x g for 5 min. The supernatant was aspirated, the cell pellet was resuspended in pre-warmed culture media, and the appropriate volume of cells was moved to a new culture flask to achieve the desired density.

For long-term storage of cell lines, aliquots of cells in FBS with 10% DSMO were frozen at -80 °C in an isopropanol-filled 'Mr Frosty' freezing container (Thermo Scientific, 5100-0001). After 24 hours, frozen vials were moved to liquid nitrogen storage.

2.2.11.2. Cell counting

Established cell lines were harvested from their flask and resuspended, typically in 10 ml culture media (**Section 2.2.11.1**). 10 µl of cell suspension was applied to the side of a glass haemocytometer with a glass cover slip placed on top, such that the liquid was pulled across the haemocytometer's central etched counting square by capillary action. When viewed under a light microscope the cells were visible in a near 2D field in the haemocytometer's 4 counting squares; the mean cell counts of these 4 squares multiplied by 10,000 provided the concentration of cells in the original suspension in cells/ml. From this, volumes needed to seed a specific number of cells were calculated.

2.2.12. CRISPR-mediated deletion of the LRIG2 SVA in SH-SY5Y cells

The CRISPR-Cas9 deletion strategy utilised here was adapted from the methodology used by Ran *et al.* 2013 and made use of non-homologous end joining (NHEJ) [199]. NHEJ-based repair is sufficient for non-exonic genome modifications where high fidelity repair is not necessary, and is less complicated than a homology-directed repair strategy in which a repair template must also be provided. The approach in this thesis made use of a single plasmid to deliver both the Cas9 protein machinery and the gRNA within an RNA scaffold, which is necessary for gRNA association with the Cas9 and genomic targeting of the protein (**Figure 2.2**). For excision of SVA elements two pSp-Cas9 constructs containing different gRNA sequences would be necessary in order to cut either side of the target element. This DSB would ideally be repaired by NHEJ, resulting in deletion of the target SVA.

2.2.12.1. Guide RNA design and oligo annealing

The gRNA design tool used here was developed by the Zhang Lab at Massachusetts Institute of Technology (http://crispr.mit.edu/). Although this tool automatically attempts to identify unique genomic regions to target and therefore avoid off-target binding and Cas9-mediated cleavage, the process was streamlined by prior screening of sequences for repetitive regions as identified in the latest release of the RepeatMasker track available on UCSC genome browser (hg38). Suitable gRNAs were identified that were immediately 5' of the Cas9 protospacer adjacent motif (PAM) sequence 5'-NGG-3' (where N is any nucleobase), which is required for Cas9 binding. These guides were scored by the software based on predicted off-target binding, and

after checking predicted thermodynamic properties for minimal formation of secondary structures (ΔG value between –0 kcal/mole and –6 kcal/mole per structure, see **Section 2.2.2**) the top three scoring guides 5' and 3' of the target were taken forward. These single-stranded guide sequences needed to be incorporated into a double-stranded DNA molecule with cohesive 5' and 3' overhangs for the downstream 'Golden Gate' cloning strategy. Therefore, a sense and antisense oligonucleotide with the necessary end modifications were designed, as in the following example:

```
gRNA 5'- CACCGTCTGGTAAGAAATCCGGCAT      -3'
     3'-     CAGACCATTCTTTAGGCCGTACAAA  -5'  (Complement)
```

Here, the red sequence in the top strand was the desired sequence for the guide RNA molecule while the red bottom strand sequence was its complement. Black nucleotides denote additional bases required for incorporation into the vector used in Golden Gate cloning. For the gRNAs designed to target the LRIG2 SVA, the sense oligonucleotide of each complementary pair (relative to the pSpCas9(BB)-2A-GFP vector) is listed in **Table 2.9** (only considering the gRNA targeting sequence i.e. the red sequence described above):

Relative to LRIG2 SVA	Oligo #	Sequence	Cas9 cut site
5’	2	GTCCCGAGGTAAGGAGATAT	Chr1: 113,068,454
	3	TTGCAAAGAGTAAAGTCCCG	Chr1: 113,068,440
3’	4	TCTGGTAAGAAATCCGGCAT	Chr1: 113,071,854
	5	CCACTTACTGCGGAGGATAC	Chr1: 113,071,748
	6	CCTGTATCCTCCGCAGTAAG	Chr1: 113,071,717

Table 2.9 – gRNA sequences and cut sites of for CRISPR-Cas9 targeting of the LRIG2 SVA.

Prior to the Golden Gate cloning process, complementary oligonucleotides were annealed together. 5 µg of each of the sense and antisense oligonucleotides were mixed with 5 µl T4 DNA ligase buffer and made up to 100 µl in NFW, heated to 95 °C for 5 min and then allowed to cool to room temperature for 1 hour.

2.2.12.2. Golden Gate cloning

‘Golden Gate’ cloning provides a streamlined and efficient strategy to insert sequences into vectors such as pSpCas9(BB)-2A-GFP. Type IIS restriction endonucleases such as BbsI recognise asymmetric sequences and cleave DNA outside of their recognition site. The gRNA scaffold region of pSpCas9(BB)-2A-GFP contains two divergent BbsI sites such that a short sequence is excised when digested with BbsI, which then competes with the modified gRNA double-stranded oligonucleotide for insertion into the vector during ligation. Successful insertion of the gRNA oligo destroys the BbsI recognition sites while reinsertion of the excised sequence regenerates them. Therefore, the desired non-digestible construct will accumulate over multiple rounds of digestion and ligation (**Figure 2.3**).

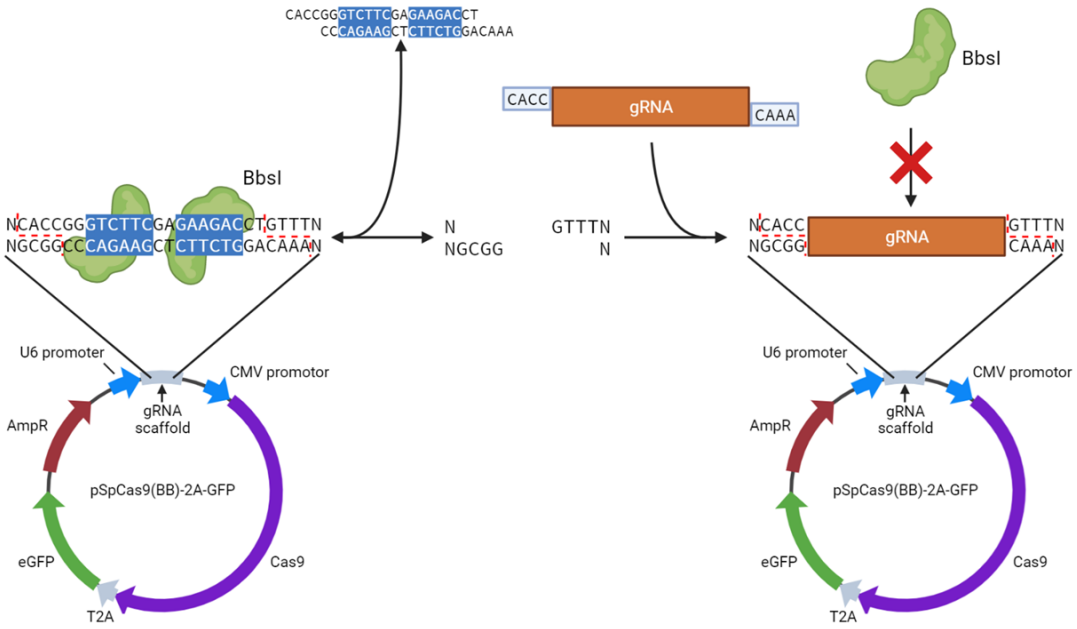


Figure 2.3 – Schematic of Golden Gate cloning strategy. The pSpCas9(BB)-2A-GFP vector is digested with BbsI, excising a short nucleotide sequence. During ligation the excised fragment and the gRNA oligo insert compete for ligation into the vector. Insert of the gRNA removes BbsI recognition sites and prevents re-excision, so the reaction is driven to the righthand side over several cycles of digestion and ligation.

The Golden Gate cloning reaction mixture for pSpCas9(BB)-2A-GFP was as follows, per reaction:

Reagent	Volume (µl)	Final quantity
Vector (100 ng/µl)	1.5	150 ng
Annealed oligos (100 ng/ µl)	3	150 ng
Ligase buffer, 10X	2	1X
T4 DNA Ligase (400 U/µl)	1	400 U
BbsI-HF (NEB, R3539, 20 U/µl)	1	20 U
NFW to 20 µl	11.5	-

This mixture was then incubated in a thermocycler under the following conditions:

Temp (°C)	Time (min)	Cycles
37	5	10
16	10	
37	30	1
80	20	1

2 µl of this reaction was then taken forward to transform chemically competent *E. coli* strain DH5α, as described previously (**Section 2.2.9.4**). After colony picking and miniprep purification of plasmid DNA (**Section 2.2.9.5**), candidate constructs were digested with BbsI (**Section 2.2.9.2**) and visualised on agarose gels since successful insertion should prevent linearisation of the plasmid. Finally, insertion of gRNA sequence was confirmed for each construct via Sanger sequencing using the U6 primer (5'-GAGGGCCTATTTCCCATGATT-3') which sequences from the vector's U6 promoter and across the gRNA scaffold (**Figure 2.2**). Positive constructs were then purified at greater quantity by maxiprep (**Section 2.2.9.5**).

2.2.12.3. Transfection of CRISPR plasmids and clonal cell line isolation

The workflow in **Figure 2.4** outlines the following protocol. The established cell line SH-SY5Y was seeded at 100,000 cells per well in 24-well plates in culture media free of penicillin-streptomycin, for a total or 4 or 8 wells (depending on number of cells available) and incubated for 24 hours. pSpCas9(BB)-2A-GFP CRISPR plasmid constructs were delivered into the cells using Lipofectamine 3000 transfection reagent (Invitrogen, L3000) in combination with Opti-MEM (Gibco, 11058-021)

according to manufacturer's instructions. Briefly, 1 µg each of two plasmids with gRNAs targeting 5' and 3' of the SVA to be deleted were added to the cell culture media of each well as part of the following mixture:

Component	Volume (µl)
Opti-MEM	50 µl
5'-targeting plasmid (500 ng/µl)	2
3'-targeting plasmid (500 ng/µl)	2
P3000 reagent	1
Lipofectamine 3000 reagent	1.5

Positive transformants were visualised for GFP fluorescence under 395 nm UV light to qualitatively gauge transformation efficiency. After 48 hours cells were trypsinised and dissociated, counted, and re-seeded (**Section 2.2.11.**) at a low density of 1,000 cells in a 10 cm cell culture dish with typical SH-SY5Y culture media containing penicillin-streptomycin (**Section 2.1.4**). The cells were cultured for 1–2 weeks until individual cells had grown into colonies visible by eye. 400–600 colonies were then mechanically picked from the culture dish on the end of a disposable pipette tip, transferred to 96-well plates, and cultured until 70–90% confluent. Cells were then split into two duplicate 96-well plates, with 75% of cells in one plate and the remaining 25% of cells into the other. The more confluent of the two plates was used for genotyping of the SVA locus while the other was cultured until successful deletions had been identified.

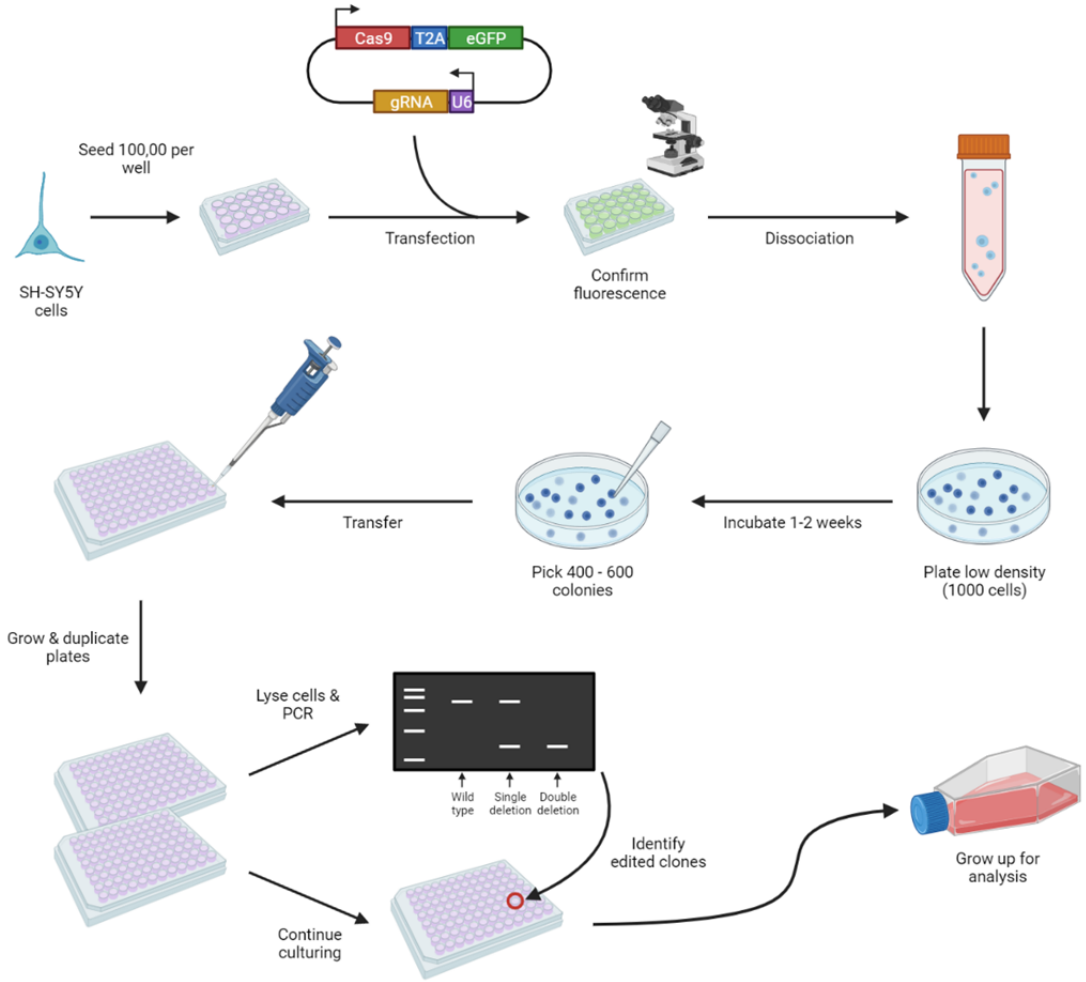


Figure 2.4 – Schematic of CRISPR-Cas9 workflow for transfection and clonal isolation workflow in deletion of SVA sequence in SH-SY5Y cell line. SH-SY5Y cells were seeded in antibiotic-free media and incubated for 24 hours before transfection with CRISPR plasmids. 48 hours later cells were trypsinised, harvested and seeded at low density. After 1 or 2 weeks visible colonies were manually picked using a pipette tip and transferred to 96-well plates for outgrowth. Once 70 – 90% confluent, cells were split 1:3 into duplicate plates and the more confluent plate was used for genotyping of the deleted region. Successful edits were mapped back to the lower density plate and taken forward for analysis.

2.2.12.4. PCR screening of candidate edited clonal cell populations

DirectPCR (cell) lysis reagent (VIAGEN Biotech, 302-C) with 1 mg/ml proteinase K (Sigma, P4850) was used as a crude lysis solution, in preparation for PCR. The more

confluent plate from each pair of clonal isolate duplicates was washed with warm PBS and 50 µl of lysis solution was added to each well. This plate was incubated for 2 hours at 55 °C in a rotating hybridisation oven (Hybaid) and then the lysates were transferred to a 96-well PCR plate and heat inactivated for 30 min at 85 °C. 1 µl of lysate was used directly as input for PCR with KOD Xtreme Hot Start Polymerase (**Section 2.2.4.**), as this enzyme is capable of efficient amplification of DNA in non-purified samples. PCR products were visualised with agarose gel electrophoresis as previously described, and when successfully edited clones were identified they were harvested from the corresponding well in the remaining duplicate plate and taken forward (**Figure 2.4**).

Chapter 3 Investigating the cis-regulatory roles of an SVA RIP in a gene promoter region

3.1. Introduction

It is postulated that SVA retrotransposons, as human-specific and contemporarily active elements, might be importance sources of genetic variation that modulate genome regulation by introducing novel elements such as TF binding sites, CpG islands, splice sites or chromosomal rearrangements. While there are a few examples of SVA insertions producing highly significant changes in gene expression, such as the *TAF1* and *CAPS8* insertions described previously (**Section 1.2.9**), in general interpersonal gene expression differences arising from TEs could be small and numerous, being cumulatively important for a phenotypic outcome but individually difficult to ascribe significance. This may be relevant for genetically complex diseases such as PD, in which retrotransposons are poorly characterised and GWAS have otherwise largely failed to identify causal variants.

However, before focussing on retrotransposons in PD, establishment of a readily accessible model system to investigate the general influence of retrotransposons insertion polymorphisms (RIPs) at gene regions was explored. It was observed that on chromosome 1 there was a 2.4 kb long fully intact SVA F1 situated ~2 kb upstream of the TSS of leucine-rich repeats and immunoglobulin-like domains 2 (*LRIG2*) (**Figure 3.1**), a protein that modulates epidermal growth factor signalling. The LRIG2 protein is ubiquitously expressed but deleterious mutations are associated with the central nervous system disorder urofacial syndrome [200, 201], characterised by perturbed nerve control of the face and bladder [202]. Specifically, *LRIG2* appears to be important for correct patterning of nerve cells during development of the urinary

tract [203]. The *LRIG2* gene shares its bidirectional promoter region with a non-coding divergent transcript, *LRIG2-DT*, the expression of which could be co-regulated [204, 205] (**Figure 3.1**). The SVA proximal to this *LRIG2* promoter region (herein 'the LRIG2 SVA') is a RIP, meaning that it was inserted relatively recently in evolutionarily terms, and was hypothesised to be a transposition event involved in human-specific modulation of gene expression. This means DNA samples from the general populace represent a naturally occurring model to address the influence of presence or absence of this RIP on nearby genome structure or gene regulation. Indeed, the presence of the LRIG2 SVA is assigned an allele frequency of 0.422 in the Database of RIPs in Humans (dbRIP, accession RIP3000013) [206], meaning that its presence or absence at the locus should be observed almost equally often. Furthermore, it was noted that the LRIG2 SVA had a GC content of around 70% and contains 170 CpG dinucleotides as identified in the UCSC genome browser, representing a sizeable CpG island. By contrast, the CpG island associated with the *LRIG2* promoter region is 669 bp long, is approximately 65% GC content and is made up of 57 CpGs (**Figure 3.1**). It could be speculated that this promoter CpG island might be sensitive to methylation changes induced by the much larger one within the LRIG2 SVA.

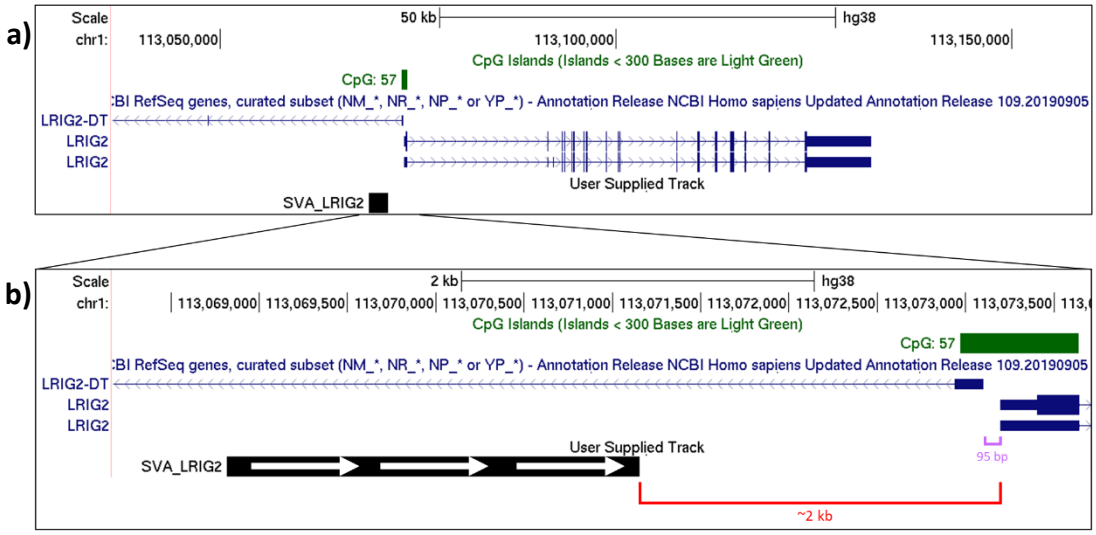


Figure 3.1 – The *LRIG2* locus in hg38 as shown on UCSC genome browser. **a)** Displayed are two validated isoforms of *LRIG2* and the first exon of *LRIG2-DT* from the RefSeq genes curated subset, the LRIG2 promoter-associated CpG island, and the LRIG2 SVA (SVA_LRIG2). **b)** A closer view of the LRIG2 SVA and *LRIG2* promoter region is displayed. Distance to locus TSS highlighted in red, and distance between LRIG2 and LRIG2-DT transcriptional start sites highlighted in purple. 5' to 3' orientation of the LRIG2 SVA is indicated by white arrows.

Altogether, the SVA F1 at the LRIG2 locus is an ideal candidate for a model system in which to study how endogenous SVA RIPs might result in interpersonal differences in gene regulation in humans. The LRIG2 SVA was therefore examined in a sample cohort from the general populace in which endogenous differences in RIP genotype can be compared to molecular phenotypes. Additionally, trends associated with LRIG2 SVA genotype were separately assessed via a gene editing approach in which the SVA is exogenously removed. It was expected that this would elucidate how SVA retrotransposons influence gene regulation in contexts besides highly deleterious disease-associated insertions, and would inform future GWAS on the importance of

incorporating available TE datasets when assessing many smaller contributions to gene regulation.

3.1.1. Aims

To characterise the influence of the SVA F1 at the *LRIG2* locus by:

- Determining frequencies of the LRIG2 SVA for RIP genotype (i.e., presence versus absence) and any length polymorphisms in the North American Brain Expression Consortium (NABEC) cohort, a resource of DNA samples from neurologically normal individuals with corresponding WGS, transcriptomic and methylation data.
- Stratifying NABEC transcriptomic and methylation datasets on these genotypes to assess the impact of SVA allele dosage.
- Generating a cell line model in which the endogenous LRIG2 SVA is deleted on one or both chromosomes, and assessing how this impacts expression and methylation patterns in an otherwise genetically identical background.

3.2. Results

3.2.1. Primers were designed to PCR amplify the LRIG2 SVA with and without flanking regions, and to address polymorphism within specific domains of the SVA

In characterising the LRIG SVA in genomic DNA samples it was first necessary to design primers that would specifically amplify it. This is a process that can require multiple iterations, as SVAs possess large repetitive elements and high GC content which can form secondary structures recalcitrant to PCR.

To determine SVA presence or absence, primers were required for an 'empty-site' PCR in which the regions flanking the element of interest were amplified along with the target. The inclusion of these flanks produces a relatively large PCR product when the SVA is present at the locus and a smaller, but still sizable, product when it is absent. The generation of two large but distinctly sized PCR fragments allows SVA RIP genotype to be readily determined following agarose gel electrophoresis and visualisation. Accordingly, DNA sequences for the LRIG2 SVA (coordinates defined by RepeatMasker, hg38) and the 2 kb upstream and downstream of the element were downloaded from UCSC genome browser (hg38) in plain text format. This was used as input for NCBI Primer-Blast to produce a list of oligonucleotides that would amplify the SVA efficiently and specifically (**Section 2.2.2**), with a minimum amplicon size of 3.5 kb specified so that even when the SVA was absent a product of at least 1 kb would be visible. Oligonucleotide properties were confirmed using OligoAnalyzer (https://eu.idtdna.com/pages/tools/oligoanalyzer). The candidate primer pair with the best predicted thermodynamic properties (detailed in **Section 2.2.2**), termed

'LRIG2 SVA + Flanks', annealed 1.1 kb upstream and 0.8 kb downstream of the genomic site of the LRIG2 SVA and were predicted to produce a 1.9 kb product when the SVA was absent.

These primers underwent PCR to amplify the SVA with a range of annealing temperatures (conditions **Section 2.2.4**). It was determined that the best compromise between primer binding specificity and amplification efficiency was achieved with a 56 °C annealing step (**Figure 3.2a**). However, in this test only the predicted 'empty site' product at 1.9 kb was visible, indicating that the LRIG2 SVA insertion was homozygous absent in the HEK293 gDNA used. Therefore, the selected annealing temperature of 56 °C was further tested in four other cell lines to ensure that the amplicon containing the SVA would also amplify efficiently. MCF-7, HAP1, SKNAS and SH-SY5Y gDNA underwent PCR using the 'LRIG2 SVA + Flanks' primer pair and it was observed that the 4.3 kb 'filled site' PCR product containing the SVA was produced with specificity and efficiency comparable to the empty amplicon in SKNAS and SH-SY5Y DNA (**Figure 3.2b**). This PCR was thus able to differentiate between LRIG2 SVA presence and absence genotypes, demonstrating that the SVA was homozygous absent in MCF-7 and HAP-1 cells, heterozygous in SKNAS and homozygous present in SH-SY5Y (primer binding sites and amplicons summarised in **Figure 3.4**).

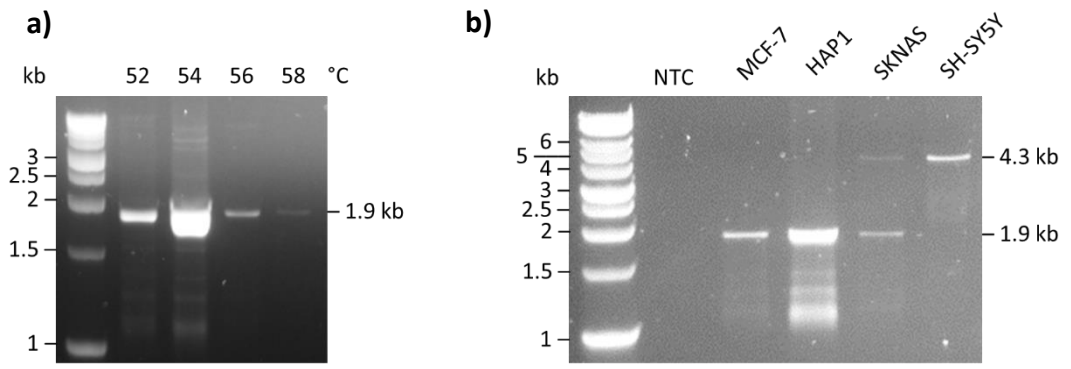


Figure 3.2 – Optimisation of 'LRIG2 SVA + Flanks' primer pair. **a)** The primers were tested in a standard PCR reaction using KOD Hot Start Polymerase, 40 cycles, 10 ng HEK293 gDNA template, and a range of annealing temperatures. PCR products were run on a 1% agarose gel at 120 V for 2 hours. **b)** Primers were used in the same PCR conditions but with 10 ng input of MCF-7, HAP1, SKNAS or SH-SY5Y gDNA at the nominated annealing temperature of 56 °C. PCR products were run on a 1% agarose gel at 140 V for 1 hour. **a & b)** NTC = No template control. Predicted amplicon sizes shown to the right of each figure.

Genomic VNTRs have been shown to represent regions of TF binding sites and epigenetic marks that can be differentially responsive to cellular stimuli depending upon repeat length polymorphisms [182]. Furthermore, it has been demonstrated that the central VNTR of SVA Ds can be bound by the chromatin architectural protein CTCF (**Figure 1.8**) [180]. Binding of TFs such as CTCF may conceivably be altered by length polymorphisms associated with the central VNTR, which might lead to hidden heterogeneity if these variants are grouped together as 'LRIG2 SVA present'. It was therefore also prudent to characterise the central VNTR of the LRIG2 SVA in addition to the element's RIP genotype in order to stratify any associations made at the locus. While the CT hexamer repeat within many SVAs can be considered a VNTR with

similar regulatory potential to the central VNTR, the 5' CT element has been lost in the SVA F1 family owing to splicing of the *MAST2* exon 1 to the SVA *Alu*-like region, thereby displacing the CT hexamer (**Figure 1.8**). Characterisation of the CT region was therefore not required for the LRIG2 SVA, which belongs to the F1 subfamily.

Primers targeting the VNTR within the SVA would be predicted to anneal to many other SVAs in the genome and thereby produce ambiguous result in PCR. Therefore, the VNTR was investigated using a 'nested' approach in which the SVA was first amplified using primers that annealed close to the SVA element, and this LRIG2 SVA-enriched template mixture was used as input for amplification with VNTR-targeting primers. To determine primers that would amplify the only the LRIG2 SVA or its central VNTR, sequences were input into Primer-Blast with minimal (<100 bp) flanks included in each direction (**Section 2.2.2** for primer design principles). The top candidate primers that annealed close to the genomic site of the SVA, herein 'LRIG2 SVA Proximal', were used in an annealing temperature gradient PCR similar to that in **Figure 3.2a** with SH-SY5Y. The observed 2.6 kb PCR product corresponded to the predicted size of the 'LRIG2 SVA Proximal' amplicon and an annealing temperature of 63 °C was selected for balance between amplification efficiency and specificity (**Figure 3.3a**). Subsequently, 'LRIG2 SVA VNTR' primers were tested in a 'nested PCR' in which SH-SY5Y gDNA underwent 20 cycles of amplification with 'LRIG2 SVA Proximal' primers, and a 1 µl aliquot of this LRIG2 SVA-enriched mixture was used as input for 15 cycles of amplification with 'LRIG2 SVA VNTR' primers with an annealing

temperature gradient. Good specificity was observed at all temperatures tested (**Figure 3.3b**), and a 60 °C annealing step was selected.

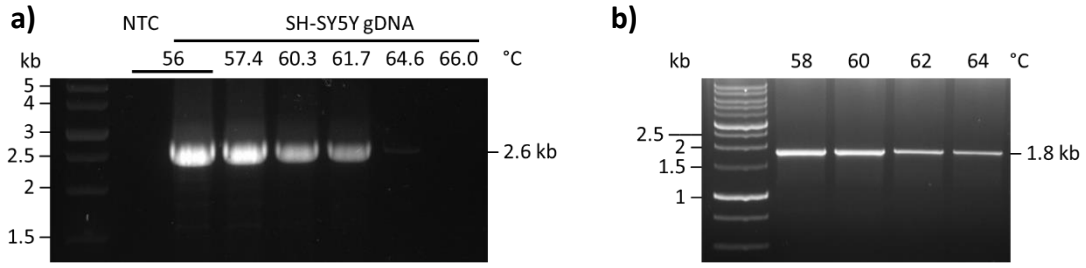


Figure 3.3 – Optimisation of 'LRIG2 SVA Proximal' and 'LRIG2 SVA VNTR' primer pairs. **a)** The 'LRIG2 SVA Proximal' primers were tested in a standard PCR reaction using KOD Hot Start Polymerase, 40 cycles, 10 ng SH-SY5Y gDNA template, and a range of annealing temperatures. 12 µl of PCR products were run on a 1% agarose gel at 90 V for 2 hours. **b)** In a test of a nested PCR, 5 ng SH-SY5Y gDNA template was used as input for a 20 cycle PCR with 'LRIG2 SVA Proximal' primers using KOD Hot Start Polymerase. 1 µl of this was used as input for 15 cycles of amplification with the 'LRIG2 SVA VNTR' primers, with a range of annealing temperatures. 8 µl of PCR products were loaded onto a 0.8% agarose gel and ran at 120 V for 1 hour. **a & b)** NTC = No template control. Predicted amplicon sizes shown to the right of each figure.

The primer annealing sites and amplicon sizes for the 'LRIG2 SVA + Flanks', 'LRIG2 SVA Proximal' and 'LRIG2 SVA VNTR' oligonucleotides are summarised below in **Figure 3.4**:

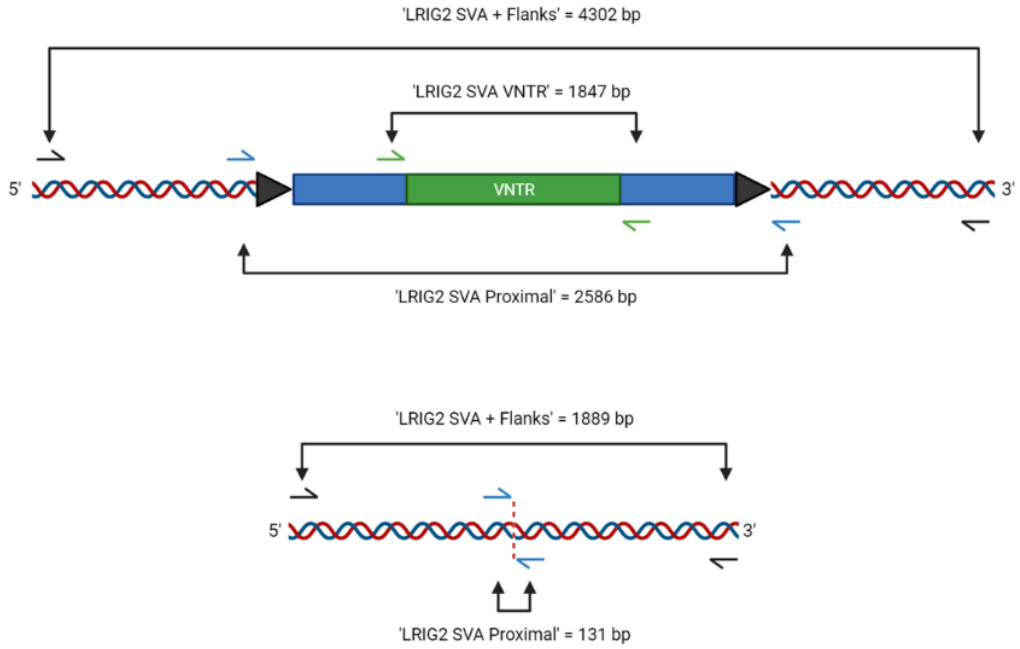


Figure 3.4 – Illustration of LRIG2 SVA primer binding sites and amplicon sizes when the LRIG2 SVA is present (top) and absent (bottom).

3.2.2. The LRIG2 SVA is a common RIP with four VNTR length variants in a North American cohort

The LRIG2 SVA was characterised in DNA samples from the North American Brain Expression Consortium (NABEC) that were provided by collaborators ([194-196], full link to study provided in **Section 2.1.3**). NABEC samples were from neurologically normal individuals and corresponding WGS, total RNA sequencing and CpG methylation datasets were available. Since repetitive DNA elements such as retrotransposons are routinely filtered out of short-read WGS data due to the inherent difficulties in mapping them back to the reference genome [207], it was necessary to genotype the LRIG2 SVA in the available samples from the NABEC cohort. 96 DNA samples were genotyped for LRIG2 SVA RIP genotype using standard

PCR with 'LRIG2 SVA + Flanks' primers (see **Table 2.5** for PCR conditions). Examples of the resulting genotypes in NABEC are displayed in **Figure 3.5a**. Of the 96 individuals, the LRIG2 SVA was absent in 14 samples (genotype −/−), 43 samples had 1 copy present (+/−, i.e., the SVA was present on a single chromosome), and 39 had 2 copies present (+/+, i.e., the SVA was present on both chromosomes). Notably, this yields an allele frequency of 0.37 for the LRIG2 SVA insertion, approximately recapitulating the 0.422 allele frequency listed on dbRIP.

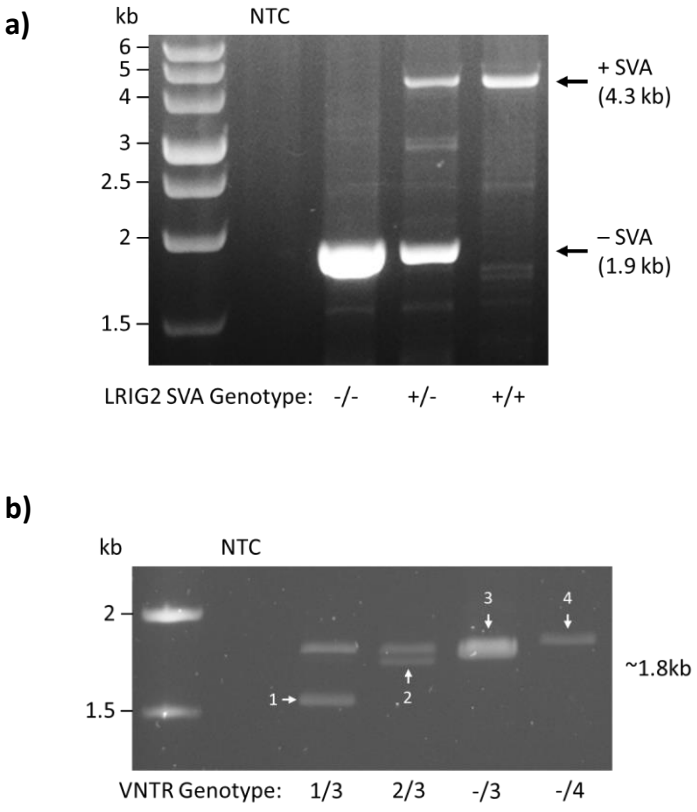


Figure 3.5 – Genotyping the LRIG2 SVA in NABEC frontal cortex DNA. **a)** Representative 'empty site' PCR of LRIG2 SVA in NABEC DNA samples using 'LRIG2 SVA + Flanks' primers, which flank the SVA by almost 1 kb each side. A 1.9 kb PCR product corresponds to SVA_LRIG2 being absent from the locus and a 4.3 kb PCR product corresponds to SVA_LRIG2 presence. 10 ng DNA input, and from a 24 µl

reaction mixture 12 µl was loaded onto a 0.8% agarose gel and run at 120 V for 90 min. SVA RIP genotypes are displayed along the bottom. **b)** Representative nested PCR of the LRIG2 SVA VNTR region. 10 ng DNA input and 20 cycles in first PCR reaction, with 2 µl (quantity doubled from optimisation in **Figure 3.3b** to obtain brighter bands, as NABEC DNA was found to amplify less efficiently than gDNA prepared from cell lines) of this used as input in second reaction which used 15 amplification cycles. 8 µl loaded onto a 0.8% agarose gel and ran at 100 V for 4 hours. White numbers indicate VNTR alleles in ascending order of length, and VNTR genotype is displayed along the bottom. **a & b)** NTC = No template control.

These NABEC samples were then examined for LRIG2 SVA VNTR genotype using the nested PCR approach described in **Section 3.2.1** (**Table 2.5**). Following agarose gel electrophoresis and visualisation it was observed that the NABEC cohort contained 4 VNTR length variants, denominated 1 – 4 in ascending order of size; representative PCR products are displayed in **Figure 3.5b**. Chromosomes lacking the LRIG2 SVA (previously determined via PCR with 'LRIG2 SVA + Flanks' primers, as in **Figure 3.5a**) were denoted VNTR genotype '-'. The VNTR genotyping results, along with the previously described RIP genotyping, are summarised in **Table 3.1**.

LRIG2 SVA RIP Genotype	Count	%	VNTR Genotype	Count	%
-/-	14	14.6	-/-	14	14.6
+/-	43	44.8	-/3	42	43.8
			-/4	1	1.0
+/+	39	40.6	1/3	3	3.1
			2/3	2	2.1
			3/3	34	35.4

Table 3.1 – Counts of SVA LRIG2 RIP and VNTR genotypes in available DNA samples.

Having identified 4 VNTR alleles within the LRIG2 SVA in NABEC, sequencing of these regions was attempted in the expectation that this might reveal differing potential for genomic regulation – for example, longer VNTRs with higher copy numbers of a given DNA repeat would be anticipated to contain greater numbers of any TF binding sites. Each VNTR allele was amplified in a nested PCR (**Section 2.2.5**) and cloned into the pCR-Blunt plasmid (**Section 2.2.9.1**). PCR mixtures produced by amplification of heterozygous VNTR genotypes '1/3' and '2/3' yielded a mixture of constructs containing each allele, but the incorporated allele was readily identified by PCR using the 'LRIG2 SVA VNTR' primer pair. Thus, pCR-Blunt constructs containing each of the 4 VNTR alleles were generated in this way. These then underwent Sanger sequencing (**Section 2.2.10**) using the forward and reverse 'LRIG2 SVA VNTR' primers separately. As Sanger sequencing typically achieves good quality sequence data for ~1.2 kb of DNA, it was expected that sequence reads initiating from upstream and downstream of the VNTR using these primers would result in overlap in the centre of the element, enabling reconstruction of even the largest VNTR allele. However, sequence reads were shorter than expected, possibly due to formation of secondary structure in the

repetitive VNTR template, and only the full-length sequence for the shortest VNTR, allele 1, could be reconstructed (alignments performed using MAFFT v7 [208]). It was therefore not possible to directly compare the DNA sequences of VNTR alleles to study differences in regulatory potential.

3.2.3. LRIG2 SVA proxy SNP generation

TEs are not automatically included in short-read WGS processing and genotype calling, instead being filtered out due to poor mappability and then added back in by software that recognises motifs or insertion hallmarks with varying degrees of accuracy [189-193]. However, it is assumed that these variants may occur within haplotype blocks shared with variants included in WGS through LD. Therefore, the genotype of uncalled variants of interest may be inferred through 'proxy' (or 'tagging') SNPs that are in high LD [209]. In this way, the genotype of the LRIG2 SVA was imputed in the WGS data of NABEC individuals for whom DNA was not available in our lab. To fully capture the genetic diversity of the LRIG2 SVA, proxy SNPs were generated for each of the VNTR alleles (see **Section 2.2.1.2**). This generated a list of 29 proxy SNPs, and the SNPs with highest linkage disequilibrium r^2 and D' values were taken forward, as shown in **Table 3.2**:

VNTR allele	Proxy SNP	r^2	D’
1	rs114767321	1	1
2	rs183751190	1	1
3	rs12744009	0.894	1

Table 3.2 – Selected LRIG2 SVA VNTR proxy SNPs and their LD values.

Proxy SNPs for VNTR allele 4 were excluded after filtering for standard genotype missingness of <0.1 (in other words, less than 90% of NABEC samples had genotypes for putative VNTR 4 proxy SNPs, a threshold below which it is standard to discard SNP data). These chosen proxy SNPs recapitulated the 96 PCR-validated LRIG2 SVA VNTR genotypes with 97.4% accuracy (187 out of 192 alleles, data not shown). When added to the existing validated genotyping, this yielded a total of 329 individuals (**Table 3.3**). To determine the overall LRIG2 SVA RIP genotype, these VNTR alleles were grouped into the genotypes −/−, +/− and +/+.

LRIG2 SVA RIP Genotype	Count	%	VNTR Genotype	Count	%
-/-	54	16.4	-/-	54	16.4
+/-	149	45.3	-/3	149	45.3
+/+	126	38.3	1/3	4	1.2
			2/3	5	1.5
			3/3	117	35.5

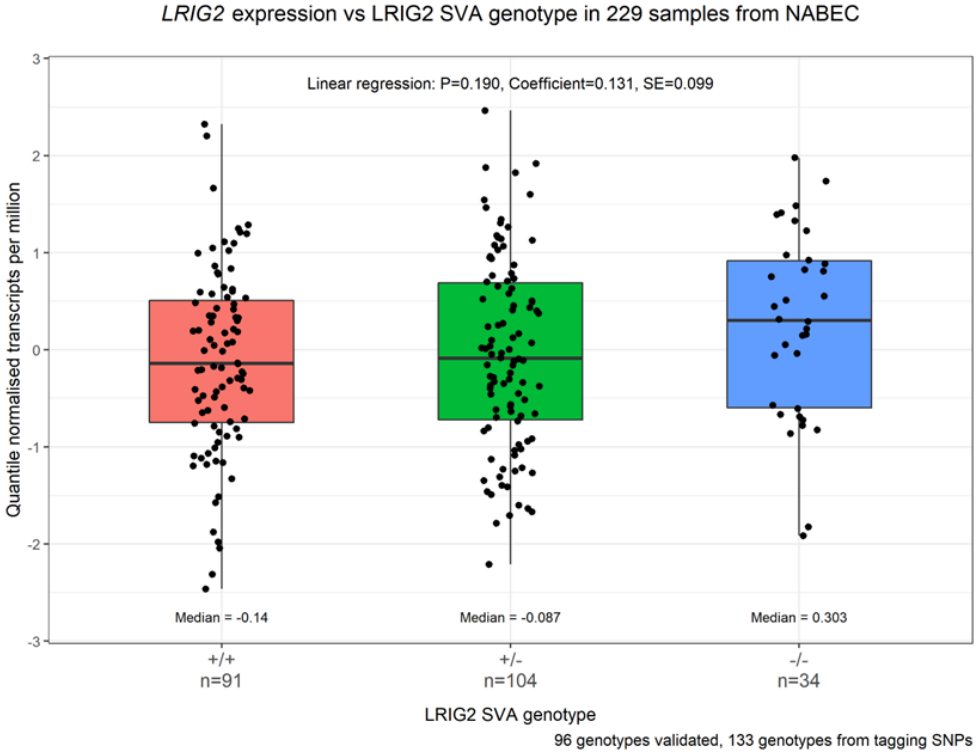
Table 3.3 – Total counts of validated and imputed LRIG2 SVA RIP genotypes in NABEC.

3.2.4. Decreased allele dosage of the LRIG2 SVA is correlated with increased transcription from the LRIG2 locus

Having determined the LRIG2 SVA RIP genotype in the wider NABEC cohort, this was compared to gene expression of *LRIG2* and *LRIG2-DT*. Frontal cortex RNA-seq data were available for 229 of the 329 individuals in NABEC with validated or imputed LRIG2 SVA genotypes. Expression values for *LRIG2* and *LRIG2-DT* were extracted, expressed as quantile normalised transcripts per kilobase million (TPM). These expression data were stratified by LRIG2 SVA genotype, producing a group of 91 individuals with the reference genotype +/+, 104 that were +/−, and 34 that were −/− for the LRIG2 SVA. Compared to individuals with the LRIG2 SVA genotype +/+, individuals with the genotype +/− displayed a median expression of *LRIG2* that was 0.390 standard deviations (SDs) higher and those with LRIG2 SVA −/− were 0.443 SDs higher in the quantile normalised data (**Figure 3.6a**). Similarly, median expression of *LRIG2-DT* was 0.228 SDs higher in LRIG2 SVA −/− individuals compared to those of genotype +/+ (**Figure 3.6b**). Notably, the relationship appeared to be non-linear for *LRIG2-DT* as the LRIG2 SVA genotype +/− exhibited the lowest levels of expression (0.196 SDs lower than genotype +/+ and 0.451 SDs lower than genotype −/−). A simple linear regression was used to assess whether the LRIG2 SVA was an expression quantitative trait locus (eQTL) in neurologically normal frontal cortex, i.e., whether allele dosage of the LRIG2 SVA correlated with differential expression of *LRIG2* or *LRIG2-DT* in the NABEC RNA-seq data. The linear model included the known covariates gender, age, ethnicity, RNA integrity number and originating brain bank (**Section 2.2.1.3**). Alpha significance level (the threshold below which a statistical test's P value must fall for the null hypothesis to be rejected) was set at 2.94E-3 using

Bonferroni correction for multiple comparisons (**Section 2.2.1.3**). Although it was observed that LRIG2 SVA RIP genotype was not significantly associated with expression of *LRIG2* or *LRIG2-DT* (**Figure 3.6a & b**, P = 0.190 and P = 0.477, respectively), the model indicated a negative trend between LRIG2 SVA allele dosage and expression of both transcripts (**Figure 3.6**, positive coefficient values as allele dosage decreases), particularly for *LRIG2*.

a)


LRIG2 expression vs LRIG2 SVA genotype in 229 samples from NABEC
Linear regression: P=0.190, Coefficient=0.131, SE=0.099
Quantile normalised transcripts per million
3
2
1
0
-1
-2
-3
Median = -0.14
Median = -0.087
Median = 0.303
+/+
n=91
+/-
n=104
-/-
n=34
LRIG2 SVA genotype
96 genotypes validated, 133 genotypes from tagging SNPs

b)

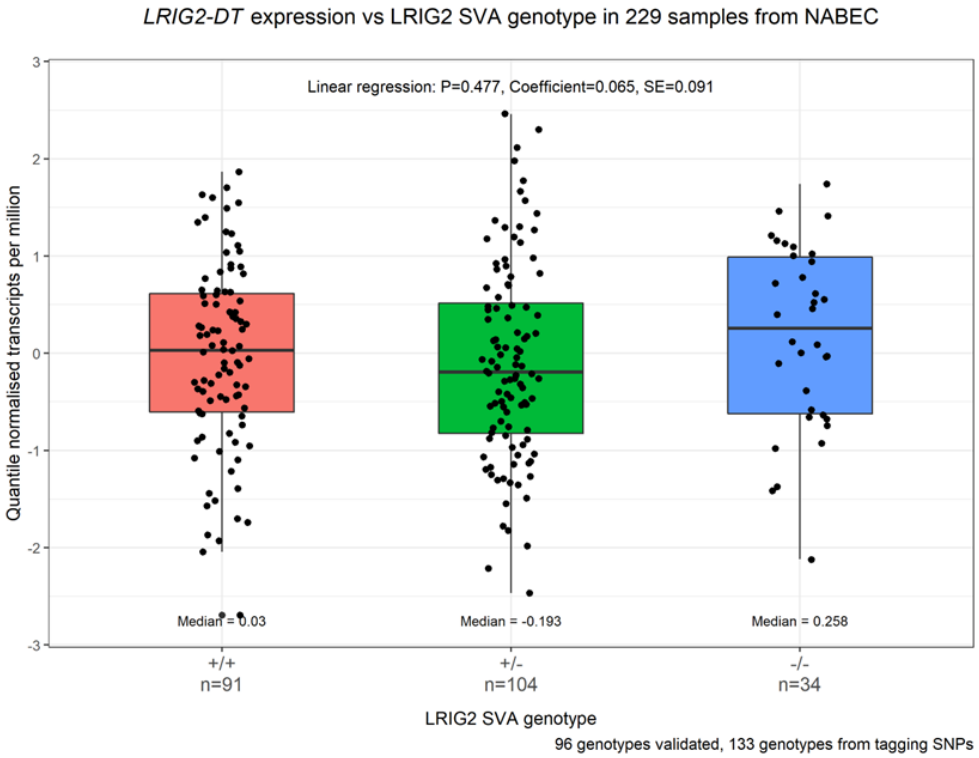
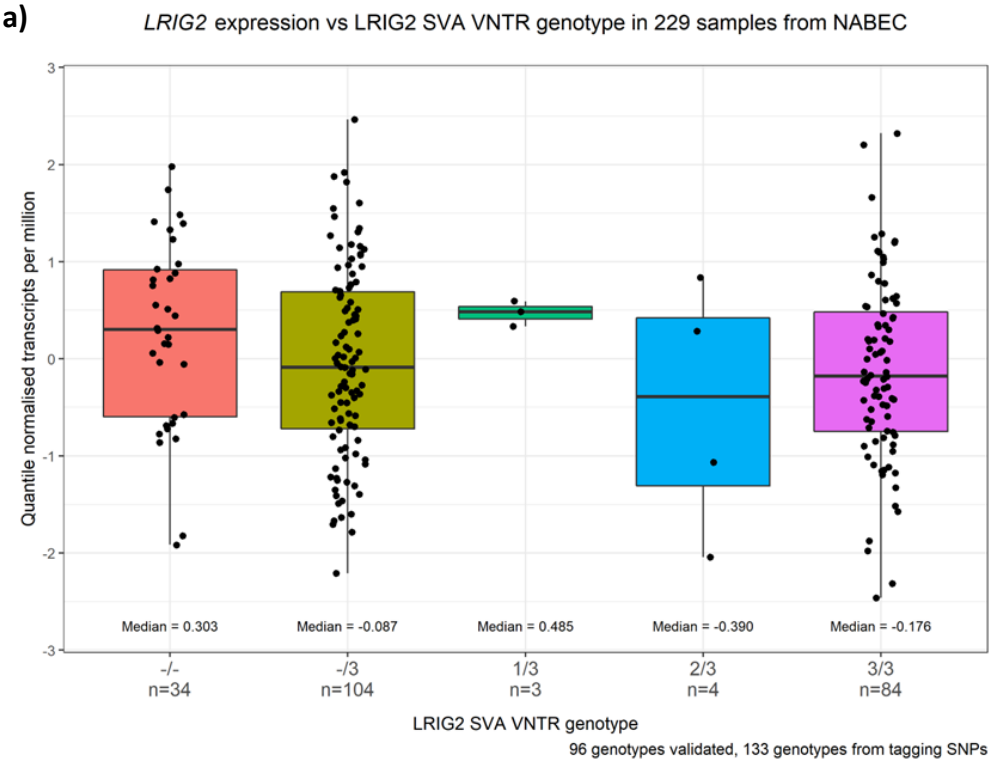

LRIG2-DT expression vs LRIG2 SVA genotype in 229 samples from NABEC
Linear regression: P=0.477, Coefficient=0.065, SE=0.091
Quantile normalised transcripts per million
3
2
1
0
-1
-2
-3
Median = 0.03
Median = -0.193
Median = 0.258
+/+
n=91
+/-
n=104
-/-
n=34
LRIG2 SVA genotype
96 genotypes validated, 133 genotypes from tagging SNPs

Figure 3.6 - LRIG2 SVA RIP genotype versus frontal cortex total RNA-seq data for *LRIG2* and *LRIG2-DT* (ENSG00000198799.12 and ENSG00000238198.2, respectively) in 229 NABEC individuals. 96 genotypes were PCR validated and 133 were imputed for a total of 229 genotypes. **a)** *LRIG2* expression. **b)** *LRIG2-DT* expression. RNA-seq data expressed as quantile normalised transcripts per kilobase million (TPM). Standard deviations from the mean of the normalised data are displayed on the y-axis. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

The NABEC RNA-seq dataset described here can be further broken down by LRIG2 SVA VNTR genotype as determined by proxy SNPs (**Table 3.3**) – indeed, these VNTR-specific proxy SNPs were used to determine the LRIG2 SVA RIP genotype originally. Doing so in the 229 samples with RNA-seq data available produces two additional genotypes for the LRIG2 SVA VNTR allele combinations 1/3 and 2/3, containing 3 and 4 individuals, respectively. These VNTR genotypes were considered in order of increasing allele dosage for VNTR repeat length (**Figure 3.7**, left to right on x-axis). As with the LRIG2 SVA RIP genotypes, the VNTR genotypes were compared to expression of *LRIG2* and *LRIG2-DT* in linear models that included donor and sample covariates (**Section 2.2.1.3**). The coefficients produced by models indicated that there was essentially no linear relationship between increasing VNTR repeat length dosage and expression of *LRIG2* (coefficient = -0.05, P = 0.22) or *LRIG2-DT* (coefficient = -0.02, P = 0.60) (**Figure 3.7a** & **b**, respectively). Moreover, post-hoc ANOVA indicated no significant differences between VNTR dosage groups when considered non-linearly (*LRIG2* P = 0.35, *LRIG2-DT* P = 0.34).

a)


LRIG2 expression vs LRIG2 SVA VNTR genotype in 229 samples from NABEC
Quantile normalised transcripts per million
3
2
1
0
-1
-2
-3
Median = 0.303
Median = -0.087
Median = 0.485
Median = -0.390
Median = -0.176
-/-
n=34
-/3
n=104
1/3
n=3
2/3
n=4
3/3
n=84
LRIG2 SVA VNTR genotype
96 genotypes validated, 133 genotypes from tagging SNPs

b)

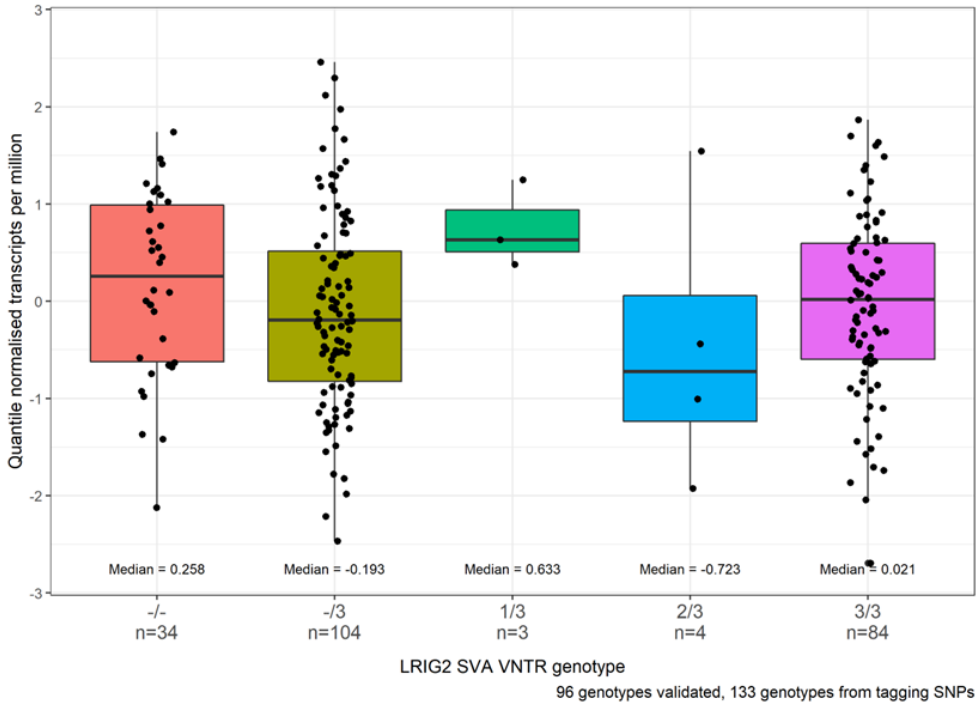

LRIG2-DT expression vs LRIG2 SVA VNTR genotype in 229 samples from NABEC
Quantile normalised transcripts per million
3
2
1
0
-1
-2
-3
Median = 0.258
Median = -0.193
Median = 0.633
Median = -0.723
Median = 0.021
-/-
n=34
-/3
n=104
1/3
n=3
2/3
n=4
3/3
n=84
LRIG2 SVA VNTR genotype
96 genotypes validated, 133 genotypes from tagging SNPs

Figure 3.7 - LRIG2 SVA VNTR genotype versus frontal cortex total RNA-seq data for *LRIG2* and *LRIG2-DT* (ENSG00000198799.12 and ENSG00000238198.2, respectively) in 229 NABEC individuals. 96 genotypes were PCR validated and 133 were imputed for a total of 229 genotypes. **a)** *LRIG2* expression. **b)** *LRIG2-DT* expression. RNA-seq data expressed as quantile normalised transcripts per kilobase million (TPM). Standard deviations from the mean of the normalised data are displayed on the y-axis. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

3.2.5. Decreased LRIG2 SVA allele dosage is associated with decreased methylation of the nearest 450K methylation probe, cg23932873

SVAs have been previously described as mobile CpG islands and DNA methylation is known to spread along adjacent DNA [16,18]. It is also known that KRAB-ZFP-mediated silencing of TEs can influence nearby gene expression through changes such as induction of hypermethylation [139, 149, 150]. Potential associations between LRIG2 SVA RIP genotype and methylation at the locus were therefore investigated. Of the 329 NABEC individuals with validated or imputed LRIG2 SVA genotypes, frontal cortex 450K CpG methylation data were available for 165. Methylation data were again stratified on the basis of LRIG2 SVA RIP genotype, resulting in 66 individuals of the SVA genotype +/+, 78 that were +/−, and 21 that were −/−. Publicly available ENCODE Methylation 450K Bead Array data list 15 CpG methylation probes in a 20 kb window around SVA LRIG2, 13 of which are within 3 kb of the *LRIG2* promotor region (probes listed in **Section 2.2.1.3**). As with the RNA-seq data described previously, linear regression was used to determine whether the SVA was a methylation QTL (mQTL) for these 15 probes. This linear regression model

included the known covariates gender, age, ethnicity and originating brain bank (**Section 2.2.1.3**). When methylation levels of these probes were correlated against LRIG2 SVA genotype only the CpG dinucleotide probe nearest to the SVA, cg23932873, was found to pass the Bonferroni-adjusted alpha level of 2.94E-3 (probe displayed in **Figure 3.8a**). Compared to individuals with the reference LRIG2 SVA genotype +/+, the median proportion of cg23932873 that was methylated was found to be 0.028 (2.8%) lower in those with the genotype +/− and 0.055 (5.1%) lower in those with genotype −/− (**Figure 3.8b**). The linear regression model yielded a *p* value of 5.1E-4 and coefficient of -0.022, indicating that there was a significant association between the decreasing SVA allele dosage and decreasing methylation of cg23932873. In other words, the LRIG2 SVA is a significant mQTL for a CpG dinucleotide at the *LRIG2* locus.

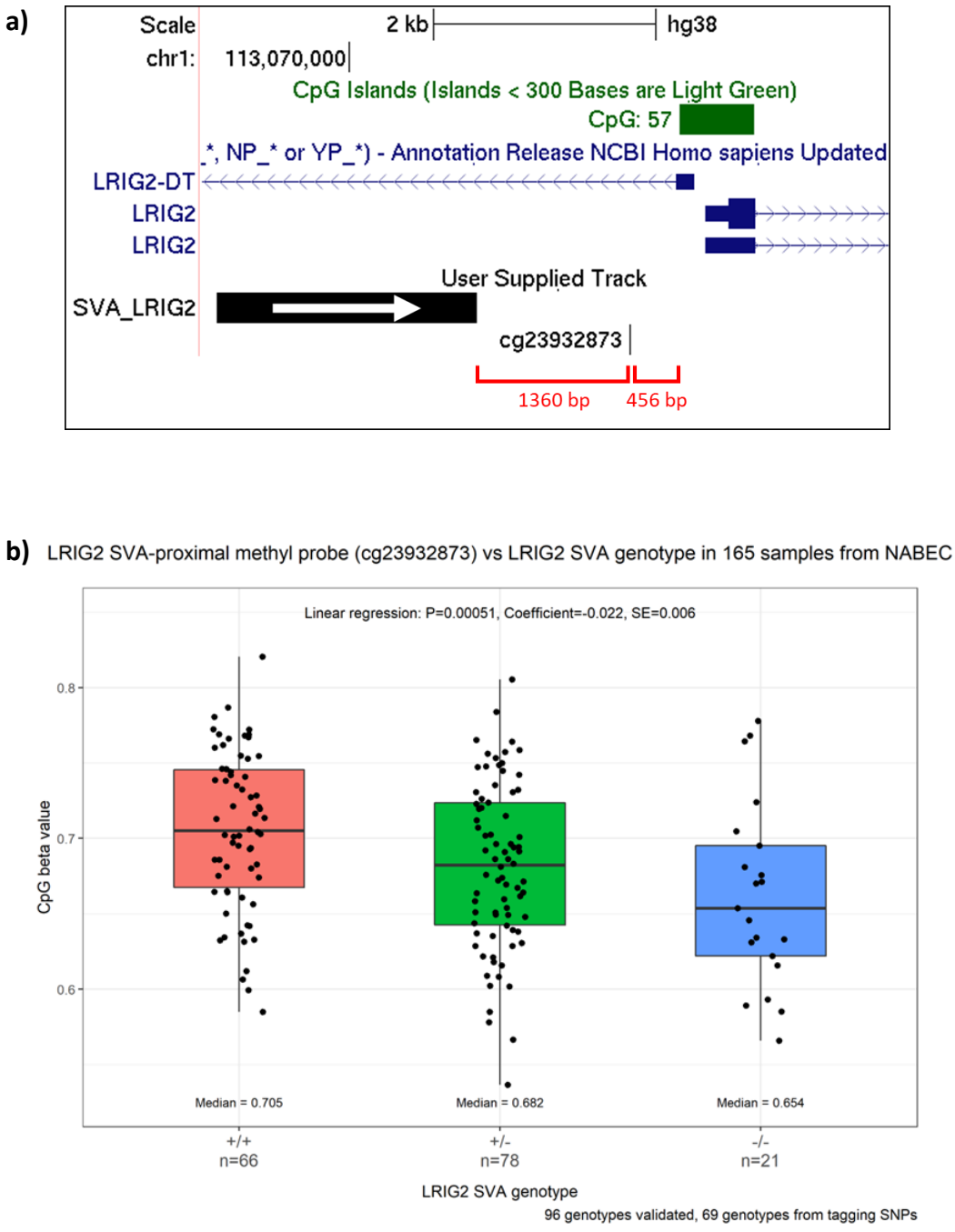


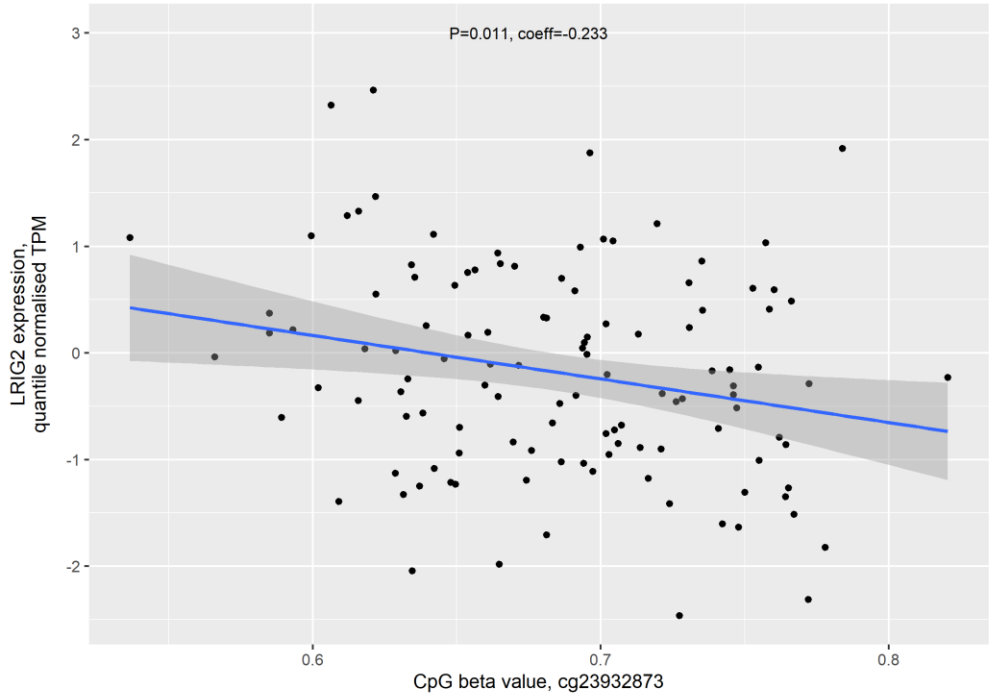
Figure 3.8 - LRIG2 SVA RIP genotype versus CpG methylation data at LRIG2 locus. **a)** The LRIG2 promoter region as displayed on UCSC genome browser hg38, with two validated LRIG2 isoforms and first exon of LRIG2-DT shown by the RefSeq genes curated subset. The position of the LRIG2 SVA is displayed in black. The position of cg23932873, the CpG probe closest to the LRIG2 SVA, is shown in

black and the LRIG2 promoter-associated CpG island is shown in green. The distances in base pairs from cg23932873 to LRIG2 SVA and the CpG island are shown in red. 5' to 3' orientation of the LRIG2 SVA is indicated by a white arrow. **b)** Frontal cortex CpG 450K methylation data for probe cg23932873 in 165 NABEC individuals grouped by LRIG2 SVA RIP genotype. 96 genotypes were PCR validated and 69 were imputed for a total of 165 genotypes. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

3.2.6. Decreased expression of *LRIG2* is weakly correlated with increased methylation of cg23932873

DNA hypermethylation is known to generally repress gene expression through mechanisms including recruitment of proteins that confer repressive histone modifications and impairment of transcription factor binding [210]. Therefore, having observed that increased LRIG2 SVA allele dosage is associated with both decreased transcription from the *LRIG2* locus and increased methylation of the CpG probe closest to the SVA, it was hypothesised that the two may be inversely correlated. When NABEC samples with RNA-seq data available were overlaid with those with available CpG methylation data and outliers were removed, 118 samples remained for the comparison involving *LRIG2* and 119 remained for *LRIG-DT*. Normalised TPM values for *LRIG2* or *LRIG2-DT* were plotted against CpG beta values for cg23932873, and a Pearson correlation coefficient was determined. A weak yet significant inverse correlation was observed between *LRIG2* expression and cg23932873 (**Figure 3.9a**), and no correlation was observed between *LRIG2-DT* expression and cg23932873 (**Figure 3.9b**).

a) *LRIG2* expression vs methyl 450k CpG probe cg23932873 in 118 samples from NABEC



b) *LRIG2-DT* expression vs methyl 450k CpG probe cg23932873 in 118 samples from NABEC

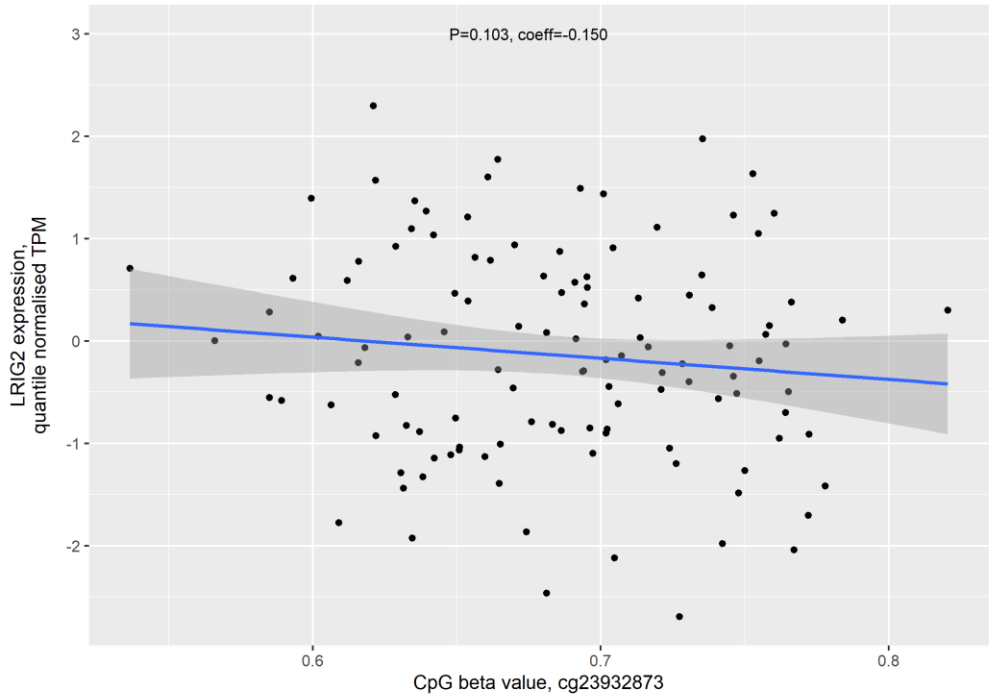


Figure 3.9 – Expression from the *LRIG2* promoter locus versus methylation of CpG 450K probe cg23932873 in NABEC frontal cortex samples. **a)** *LRIG2* (ENSG00000198799.12) in 118 individuals. **b)** *LRIG2-DT* (ENSG00000238198.2) in 118 individuals. Blue line indicates trend line; dark grey zone indicates 95% confidence interval. Displayed are Pearson correlation coefficients and corresponding *p* values.

3.2.7. Multiple gRNAs were tested for CRISPR-Cas9-medited deletion of the LRIG2 SVA

Thus far it had been observed that the LRIG2 SVA allele dosage was indicative of lower *LRIG2* expression and was significantly associated with increased proximal DNA methylation. To probe these differences further the SVA was deleted in a cell line using CRISPR-Cas9-mediated double-stranded breaks (DSBs), and these expression and methylation characteristics were compared to cells retaining the SVA in an otherwise genetically identical background. The established cell line SH-SY5Y was selected as a starting point, as it was previously demonstrated that these cells are homozygous for presence of the SVA (**Figure 3.2b**) and are karyotypically normal at the locus [211]. The strategy employed relied upon two different gRNA molecules which would each associate with Cas9 enzymes and guide them to sites upstream and downstream of the LRIG2 SVA, where DSBs would be induced and the SVA excised. As homology-directed repair is largely offline in interphase the majority of DSBs are expected to be repaired by non-homologous end joining (NHEJ), resulting in deletion of the LRIG SVA on one or both chromosomes (**Figure 3.10**).

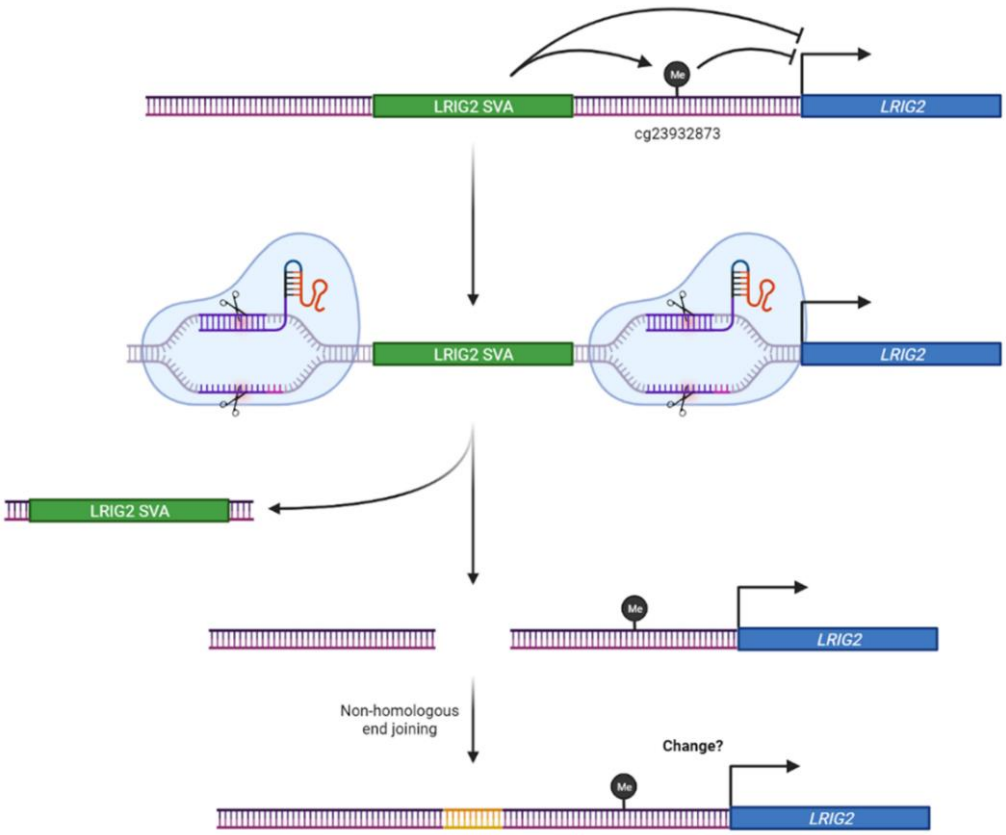


Figure 3.10 – Outline of LRIG2 SVA deletion strategy with CRISPR-Cas9. Putative mechanisms for influence of LRIG2 SVA on *LRIG2* expression via the cg23932873 CpG pictured top. DSBs are introduced 5' and 3' of the SVA by Cas9 enzymes targeted by separate gRNA molecules. After excision the SVA, repair by NHEJ results in permanent deletion.

Details of the gRNA design pipeline are provided in **Section 2.2.12**. Briefly, a shortlist of oligonucleotides was generated using software developed by the Zhang Lab at MIT (http://crispr.mit.edu/) which searches for unique ~20 bp genomic sequences with correctly spaced PAM sequences, required for Cas9 binding, in the sequence provided. The top 3 scoring nominated oligonucleotides that targeted 5' and 3' of the LRIG2 SVA, respectively designated #1–3 and #4–6, were taken forward and inserted into the pSpCas9(BB)-2A-GFP plasmid as described in **Section 2.2.12.2**. This plasmid

expresses the gRNA within the correct nucleotide framework for association with Cas9, along with the Cas9 protein itself. At this stage it was noticed that a mistake was made in the design of gRNA #1, and it was discarded. The remaining gRNAs target regions ~300 bp upstream and ~750 bp downstream of the LRIG2 SVA, as shown in **Figure 3.11a**. To assess how a genomic deletion mediated by these gRNAs might affect regulatory elements at the locus, the gRNA-targeted sites were also visualised alongside predicted enhancer and promoter regions from the GeneHancer database (https://www.genecards.org/Guide/GeneCard#enhancers), histone modifications as determined by ChIP-seq of 7 cell lines from the ENCODE database (https://genome.ucsc.edu/ENCODE/index.html), predicted cis-regulatory elements from ENCODE, and the curated list of NCBI RefSeq Functional Elements (https://www.ncbi.nlm.nih.gov/refseq/functionalelements/). In **Figure 3.11a** the GeneHancer database unsurprisingly identifies the region around the *LRIG2* TSS as a promoter (red block) but listed no enhancer activity in the region (grey blocks in this track – none displayed), the ENCODE histone track listed H3K4Me1 overlapping the 3' gRNA sites only in K562 cells (purple peaks), the ENCODE cis-regulation track included a single predicted CTCF binding site overlapping the 5' gRNA sites (light blue block), and the curated list of 'functional elements' from NCBI RefSeq did not contain any regulatory regions at the locus (bottom of image). Altogether, this was interpreted to mean that the flanking regions to be excised along with the LRIG2 SVA contained minimal, if any, validated regulatory elements that might influence nearby gene expression.

The 5'- and 3'-targeting gRNAs were then tested in different combinations to determine the best performing pair for deletion of the LRIG2 SVA. These gRNA-containing CRISPR plasmids were transfected into SH-SY5Y cells alongside a transfection agent-only control using the technique described in **Section 2.2.12.3** (However, clonal isolation steps were not used in this optimisation – Cells were examined at the population level as clonal isolation is time- and labour-intensive, and not necessary for qualitative assessment of efficacy). 48 hours post-transfection genomic DNA was extracted and underwent PCR with the 'LRIG2 SVA + Flanks' primers followed by gel electrophoresis to assess SVA deletion efficacy. The 'LRIG2 SVA + Flanks' PCR products associated with LRIG2 SVA deletion (herein ΔLRIG2 SVA) are 0.8 – 1 kb in size, corresponding to the CRISPR-mediated removal of ~3.4 kb of central sequence (**Table 2.5** for PCR conditions). From these PCRs it was observed that the 4.3 kb bands corresponding to the unedited PCR product were of similar intensity, suggesting approximately equal DNA input to the PCR, while the amplicon corresponding to the ΔLRIG2 SVA allele was markedly brighter in cells which received gRNAs #3 and #4 (**Figure 3.11b**). This indicated that these cells had greater quantities of the modified ΔSVA locus, and that this combination of gRNAs was the most

efficient in deletion of the LRIG2 SVA. gRNAs #3 and #4 were therefore taken forward for deletion of the LRIG2 SVA.

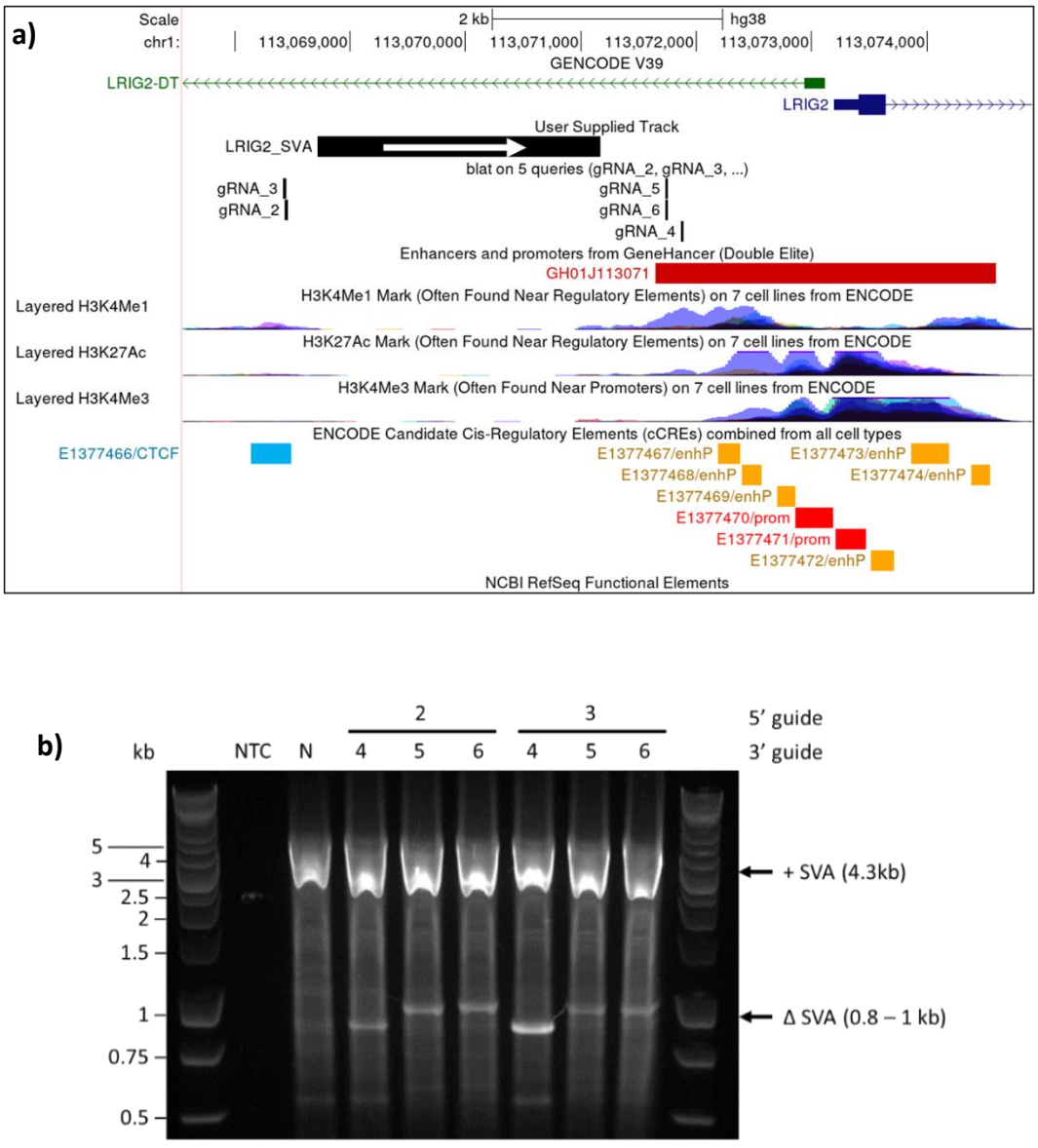


Figure 3.11 – Multiple combinations of gRNAs were tested for deletion of the LRIG2 SVA. **a)** The LRIG2 SVA and *LRIG2* promoter as displayed on UCSC genome browser hg38 with the binding sites of gRNAs #2-6 shown. Also displayed are predicted enhancer and promoter regions from the GeneHancer database (red blocks denote promoters, grey blocks denote enhancers), histone modifications as determined by ChIP-seq of 7 cell lines from the ENCODE database (blue/purple peaks), predicted *cis*-

regulatory elements from ENCODE (light blue indicates CTCF binding, orange indicates an enhancer-proximal region, red indicates a promoter), and the curated list of NCBI RefSeq Functional Elements (no hits in this genomic window). A white arrow indicates 5' to 3' orientation of the LRIG2 SVA. **b)** Following transfection with a combination of a 5'- and 3'-targeting gRNA, DNA was extracted and amplified with the 'LRIG2 SVA + Flanks' primer pair. 10 µg gDNA was used as input with KOD Hot Start Polymerase with 40 cycles. 12 µl was loaded onto a 1% agarose gel and run for 90 min at 140 V. NTC = No template control. N = Negative control (Lipofectamine added with no CRISPR plasmid).

3.2.8. Optimisation of *LRIG2* qPCR and cg23932873 pyrosequencing

The objective of deleting the LRIG2 SVA via CRISPR-Cas9 was to replicate the observations made regarding expression and methylation trends correlated with LRIG2 SVA genotype in the NABEC cohort. It was therefore necessary to perform qPCR and pyrosequencing in the ΔLRIG2 SVA SH-SY5Y cell lines generated here – approaches which require prior optimisation.

3.2.8.1. *LRIG2* qPCR optimisation

For qPCR, primers that satisfy MIQE requirements [198] by avoiding known SNPs and off-target binding were ordered from the 'KiCqStart' range by Sigma Aldrich. Two primers pairs for each of *LRIG2* and *ACTB* (a standard housekeeping gene used for normalisation of qPCR) were ordered – no primers targeting *LRIG2-DT* were available from Sigma Aldrich. The *LRIG2* and *ACTB* primers were tested for amplification efficiency, a key criterion for publication according to MIQE guidelines. In short, a dilution series of cDNA underwent qPCR with the trialled primers and the difference in Cq with each dilution was used to calculate primer efficiency – see **Section 2.2.7.2**

for details. Additionally, visual appraisal of amplification and melt curves allowed assessment of qPCR efficiency and specificity. These qPCR reactions were carried out using a standard set of reaction conditions provided in **Table 2.7** with a dilution series starting at 22.5 ng/µl cDNA from wildtype SH-SY5Y cells (**Section 2.2.7.2**). For the *LRIG2* primers, a 5-fold dilution series was performed to enable enough datapoints to be collected before amplification-associated fluorescence fell below the detection threshold. It was observed that pair #1 failed to meet the 90% minimum threshold for efficiency set out by MIQE while primer pair #2 performed considerably better, with an efficiency approaching 100% (**Figure 3.12a**). Furthermore, the melt curve for *LRIG2* #2 had fewer absorbance peaks away from the main peak than set #1, suggesting a more specific amplification. A 10-fold dilution series was performed for the *ACTB* primers, since housekeeping genes are expected to be constitutively expressed at high levels. As with the *LRIG2* primers only the *ACTB* #2 primer set passed the requirement for 90% amplification efficiency, with these primers showing less variability in their melt curve than the #1 set (**Figure 3.12b**). The *LRIG2* #2 and *ACTB* #2 primer sets, herein simply *LRIG2* and *ACTB* qPCR primers, were taken forward for qPCR analysis of *LRIG2* expression.

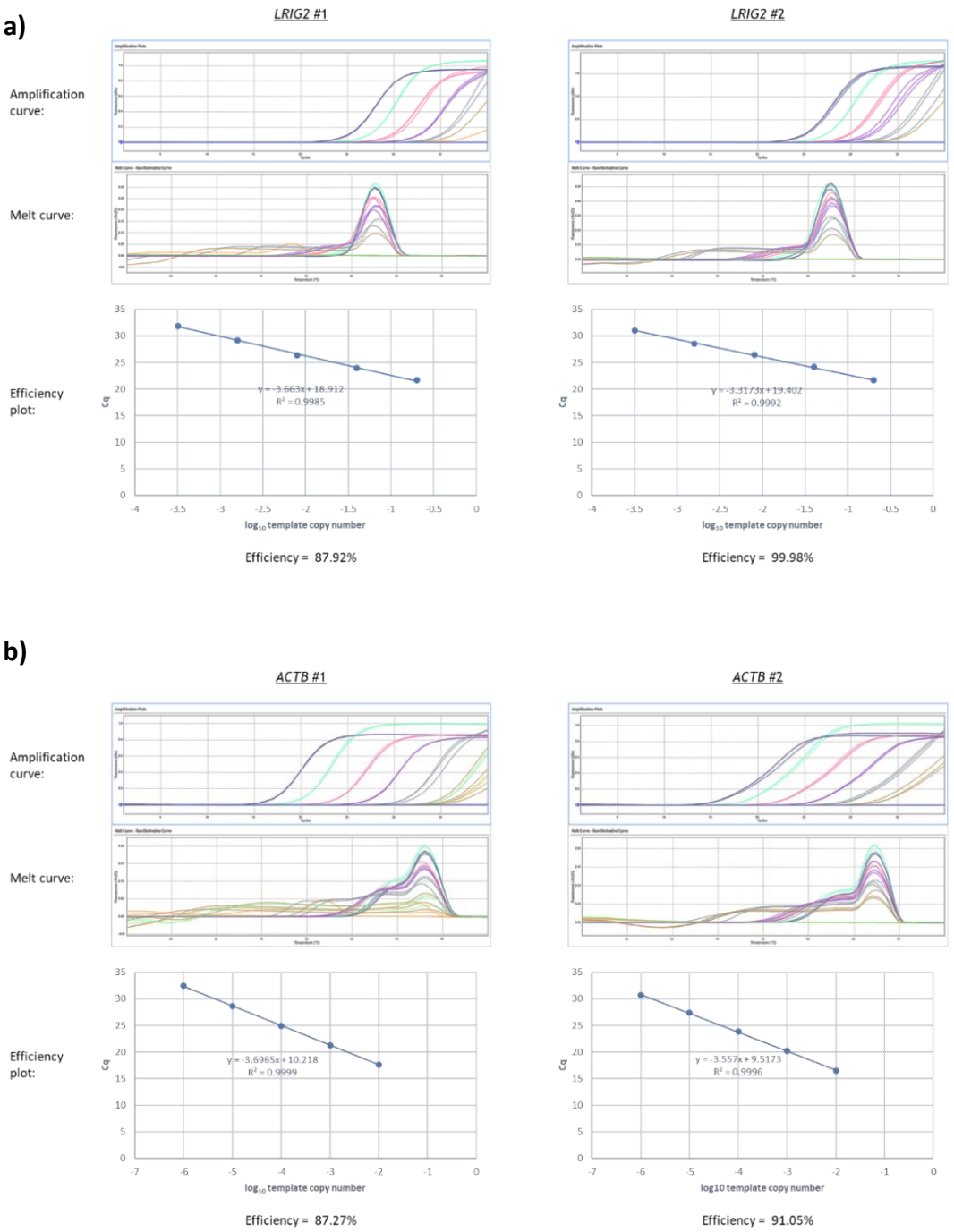


Figure 3.12 – qPCR dilution series and efficiency plots for tested primers. 22.5 ng/µl SH-SY5Y cDNA was serially diluted and underwent PCR with the Promega GoTaq qPCR Master Mix, 1 mM of each primer, and a combined annealing and extension step at 60 °C – see **Section 2.2.7.1** for cycling conditions. Amplification and melt curves are displayed, with colours corresponding to sequential dilutions. Efficiency plots are shown, featuring Cq versus corresponding $\log_{10}$ values of template copy number input (Dilution series starting from copy number of '1'). **a)** *LRIG2* #1 and #2 primer sets

amplifying a 5-fold dilution series of cDNA. **b)** *ACTB* #1 and #2 primer sets amplifying a 10-fold dilution series of cDNA.

3.2.8.2. cg23932873 pyrosequencing optimisation

Prior to pyrosequencing of a CpG dinucleotide of interest, bisulphite converted DNA must first have the region containing the CpG amplified (see **Section 2.2.8** for full details on pyrosequencing pipeline). Using PyroMark Assay Design Software 2.0.2 (QIAGEN), a 'cg23932873 Amplification' primer pair with a 5'-biotin-labelled forward primer were designed, along with a 'cg23932873 Sequencing' oligonucleotide (**Table 2.5**). As with standard PCRs, amplification of this region with the Pyromark PCR Kit (from QIAGEN) requires optimisation to find reaction conditions that ensure efficient and specific amplification. 500 ng of wildtype SH-SY5Y gDNA was bisulphite converted in 10 replicates which were then pooled (**Section 2.2.8**). Aliquots of this were amplified with the 'cg23932873 Amplification' primers using the Pyromark kit at a range of annealing temperatures (conditions **Table 2.8**), in duplicate – with and without QIAGEN's 'Q solution', which is purported to potentially improve efficiency of difficult PCR templates. It was found that without Q solution the PCR worked efficiently at all temperatures tested, but the addition of Q solution drastically reduced efficiency (

Figure 3.13a). The 'cg23932873 Amplification' reaction carried out at 55 °C without Q solution was therefore selected for further testing of efficacy. The PCR was repeated with a 55 °C annealing temperature along with a non-template control, and the two went through the streptavidin pulldown purification described in **Section 2.2.8**. The streptavidin-purified non-template control and 'cg23932873 Amplification' amplicon then underwent pyrosequencing with the 'cg23932873 Sequencing' oligonucleotide alongside a water negative control. As the sequencing progresses and reaches the C/G base of interest (depending on whether top or bottom strand was sequenced) there will be an attempt to incorporate this nucleotide followed by an A or T, which is the result of bisulphite conversion of an unmethylated cytosine residue. The light signal emitted in each case is proportional to the amount of nucleotide incorporated, and can be used to calculate a percentage of cytosine methylation. In this optimisation run both the non-template control and 'blank' inputs exhibited no

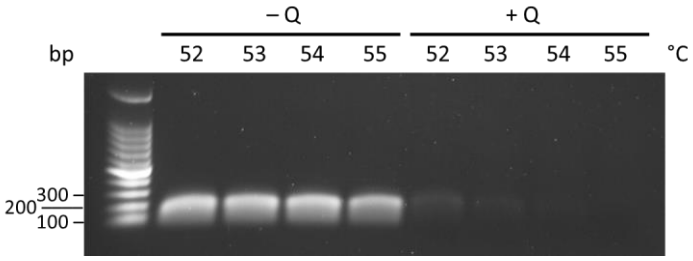
light emittance while the predicted nucleotide incorporation was observed for the 'cg23932873 Amplification' amplicon, indicating the signal observed during sequencing was specific and contaminant-free (

Figure 3.13b). In this test run, it was found that 85% of C/G residues at the target position in the SH-SY5Y gDNA were not converted to T/A (comparison of peaks within blue region of middle pyrogram in

Figure 3.13b), indicating that this CpG was mostly methylated in this sample. Overall bisulphite conversion efficiency may be qualitatively assessed using this pyrogram by comparing the G residue highlighted in yellow, corresponding to a non-CpG cytosine in the opposite strand of the Pyromark PCR product, to the preceding A residue incorporation. This C residue at the yellow-highlighted position is expected to be unmethylated and therefore completely converted to T, resulting in the incorporation of A instead of G in the sequenced (opposite) DNA strand.

Figure 3.13b indicates that this is indeed the case, suggesting high overall bisulphite conversion efficiency.

a)



b)

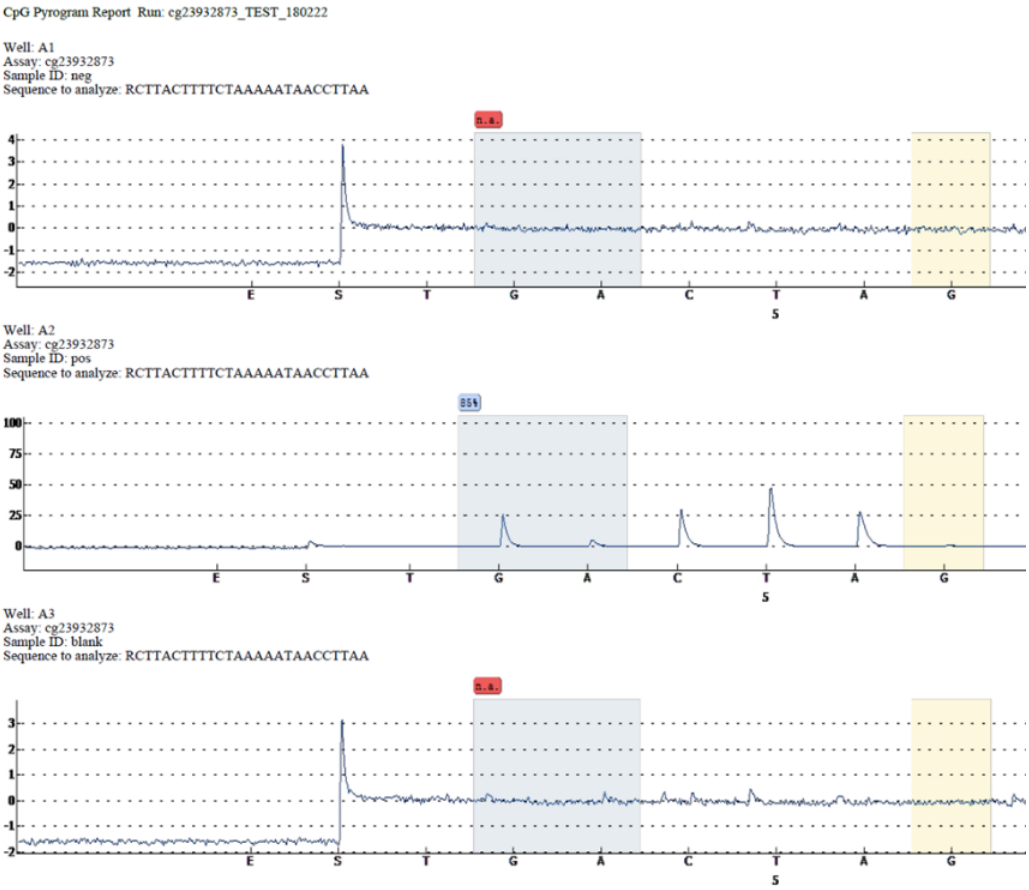


Figure 3.13 – Optimisation of Pyromark PCR and pyrosequencing. **a)** 50 ng of bisulphite converted DNA underwent PCR using the 'cg23932873 Amplification' primers, featuring a 5'-biotinylated forward primer, using the Pyromark PCR kit at a range of annealing temperatures (displayed) (conditions **Table 2.8**). Parallel temperature gradients were run with and without Q solution. **b)** Pyrogram for sequencing of cg23932873. A 'cg23932873 Amplification' Pyromark PCR product underwent streptavidin pulldown

followed by pyrosequencing with the 'cg23932873 Sequencing' oligonucleotide (middle pyrogram) (**Section 2.2.8**). This was sequenced alongside a non-template PCR control (top) and a water 'blank' (bottom). Percentage of light emitted at the base of interest when a C/G is incorporated versus A/T incorporation shown in blue – 'n.a.' shown in red when no light detected. Yellow box indicates 'bisulphite control' residues which are expected to be fully converted to an A/T pair.

3.2.9. Deletion of the LRIG2 SVA results in a modest increase in *LRIG2* expression and decrease in cg23932873 methylation in SH-SY5Y

With gRNAs chosen for the Cas9-directed deletion of the LRIG2 SVA and measurement of expression and methylation at the *LRIG2* promoter region optimised, clonal populations of ΔLRIG2 SVA-edited SH-SY5Y cells could be generated and characterised. Using the pipeline described in **Section 2.2.12.3**, wildtype SH-SY5Y cells were transfected with the pSpCas9(BB)-2A-GFP plasmids containing the LRIG2 SVA-targeting gRNAs #3 and #4, clonal populations were isolated and genotyped, and cells in which the SVA had been deleted were propagated. Additionally, several 'unedited' populations of cells that had gone through this selection process but had retained the wildtype +/+ genotype for the LRIG2 SVA were taken forward as control cell lines. In total 4 clonal populations with biallelic deletions (genotype Δ/Δ) and 3 clonal populations with monoallelic deletions (+/Δ) for the LRIG2 SVA were produced in the SH-SY5Y line along with 5 unedited lines (+/+) (**Figure 3.14**). In all but one case the ΔLRIG2 SVA amplicon was the same size, indicating near-identical NHEJ repair outcomes, but for 'Biallelic edit' #3 a second larger PCR product is visible. This corresponds to a smaller, incomplete deletion in which some of the region to be deleted has instead been retained, and may be the result of an aborted attempt at

homologous recombination-mediated repair. Based on the gel image it can be estimated that 200–300 bp of genomic sequence was retained. Since the LRIG2 SVA-targeting gRNAs induce DSBs ~300 bp upstream and ~750 bp downstream of the SVA, it was determined that it was unlikely that any meaningful amount of LRIG2 SVA sequence was retained on this partially edited allele.

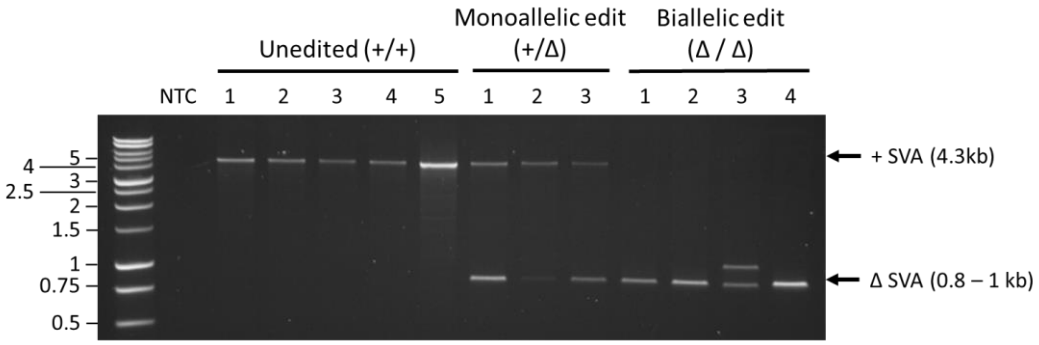
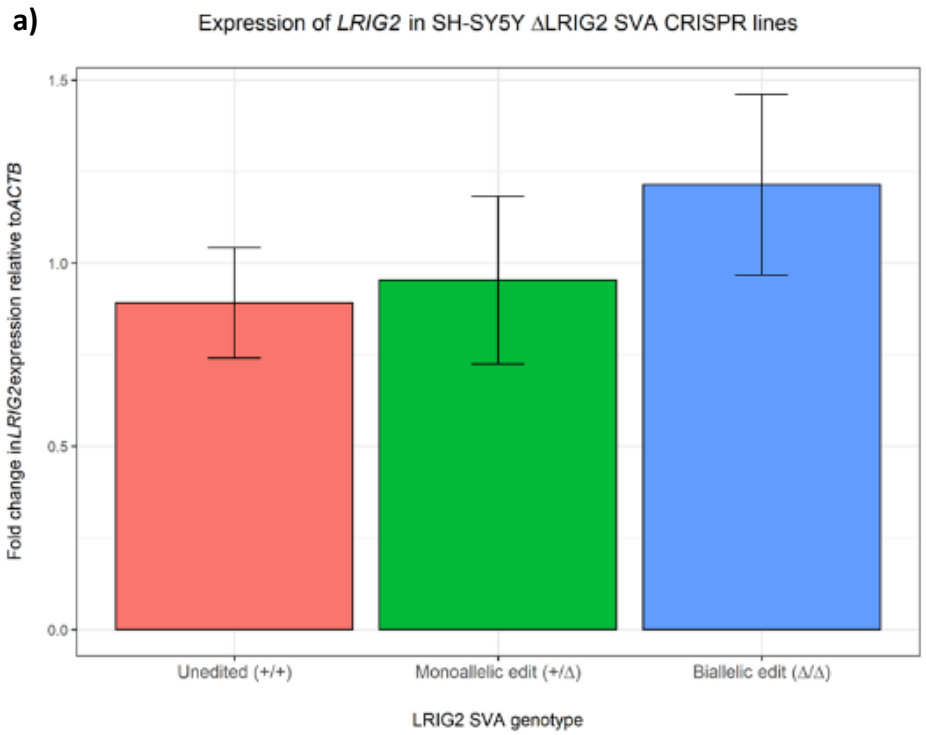


Figure 3.14 – LRIG2 SVA genotypes of CRISPR-edited clonal SH-SY5Y populations. After propagation of selected clonal lines, gDNA was harvested from cells grown in T75 flasks and underwent PCR with the 'LRIG2 SVA + Flanks' primers. KOD Hot Start Polymerase was used to amplify 5 µg gDNA input with 57.5 °C annealing temperature and 32 cycles. 6 µl of samples was loaded on 1% agarose and run at 100 V for 90 min.

Expression and methylation patterns at the *LRIG2* promoter locus were then examined in these clonal SH-SY5Y lines. To ensure consistency in growth conditions at the point of sample collection, all were seeded at 600,000 cells per well in 6-well plates and incubated for 48 hours in the same batch of growth media (**Section 2.2.11**). Cells were then trypsinised, split into two aliquots and pelleted. From the first aliquot of each cell pellet, RNA was extracted and converted to cDNA according to protocols described in **Sections 2.2.3.2 and 2.2.7.1**. *LRIG2* expression was examined in via qPCR using the selected *LRIG2* and *ACTB* qPCR primer sets (**Section 2.2.7.1**, primer details **Table 2.5),** with fold change

in gene expression relative to a randomly chosen 'unedited' SH-SY5Y line calculated using the ΔΔCT method (**Section 2.2.7.3**). There was no amplification detected in a parallel non-template control for the amplification (not shown). It was observed that as allele dosage of LRIG2 SVA decreased, expression of *LRIG2* increased; relative to the mean *LRIG2* expression in unedited SH-SY5Y lines (SVA +/+), a single LRIG2 SVA deletion (SVA genotype +/Δ) is associated with a 6.2% increase in gene expression and deletion of both alleles (genotype Δ/Δ) yields a 36.2% increase (**Figure 3.15a**). However, these differences were found to be non-significant when examined in one-way ANOVA (p = 0.104). From the second aliquot of cell pellets gDNA was extracted and bisulphite converted as described in **Sections 2.2.3.1 and 2.2.8**. This DNA underwent Pyromark PCR with the 'cg23932873 Amplification' primers and optimised conditions identified in

Figure 3.13, and the resulting PCR reaction mixture was streptavidin purified and pyrosequenced with the 'cg23932873 Sequencing' oligonucleotide (**Section 2.2.8**, primer details **Table 2.5**). No nucleotide incorporation was observed in a water 'blank' or a non-template Pyromark PCR control (not shown). In the LRIG2 SVA CRISPR-edited SH-SY5Y lines it was found that deletion of one allele of the SVA (genotype +/Δ) produced a 0.8% reduction in the methylation levels of the CpG dinucleotide cg23932873 when compared to the unedited cells (SVA +/+), while deletion of both alleles (SVA Δ/Δ) resulted in a 2.1% decrease (**Figure 3.15b**). A Shapiro-Wilk test determined that these data were non-normal, and so the methylation levels for the three genotypes were assessed by Kruskal-Wallace test; this indicated that they were not significantly different (p = 0.106).



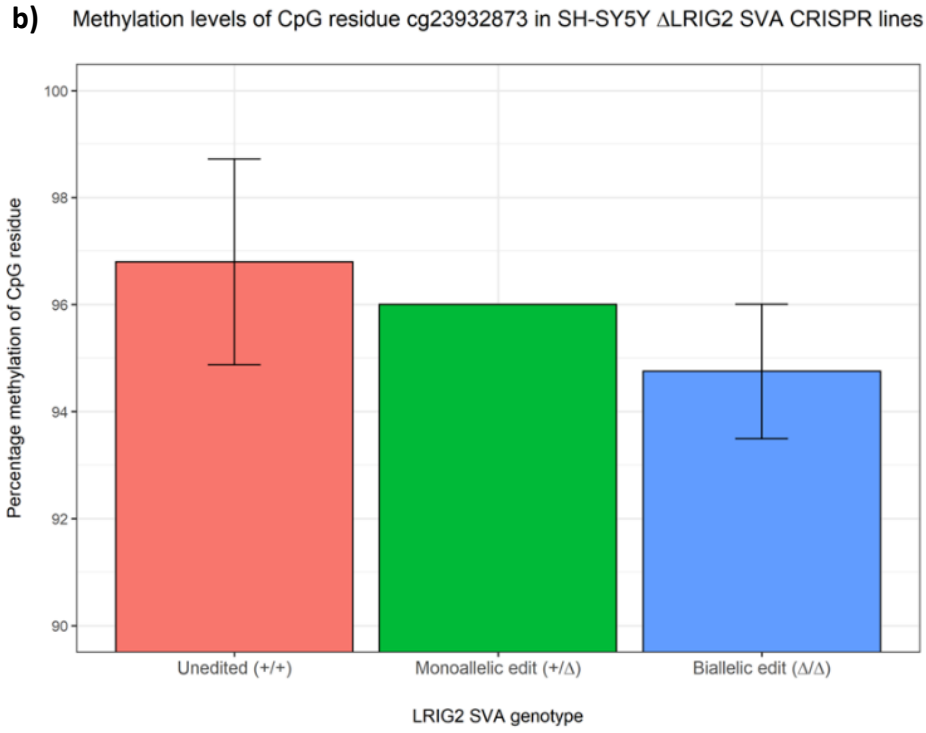


Figure 3.15 – Expression and methylation at the *LRIG2* promoter locus in ΔLRIG2 SVA SH-SY5Y cell lines. ΔLRIG2 SVA SH-SY5Y clones were seeded at 600,00 cells per well and incubated 48 hours. **a)** 5ng cDNA underwent qPCR with *LRIG2* and *ACTB* primer sets in technical triplicate (**Section 3.2.8**) using

the Promega GoScript Reverse Transcription System. Fold change in *LRIG2* expression was calculated relative to a randomly chosen 'unedited' SH-SY5Y line and normalised to *ACTB* expression, using the ΔΔCT method. **b)** 75 ng bisulphite converted gDNA amplified by Pyromark PCR with 'cg23932873 Amplification' primers. These biotinylated amplicons were purified by streptavidin pulldown and pyrosequenced with the 'cg23932873 Sequencing' primer. CpG methylation percentages are taken directly from pyrogram outputs. **a & b)** '+/+' n=5, '+/Δ' n=3, 'Δ/Δ' n=4. Error bars represent standard deviation for each genotype.

3.2.10. *LRIG2* expression and cg23932873 methylation are moderately but non-significantly inversely correlated in ΔLRIG2 SVA SH-SY5Y cell lines

It was then investigated whether the observed changes in *LRIG2* expression and cg23932873 methylation that occur with LRIG2 SVA deletion might be correlated, as this may indicate a functional relationship between the two. The expression and methylation data from **Figure 3.15** were plotted together and, as pyrosequencing data were previously shown to be non-parametric, a Spearman's Rank correlation coefficient was determined. It was observed that while that there was a moderate inverse relationship between *LRIG2* expression and cg23932873 methylation in the CRISPR-edited lines (**Figure 3.16**, trend line and negative rho coefficient), this was not statistically significant.

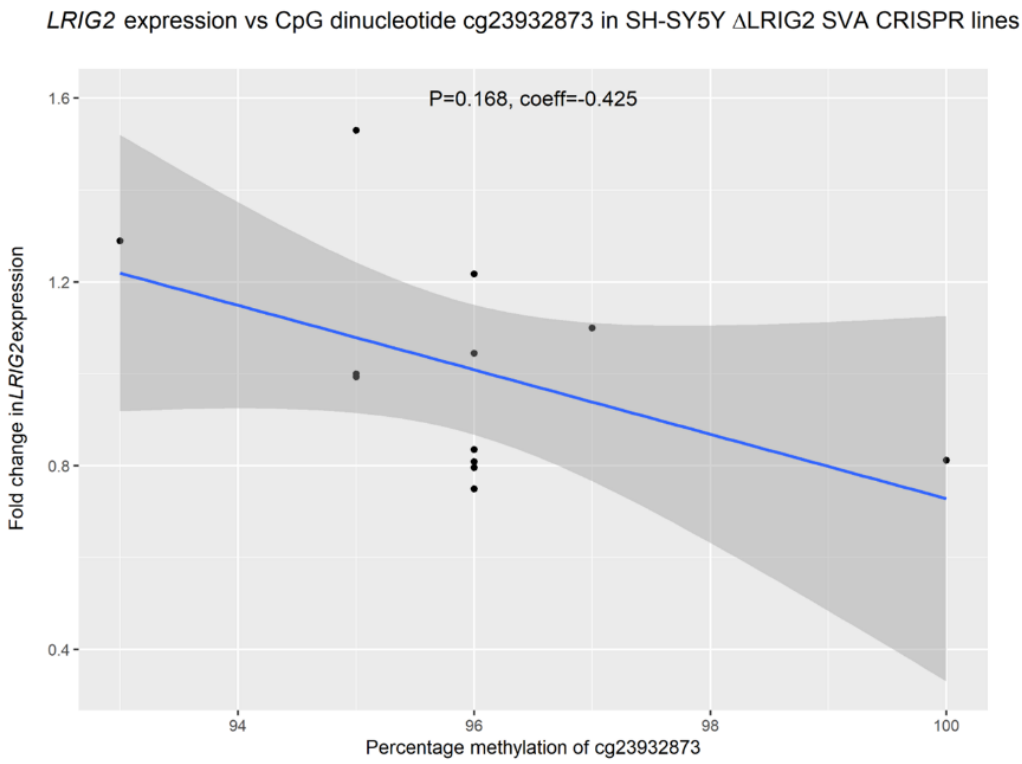


Figure 3.16 - Expression from the *LRIG2* promoter locus versus methylation of CpG cg23932873 in ΔLRIG2 SVA SH-SY5Y cell lines. Blue line indicates trend line; dark grey zone indicates 95% confidence interval. Displayed is Spearman correlation coefficient and corresponding *p* value.

3.3. Discussion

In this chapter the SVA F1 located 2 kb upstream of the *LRIG2* gene, which is a RIP, was studied as a model for how presence or absence of an SVA might influence gene expression, particularly in the context of an insertion upstream of a promoter. It was speculated that the LRIG2 SVA could promote increased expression from the *LRIG2* promoter region 2 kb away by recruiting activating TFs to the locus or via read-through transcription originating at the SVA's internal promoter; equally, it was hypothesised that the LRIG2 SVA could instead lead to transcriptional repression due to targeting by KRAB-ZFPs and spread of heterochromatin at the locus. This disparate regulatory potential is even more stark for an SVA of the F1 subfamily, since the 5' *MAST2* exon 1 transduction that replaces the CT element has been shown to possess promoter activity but increases the CpG content associated with the SVA [179]. To investigate the regulatory influences of the SVA, it was first confirmed that the LRIG2 SVA is a RIP in the NABEC cohort (**Figure 3.5a**) – an immensely useful resource for which DNA samples with corresponding high-throughput phenotypic data were available. Proxy SNPs were identified that were in high LD with the LRIG2 SVA, therefore allowing its RIP genotype to be inferred in NABEC samples for which WGS data was available but not the DNA itself. It was also observed that the LRIG2 SVA central VNTR exhibited length polymorphisms, with a total of 4 identified in NABEC (**Figure 3.5b**). It was not possible to reconstruct sequences for VNTR alleles from Sanger sequencing data for VNTRs besides the smallest allele, preventing direct comparison of DNA sequence and regulatory potential (**Section 3.2.2**). Although stratifying the dataset by these alleles was uninformative as their frequencies were too low (**Figure 3.7**), with alleles 1 and 2 occurring in 1 and 3% of samples, they were

used to build a more accurate picture of the locus in the wider NABEC cohort by producing VNTR allele-specific proxy SNPs. VNTR length is likely an important factor in deciding the exact effect an SVA has at a locus, for example by providing more or less copies of binding motifs for factors such as CTCF [180]. Therefore, future work characterising the LRIG2 SVA would greatly benefit from increased sample sizes in order to robustly characterise the gene expression influences of these rare VNTR alleles.

Nevertheless, by combining VNTR-specific alleles it was possible to build up accurate RIP genotypes for the LRIG2 SVA in NABEC amounting to 329 genotypes (**Table 3.3**). Although not achieving statistical significance in the linear model generated, an inverse correlation between LRIG2 SVA allele dosage and expression of the *LRIG2* transcript was observed (**Figure 3.6**). The relationship between SVA allele dosage and the divergent transcript from the same promoter, *LRIG2-DT*, was more complicated: compared to the reference +/+ LRIG2 SVA genotype, the absence of one SVA at the locus (genotype +/-) was associated with decreased *LRIG2-DT* expression while absence of the SVA on both chromosomes (-/-) was associated with an increase. As with *LRIG2* expression, there was no statistically significant relationship between LRIG2 SVA genotype and *LRIG2-DT* expression. It is worth noting, however, that both SVA genotypes +/- and +/+ exhibited lower *LRIG2-DT* expression that in the absence of the SVA. In other words, presence of at least one LRIG2 SVA was associated with lower divergent transcript expression from the locus than when it was absent. Taken together, these data suggest that the LRIG2 SVA may act as a weak repressor of

transcription at the *LRIG2* promoter locus. In the case of *LRIG2-DT* the SVA might alternatively modify splicing of the mRNA, as the SVA sits within the transcript's first exon. It has been demonstrated that intronic SVAs are associated with intron retention in mRNA [184, 187], which may trigger various nuclear degradation mechanisms and result in decreased overall transcript levels [212].

In the same samples, methylation levels of CpG probes at the locus were examined and it was found that the LRIG2 SVA was a significant mQTL for the nearest CpG 450K methylation probe, cg23932873 (**Figure 3.8**). It can be speculated that repressive targeting of the LRIG2 SVA by heterochromatin-forming KRAB-ZFPs resulted in hypermethylation of the surrounding genome – including cg23932873. Indeed, recruitment of KRAB-ZFPs is an established mechanism through which TEs may influence gene expression [139, 149, 150]; it was therefore examined whether there might be a correlation between methylation of cg23932873 and expression from the *LRIG2* locus. When correlation coefficients were determined it was found that there was a weak but statistically significant correlation between *LRIG2* expression and cg23932873 methylation levels, but no such association was seen for *LRIG2-DT* (**Figure 3.9**). Taken together, these data suggest that the LRIG2 SVA exerts subtle influences on gene expression at the *LRIG2* locus, potentially through the induction of local hypermethylation which in turn contributes to a transcriptionally repressive environment. In support of this, it has previously been demonstrated that proteins associated with transcriptional repressor complexes can recognise methylated CpG dinucleotides [213], and that DNA methyltransferases can cooperate with enzymes

that add methylation or remove acetylation at histones [214-216] – both of which are repressive changes. It is possible that the LRIG2 SVA also has more direct influences upon the *LRIG2* promoter, for example via the direct deposition of silencing chromatin marks at the promoter region by factors such as KRAB-ZFPs recruited to the SVA, but investigating this was beyond the scope of the data available in NABEC.

After observing potential changes in expression and methylation around the *LRIG2* promoter region associated with LRIG2 SVA RIP genotype in the general populace, this relationship was investigated in an otherwise genetically identical background. The established cell line SH-SY5Y was shown to be homozygous for presence of the LRIG2 SVA insertion and is karyotypically normal at its insertion site on chromosome 1p. SH-SY5Y was therefore selected for CRISPR-Cas9-mediated deletion of the SVA and the generation of a model system for study of its influence at the locus. A total of 3 clonal populations with the SVA deleted on a single chromosome (SVA genotype +/Δ) were generated and 4 were generated with the LRIG2 SVA deleted on both chromosomes (Δ/Δ). Additionally, 5 'unedited' clonal populations that went through the deletion process but were unchanged at the LRIG2 SVA (+/+) were retained. To ensure that these clonal populations experienced similar growth conditions prior to harvesting of material, they were each seeded at the same density and grown in the same culture medium in parallel. Examination of basal *LRIG2* expression in the ΔLRIG2 SVA SH-SY5Y lines indicated that deletion of a single copy of the LRIG2 SVA (+/Δ) was associated with a mean increase of 6.2% in *LRIG2* expression compared to the wildtype 'unedited' SH-SY5Y lines, while deletion of the SVA on both chromosomes

(Δ/Δ) was associated with a mean increase of 36.2% (**Figure 3.15a**). Pyrosequencing of the same samples at the cg23932873 CpG proximal to the LRIG2 SVA yielded a 0.8% mean reduction in methylation the LRIG2 SVA +/Δ lines and a 2.1% reduction in the Δ/Δ lines compared to the unedited SVA +/+ cells (**Figure 3.15b**). Finally, it was determined that *LRIG2* expression and cg23932873 methylation were moderately anticorrelated in the ΔLRIG2 SVA SH-SY5Y lines (**Figure 3.16**).

Although these observations all fell short of achieving statistical significance at the sample sizes obtained here, it is notable that CRISPR-Cas9-mediated deletion of the LRIG2 SVA recapitulates the trends seen in the NABEC datasets: namely, that the LRIG2 SVA is associated with decreased expression at the *LRIG2* locus and with increased methylation at a nearby CpG, and that these two phenotypes are inversely correlated. This agreement between the two contexts – general populace observations and transgenic cell model – confers extra weight to the proposal that the LRIG2 SVA is influencing the transcriptional environment at the locus.

It should be noted that the regulatory influences of an SVA need not be dramatic to be biologically important, particularly for individuals of a certain genetic background; in the study of complex diseases it is generally accepted that disease can result from the cumulative effect of many low-contribution variants [16]. Therefore, it is perhaps unsurprising that examination of the LRIG2 SVA under basal conditions (both in NABEC and the ΔLRIG2 SVA SH-SY5Y cells) does not reveal immediately statistically significant effects as its influence may be too subtle to be detected at the sample

sizes examined here. This is particularly true for the observational study performed in the NABEC datasets – this was an assessment of 229 individuals in the expression dataset, for example, whereas eQTL analyses in contemporary GWAS of complex disease now make use of tens of thousands of disease cases in tandem with millions of controls. Furthermore, the NABEC frontal cortex data is derived from a heterogenous mixture of cell types including neurons and glial cells, which may act to convolute expression and methylation patterns.

There are several obvious routes through which the work presented in this chapter could be readily expanded upon. Firstly, the eQTL analysis of effects at the *LRIG2* promoter using proxy SNPs could be repeated in a cohort larger than NABEC, as larger sample sizes might confer statistical significance to linear models of *LRIG2* expression versus LRIG2 SVA dosage, for example. Such an analysis would not even require additional genotyping to be performed, as the same or equivalent SNPs would likely be available in the larger cohort. Indeed, use of the proxy SNPs specific to the LRIG2 SVA VNTR that were generated here in a sufficiently large WGS dataset might allow the effect of rarer VNTR alleles to be robustly detected, thereby extending the preliminary analysis outlined here. As mentioned previously, contemporary cohort studies feature data from thousands or even millions of individuals, potentially making this extension readily achievable. Moreover, it may be revealing if this second cohort makes available RNA-seq data for individual transcript isoforms. This was data not accessible in NABEC, as all transcript isoforms were aggregated into a single expression value for each gene. Transcript-specific influences of the LRIG2 SVA may

therefore have been missed in this examination of NABEC RNA-seq data. This is particularly true of the *LRIG2-DT* transcript, as the SVA is situated within its first exon.

Similarly, increasing the number of ΔLRIG2 SVA SH-SY5Y cell lines studied may confer significance to the observations made there. The CRISPR-Cas9 editing pipeline described in **Section 2.2.12.3** was relatively inefficient since a method of screening for modified cells was not incorporated, besides PCR analysis of the LRIG2 SVA locus after the labour- and time-intensive clonal isolation process. This could have been improved if a readily selectable marker had been introduced during CRISPR. For example, during the induction of DSBs to delete the SVA a repair template could be provided which contained a marker such as the gene red fluorescent protein (RFP; GFP was already present as part of the CRISPR plasmid). In the event of homology-directed repair the gene for RFP would be incorporated in place of the deleted SVA, and cells that were RFP^+ – and therefore ΔSVA on at least one chromosome – could be selected by fluorescence-associated cell sorting (FACS) with relative ease. While it is true that insertion of the RFP gene would not be an accurate replication of the endogenous '–SVA' genotype, the ΔSVA genotype resulting from CRISPR is not a perfect replication either – here, deletion of the LRIG2 SVA was necessarily accompanied by the deletion of ~1 kb of flanking region, which may have consequences for the locus. Considering this, it is pertinent that the NHEJ-dependent CRISPR strategy employed here yielded variable repair outcomes, producing one ΔLRIG2 SVA allele that was larger than the other repaired loci. This was speculated to have resulted from an aborted attempt at homology-directed repair, and highlights

that even simply deleting region can have diverse repair outcomes that might have different effects at the locus. While the alternative strategy such as insertion of RFP (or other marker) via homology-directed repair might produce more specific and consistent outcomes, this would still need to be carefully designed to minimise unintended effects. Furthermore, a selection process involving FACS would require the identification of a cell line that was amenable to it and was endogenously +/+ for the LRIG2 SVA. Indeed, GFP^+ selection (to enrich for cells transfected with the CRISPR plasmid) by FACS was initially attempted in the SH-SY5Y cell line used here but it was found that cells very rarely survived. Therefore, while this suggested selection process would require potentially extensive genotyping of cell lines that were not available here, it should be feasible to produce a higher-throughput pipeline for generation of additional ΔLRIG2 SVA cell lines for validation of the cellular effects hinted at in this chapter.

In summary, in this chapter the LRIG2 SVA, a common RIP, has been explored as a model for how presence or absence of an SVA near a promoter region may influence local gene transcription. In both the general populace and a transgenic cell line model allele dosage of the SVA was associated with decreased expression from the locus and increased methylation of a proximal CpG dinucleotide, the two of which were inversely correlated and may be functionally linked. Although most of these observations did not achieve statistical significance, it was outlined how these analyses might easily be expanded to increase their power. Even without statistical significance, the concordance between population study and cell line models is

encouraging and lends credence to the suggestion that RIPs may result in differential modulation of transcription. Moreover, this supports the idea that RIPs may represent a relatively underappreciated source of interpersonal genomic variation in the 'common disease, common variant' hypothesis in GWAS, as they are often poorly mapped by TE detection tools at present. This problem of TE mappability in WGS is a result of the upper limit for read length of currently widely used short-read sequencing technologies, which is around 300 bases and therefore falls short of sequencing most intact and full-length TEs. By contrast, recently developed 'long-read' sequencing often produces reads exceeding 10 kb [217], meaning repetitive or large structural variants such as TEs may be readily detected in WGS. Excitingly, technical advances and reduced costs associated with long-read sequencing may soon permit routine implementation [213], enabling precise detection of TEs in WGS data and a greater understanding of how RIPs shape the contemporary human genome.

Chapter 4 Investigating the influences of a non-reference genome SVA RIP at the *MAPT* locus

4.1. Introduction

Investigation of the LRIG2 SVA suggested that SVA RIPs may result in subtle differences in gene expression at a locus. It is logical to postulate that this might be relevant to disease, in addition to highly deleterious disease-associated SVA insertions into gene regions – as occurs in Fukuyama muscular dystrophy and XDP [83, 184]. As has been posited in the previously discussed 'common disease, common variant' hypothesis of genetically complex disease, manifestation of such diseases can occur as the cumulative product of many alleles of small effect size [16]. SVA RIPs acting in a similar manner to the LRIG2 SVA could be exactly this kind of small effect allele, with different RIP genotypes thereby contributing to differences in disease risk. Recent approaches in the characterisation of the genetic basis of complex disease have centred heavily upon GWAS utilising short read sequencing, but are likely to incompletely capture repetitive and poorly mapped DNA elements such as SVAs and other TEs [189, 193] – and may therefore have missed a source of human-specific variation which may be important to the genetic burden of human-specific disease. The genetic basis of PD is a field of study in which this hypothesis may be relevant, since GWAS have identified increasing numbers of risk loci but few functional relationships have been delineated [23].

As part of collaborative efforts to address this question, collaborators at the National Institutes of Health (NIH), USA, shared with our laboratory the results of their recent analysis of NABEC WGS data using the Mobile Element Locator Tool (MELT), a widely implemented program for TE annotation (https://melt.igs.umaryland.edu/, [218]).

Their work had identified 3 novel SVA RIPs at the PD-relevant *MAPT* locus on chromosome 17 that were not included in the reference genome but were relatively common (**Figure 4.1**: SVA_704, SVA_705 and SVA_706) (Dr Kimberley Billingsley, NIH, personal correspondence). The *MAPT* locus, named for the prominent association of microtubule-associated protein tau (MAPT) with neurodegenerative diseases including Alzheimer's Disease [219], has been consistently associated with differential risk of PD by GWAS [17, 19-22]. Numerous genes in the region have been proposed to play a role in PD. Although PD has not historically been considered to involve pathology of tau, the product of *MAPT*, the *MAPT* gene has been nominated solely due its involvement in other neurodegenerative diseases [17]. Furthermore, there is now accumulating evidence of tauopathy and formation of neurofibrillary tangles (tau protein aggregates associated with Alzheimer's Disease) in PD brains [220]. By contrast, more recent eQTL analysis has suggested *WNT3* may be among the most important [23]. Investigation of the region is complicated by a ~1 megabase inversion polymorphism that occurs in 2 distinct non-recombinant haplotypes [221], denoted H1 (canonical) and H2 (inverted), with the H1 haplotype exhibiting a number of sub-haplotypes (**Figure 4.1**, alternative haplotypes visualised at the top in red) [222]. Despite the identification of disease-associated SNPs at the *MAPT* locus (**Figure 4.1**, black rs numbers along the bottom of the figure), the presence of extended haplotypes increases the difficulty in ascribing function to these genetic signals, since the signals are associated with a ~1 Mb block of genetic variants that are coinherited [221]. By contrast, several reports have indicated that haplotype blocks are typically in the region of 5–20 kb [223]. The H2 inverted haplotype is likely to be exclusively Caucasian in origin and occurs at an allele frequency of ~25% in this group, with

frequencies of ~5% in central Asian and almost 0% in other populations likely representing admixture [224]. This makes detailed investigation of the region via GWAS problematic, as these datasets are largely derived from Caucasians – as of January 2019, >78% of individuals in the GWAS catalogue (https://www.ebi.ac.uk/gwas/) were of European descent [225]. Accordingly, haplotype-specific variants causative of gene expression changes are yet to be identified at the *MAPT* locus, despite a well-established overrepresentation of the H1 haplotype in PD [17, 226], as well as in diseases such as Alzheimer's Disease [227] and progressive supranuclear palsy [228].

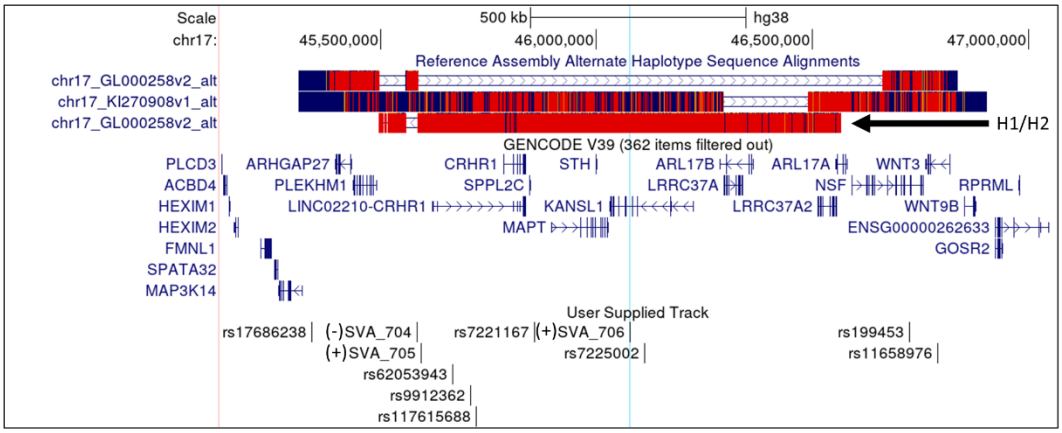


Figure 4.1 – The *MAPT* locus as shown on the UCSC Genome Browser, hg38. Genes from the RefSeq curated subset are displayed in blue. Alternate H1 and H2 haplotypes at the locus are displayed in red along the top: The top two red and blue regions are H1 sub-haplotypes, while the bottom band (arrow) corresponds to the H1/H2 inversion region. PD-associated SNPs [23] and the non-reference SVA RIPs identified by MELT are displayed in black along the bottom. Putative orientations of MELT-annotated

SVA insertions are displayed (+ for sense, - for antisense). The location of the SVA_706/KANSL1 SVA is highlighted in light blue.

It was noted that of the 3 SVA RIPs discovered via MELT by our collaborators at NIH, one of them, SVA_706, was located within intron 4 of *KANSL1* (**Figure 4.1**, insert location highlighted in light blue). This SVA RIP was associated with the non-PD risk H2 haplotype and was 35 kb from the PD risk SNP rs7225002 located within *KANSL1*, which is a gene that was recently identified via functional screens as a regulator of PINK-1-regulated mitophagy [229] – a mitochondrial quality control process known to be dysregulated in some familial cases of PD [230]. It has been shown that *KANSL1* knockdown leads to a reduction in phosphorylation of ubiquitin, a mitophagy marker mediated by PINK-1, and a reduction in expression of PINK-1 [229]. Furthermore, the PD risk-associated *MAPT* H1 haplotype has been associated with decreased expression of *KANSL1* [229]. Notably, knockdown of 30 other transcripts from the locus including *MAPT* did not replicate this disruption of ubiquitin phosphorylation, and so the authors proposed that variation influencing *KANSL1* may be the driver of PD risk at the locus.

Considering this, the *KANLS1*-intronic SVA_706, herein 'KANSL1 SVA', became a focus of study during this thesis. In light of the lack of clear annotation of functional variants at the *MAPT* locus it was postulated that previously unstudied SVAs on each haplotype may be contributing differences in their association with PD. Specifically, it was hypothesised that the H2-associated KANSL1 SVA may contribute to changes

at the locus that are protective against PD, thereby explaining some of the lower risk of PD conferred by the H2 haplotype. Furthermore, the reported haplotype specificity of the KANSL1 SVA suggests that this SVA insertion occurred after the *MAPT* locus chromosomal inversion that gave rise to the H1 and H2 haplotypes. If the insertion was relatively evolutionarily recent it is therefore possible that the KANSL1 SVA is not present within every H2 haplotype within the global populace, which may have consequences for the predicted PD-protective effects of the haplotype and its utility as a PD biomarker. Furthermore, sequence variation within the SVA may have arisen after insertion via replication slippage or recombination at repetitive CT and VNTR elements, further complicating association of the KANSL1 SVA with PD.

4.1.1. Aims

To explore any relationship between the KANSL1 SVA and haplotype-specific gene expression patterns at the *MAPT* locus by:

- Genotyping presence vs absence for the KANSL1 SVA along with any repeat length polymorphisms
- Generating proxy SNPs for the KANSL1 SVA to assess its association with the H1/H2 haplotype, PD risk SNPs, and gene expression at the *MAPT* locus in both NABEC and PD sample cohorts
- Assessing whether the KANSL1 SVA can be characterised in *in vivo* using CRISPR-Cas9, as with the LRIG2 SVA, or *in vitro* via an approach such as a luciferase reporter assay

4.2. Results

4.2.1. Primers were designed to amplify the KANSL1 SVA and its flanking region, which then suggested inaccuracies in the MELT prediction of SVA length

Much like the LRIG2 SVA, 'empty site PCR' primers targeting the KANSL1 SVA and its flanking regions needed to be designed *de novo* to validate the MELT predictions of insertion coordinates and size. The 2 kb flanks upstream and downstream of the KANSL1 SVA insert site predicted by MELT were downloaded from UCSC Genome Browser (hg38) and submitted to NCBI Primer-Blast with the standard primer design parameters described previously (**Section 2.2.2**). Since the reference genome sequence does not include the KANSL1 SVA, a minimum amplicon size of 1 kb was specified so that this 'empty site' PCR product lacking the SVA insertion would be easily visualised on a gel. The properties of output oligonucleotides were double-checked using OligoAnalyzer (**Section 2.2.2**). The selected primers were named 'KANSL1 SVA + Flanks' and were predicted to anneal ~650 bp upstream and ~820 bp downstream of the putative KANSL1 SVA insertion, and therefore produce a ~1.5 kb 'empty site' PCR product (primer details **Table 2.5**). MELT had predicted that the KANSL1 SVA was 1313 bp in length likely reflecting a truncated SVA insertion (Dr Kimberley Billingsley, NIH, personal correspondence), and so a ~2.8 kb 'filled site' PCR product was expected when the element was present.

MELT analysis of NABEC WGS data provided predicted KANSL1 SVA RIP genotypes. To assess how efficiently the 'KANSL1 SVA + Flanks' amplified the empty and filled site amplicons a predicted heterozygous DNA sample (RIP genotype +/-), arbitrarily

designated sample '#1' here, underwent PCR with these primers at a range of annealing temperatures (**Section 2.2.4**). It was observed that a single PCR product corresponding to the predicted 1.5 kb 'empty site' amplicon was produced at 56–58 °C annealing temperatures, but the 2.8 kb 'filled-site' product was missing (**Figure 4.2a**). To test whether this was the result of inefficient amplification of the SVA-containing amplicon or mis-annotation by MELT, two additional predicted heterozygous samples labelled '#2' and '#3' underwent PCR (**Figure 4.2b**). In these subsequent reactions the cycle number was increased from 30 to 35 and the range of annealing temperatures was lowered to favour increased product formation, potentially at the expense of specificity. For both samples '#2' and '#3' a single amplicon of ~4 kb was visible at all temperatures, albeit at different intensities, while the anticipated 1.5 kb 'empty site' product was absent (**Figure 4.2b**). Although the ~4 kb PCR product was larger than expected for the 'filled site' amplicon, its absence in the same reaction with DNA sample #1 indicated that it was not a non-specific product. Rather, it was assumed that this represented a 'filled site' product containing a full-size ~2.5 kb SVA, which would yield a 4 kb amplicon when amplified along with 1.5 kb flanking region captured by the 'KANSL1 SVA + Flanks' primer set. It was notable that while MELT had apparently correctly identified the location of an SVA RIP, as evidenced by the ~2.5 kb discrepancy between 'KANSL1 SVA + Flanks' PCR products from different samples, it had made errors predicting the size and sample genotypes of the element. This is in line with the previously discussed limitations of the tools currently used to query WGS data and highlights the need for validation of their outputs.

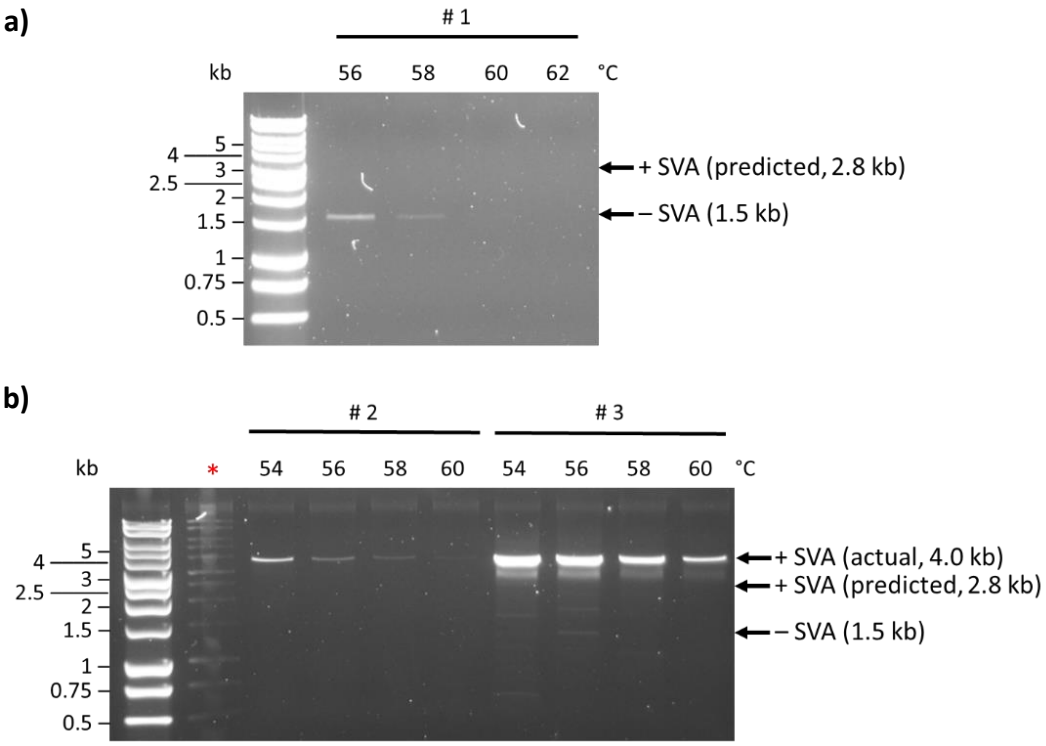


Figure 4.2 – Optimisation of 'KANSL1 SVA + Flank' primers. **a)** The primers were tested in a standard PCR reaction using KOD Hot Start Polymerase, 30 cycles, 10 ng NABEC sample '#1' gDNA template, and a range of annealing temperatures. 8 µl of PCR products were run on a 0.8% agarose gel at 110 V for 1 hour. **b)** Primers were used with the same PCR conditions but with 10 ng input of gDNA from samples '#2' and '#3' with 35 cycles and adjusted range of annealing temperatures. 8 µl of PCR products were run on a 0.8% agarose gel at 120 V for 1 hour. Red asterisk indicates lane containing spill-over from the DNA ladder (leftmost lane). **a & b)** Amplicon sizes for empty, predicted filled and observed filled sites are shown to the right.

4.2.2. Primers annealing proximal to the KANSL1 SVA confirm via sequencing that the SVA is full size and of the F subclass

Confirmation of KANSL1 SVA size (and determination of features such as subclass) required sequencing of the element, which necessitated design of primers which annealed close to the putative SVA insert site since the 'KANSL1 SVA + Flanks' primer

pair annealed to the genomic locus outside of the effective range of Sanger sequencing (~ 1 kb). NCBI Primer-Blast was used to design a primer pair that annealed proximal to the SVA insert site and incorporated ~250 bp of total flanking sequence (**Section 2.2.2**). These 'KANSL1 SVA Proximal' primers were tested in a temperature gradient PCR with 'NABEC sample #3', putatively shown to be KANSL1 SVA +/+ in **Section 4.2.1**, as template using KOD Xtreme Hot Start polymerase for its high processivity and product yields (**Section 2.2.4**). Based on the observation in **Figure 4.2b** that presence of the KANSL1 SVA contributes 2.5 kb to the size of an amplicon it was expected that the 'KANSL1 SVA Proximal' primers would produce a ~2.75 kb product. These primers exhibited good specificity and products of the anticipated size were detectable at PCR annealing temperatures 59–63 °C (**Figure 4.3**), and a 60 °C annealing step was decided upon for future amplifications.

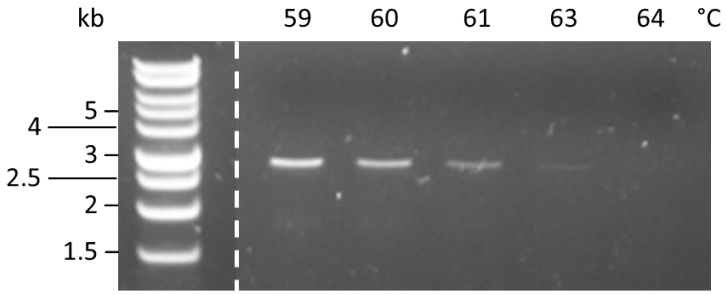


Figure 4.3 – Annealing temperature gradient PCR of 'KANSL1 SVA Proximal' primers. **)** The primers were tested in a standard PCR reaction using KOD Xtreme Hot Start Polymerase, 30 cycles, 5 ng NABEC sample '#3' gDNA template, and a range of annealing temperatures. 20 µl of PCR products were run on a 0.8% agarose gel at 110 V for 1 hour. White dashed line indicates lanes where unsuccessful tests of other primer pairs were run (not described here) and cropped out of the image.

With the gel image indicating that the 'KANSL1 SVA Proximal' primers yield a single PCR product, these were used to amplify the SVA for sequencing. The KANSL1 SVA was amplified with the proximal primers with 5 ng 'NABEC sample #3' gDNA as template for 40 cycles with a 60 °C annealing step, and the resulting PCR mixture was found by spectroscopy to contain 86.6 ng/µl DNA (**Section 2.2.3.3**). To increase quantity of this 'KANSL1 SVA Proximal' amplicon, it was blunt-end ligated into the pCR-Blunt plasmid at a ratio of 10:1 insert:vector with 25 ng of vector (reaction details **Section 2.2.9.1**). The resulting construct was transformed into chemically competent *E. coli*, grown out and extracted via miniprep (**Section 2.2.9.5**). The presence of the 'KANSL1 SVA Proximal' amplicon insert in the purified plasmid was verified by PCR with the same primers (not shown). Sanger sequencing of this construct was conducted by Source Bioscience externally using each of the forward and reverse 'KANSL1 SVA Proximal' primers (**Section 2.2.10**). Sanger sequencing has an effective range of 1–1.5 kb, and so by sequencing from each end it was expected that the whole SVA could be constructed from the central overlap. This was indeed the case, and the sequence confirmed that the insertion present within the 'KANSL1 SVA Proximal' amplicon in 'NABEC sample #3' was a 2323 bp SVA retrotransposon (**Figure 4.4**). Comparison of the KANSL1 SVA SINE-R region, described previously as the determinant of SVA subclass [66], to that of all SVA subclass consensus sequences using NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) indicated that the closest match was that of the SVA F class, with 99.19% sequence identity. Notably, the KANSL1 SVA lacks an appreciable poly-A tail (**Figure 4.4**, purple sequence); while poly-A tails are classically considered to be 150-250 nucleotides long, more recent findings have suggested that they can be as short as 50 residues [231, 232] – but the total

absence of a poly-A sequence here is suggestive of a small 3' truncation event upon SVA insertion.

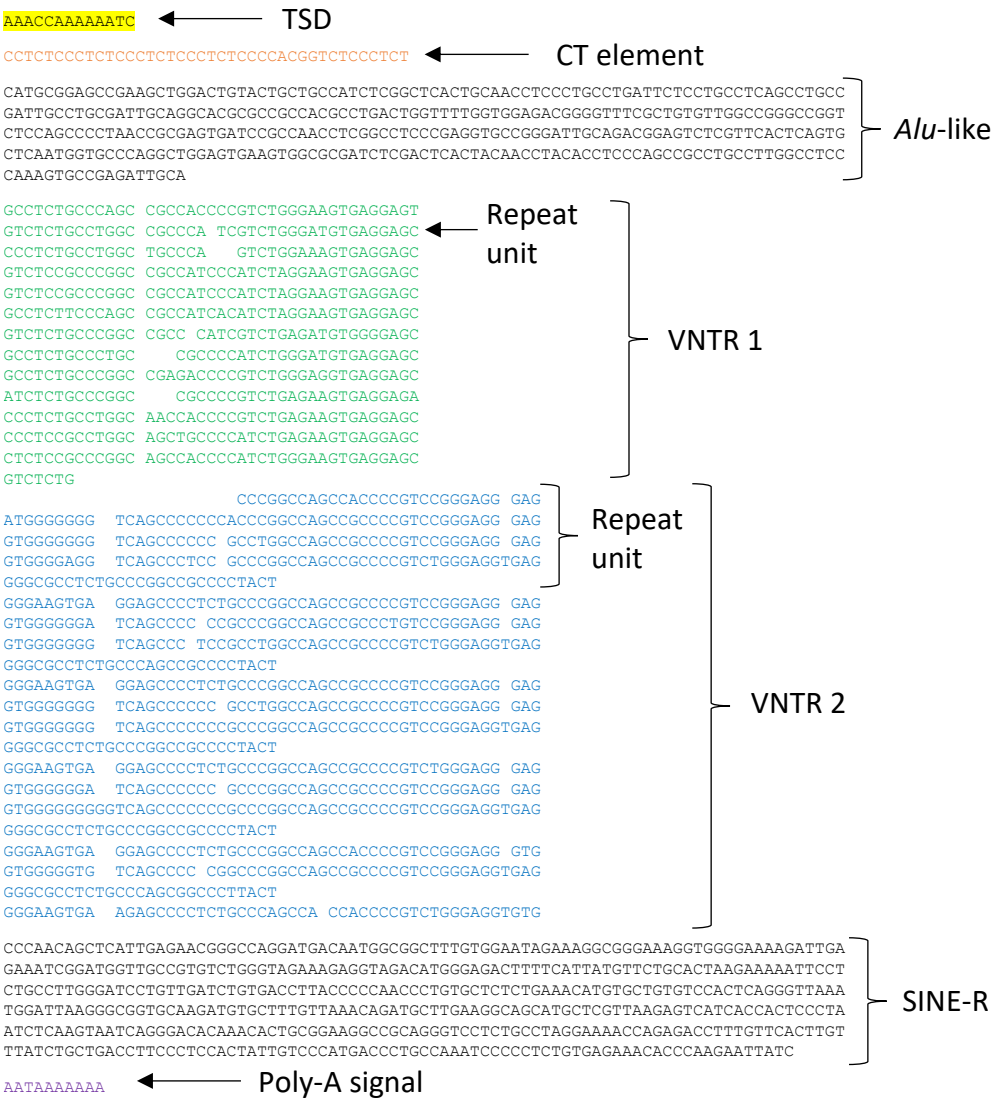


Figure 4.4 – Sequence of the KANSL1 SVA insert in 'NABEC sample #3'. Target site duplications (TSDs) highlighted in yellow, CT element in orange, VNTR 1 in green, VNTR 2 in blue, poly-A signal in purple. VNTR regions are broken down into approximate repeat monomers and aligned. Sequence is sense relative to the genome presented on UCSC genome browser (hg38).

Having determined the DNA sequence of the KANSL1 SVA, primers targeting internal components could be devised (**Section 2.2.2**). Primers were designed to amplify the entire central VNTR region (**Figure 4.4**, green and blue regions) as well as the two VNTR regions separately for use in a nested PCR with prior amplification of the whole SVA using the'KANSL1 SVA Proximal' primer set (**Section 2.2.5**) – the same strategy employed to amplify the LRIG2 SVA (**Section 3.2.1**). Primers targeting the CT element and poly-A tail were also designed, but their respective positions at the 5' and 3' ends of the SVA meant that a primer could be placed outside of the insertion (i.e., in genomic DNA that is more likely to be unique) and a nested PCR was therefore unnecessary. Using 'NABEC sample #3' as template, these primers were tested in PCRs with annealing temperature gradients. The 'KANSL1 SVA Combined VNTR' primers exhibited good specificity for an amplicon of the expected size at all temperatures tested (**Figure 4.5a**), and a 58 °C annealing step was selected for future amplifications. The 'VNTR 1'- and 'VNTR 2'-specific primer pairs, however, produced many off-target PCR products at all temperatures tested despite use of a nested PCR approach which was intended to minimise off-target genomic binding by enriching the KANSL1 SVA (**Figure 4.5b**). The reverse and forward oligonucleotides of the 'VNTR 1'- and 'VNTR 2'-specific primer pairs, respectively, had targeted the junction between VNTR regions 1 and 2 (**Figure 4.4**, where green and blue regions meet) and were expected to yield a single specific amplicon since this sequence is unique within the element. However, the presence of multiple PCR products suggests that these junction-targeting primers may have possessed sufficient sequence similarity to repeats within the VNTR, resulting in annealing in several locations. Indeed, the 50-100 bp step change observed for the off-target amplicons in **Figure 4.5b** would be

consistent with a given primer annealing at sites separated by 1 or 2 of the 40-50 bp repeats within the KANSL1 SVA VNTR region (**Figure 4.4**). Iterative rounds of adjusted PCRs did not produce specific products (not shown), so amplification of the separate VNTR regions was abandoned. The 'KANSL1 SVA CT' and 'KANSL1 SVA Poly-A' primers each efficiently produced a single amplicon of the expected sizes (**Figure 4.5c & d**), and a 62 °C annealing step was chosen for both.

The binding sites for selected oligonucleotides targeting the whole KANSL1 SVA with and without its flanking regions, VNTR region, CT element and poly-A signal are summarised in **Figure 4.6**.

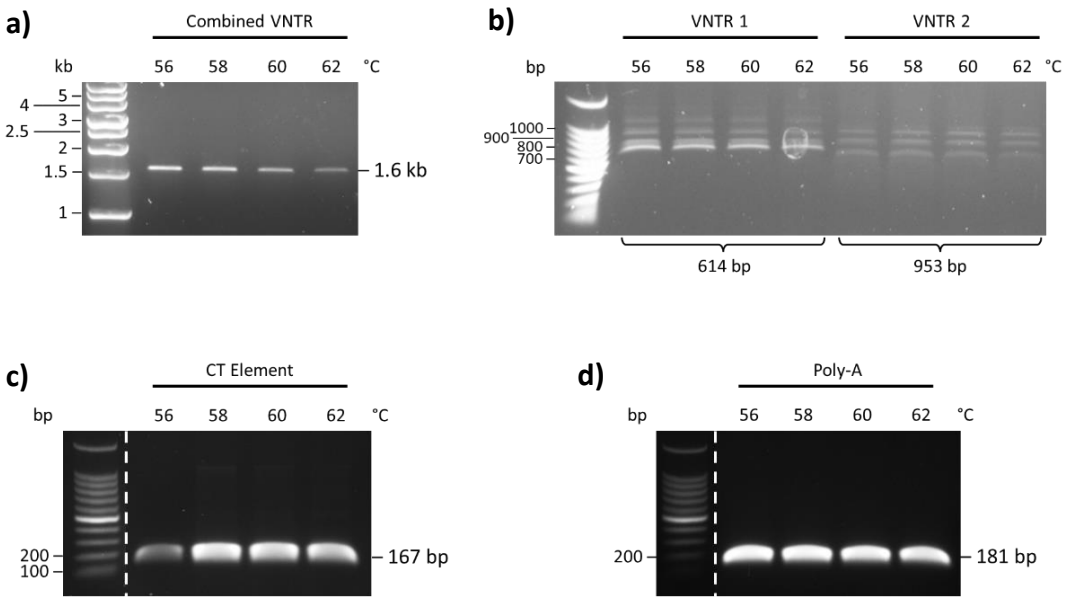


Figure 4.5 – Temperature gradient PCRs of primers targeting KANSL1 SVA internal components. All PCRs used 'NABEC sample #3' as template. **a)** 5 ng DNA underwent 20 cycles of amplification with 'KANSL1 SVA Proximal' primers, and then 2 µl of this PCR product was used as input for PCR using 'KANSL1 SVA Combined VNTR' primers with 20 amplification cycles and a range of annealing temperatures. 10 µl of each final PCR product was loaded in a 1% agarose gel and run at 110 V for 1 hour. **b)** 10 ng DNA underwent 25 cycles of amplification with 'KANSL1 SVA Proximal' primers, and then 1 µl of this PCR product was used as input for PCR using 'KANSL1 VNTR 1' or 'KANSL1 VNTR 2' primers with 20 amplification cycles and a range of annealing temperatures. 10 µl of each final PCR product was loaded in a 1% agarose gel and run at 110 V for 1 hour. **a & b)** Used KOD Hot Start Polymerase. **c)** The 'KANSL1 SVA CT' primers were tested in a standard PCR reaction using GoTaq G2 Hot Start Polymerase, 35 cycles, 10 ng DNA template, and a range of annealing temperatures. 12 µl of PCR products were run on a 1% agarose gel at 100 V for 1 hour. **d)** Conditions were the same as for (**c**) but using the 'KANSL1 SVA Poly-A' primers.

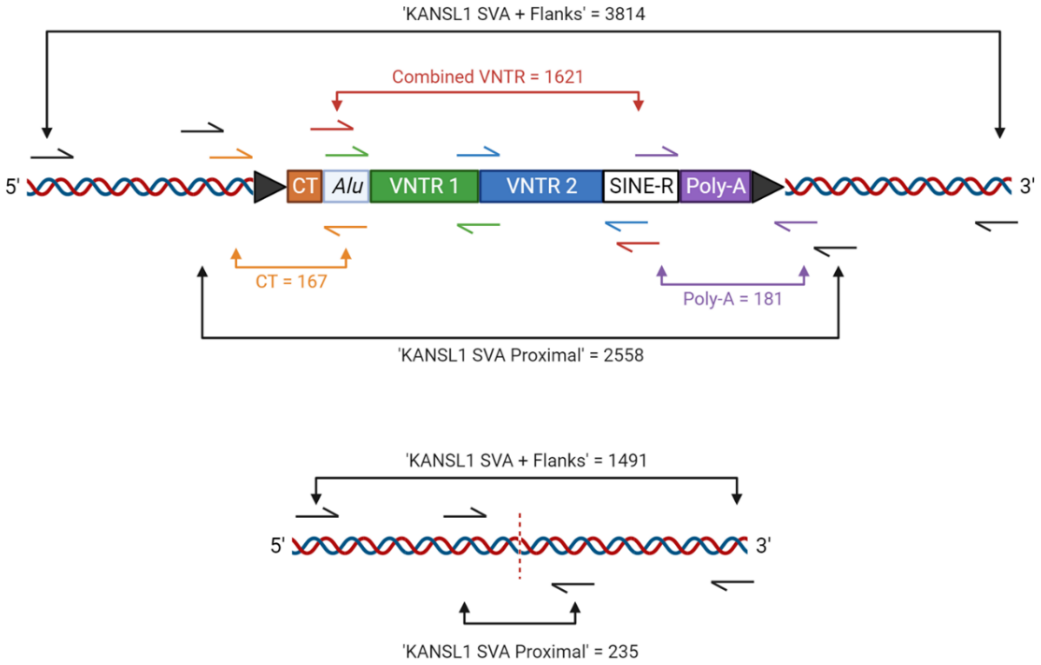


Figure 4.6 – Illustration of KANSL1 SVA primer binding sites and amplicon sizes when the KANSL1 SVA is present (top) and absent (bottom). Primer binding sites for the unsuccessful designs targeting the SVA VNTR 1 and VNTR 2 regions are also shown (green and blue arrows, respectively).

4.2.3. It was confirmed that the KANSL1 SVA was a RIP in NABEC DNA samples but not in the genotypes predicted by MELT

With primers designed and optimised for amplifying the KANSL1 SVA and its individual components, the SVA was characterised in the 96 available NABEC DNA samples. Initially, the KANSL1 SVA RIP was genotyped for presence versus absence at the locus by PCR with the 'KANSL1 SVA + Flanks' primers. As gDNA samples can be of varying quality, conditions favouring abundant product formation were used to ensure that enough of the KANSL1 SVA 'filled' and 'empty' site amplicons were produced for visualisation – with the caveat that this can reduce PCR specificity. Specifically, the 56 °C annealing temperature used here was at the lower end of those

tested (**Section 4.2.1**). As predicted, these PCR conditions produced a gel image with visible secondary bands (**Figure 4.7**, bands smaller than the 1.5 kb amplicon corresponding to '–SVA' genotype) but these were easily distinguished from the major PCR products corresponding to on-target amplicons (**Figure 4.7**, brighter bands). Across 96 DNA samples 46 were KANSL1 SVA -/- (reference genotype), 41 were +/-, 7 were +/+, and 2 did not produce detectable PCR products – a representative gel image in shown in **Figure 4.7**. For the 94 samples that yielded amplicons, these genotype frequencies are approximately in line with those expected of a variant at Hardy-Weinberg equilibrium with a minor allele frequency of 25% (the frequency of the H2 haplotype, with which the KANSL1 SVA is thought to be in strong LD), as outlined in **Table 4.1**:

KANSL1 SVA genotype	Expected proportion (Hardy-Weinberg)	Expected frequency	Observed frequency
-/-	56.3%	53	46
+/-	37.5%	35	41
+/+	6.3%	6	7
Total	100%	94	94

Table 4.1 – Comparison of validated KANSL1 SVA RIP genotype frequencies with those expected from Hardy-Weinberg equilibrium of a polymorphism with minor allele frequency of 0.25 (25% occurrence of H2 haplotype).

These validated genotypes were compared to those predicted by MELT and it was found that they matched in only 33 out of 94 samples (35%). Again, this highlights the need for validation of *in silico* retrotransposon annotations.

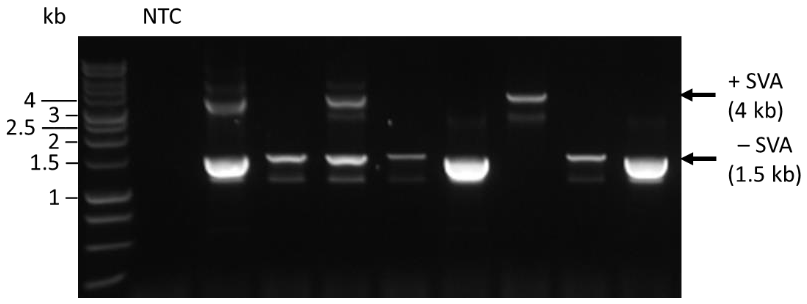


Figure 4.7 – KANSL1 SVA RIP genotyping in NABEC DNA samples. 10 ng of gDNA from available NABEC samples was amplified using KOD Hot Start Polymerase with 'KANSL1 SVA + Flanks' primers, an annealing temperature of 56 °C, and 35 amplification cycles, using KOD Hot Start Polymerase. 10 µl of PCR products were loaded onto a 0.8% agarose gel and run at 110 V for 1 hour. NTC = No template control.

4.2.4. Genotyping indicated that the KANSL1 SVA had a polymorphic CT element with a rare minor allele

Next, the 48 NABEC samples found to possess at least one copy of the KANSL1 SVA were genotyped for length polymorphisms of the internal components. The KANSL1 CT element, central VNTR and poly-A regions were amplified separately using the primers optimised previously (**Section 4.2.2**) – representative gel images of the 48 samples analysed are provided in **Figure 4.8**. It was observed that the CT element possessed two alleles for repeat length: a longer allele that was observed in every

individual, and a shorter CT allele that was observed in one sample that was KANSL1 SVA +/+ but was heterozygous for these two length variants (**Figure 4.8a**, leftmost lane). Meanwhile, the VNTR region and poly-A displayed no detectable length polymorphisms in this cohort (**Figure 4.8b & c**).

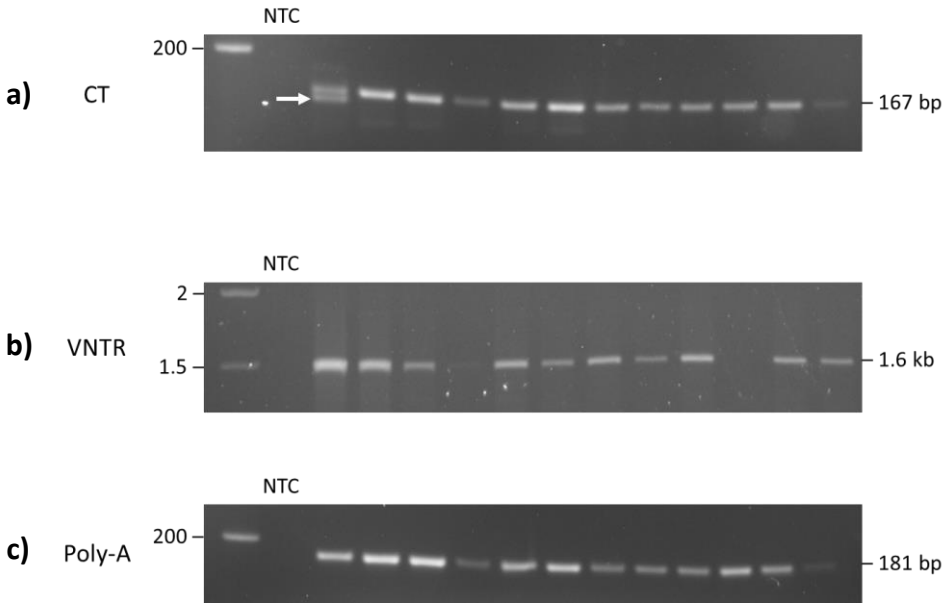


Figure 4.8 – Only the CT element of the KANSL1 SVA displays repeat length polymorphism in available NABEC DNA samples shown previously to possess at least one copy of the KANSL1 SVA (**Figure 4.7**). **a)** 10 ng gDNA underwent PCR with 'KANSL1 SVA CT' primers with 62 °C and 30 amplification cycles, using GoTaq G2 Hot Start Polymerase. 8 µl of PCR products were run on a 3% agarose gel at 100V for 4.5 hours. The 'short' CT allele is indicated by a white arrow. **b)** 10 ng gDNA underwent 20 cycles of PCR with the 'KANSL1 SVA Proximal' primers with a 60 °C annealing step. 2 µl of this PCR mixture was used an input for PCR with the 'KANSL1 SVA Combined VNTR' primers with 25 amplification cycles and 58 °C annealing temperature. Both reactions used KOD Hot Start Polymerase. 10 µl of PCR product was loaded onto a 0.8% agarose gel and run at 100 V for 4 hours. **c)** Conditions were the same as in (**a**) but instead using the 'KANSL1 SVA Poly-A' primer set.

4.2.5. Proxy SNPs were identified for the KANSL1 SVA RIP genotype

Genotyping of the KANSL1 SVA RIP enabled the generation of proxy SNPs for extrapolation of genotypes in the wider NABEC cohort (**Section 2.2.1.2**). The SVA was first considered as a simple biallelic variant with two alleles, corresponding to SVA absence (reference genotype) or presence (alternate genotype). When proxy SNPs were identified in the NABEC hg38 WGS dataset three SNPs tied as the top-performing proxy SNPs for linkage with the KANSL1 SVA (r^2=0.9744, D' = 1). One of these, rs140819255, was selected at random and taken forward to tag SVA presence versus absence. It was noted that in the 359 NABEC samples for which genotyping data was available that rs8070723G, an established proxy SNP for the H2 haplotype, was in high LD (r^2=0.9855, D'=1) with the rs140819255 allele that tagged the KANSL1 SVA. It was also found that the 94 PCR-validated KANSL1 SVA genotypes were in high LD with the H2-tagging SNP (r^2=0.9497, D'=1), which was taken to confirm the association observed in the wider cohort via proxy SNPs. Taken together, these observations reinforce the association of the KANSL1 SVA insertion with the H2 haplotype.

In light of the identification of a shorter KANSL1 SVA CT element in one NABEC sample (**Figure 4.8a**), CT element-specific proxy SNPs were also generated. This was achieved by considering the SVA insertion site as to contain two separate biallelic SNPs, representing insertion of an SVA with a short or long CT element. With an r^2 cut-off of >0.95, a single proxy SNP for the long CT allele was identified, rs150334020, while

two equivalent SNPs were identified for the short CT allele – as before, one of these, rs1209310955, was randomly selected for use.

The selected proxy SNPs for the KANSL1 SVA and its CT element are summarised in **Table 4.2** and their positions at the *MAPT* locus are shown in **Figure 4.9**. In each case the proxy SNP alternate allele was in phase with the KANSL1 SVA insertion (also considered the alternate allele).

KANSL1 SVA:	**Proxy SNP**	**r^2**	**D'**
Presence vs absence	rs140819255	0.9744	1
Short CT	rs1209310955	1	1
Long CT	rs150334020	0.9741	1

Table 4.2 – Selected proxy SNPs for the KANSL1 SVA and its CT element alleles in NABEC hg38, with corresponding r^2 and D' values.

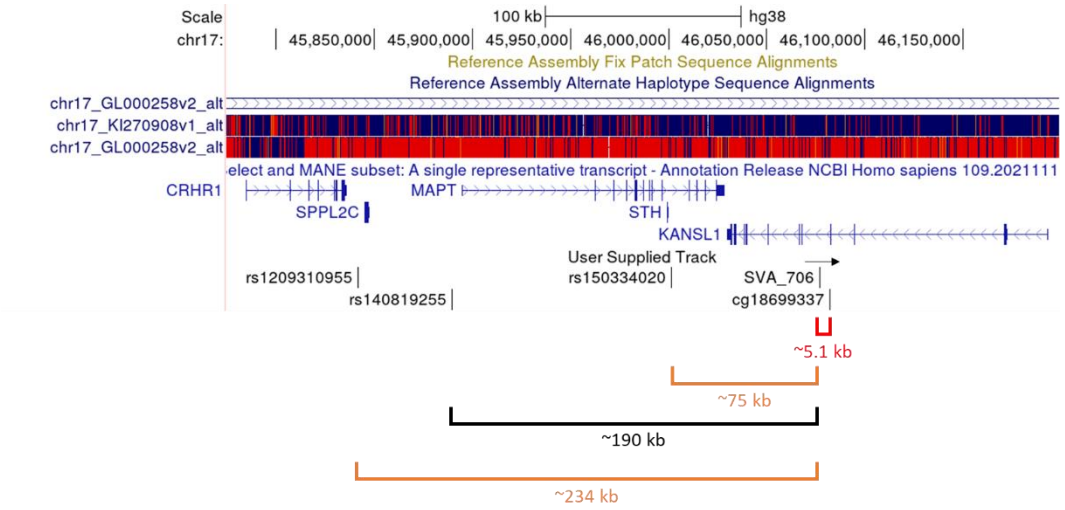


Figure 4.9 – Selected proxy SNPs for the KANSL1 SVA and its CT element alleles at the *MAPT* locus, from UCSC hg38. Linear distances for the whole SVA-tagging SNP in black and those for CT element-specific proxy SNPs in orange. Distance to the nearest CpG methylation probe from the KANSL1 SVA, cg18699337, shown in red (CpG annotation unavailable in hg38, so coordinates were lifted from hg19). 5' to 3' orientation of the KANSL1 SVA is indicated by the black arrow.

4.2.6. KANSL1 SVA RIP allele dosage is associated with expression of *KANSL1* and methylation at the nearest CpG probe in NABEC, but the two are not correlated

With KANSL1 SVA proxy SNPs selected, first rs140819255 was used to infer how SVA presence or absence might correlate with expression of *KANSL1* in NABEC RNA-seq data. As with the LRIG2 SVA, an 'extended' list of genotypes for the KANSL1 SVA was produced by supplementing genotypes ascertained by PCR with those predicted by the proxy SNP rs140819255. When combined with the available transcriptomic data, and after removal of data points from children (<15 years of age at death) and outliers, expression data was available for 243 individuals in NABEC. This was made up of 136 KANSL1 SVA RIP genotypes -/-, 92 +/- genotypes, and 15 +/+ genotypes. When KANSL1 SVA allele dosage was compared to the quantile normalised data for

KANSL1 expression a positive correlation was observed: compared to the absence of the SVA (-/-), the +/- SVA genotype was associated with a median increase in gene expression of 0.773 SDs while the +/+ genotype was associated with a further 0.763 SD increase (therefore -/- vs +/+ corresponded to a difference of 1.536 SDs) (**Figure 4.11**). Construction of a linear model for KANSL1 SVA dosage and *KANSL1* expression (**Section 2.2.1.3**) indicated that this relationship was statistically significant and corresponded to a change in gene expression of 0.691 SDs per SVA insertion (**Figure 4.11**, P = 3.03×10^{-13}).

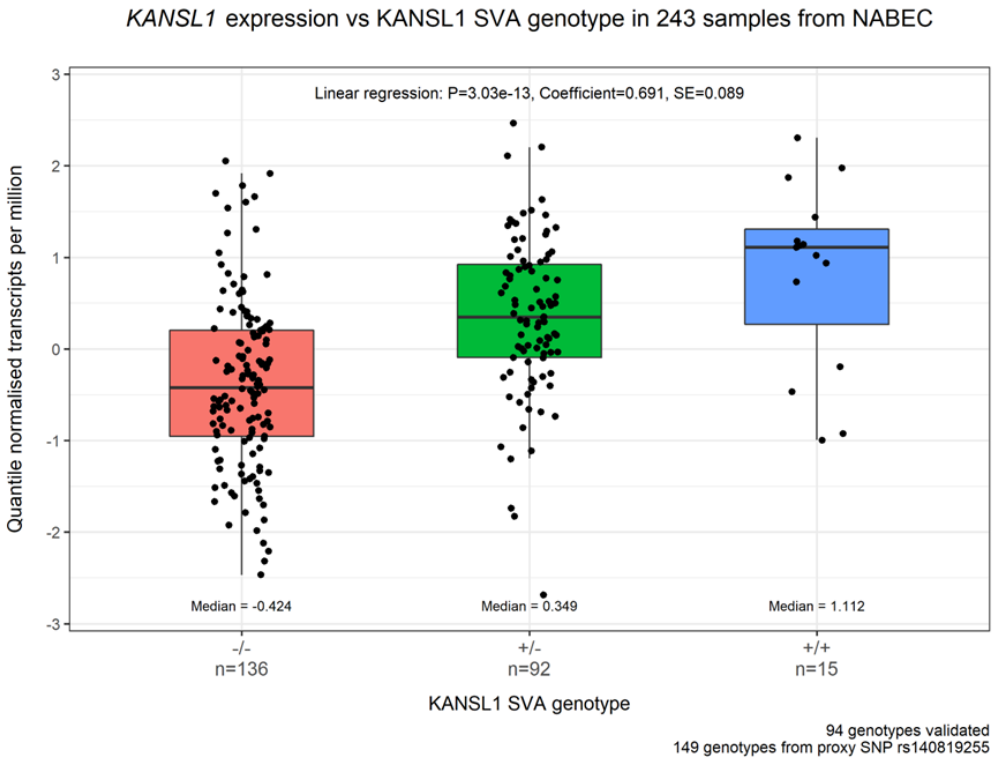


Figure 4.10 – KANSL1 SVA RIP genotype versus frontal cortex total RNA-seq data for *KANSL1* (ENSG00000120071.14) in 243 NABEC individuals. 94 genotypes were PCR validated and 149 were imputed for a total of 243 genotypes. RNA-seq data expressed as quantile normalised transcripts per

kilobase million (TPM). Standard deviations from the mean of the normalised data are displayed on the y-axis. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

Having previously observed that methylation of the closest CpG probe (as listed on UCSC hg19) to the LRIG2 SVA was associated with SVA allele dosage, it was examined whether the same was true for the KANSL1 SVA RIP. The nearest CpG probe to the KANSL1 SVA was found to be cg18699337, located ~5.1 kb away (**Figure 4.9**). Intersection of KANSL1 SVA RIP genotypes with available methylation data in NABEC produced a subset of 177 individuals, composed of 95 -/-, 71 +/- and 11 +/+ SVA genotypes. The KANSL1 SVA +/- genotype was associated with a 3.8% increase in the median proportion of cg18699337 residues that were methylated compared to the -/- genotype, while the +/+ genotype was associated with a further 6% increase (**Figure 4.11**). Linear modelling indicated that there was a significant relationship between SVA allele dosage and cg18699337 methylation in which presence of each additional allele corresponded to a 2.78% increase in methylation (**Figure 4.11**, P = $2.27x10^{-10}$).

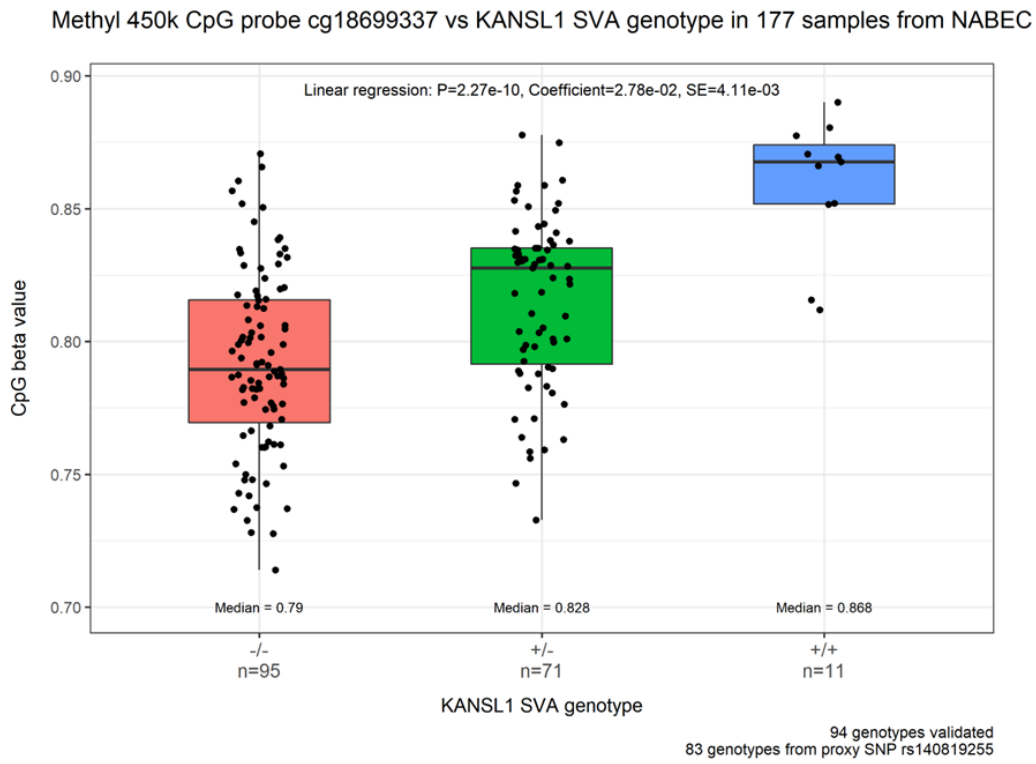


Figure 4.11 – KANSL1 SVA RIP genotype versus CpG methylation data for nearest probe. Frontal cortex CpG 450K methylation data for probe cg18699337 in 177 NABEC individuals grouped by KANSL1 SVA RIP genotype. 94 genotypes were PCR validated and 69 were imputed for a total of 165 genotypes. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

Finally, it was examined whether *KANSL1* expression and cg18699337 methylation, both statistically associated with the KANSL1 SVA, were correlated. Due to incomplete overlap in the molecular datasets available for NABEC individuals, combination of RNA-seq and 450K methylation samples for which KANSL1 SVA proxy SNPs were available produced a final group of 125 individuals. When *KANSL1* expression was compared to cg18699337 methylation and a Pearson correlation was

determined (both datasets found to be normally distributed, not shown) it was observed that there was no association between the two (**Figure 4.12**, Pearson correlation P value = 0.273).

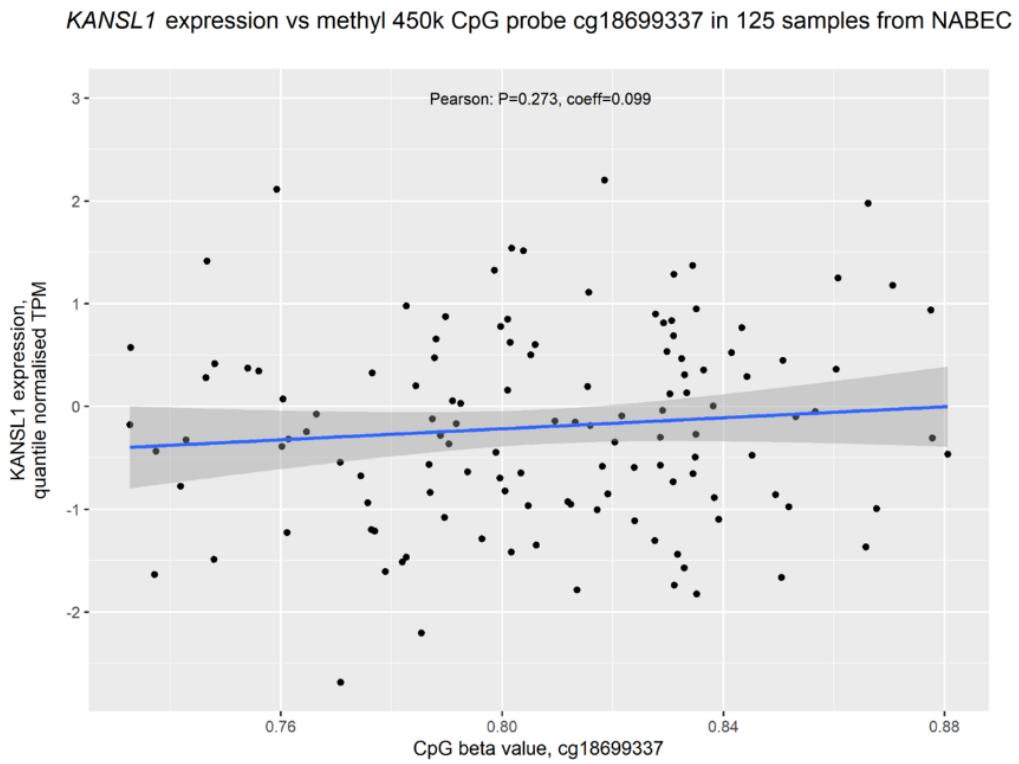


Figure 4.12 – Expression of *KANSL1* versus methylation of CpG 450K probe cg18699337 in 125 NABEC frontal cortex samples. Blue line indicates trend line; dark grey zone indicates 95% confidence interval. Displayed are Pearson correlation coefficients and corresponding *p* values.

4.2.7. CT element-specific KANSL1 SVA proxy SNPs did not predict any additional NABEC DNA samples harbouring the shorter CT allele

After using a proxy SNP that tagged presence versus absence for the KANSL1 SVA to query expression and methylation data, it was assessed whether the use of CT

element-specific proxy SNPs would reveal how variation within the SVA influences interpersonal differences in gene expression for individuals harbouring at least one copy of the KANSL1 SVA. The chosen proxy SNPs for short and long CT alleles (rs1209310955 and rs150334020, respectively) perfectly recapitulated predictions for overall allele dosage of the SVA made by the proxy SNP for presence versus absence (rs140819255). However, out of 360 NABEC individuals with available WGS (hg38) data the CT element proxy SNPs did not indicate that any other samples possessed the shorter CT repeat element besides the single individual previously identified via PCR (**Figure 4.8a**). Nevertheless, the datapoint for this heterozygous CT 'short/long' (S/L) genotype was separated out from other KANSL1 SVA genotypes, denoted '-' for absence and 'L' for presence of the long CT allele. The S/L genotype did not indicate dramatic effects on gene expression, with its *KANSL1* expression value falling within the interquartile range of the KANSL1 SVA L/L genotype (**Figure 4.13**, range of purple box). Subsequently, the SVA S/L genotype was considered as a CT element dosage between that of the -/L and L/L genotypes for the purposes of construction of a linear model. While this model indicated a significant association between CT length dosage and *KANSL1* expression (**Figure 4.13**, $P = 1.19 \times 10^{-11}$), this was a reduction in significance compared to the same analysis without partitioning of CT element alleles (**Figure 4.11**, $P = 3.03 \times 10^{-13}$). Therefore, consideration of the KANSL1 SVA CT element variant only acts to dilute the strength of observations made for the KANSL1 SVA RIP in the available data.

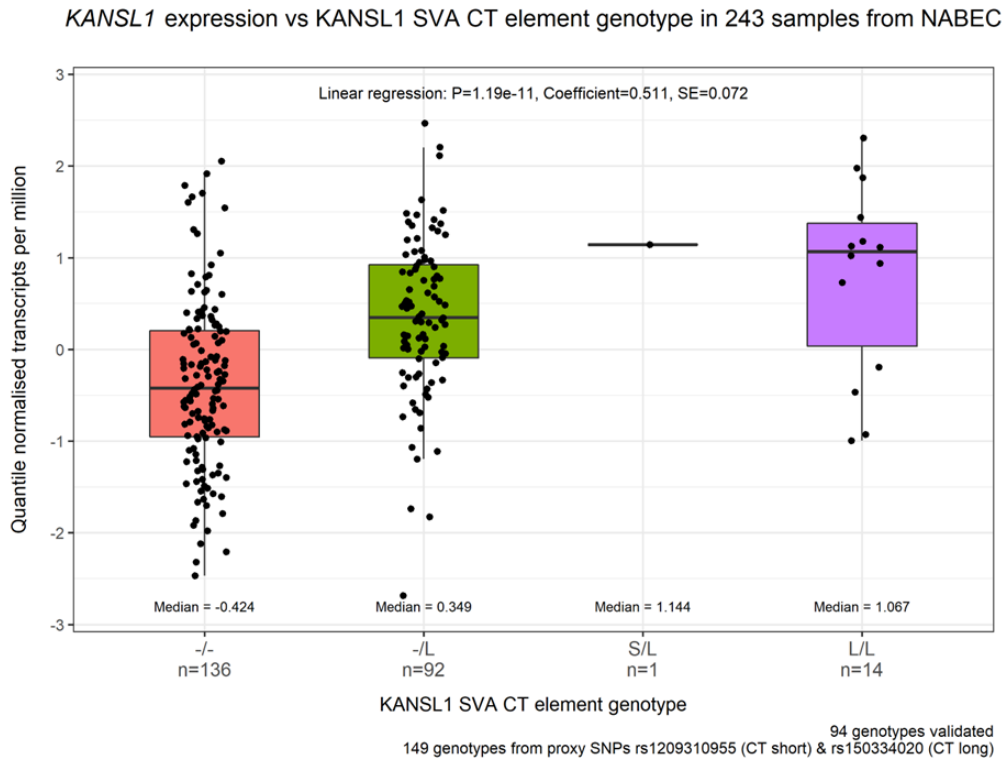


Figure 4.13 – KANSL1 SVA CT element genotype versus frontal cortex total RNA-seq data for *KANSL1* (ENSG00000120071.14) in 243 NABEC individuals. 94 genotypes were PCR validated, one of which harboured a single 'short' CT allele, and 149 genotypes were imputed using rs1209310955 (short CT) and rs150334020 (long CT) for a total of 243 genotypes. For CT element genotype, '-' indicates absence of the SVA, 'S' indicates the presence of a KANSL1 SVA with the short CT element allele, and 'L' represents presence of the long CT allele. RNA-seq data expressed as quantile normalised transcripts per kilobase million (TPM). Standard deviations from the mean of the normalised data are displayed on the y-axis. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

4.2.8. The KANSL1 SVA extended previous associations between gene expression and H1/H2 *MAPT* haplotype to predict expression of putative PD gene *WNT3*

Here the KANSL1 SVA has been interrogated as an eQTL for expression of *KANSL1*, a gene speculated to play a causal role in PD [229], on the premise that this non-reference genome SVA may be a hitherto unexplored haplotype-specific genetic variant at the *MAPT* locus that contributes to gene expression differences between H1 and H2 haplotypes. Since the chosen proxy SNP for the KANSL1 SVA was found to be in high LD with a proxy SNP for the H2 haplotype in NABEC (**Section 4.2.5**), it was reasoned that the KANSL1 SVA may also be correlated with expression of other genes at the locus that have previously been observed to be associated with a specific haplotype. Although the most recent GWAS meta-analysis was unable to associate several of the PD risk SNPs at the *MAPT* locus with gene expression changes [23], work by O'Brien *et al* on eQTLs in the developing human brain provided a useful benchmark; of the 21 genes they identified at the *MAPT* locus that were differentially expressed in association with specific SNPs, 13 could be "explained" by H1/H2 haplotype [233] (**Table 4.3**, genes highlighted in yellow). These 21 genes were taken as prime candidates for gene expression that might be altered by genetic variants, and their expression levels in the available NABEC transcription data were therefore compared to KANSL1 SVA genotype. Expression of *MAPT* was also assessed, in light of its established association with neurodegenerative disorders besides PD, along with that of *WNT3*, the QTL-nominated gene for 3 PD risk SNPs at the *MAPT* locus in the latest PD GWAS meta-analysis [23]. 6 of the transcripts with expression associated with *MAPT* haplotype identified by O'Brien *et al* were pseudogenes that were not available in the NABEC expression data (DND1P1, RN7SL199P, AC138645.1,

AC005670.2, AC091132.5, NSFP1). The remaining 17 transcripts were individually incorporated into linear models to assess the contribution of the KANSL1 SVA and sample covariates to their expression (**Section 2.2.1.3**). From these 17 transcripts expression levels for 12 were found to be associated with KANSL1 SVA RIP genotype (**Table 4.3**, lefthand P values highlighted in green), 10 of which were previously linked to *MAPT* haplotype by O'Brien *et al*. It is perhaps unsurprising that all of the eQTLs that O'Brien *et al* explained via *MAPT* haplotype were also significantly associated with KANSL1 SVA RIP genotype, given the high linkage between H2 haplotype and SVA presence proxy SNPs (r^2=0.9855, D'=1). However, it is notable that 2 gene transcripts associated with KANSL1 SVA RIP genotype, *ARL17B* and *WNT3*, were not previously linked to *MAPT* locus haplotypes. This suggested that the KANSL1 SVA RIP genotype may possess greater predictive power for gene expression at the *MAPT* locus than haplotype-associated SNPs, possibly due to imperfect linkage between H1/H2 haplotypes and absence/presence of the SVA. To investigate this further, the strength of associations in NABEC expression data between the 17 transcripts under investigation and KANSL1 SVA genotype were compared to those of linear models instead incorporating the H2-tagging SNP rs8070723G. It was found that for the 10 H2-associated transcripts identified by O'Brien *et al* that the H2 proxy SNP was a superior eQTL than the KANSL1 SVA, yielding more significant associations in the linear models constructed and finding an additional link to *AC091132.4* expression (**Table 4.3**, righthand P values highlighted in green; comparisons with KANSL1 SVA associations summarised in 'Best eQTL' column). However, the strength of these gene expression associations were broadly similar for the KANSL1 SVA and H2 proxy SNP, with linear model P values for 6 genes differing by less than an order of magnitude

and a further 3 genes differing by less than 2 orders of magnitude. Importantly, the significant association between the nominated PD gene *WNT3* and KANSL1 SVA dosage was not recapitulated by the H2 proxy SNP. Although it was observed in NABEC data that expression of *ARL17B* was significantly associated with the H2 proxy SNP, in contrast with observations made by O'Brien *et al* [233], this association was weaker than that exhibited with the KANSL1 SVA. Taken together, these data indicate that while the KANSL1 SVA is a weaker predictor of individual expression of genes at the *MAPT* locus than an established H1/H2 proxy SNP, its predictive power is largely similar and additionally captures expression of another putatively important PD gene, *WNT3*.

	Linear regression with:				
	KANSL1 SVA		rs8070723G (H2 proxy)		
Gene	**Coefficient**	**P value**	**Coefficient**	**P value**	**Best eQTL**
LRRC37A4P	-1.1014	9.03E-44	-1.1250	5.78E-49	H2
KANSL1-AS1	1.0695	8.79E-39	1.0428	3.77E-39	H2
LRRC37A	1.0820	2.64E-34	1.0418	5.82E-35	H2
MAPK8IP1P2	3.2090	4.35E-34	3.0887	2.53E-35	H2
LRRC37A2	1.0596	9.81E-31	1.0373	2.06E-33	H2
AC091132.6	0.8273	1.34E-22	0.8011	2.42E-23	H2
RN7SL656P	0.8273	1.34E-22	0.8011	2.42E-23	H2
ARL17B	0.7606	7.18E-16	0.7051	2.86E-15	KANSL1 SVA
KANSL1	0.6914	3.03E-13	0.6724	2.12E-14	H2
AC091132.3	0.6587	4.59E-13	0.6770	2.74E-15	H2
Metazoa_SRP	-1.3048	3.86E-06	-1.3624	3.67E-07	H2
WNT3	0.3018	0.00209	0.2359	0.01200	KANSL1 SVA
AC091132.4	-0.2653	0.00558	-0.2926	0.00118	H2
MAPT-AS1	-0.1360	0.17139	-0.1587	0.08846	H2
NSF	0.1181	0.21706	0.1031	0.25266	KANSL1 SVA
ARL17A	-0.0131	0.89539	-0.0557	0.55426	H2
MAPT	-0.0009	0.99209	-0.0434	0.62498	H2
DND1P1	Not in NABEC				
RN7SL199P					
AC138645.1					
AC005670.2					
AC091132.5					
NSFP1					

Table 4.3 – Comparison between gene expression associated with the KANSL1 SVA and previously identified *MAPT* haplotype eQTLs. Genes listed are those at the *MAPT* locus with eQTLs identified by O'Brien *et al* 2018 [233], with genes they associated with H1/H2 haplotype highlighted in yellow. Additional genes that were not part of this previous analysis are highlighted in blue. Coefficients and P values are from individual linear regressions of gene expression the NABEC cohort against KANSL1 SVA RIP genotype and sample covariates. P values that passed the Bonferroni-adjusted alpha significance level (0.05/17=0.00294) are highlighted in green.

4.2.9. KANSL1 SVA RIP allele dosage is associated with *KANSL1* expression in the AMP-PD cohort, but SNP-inferred genotypes were not at expected frequencies

All analysis of the potential influence of the KANSL1 SVA so far was performed in the NABEC cohort, which is composed entirely of healthy (or at least free from overt disease) neuronal samples. The *MAPT* locus and H1/H2 haplotypes are implicated in PD [17, 226], and therefore a logical next step was to examine the KANSL1 SVA in a PD dataset. Access was gained to the Accelerating Medicines Partnership – Parkinson's Disease (AMP-PD, https://amp-pd.org/) cohort, which is a combined and harmonised dataset of 8 PD cohorts with genotyping and phenotypic data (**Section 2.1.4**). It is important to note that while the NABEC genotyping and transcriptomic data were obtained from post-mortem neuronal tissue, the same data in the AMP-PD combined cohort was derived from whole blood – meaning that gene expression patterns will likely inherently vary between the two datasets. Nevertheless, initial exploration of associations with KANSL1 SVA genotype was undertaken in the AMP-PD cohort. It was found that the proxy SNP for the KANSL1 SVA used in the hg38 NABEC WGS data, rs140819255, was unavailable in the AMP-PD 'v1 release' as this dataset was annotated relative to the hg19 genome. To identify a proxy SNP for use in AMP-PD, the validated KANSL1 SVA RIP genotypes were merged with the hg19 annotation of NABEC WGS data and proxy SNPs were generated (**Section 2.2.1.2**). In this version of the NABEC genotyping data the top performing SNP was rs200610218 (r^2=1, D'=1), and therefore its genotype data was downloaded along with RNA-seq data for *KANSL1*. Expression of *KANSL1* was stratified by KANSL1 SVA RIP genotype, and after removal of gene expression outliers this yielded 2698 datapoints made up of 1921 KANSL1 SVA -/-, 771 +/-, and 6 SVA +/+ genotypes. It was observed that gene

expression was significantly associated with KANSL1 SVA allele dosage (**Figure 4.14**, linear model P<2e16). However, cohort sample covariates (such as age, gender, ethnicity) could not be retrieved for AMP-PD v1 release, and so the value reported by this statistical test should be interpreted with caution. Furthermore, it was surprising that the tagging SNP rs200610218 indicated that only 6 samples out of 2698 harboured the KANSL1 SVA genotype +/+, as ~170 individuals with this genotype are expected (6.3%, based on ~25% of alleles harbouring the H2 haplotype which is in tight LD with the KANSL1 SVA) if the population is at Hardy-Weinberg Equilibrium. At the time of writing, it was not possible to determine how this had occurred due to data access limitations. For these reasons, further analysis was not conducted in the 'AMP-PD v1' dataset, including comparison of data from control and PD individuals. Regardless, the trend observed here supports the previous association of the KANSL1 SVA with increased expression of *KANSL1* made in the NABEC cohort (compare **Figure 4.14** with **Figure 4.10**). This finding is striking in consideration of the highly dissimilar cell types from which these cohorts were derived, and suggests a more ubiquitous association between the KANSL1 SVA and local gene expression.

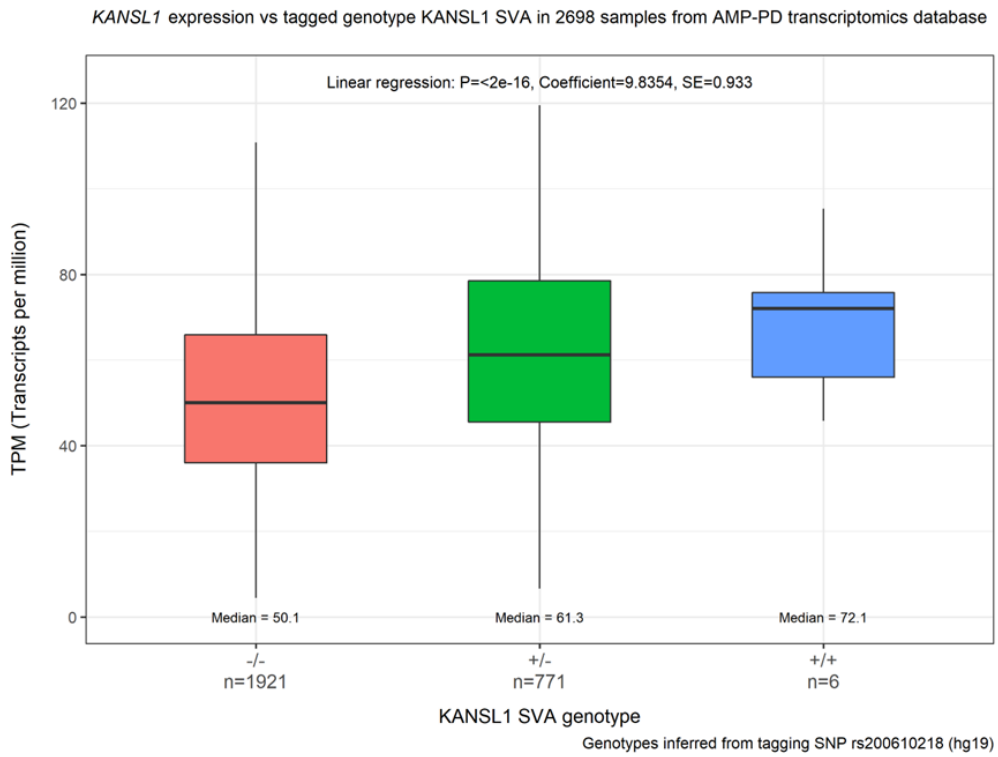


Figure 4.14 – KANSL1 SVA RIP genotype versus RNA-seq data for *KANSL1* in 2698 individuals from the AMP-PD cohort. RNA-seq data (ENSG00000120071.14) expressed as transcripts per kilobase million (TPM). KANSL1 SVA RIP genotype was inferred in hg19 genotyping data from the proxy SNP rs200610218. Linear regression analysis is shown, reporting p value of association analysis (P), model coefficient and standard error (SE).

4.2.10. The KANSL1 SVA was not homozygous present in available cell lines, preventing CRISPR-Cas9-mediated deletion

Thus far the genomic impact of the KANSL1 SVA has been examined in the genotypic and phenotypic data available from NABEC and AMP-PD cohorts. These findings have been entirely observational, and studies of the general populace can be influenced by myriad confounders that are not captured as covariates, such as lifestyle or

undiagnosed disease. Moreover, the high LD between the KANSL1 SVA and the wider H2 haplotype at the *MAPT* locus makes it difficult to directly ascribe function to the SVA based on WGS data alone. Therefore, it was investigated whether the KANSL1 SVA in isolation was amenable to characterisation in the laboratory.

First, the genotype of the KANSL1 SVA RIP was determined in established cell lines available in the laboratory. If any were identified as harbouring an SVA +/+ genotype then they would be taken forward as candidate cell lines in which to delete the KANSL1 SVA via CRISPR-Cas9 and subsequently phenotype changes would be measured, as was done for the LRIG2 SVA. The readily available cell lines in our lab were HeLa, JAR, MCF-7, SKNAS, HEK293 and SH-SY5Y. These 6 lines were grown to confluency in T75 flasks and gDNA was harvested (**Section 2.2.3.1**). These samples underwent PCR (**Section 2.2.4**) using the 'KANSL1 SVA + Flanks' primer pair and it was observed that only JAR cells were of the +/- genotype, while the other 5 cell lines did not harbour any KANSL1 SVA insertions (**Figure 4.15**). Based on this observation it was decided that CRISPR-based deletion of the single KANSL1 SVA in JAR cells would not be an effective use of time, considering the substantial requirements of the deletion pipeline and that information would be missed by being unable to compare to an endogenous SVA +/+ genotype. Furthermore, JAR cells have been shown to be near triploid [234] while ATCC describes their karyotype as "extremely complex" (https://www.atcc.org/products/htb-144); the increased ploidy of JAR cells would complicate resolution of ΔKANSL1 SVA genotypes, which reinforced the decision to not excise the SVA in this cell line.

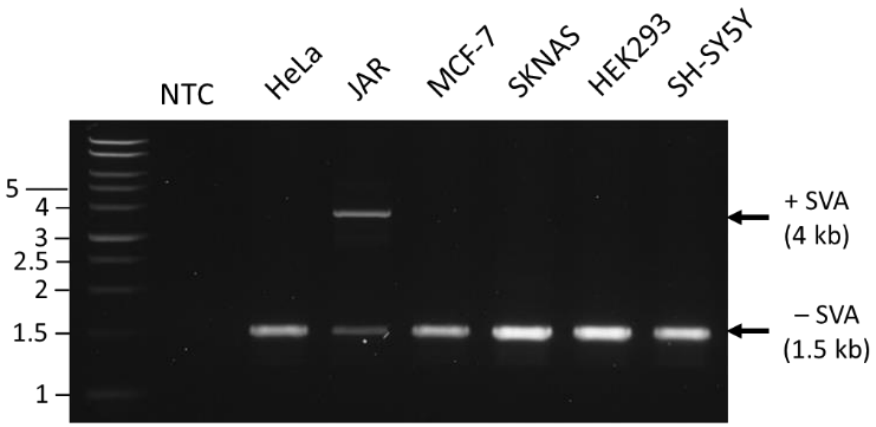


Figure 4.15 – KANSL1 SVA RIP genotypes in established cell lines available in the laboratory. 5 ng of gDNA from each cell line underwent PCR with KOD Hot Start polymerase using the 'KANSL1 SVA + Flanks' primer pair with 35 cycles and an annealing temperature 62 °C. 10 µl of PCR mixture was loaded onto a 0.8% agarose gel and run at 100V for 1 hour 30 min.

4.2.11. The KANSL1 SVA was cloned into the luciferase reporter pGL3P in a single orientation in the promoter region

As an alternative to studying the effect of removal of endogenous KANSL1 SVA, it was decided that SVA *cis*-regulatory functions would instead be measured via insertion into a reporter gene plasmid. This approach has been previously used by Savage *et al* to demonstrate that SVAs (and their separate internal component regions) from the *PARK7* and *FUS* gene loci can regulate gene expression *in vitro* [173, 181]. The same strategy would be employed as Savage *et al*: the regulatory capabilities of the KANSL1 SVA would be assessed in the pGL3-Promoter vector (pGL3P, from Promega) via insertion upstream of the plasmid's minimal promoter which drives expression of *luc+*, a modified Firefly Luciferase gene. Subsequently luciferase activity would be measured as a proxy for *luc+* expression. Since effects of the KANSL1 SVA at the *MAPT*

locus may be bidirectional, the SVA would be inserted in both sense and antisense orientations relative to the *luc+* reporter gene (**Figure 4.16**).

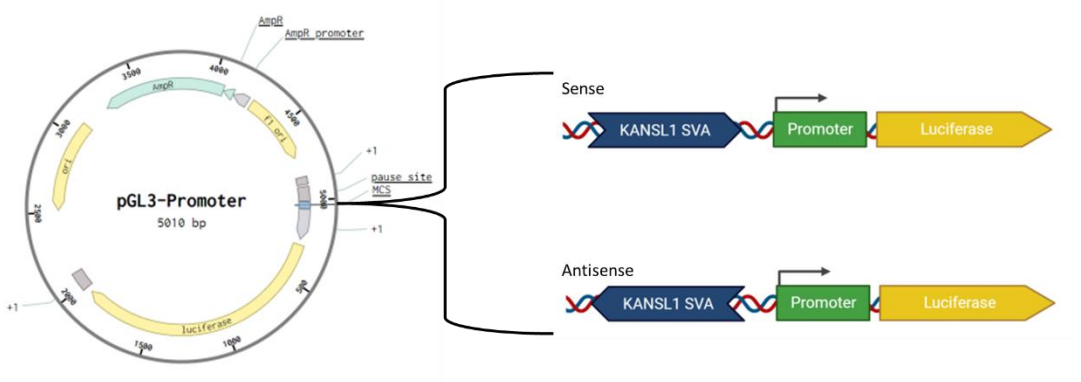


Figure 4.16 – Schematic depicting the pGL3P reporter gene plasmid and KANSL1 SVA insertions in the minimal promoter region, in both sense and antisense orientations.

A cloning strategy was devised in which the KANSL1 SVA would be PCR amplified, ligated into pCR-Blunt and subsequently subcloned into the pGL3P expression vector (overview in **Section 2.2.9**). Specifically, the SVA was amplified using the 'KANSL1 SVA Proximal' primers using 'NABEC sample #3' as it was previously shown to harbour 2 copies of the KANSL1 SVA (**Section 4.2.1**). This PCR was carried out with KOD Xtreme Hot Start polymerase, as this enzyme is highly processive and large quantities of PCR product were desirable (PCR conditions: 5ng gDNA input in 20 µl, 40 cycles, 61 °C annealing temperature). The resulting PCR mixture enriched for the 'KANSL1 SVA Proximal' amplicon was found by spectroscopy to contain 86.6 ng/µl DNA, and this underwent a ligation reaction with 25 ng pCR-Blunt and a 10:1 ration of insert:vector (**Section 2.2.9.1**). The ligation mixture was used to transform chemically competent

E. coli, and following antibiotic selection, bacterial outgrowth, plasmid extraction (**Sections 2.2.9.4 & 2.2.9.5**) and Sanger sequencing (**Section 2.2.10**) it was determined that pCR-Blunt vectors containing the KANSL1 SVA in both sense (KANSL1 SVA-S) and antisense (KANSL1 SVA-AS) orientation had been generated (sequencing not shown).

The type II restriction endonucleases SacI and XhoI were selected for subcloning because recognition sites for these were located either side of the KANSL1 SVA insert in the pCR-Blunt constructs and within the pGL3P MCS. Excision of the KANSL1 SVA with SacI and XhoI therefore permits direct insertion into pGL3P backbone linearised with the same enzymes (**Figure 4.17**). Importantly, these endonucleases produce non-compatible DNA overhangs upon digestion which will result in the insertion of the KANSL1 SVA into pGL3P a single direction of insertion (**Figure 4.17**). Incidentally, this retains the orientation that the SVA insert had within pCR-Blunt – i.e., subcloning of a sense-oriented insert from pCR-Blunt will produce an insert in the sense orientation relative to the luciferase reporter gene of pGL3P, and vice versa.

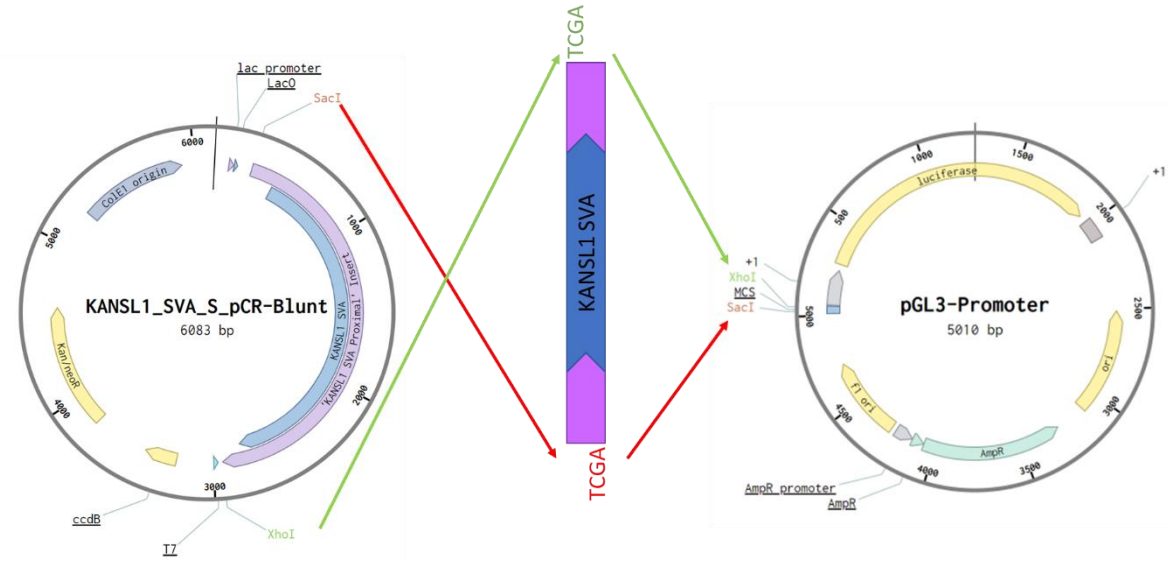


Figure 4.17 – Illustration of strategy for subcloning the KANSL1 SVA from pCR-Blunt to pGL3P. Subcloning from the pCR-Blunt construct containing the 'KANSL1 SVA Proximal' amplicon insertion (purple segment) in the sense orientation (KANSL1_SVA_S_pCR_Blunt) is shown. KANSL1 SVA sequence without any flanking region shown in blue. Cut sites and overhangs for SacI (red) and XhoI (green) are shown, along with arrows depicting movement of associated DNA overhangs.

In the first iteration of this subcloning strategy, the SacI/XhoI-excised KANSL1 SVA insert underwent agarose gel electrophoresis and was cut out of the gel: 5 µg of each of the KANSL1 SVA-S and KANSL1 SVA-AS pCR-Blunt constructs and pGL3P were digested with 10 U of SacI and XhoI at 37 °C for 1 hour with a 20 min 65 °C inactivation step (**Section 2.2.9.2**). The entire reaction volume of each digest was run on a 0.8% agarose gel at 120 V for 2 hours 30 min. The bands corresponding to SacI/XhoI-restricted KANSL1 SVA-S, KANSL1 SVA-AS and SacI/XhoI-linearised pGL3P were excised from the gel, and the DNA fragments were purified (**Section 2.2.6.1**). DNA concentrations of each were determined, and the KANSL1 SVA-S and KANSL1 SVA-AS fragments underwent a standard ligation reaction with 50 ng of pGL3P fragment

(specific conditions in **Section 2.2.9.1**). This ligation mixture was then used to transform chemically competent *E. coli*, which were selected by plating on ampicillin-containing agar. However, after two attempts at plating no colonies were formed despite successful plating of a control transformation plasmid (pUC19), indicating unsuccessful ligation.

It was suspected that the process of gel band excision was damaging the single-stranded overhangs from SacI/XhoI digestion, disrupting annealing of KANSL1 SVA and pGL3P sticky ends. The subcloning strategy was therefore attempted without excision of target fragments after restriction digests: 1 µg of KANSL1 SVA-S in pCR-Blunt, KANSL1 SVA-AS in pCR-Blunt and pGL3P were digested with 10 U of SacI and XhoI as before. To prevent reinsertion of the small SacI/XhoI-excised fragment from the pGL3P MCS back into the vector backbone, exposed DNA 5'-ends were dephosphorylated in 500 ng of this restriction digest mixture using Antarctic phosphatase (**Section 2.2.9.3**). The dephosphorylated linearised pGL3P was then ligated to KANSL1 SVA-S and KANSL1 SVA-AS inserts at 10:1 insert:vector for 3 hours at room temperature (**Section 2.2.9.1**). 2 µl of each ligation mixture was used to transform competent *E. coli,* and after plating on antibiotic selective agar 11 colonies were retrieved: based on the pCR-Blunt backbone they were excised from, 6 harboured the KANSL1 SVA-S pGL3P construct and 5 harboured the KANSL1 SVA-AS pGL3P construct. The candidate transformed *E. coli* colonies were grown out in LB broth and plasmid DNA was extracted by miniprep (**Section 2.2.9.5**).

In the predicted KANSL1 SVA pGL3P construct sequences it was observed that there were 3 recognition sites for the restriction endonuclease BamHI: one within the pGL3Pbackbone, one within the MCS sequence carried over from pCR-Blunt during subcloning, and one located near the 3' end of the KANSL1 SVA (**Figure 4.18a**). Thus, the alternate orientations of the KANSL1 SVA within pGL3P will yield DNA fragments of distinct sizes when digested with BamHI (**Figure 4.18a**) which will enable SVA presence and orientation to be easily determined when these fragments are separated by electrophoresis (predicted banding patterns in **Figure 4.18b**). As such, 200 ng of each of the 11 putative KANSL1 SVA-pGL3P constructs was digested with 1 U of BamHI under standard conditions (**Section 2.2.9.2**) and underwent agarose gel electrophoresis. Surprisingly, upon BamHI digest only 1 of the plasmids produced the expected banding pattern (**Figure 4.18b**): a KANSL1 SVA-AS pGL3P construct, sample #11 (**Figure 4.18c**, lane marked with a green tick). The presence of the KANSL1 SVA upstream of the pGL3P minimal promoter in the antisense orientation was confirmed by Sanger sequencing (**Section 2.2.10**).

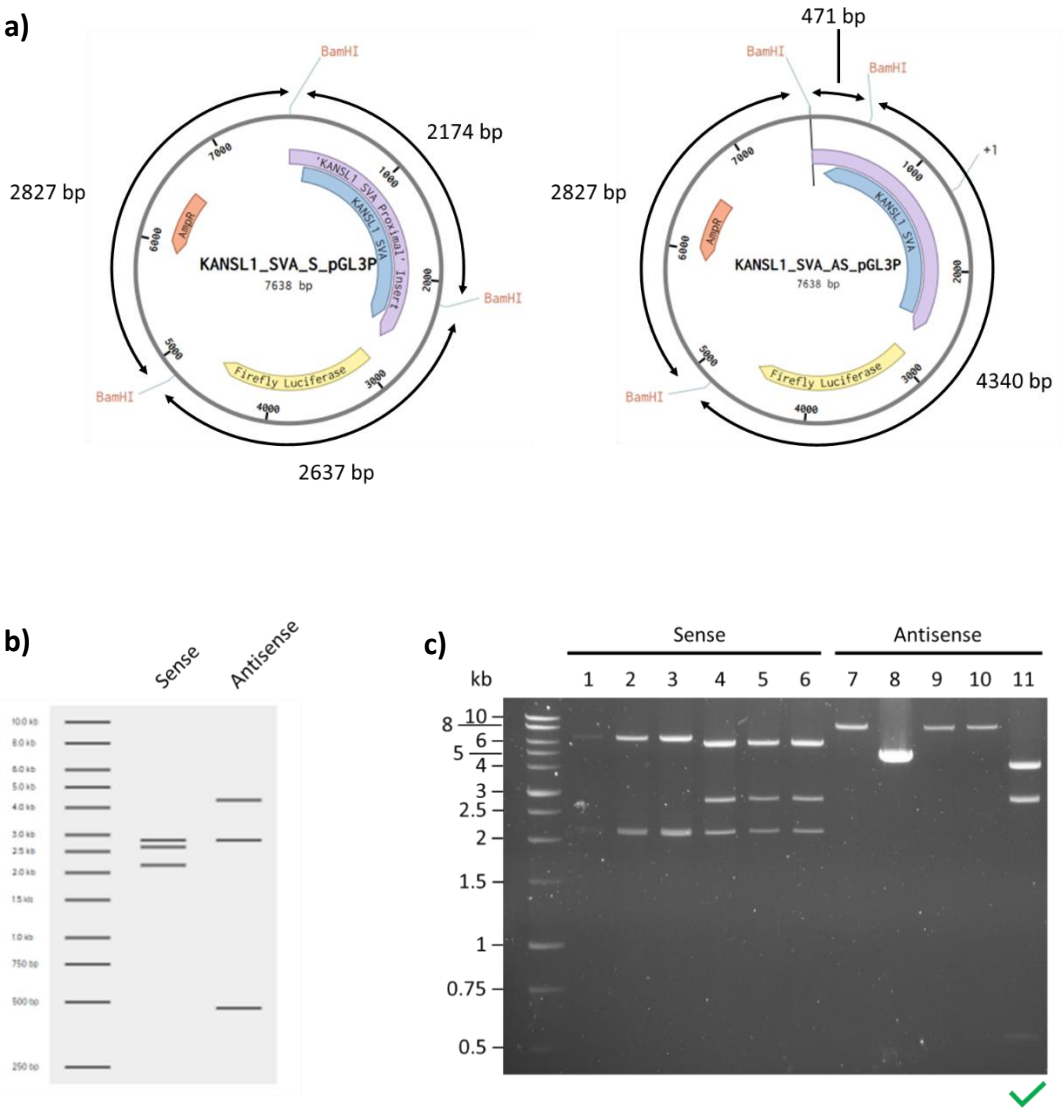


Figure 4.18 – Restriction mapping of putative KANSL1 SVA sense and antisense pGL3P constructs. **a)** Plasmid maps of pGL3P with the KANSL1 SVA inserted in the sense (left, KANSL1_SVA_S_pGL3P) and antisense (right, KANSL1_SVA_AS_pGL3P) orientations relative to the luciferase reporter gene with BamHI cut sites and digestion product sizes. **b)** Predicted gel image of BamHI digest products of pGL3P containing the KANSL1 SVA sense (KANSL1_SVA_S_pGL3P) and (KANSL1_SVA_AS_pGL3P) insertions. **c)** 200 ng of each of the putative KANSL1 SVA-pGL3P constructs was digested with 1 U of BamHI (BamHI-HF, NEB) at 37 °C for 1 hour followed by a 20 minute heat inactivation at 65 °C. 10 µl of digests were mixed 5:1 with loading dye and loaded onto a 1% agarose gel and ran at 100 V for 2 hours.

It was assumed that plasmids #7–10 were pGL3P vectors that had re-linearised during ligation despite the inclusion of a phosphorylation step, as unmodified pGL3P possesses a single BamHI cut site and digestion therefore produces a single band corresponding to linearised vector. On the other hand, it was not obvious why the supposed KANSL1 SVA-S pGL3P constructs, plasmids #1–6, did not produce the expected pattern of bands. To examine this, constructs #3 and #4 were sent for Sanger sequencing using 'KANSL1 SVA CT' reverse and 'KANSL1 SVA Poly-A' forward primers (**Figure 4.6**) – these anneal within the SVA at its 5' and 3' ends, respectively, and prime sequencing 'outwards' and into the vector. This revealed that in both constructs the KANSL1 SVA along with the pCR-Blunt backbone had been inserted into the MCS of pGL3P, with this SVA-pCBR-Blunt block oriented antisense to pGL3P in plasmid #3 and sense in plasmid #4 (not shown). Given that the pCR-Blunt vector was located 3' of the KANSL1 SVA insert in these spurious constructs, it can be determined that these inappropriate insertions occurred due to inefficient digestion by XhoI leading to linearisation of the KANSL1 SVA-S pCR-Blunt instead of excision of the SVA.

Due to time constraints it was not possible to devise a new subcloning strategy to insert the KANSL1 SVA into pGL3P, or to perform luciferase assays using the successful KANSL1 SVA-S pGL3P construct. However, the generation of the KANSL1 SVA-S pCR-Blunt and KANSL1 SVA-AS pGL3P represents a pump-priming of the functional assessment of the KANSL1 SVA and completion of this project should be readily achievable.

4.3. Discussion

In this chapter a MELT annotation in NABEC WGS of a novel SVA RIP within *KANSL1* at the *MAPT* locus was characterised, on the basis that it was associated with the region's H2 1 Mb inversion haplotype; while this haplotype has been linked to a decreased risk of PD [17, 226], causal genetic variants are yet to be identified and it was speculated that the SVA insertion may at least partially underpin gene expression changes associated with H2. In addition to influencing wider changes at the *MAPT* locus, it was specifically anticipated that the KANSL1 SVA would be associated with expression changes of *KANSL1*, the gene in which it resides. *KANSL1* is a regulator of PINK-1-regulated mitophagy, which is a process dysregulated in some familial cases of PD, and reduction of *KANSL1* expression has been associated with H1 haplotype [229, 230]. It was postulated here that the *cis*-regulatory influence of the haplotype-specific KANSL1 SVA insertion was the driver of increased expression of *KANSL1* observed with the H2 haplotype, resulting in up-regulated mitophagy, increased mitochondrial quality control and reduced risk of PD.

By designing primers that annealed flanking the putative SVA insertion site it was confirmed by PCR and agarose gel electrophoresis that the region harboured a RIP (**Figure 4.2**). Subsequently the element was sequenced which revealed that it was a 2.3 kb (full length) SVA F retrotransposon, although the lack of a long region of A nucleotides following the elements' poly-A signal suggests a minor truncation event (**Figure 4.4**). Not only had MELT predicted that the SVA was ~1.3 kb in length but it was also observed via 'empty site' PCR that the program appeared to have incorrectly

predicted the KANSL1 SVA RIP genotype in 65% of NABEC individuals. Despite the inaccuracy of genotypes predicted by MELT, generation of a KANSL1 SVA proxy SNP from PCR-validated genotypes confirmed that the SVA was indeed closely associated with the H2 haplotype in the wider NABEC cohort, via its high LD with the H2-tagging SNP rs8070723 (r^2=0.9855, D'=1). It was noted that it was unusual for predicted and actual genotypes to diverge so drastically and yet still be associated so robustly with the H2 haplotype. Despite considerable troubleshooting it remains unclear how this discrepancy has arisen. Altogether, this highlights that bioinformatic annotations of repetitive elements in short-read sequencing data, while powerful, may require validation of their accuracy before use in downstream applications. Nevertheless, the close linkage of KANSL1 SVA RIP genotype with H1/H2 haplotype was interpreted as validation of the accuracy of the genotyping undertaken. In other words, a systemic error such as sample mix-up is unlikely to have occurred as this would not be expected to enable the recapitulation of high levels of LD with a target SNP. Therefore, KANSL1 SVA proxy SNPs were taken forward.

As with the LRIG2 SVA, the KANSL1 SVA proxy SNP permitted evaluation of gene expression stratified by RIP genotype in the NABEC dataset. It was observed that increased allele dosage of the SVA was closely associated with increased *KANSL1* expression (**Figure 4.10**). This was largely to be expected given prior associations between H2 haplotype and increased *KANSL1* expression (and vice versa) [229, 233], but acts to confirm that effects of the KANSL1 SVA RIP do not deviate dramatically from those associated with *MAPT* haplotype. In other words, this was a step towards

validation of the KANSL1 SVA as a candidate contributor to the gene expression patterns associated with the H2 haplotype. It was also found that KANSL1 SVA RIP genotype was positively associated with methylation levels of cg18699337 (**Figure 4.11**), a 450K methylation probe ~5.1 kb away that is nonetheless the closest CpG probe to the KANSL1 SVA insertion site listed on the UCSC genome browser (hg19). Conceptually, this is in line with the observation of increased expression of *KANSL1* with increased SVA allele dosage, as hypermethylation within gene bodies has previously been associated with increased gene expression [95, 96]. In other words, the KANSL1 SVA may contribute to increased methylation within the *KANSL1* gene which leads to increased expression. However, when *KANSL1* expression was compared to cg18699337 methylation levels a correlation between the two was not apparent (**Figure 4.12**). It should be noted that this lack of correlation was established using the 125 NABEC samples for which hg38 WGS, expression data and methylation data were all available – which is around half of the datapoints available within the expression data (~250) or one third of the samples included in the wider WGS dataset (~360). The 125 samples in the overlap of expression and methylation datasets therefore represents a relatively small fraction of the NABEC cohort, and it is reasonable to speculate that future expansion of this comparison into considerably larger datasets may reveal an association between the two variables.

Additionally, it was observed via PCR that a single NABEC DNA sample harboured a KANSL1 SVA insertion allele with a shorter CT element than the others (**Figure 4.8a**), and it was possible to generate short and long allele-specific proxy SNPs in very high

LD (r^2>0.95) for each (**Table 4.2**). It was presumed that this might unveil any gene expression heterogeneity among those harbouring the H2 haplotype arising from previously unmeasured sequence variation within the KANSL1 SVA CT element. For instance, CT repeat length variants may carry different of TF binding site copy numbers [180]. Alternatively, if the KANSL1 SVA mediates aberrant splicing of *KANSL1* then CT element length may be a crucial predictor of overall gene expression levels, as is the case for the disease-causative intronic SVA insertion within *TAF1* in XDP [183-185]. However, isoform-specific transcriptomic data was not available for NABEC, and therefore assessment of CT element or whole KANSL1 SVA influences on splicing could not be undertaken. For examination of changes in overall transcription, CT-specific proxy SNPs predicted that in the wider NABEC cohort the only occurrence of the short allele was the one already detected by PCR, and separation of this datapoint from the rest of the NABEC expression data did not suggest any obvious transcriptional changes associated with this genotype (**Figure 4.13**). At this sample size a role for KANSL1 SVA CT repeat length variants in modulation of gene expression cannot be ruled out, but armed with the proxy SNPs identified here it should be relatively easy to repeat this investigation in larger cohorts with WGS and expression data. Similarly, access to a greater number of DNA samples with matched genotyping and phenotypic data might allow for identification of KANSL1 SVA central VNTR or poly-A length variants and relevant proxy SNPs, which may prove informative.

Sample size has been suggested consistently to be a limitation of the NABEC datasets, and it was therefore investigated whether the analyses performed in NABEC using

proxy SNPs might be repeated in a larger cohort. Access was gained to the multi-cohort PD initiative AMP-PD, which contained ~2700 datapoints with expression and WGS data in its 'v1 release' and represented a significant expansion in size over NABEC. It was initially investigated whether KANSL1 SVA presence or absence could be inferred in AMP-PD WGS using proxy SNPs identified in NABEC DNA samples and WGS, and whether these would allow observations made in NABEC to be recapitulated. In order to infer KANSL1 SVA presence in the hg19-annotated AMP-PD genotyping data, a new presence/absence proxy SNP in hg19 was necessary; this led to the identification of a top performing SNP, rs200610218, which was in perfect LD with the KANSL1 SVA in NABEC hg19 WGS data. As an initial analysis all *KANSL1* expression data that was available for AMP-PD, regardless of PD diagnosis, was stratified based on inferred KANSL1 SVA genotype. This suggested that *KANSL1* expression was positively associated with KANSL1 SVA allele dosage, just as it was in NABEC (**Figure 4.14**). However, the chosen proxy SNP predicted that out of nearly 2700 individuals only 6 of them harboured the KANSL1 SVA +/+ RIP genotype. Since the SVA insertion was in high LD with the H2 haplotype which has an allele frequency of ~25%, it had been expected that this haplotype and presence of the KANSL1 SVA would occur in around 170 samples. This raised questions regarding the accuracy of this proxy SNP in the AMP-PD genotype dataset and data access limitations prevented troubleshooting is this disparity, so these data were therefore not used for further analyses. However, AMP-PD should not be disregarded as a resource for investigation of the effect the KANSL1 SVA. All of the primers necessary for genotyping of the SVA RIP and the SVA component elements have been designed and optimised, and obtaining DNA samples from the AMP-PD initiative should not be

overly problematic. Therefore, genotyping of the KANSL1 SVA in a subset of AMP-PD DNA samples to generate accurate proxy SNPs for use in the wider WGS dataset represents a relatively straightforward way to improve the investigation of the SVA's influences in this cohort.

Having extensively characterised the relationship between KANSL1 SVA RIP genotype and expression of *KANSL1*, the gene in which it resides, the scope was expanded and genes in the wider *MAPT* locus were examined for influence of the SVA. Specifically, a shortlist of 21 genes at this locus with expression levels previously found to correlate with the presence of SNP alleles (eQTLs) [233], were assessed. This study was selected as a benchmark because the authors further classified these genes by whether their differential expression could be explained by the H1/H2 haplotype of each individual. By stratifying the 15 transcripts that were available in NABEC, plus the notable PD candidate genes *MAPT* and *WNT3* (*MAPT* and tau pathologies discussed in **Section 4.1**), by KANSL1 SVA RIP genotype (as determined by proxy SNPs) it was found that the list of genes with expression levels associated with H1/H2 haplotype determined by O'Brien *et al* could be reproduced (**Table 4.3**). Importantly, stratification of RNA-seq by SVA genotype identified 2 additional differentially expressed genes: *WNT3*, previously nominated as the PD-causative gene at the *MAPT* locus in a GWAS meta-analysis [23], and *ARL17B*, a novel association. Briefly, the Wnt proteins are a family of secreted signalling proteins that act through β-catenin and have been proposed to govern neuronal health, protection and regeneration in neurons, and this signalling pathway has been observed to be dysregulated in PD

[235]. *ARL17B*, meanwhile, is a poorly characterised gene with predicted GTP binding activity and predicted roles in intracellular transport (www.genecards.org/). Interestingly, however, expression of *ARL17B* was found to be associated with risk SNPs for Alzheimer's disease and progressive supranuclear palsy [236, 237]. To expand upon the links between H2 haplotype and local gene expression studied by O'Brien *et al*, the association between these genes and dosage of the H2-tagging SNP rs8070723G was examined here in NABEC data and compared to the findings for the KANSL1 SVA. It was found that the KANSL1 SVA was a similar but inferior predictor expression of the 10 previously H2-associated genes than the H2 proxy SNP, while this SNP failed to replicate the SVA's significant association with *WNT3*. Interestingly, while the H2 proxy SNP was found to be associated with *ARL17B* it was a weaker predictor of expression than the KANSL1 SVA. Altogether these data suggest that the KANSL1 SVA may be a primary driver of expression patterns of *ARL17B* and the potentially important PD gene *WNT3*, and may contribute to gene expression patterns across the wider H1/H2 *MAPT* haplotype polymorphism. Notably, these data add to the emerging association between *ARL17B* and neurodegenerative diseases by nominating the KANSL1 SVA as potential mediator of its expression.

The observation that the KANSL1 SVA RIP genotype displayed associations with gene expression that broadly overlapped with that of a H1/H2-tagging SNP but diverged significantly for *WNT3* expression indicated that the KANSL1 SVA may be a distinct indicator of gene expression at the *MAPT* locus. This was supported by the earlier finding that the KANSL1 SVA proxy SNP was not in perfect LD with that of a H1/H2-

tagging SNP (rs8070723) in the NABEC cohort (r^2=0.9855, D'=1), indicating that the two genotypes might meaningfully diverge in linkage. This was confirmed by comparison of SVA and H1/H2-tagging SNP genotypes in the 94 NABEC samples for which KANSL1 SVA genotype was validated by PCR; it was observed that there were two samples in which an allele carrying the H2 haplotype, as indicated by rs8070723, did not harbour a KANSL1 SVA insertion. This demonstrates that the KANSL1 SVA is present at most but not all H2 alleles in the human population. As postulated previously, this imperfect linkage may be the result of the KANSL1 SVA insertion into the H2 haplotype being more evolutionarily recent than the inversion event and not yet becoming fixed in the populace, as is typical of RIPs. Presently, no sub-haplotypes of H2 have been identified [238], and therefore the modest divergence in association between the SVA insertion and H2 haplotype are unlikely to result from an ancestral sub-haplotype-forming rearrangement. This observation that the KANSL1 SVA may rarely be found to be absent on the H2 haplotype supports the finding that there is an association between SVA genotype and *WNT3* expression that was not explained by H1/H2 haplotype, and suggests that presence or absence of the KANSL1 SVA is a previously undescribed regulator of this putative PD gene at the *MAPT* locus.

Considering this potential role for the KANSL1 SVA as functional contributor to eQTLs at the H1/H2 haplotype, steps towards *in vivo* or *in vitro* examination were taken. It was found that none of the established human cell lines readily available in the lab were KANSL1 SVA +/+ (**Figure 4.15**), and so CRISPR-mediated deletion of the element was not pursued since the important comparison between +/+ and -/- genotypes

could not be made as was done for the LRIG2 SVA (**Section 3.2.9**). Had time permitted, a larger panel of cell lines could have been explored. To this end, it was speculated that the search for a KANSL1 SVA +/+ cell line could be narrowed down considerably by first examining genotyping data of established or patient-derived lines for SNPs that tag the H2 haplotype, as any cell lines that were homozygous for such SNPs would be anticipated to be KANSL1 SVA +/+.

Instead of deletion by CRISPR it was decided that the KANSL1 SVA, amplified from a NABEC sample, would be characterised via insertion into the luciferase reporter gene construct pGL3P and its influence determined. This was met with partial success, with the SVA inserted into the subcloning vector pCR-Blunt in both orientations and upstream of pGL3P's minimal promoter region in the antisense orientation (**Figure 4.18**). However, it would be prudent to also assess the KANSL1 SVA in the sense orientation in the promoter region of pGL3P, as its regulatory influences may differ with direction. Since the subcloning of the KANSL1 SVA from pCR-Blunt into pGL3P in the sense orientation failed due to incomplete digestion by XhoI (leading to ligation of a linearised SVA-pCR-Blunt construct into pGL3P), a minor adjustment such as extension of digestion time might easily lead to success of this strategy. Furthermore, it was noted that in the pCR-Blunt construct containing the sense-oriented KANSL1 SVA (KANSL1_SVA_S_pCR_Blunt) BamHI recognition sites are located 5' of the SVA and within the 3' end of the SVA, while a BamHI cut site is found 3' of the *luc+* gene in pGL3P (**Figure 4.18a**). Therefore, it should be possible to subclone the KANSL1 SVA with a short truncation from KANSL1_SVA_S_pCR_Blunt into the region downstream

of the luciferase reporter in pGL3P. Doing so may be informative as this is more representative of the genomic context of the KANSL1 SVA insertion relative to *KANSL1*, since the SVA is located within an intron and not upstream of the gene's promoter. Indeed, a pGL3-Enhancer vector available from Promega has an SV40 enhancer inserted into this site (promega.co.uk, catalogue number E1771). The generation of the KANSL1_SVA_S_pCR_Blunt construct in this thesis should make this subcloning strategy readily achievable. This KANSL1 SVA 'enhancer region' pGL3P construct in combination with the plasmids containing promoter region SVA insertions should then permit a multifaceted *in vitro* characterisation of the KANSL1 SVA.

If the KANSL1 SVA proves to be functional in these gene reporter assays it was postulated that future work might involve a 'knock-in' strategy for *in vivo* study of the element, since it may be difficult to find a karyotypically normal established cell line that is +/+ for the KANSL1 SVA in order to pursue a knock-out model. Such a knock-in strategy would likely involve identification of a suitable cell line that is -/- for the SVA and using CRISPR to induce DSBs at the site where the element should be. Provided together with the CRISPR plasmid would be a second plasmid which contains the KANSL1 SVA along with flanking sequence homologous to the sequence flanking the genomic cut site. In the event that the genomic DSB is repaired via homology-directed mechanisms using the provided SVA-containing repair template (as opposed to NHEJ), the KANSL1 SVA will be inserted into the locus. In this way, the SVA may be introduced to the genome of a cell line and then its phenotype may be compared to

that of the parental cell line. This approach would allow the influence of the KANSL1 SVA RIP to be assessed in otherwise genetically identical backgrounds without the need for a cell line that is +/+ to be identified for CRISPR-mediated deletion. However, this strategy is not without its own hurdles – for instance, homology-directed repair is largely offline in interphase cells and may therefore be inefficient, and it may be difficult to insert a large and repetitive element such as an SVA retrotransposon.

In this chapter a predicted novel SVA RIP associated with the H2 inverted haplotype at the *MAPT* locus on chromosome 17, referred to here as the KANSL1 SVA, was validated and characterised. Its presence was found to correlate with increased expression of the gene in which it resides, *KANSL1*, which may lead to upregulation of PINK-1-mediated mitophagy and confer the protective effects against PD that have been associated with the *MAPT* H2 haplotype. KANSL1 SVA RIP genotype was further correlated with other genes previously associated with the haplotype, in addition to the genes *WNT3* and *ARL17B* that had not previously been linked to H1/H2. The greater predictive power of gene expression by SVA genotype than haplotype markers alone was supported by the observation that the KANSL1 SVA was observed to not be in perfect linkage disequilibrium with the H2 haplotype. This suggested that presence of the KANSL1 SVA was not only variable within H2, but that it may be at least partly causative of gene expression differences – potentially explaining some of the transcriptional divergence from the H1 haplotype, which is not thought to harbour the SVA. This could result from several properties associated with SVAs, such as the insertion introducing additional TF binding sites or a novel CpG island to the

locus, or via epigenetic changes around the SVA resulting from repressive targeting by KRAB-ZFPs and KAP1. To specifically quantify the regulatory influences of the KANSL1 SVA separate from the wider H2 haplotype, the groundwork was laid for insertion of the element into a reporter gene construct and a path to its completion was delineated. Further, a framework for knock-in of the SVA into cell lines for *in vivo* analysis of its genomic impact was envisaged. Altogether, this work lays the groundwork for investigation into a novel RIP that may constitute a functional regulatory element associated with the H2 inverted haplotype, which may be contributing to differences in gene expression patterns between H1 and H2. Considering the established association between the H2 haplotype and a decreased relative risk of PD, KANSL1 SVA RIP genotype might ultimately supersede H1/H2-tagging SNPs as a predictive marker for PD – and find use in PRS, for example.

Chapter 5 Leveraging genome-wide datasets to assess contributions of retrotransposons to 3D chromatin structure

5.1. Introduction

Thus far, the genomic *cis*-regulatory influences of selected SVA retrotransposons have been examined in both a generic model system and in a PD-relevant context. This was primarily accomplished in cohort datasets and utilising a CRISPR-generated cell line model via study of how SVA RIP genotype correlated with gene expression and methylation at SVA loci. As discussed previously, there is a growing body of evidence that TEs harbour TF binding sites which can be introduced to novel loci upon TE mobilisation (**Section 1.2.8**). Of particular note is TE-associated binding of the chromatin architectural protein CTCF, which has been shown to bind genomic SINE, LINE and ERV elements in humans and several other mammals [159, 161, 239]. As described previously, CTCF binding has also been demonstrated at human SVA retrotransposons [180].

Among other functions, CTCF can mediate intrachromosomal looping of chromatin. Indeed, it is well established that pairs of distant chromosome loci can come together in 3D space, making stronger interactions than with intervening loci [240]. ~30% of these chromatin loops bring promoters and enhancers together [241], which mediates changes in gene expression by bringing regulatory elements and bound TFs to the promoter region [242]. Enhancer elements can be located kilobases or even megabases from the gene they influence [243], and it has been estimated that on average a gene in any given cell type is regulated by 4 distinct enhancers at any given moment [244] – although genes important for development may be under the influence of tens of enhancers [245]. Furthermore, it has been demonstrated that

genes can become colocalised at specific 'transcription factories' within the nucleus to enable efficient coregulation via an increased concentration of TFs and RNA polymerases [246, 247]. Chromatin looping also permits formation of topologically associated domains (TADs), which are chromosomal partitions ranging from 40 kb to 3 Mb in humans within which loci interact at high frequency [241, 248]. By contrast, loci within different TADs interact at lower frequencies even if they are proximal on the linear genome (compare enhancers in **Figure 5.1c**) [248]. Consequently, enhancer-promoter interactions are largely constrained to occur within TADs [241, 248]. While enhancer-driven gene expression is largely resistant to inactivating mutations due to redundancy among multiple enhancers at a given gene [243, 249], by contrast gene dysregulation resulting from disruption of TADs has been documented to cause congenital limb malformations and cancer [250, 251]. Chromatin looping, whether to directly form enhancer-promoter interactions or TADs, is therefore a crucial determinant of gene expression patterns that is increasingly important for understanding transcription in health, development and disease.

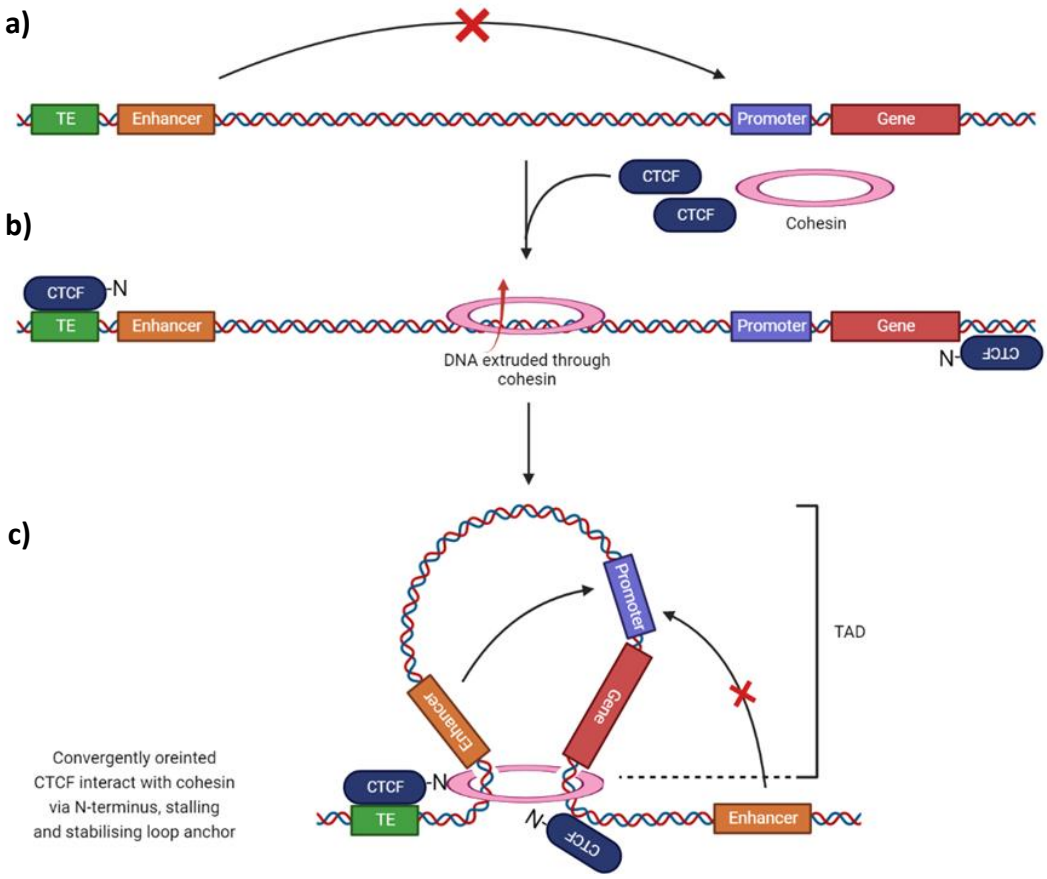


Figure 5.1 – Chromosome looping with CTCF and TEs. a) Before chromatin looping an enhancer element is too distant from a promoter to influence its activity. **b)** CTCF binds its cognate DNA motifs, one of which was introduced by a TE insertion. These CTCF binding sites are convergently oriented, meaning that the N-termini of the two CTCF proteins are directed towards each other. Cohesin is loaded onto the DNA and begins to move DNA through it in a 'loop extrusion'. **c)** DNA extrusion pauses when cohesin reaches a bound CTCF protein. Additionally, the N-termini of the CTCF proteins interact with cohesion and block binding of the cohesion unloading factor WAPL, thereby preventing unloading and stabilising the loop structure. A topologically associated domain (TAD) has been formed within which the original enhancer can now interact with the gene promoter. By contrast, an enhancer outside of the TAD is constrained from reaching the promoter.

As mentioned previously, a key effector of this 3D genomic architecture is the factor CTCF. The protein features a central DNA binding domain composed of 11 zinc finger motifs which is flanked by unstructured N- and C-terminal domains [252], and it is hypothesised that CTCF is capable of binding diverse DNA sequences using varying combinations of its 11 ZFs (**Figure 5.1**) [253]. CTCF binding sites are found at regions that undergo chromatin-chromatin interactions such boundaries of TADs and sub-TADs where it interacts with cohesin, a ring-shaped multi-protein complex that encircles double-stranded DNA [254]. It has been established that genome domain formation occurs via a 'loop extrusion' process in which a cohesin ring extrudes chromatin until it stalls at convergently oriented CTCF binding sites (**Figure 5.1b**) [255]. This ultimately results in stabilisation of the chromatin loop via antagonisation of the cohesin unloading factor WAPL [256]. It was shown that this is mediated by the N-terminal region of CTCF (**Figure 5.1c**), thereby providing a mechanistic explanation for the directional requirement of CTCF binding sites in loop anchors [256]. The CTCF C-terminal domain, meanwhile, plays a minimal role in blocking extrusion by cohesin but has been demonstrated to promote protein stability and provide RNA binding activity, which facilitates CTCF clustering [256, 257].

That CTCF has been repeatedly observed to bind TEs is therefore of great interest [159, 161, 180, 239], as this suggests that TE insertion may introduce new sites for chromatin loop anchor formation. Indeed, a considerable fraction of CTCF binding sites are derived from various classes of TEs – in one study, overlapping CTCF ChIP-seq data with TE coordinates from RepeatMasker indicated that 22.8% of CTCF

binding sites are derived from TEs in the human genome, while this proportion is 40% in mice [161]. Importantly, this included TEs from the *Alu* (listed in this study as SINEs), L1, HERV and DNA transposon families [161]. Ongoing TE mobilisation results in species-specific TE-associated CTCF binding sites – the most numerous contributors of which are HERVs in humans and B2 SINE elements in mice [159, 239]. However, the occurrence of species-specific TE-associated CTCF binding is at odds with the high level of conservation between human and mouse TAD boundaries [248], since novel TE-associated CTCF binding sites would be expected to result in formation of unique TADs. Using Hi-C (a high-throughput variation on chromosome conformation capture) and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET, a technique similar to Hi-C that incorporates immunoprecipitation) a recent study provided one explanation that reconciles these contrasting observations by demonstrating that novel CTCF binding sites introduced by TEs can replace ancient binding sites at TAD boundaries or provide additional redundant sites [258]. In this way, transposition of CTCF motif-bearing TEs can facilitate conservation of 3D genomic structure instead of disrupting existing topology with every insertion event [258]. In contrast, however, recent analysis has suggested that novel CTCF binding and domain formation associated with TE insertion does indeed occur, and moreover is more likely to involve gene-sparse regions [259]. Additionally, recent studies have demonstrated that transcription of some TEs can form strong TAD boundaries which may potentially influence gene regulation [260, 261], although this is yet to be linked to specific diseases. Taken together, these studies suggest that TEs play a role in architectural innovation of the genome that is tempered by their co-option by the host as sources of genomic

conservation, likely making the impact of TEs on the 3D genome more subtle and context sensitive.

Therefore, it was speculated that changes in genome topology associated with presence of TEs could influence intracellular processes relevant for genetically complex disease. As in previous chapters, this hypothesis postulates that TEs are a source of genomic variation relevant to PD that is not fully appreciated by current short-read DNA sequencing approaches due to difficulties in read mapping. Moreover, it is conceptually appealing for TE involvement in PD aetiology to be at least partially mediated by changes to genome structure via CTCF: it is well established that CTCF binding is precluded by DNA methylation [262], and there is growing evidence that PD is strongly associated with dysregulated methylation patterns (reviewed in [263]). Although changes in gene regulation resulting from perturbation of chromatin looping are an attractive candidate mechanism for unexplained pathogenicity of non-coding PD-associated SNPs, this has only begun to be explored relatively recently: One such integrative analysis of chromatin looping has identified putative target genes that make long distance interactions with PD-associated SNPs [264]. Building upon this, a more recent study utilising established Hi-C libraries and transcriptomic datasets found that expression of 518 genes were associated 76 PD SNPs across 49 tissues, and, importantly, that one third of these associations were mediated *in trans* by long-range chromatin interactions (defined as >1 Mb chromatin looping or between different chromosomes) [265]. Similarly, recent work comparing histone marks and chromatin conformation data in the PD

and control substantia nigra identified 656 genes that make 3D interactions with PD risk SNPs and enhancer elements that are perturbed in the disease state [266]. Taken together, these studies support an emerging role for dysregulated chromatin architecture underpinning PD. Crucially, to the best of our knowledge no study has attempted to systematically incorporate the role of TEs into this picture at the time of writing. Therefore, TEs may currently represent largely uncharacterised CTCF binding sites that might direct altered chromatin looping in PD, perhaps via acquisition of altered methylation and disrupted CTCF binding in the disease state, resulting in gene expression patterns.

To investigate chromatin looping in this context, collaborators at NIH, Maryland, USA provided a Hi-C dataset produced in iPSCs derived from PD and control individuals, before and after a dopaminergic neuronal differentiation protocol (this cell type is of particular interest in PD, as its degeneration is classically associated with the disease [3]). These data were produced as part of The Foundational Data Initiative for Parkinson's Disease (FOUNDIN-PD, **Section 2.2.1.4**). Briefly, Hi-C is an 'all-vs-all' study of chromatin looping in which interacting regions are crosslinked, sheared from surrounding genome, overhangs are filled in with a biotin tag, the two DNA fragments are ligated together, and finally purified via streptavidin pulldown. Upon sequencing, the two DNA fragments can be identified and it can be determined which distant parts of the genome were interacting in 3D space, with transient interactions filtered from stable loop anchors by virtue of detected interaction frequency [267].

These Hi-C data, which essentially list pairs of genome coordinates corresponding to chromatin loop anchors, were overlaid with the hg38 coordinates for genes and both LTR (i.e., HERVs) and non-LTR retrotransposons to assess the involvement of TEs in loops at gene loci. It was anticipated that this would yield novel insights into TE colocalisation with gene-associated loop anchors (herein GALAs), and how this may differ in control and PD iPSC lines. Importantly, the availability of Hi-C data before and after dopaminergic differentiation permits the additional study of TE association with GALAs, and how this might change with PD, in an additional temporal dimension. Aside from the context of PD, simply examining any changes in TE colocalisation with GALAs upon differentiation of these iPSCs may lead to insights into TE contribution to 3D genome structure during development and nominate loci for functional validation. Additionally, in a subset of iPSC lines a more targeted analysis was performed in which GALA coordinates were laid over *de novo* retrotransposon annotations in the WGS data of these lines. These annotations were produced by collaborators at FOUNDIN-PD project using the Mobile Element Locator Tool (MELT), which identifies novel TE insertions via features such as target site duplications in short-read WGS data that are absent in the reference genome [218]. In this way, the potential for specific chromatin loop changes arising from RIPs could be investigated.

5.1.1. Aims

The importance of TE associations with GALAs in iPSC lines will be examined by:

- Overlapping Hi-C loop anchor coordinates with coordinates of genes and those of TEs, which were derived from the reference genome (hg19, lifted to

hg38 coordinates), a database of non-reference TE insertions (hg38) and *de novo* annotations from the MELT programme (**Section 2.2.1.4**, and descriptions in relevant results section).

- Comparing occurrence of these TE-associated GALAs prior to and after dopaminergic neuronal differentiation of the iPSC lines, to assess developmental trends.

- Comparing control and PD cell lines within the iPSC sample set, to assess disease-associated changes in TE colocalisation with GALAs. Moreover, the analysis was focussed by subsequently only considering involvement of TEs in loop anchors at PD-relevant loci.

5.2. Results

5.2.1. Intersection of gene, TE and iPSC Hi-C coordinate data

Hi-C data (briefly, pairs of genomic coordinates of regions found to interact in 3D space) from 8 iPSCs lines from FOUNDIN-PD were provided by collaborators at NIH, Maryland, USA. Data for the 8 lines were for undifferentiated states (day 0) and after 65 days of a dopaminergic differentiation protocol, and were provided pre-processed and quality controlled. A general framework for assessment of colocalisation of chromatin loop anchors (as defined in Hi-C), TE and genes was established (**Section 2.2.1.4**). First, the 'intersect' function of the Bedtools genomics suite (hosted on the University of Liverpool's compute cluster) was used to overlap loop anchor coordinates with TE coordinates (overlaps visualised in **Figure 5.2**). These TE coordinates were derived from either the reference genome TE annotations provided by RepeatMasker on the UCSC genome browser, non-reference TE annotations as listed in the gnomAD-SV database of known structural genomic variants or *de novo* annotations of non-reference TEs generated in iPSC WGS using MELT – further details of these TE datasets are provided in the relevant results sections. This essentially expanded the Hi-C dataset by listing all TEs that overlapped with either loop anchor, and retained all loops with no TE overlaps at this stage. This list of chromatin loops with TEs was then overlapped with gene coordinates from the 'curated' subset of NCBI RefSeq Genes (https://www.ncbi.nlm.nih.gov/refseq/about/), which was selected as each gene was represented by a single isoform and it contained no unvalidated predicted transcripts. Details on data providence, use of Bedtools and example code are provided in **Section 2.2.1.4**. The output of this process were files that listed all chromatin loop anchors along with all overlapping TEs and genes, if any

were present (**Figure 5.2a, b & c**). It should be noted that regions flanking Hi-C, gene or TE coordinates were not included when these genomic elements were intersected, to produce an analysis more focused on regions with direct overlap with TE sequences. Moreover, Hi-C loop anchors ranged from 5 kb to 25 kb in length, and therefore it was deemed unnecessary to examine wider genomic windows.

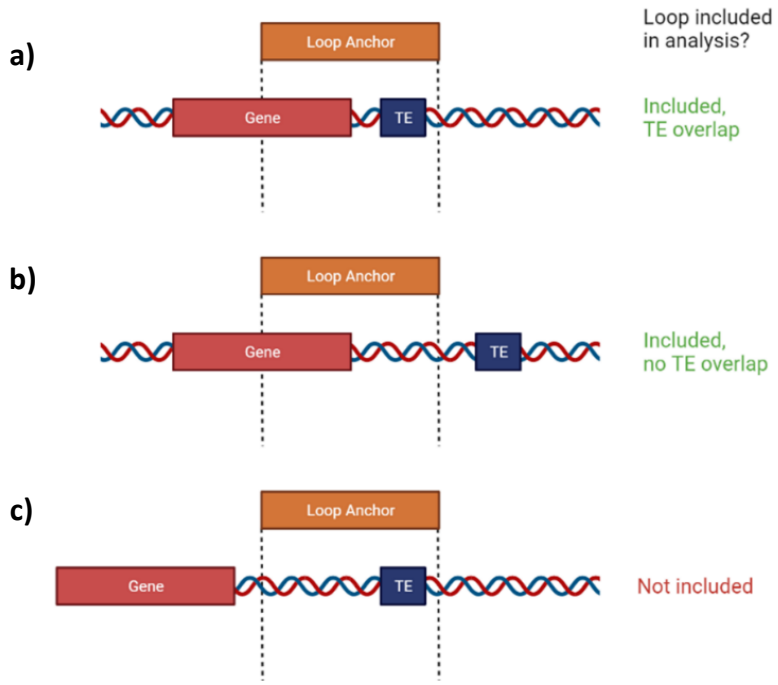


Figure 5.2 – Illustration of Hi-C, TE and gene coordinate overlaps and inclusion criteria when considering loop involvement at transcribed regions. All loop anchors that overlapped with protein-coding regions were taken forward and then separated into those with and those without TE overlap. Loop anchors that did not overlap gene coordinates were excluded, even if they included a TE, in order to focus on how TEs might influence gene-associated looping.

From this intersection of loop anchors, TEs and transcribed regions, retention of all loop anchors featuring TEs would have created an analysis that was overly broad since retrotransposons such as *Alu* are so commonplace in the genome [60]. However, discarding of all *Alu* was deemed undesirable as they could potentially contribute to chromosomal architecture. It was anticipated that by instead considering loop anchor prevalence at gene coordinates a more refined analysis could be produced, as gene-associated loop anchors (GALAs) would be restricted to a fraction of the genome with putatively meaningful genomic consequences such as bringing gene bodies and regulatory elements together in 3D space. Moreover, GALAs were expected to vary depending upon developmental and PD states due to differences in the long-range interactions of regulatory elements under these intracellular conditions. In light of the observed roles for TEs in CTCF binding and TAD formation, a key question is whether TEs are drivers of any differences in GALA formation in development and disease. Therefore, loop anchors that did not overlap with transcribed regions were discarded (**Figure 5.2c**) in order to focus on how TE-mediated looping may affect gene regulation. Subsequent analysis consequently centred on comparing the number of GALAs with TEs (**Figure 5.2a**) to GALAs without TEs (**Figure 5.2b**), and how this varied with differentiation and PD.

In summary, the proportion of GALAs that overlapped reference TEs, known coordinates of non-reference TEs and novel annotations of non-reference TEs was examined using the Hi-C data of iPSC lines from FOUNDIN-PD. Given that the iPSC lines were relevant to PD, later in this chapter these analyses were refined to only

consider loop anchors at genes which have been nominated as functionally associated with PD SNPs [23] – herein PD gene-associated loop anchors (PD GALAs). For clarity and future reference, the groups resulting from intersection of loops anchors, reference TEs, non-reference TEs, PD genes and non-PD genes are summarised in **Figure 5.3**:

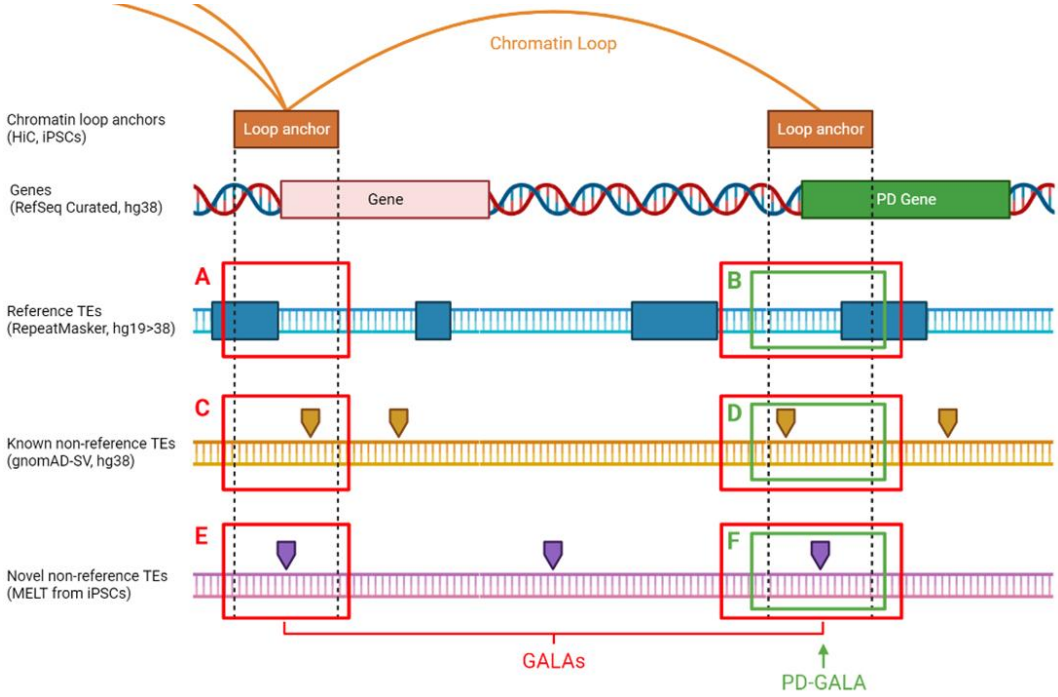


Figure 5.3 – Schematic summarising overlaps between chromatin loop anchors in iPSCs and TEs from various sources at non-PD and PD-relevant gene loci (red and green blocks, respectively). All GALAs are highlighted in red while PD GALAs are highlighted in green, such that both groups can be visualised against each TE dataset: **A)** All GALAs + reference TEs. **B)** PD GALAs + reference TEs. **C)** All GALAs + known non-reference TEs. **D)** PD GALAs + known non-reference TEs. **E)** All GALAs + novel annotations of TEs in FOUNDIN-PD iPSCs. **F)** All GALAs + novel annotations of TEs in FOUNDIN-PD iPSCs. It should

be noted that when 'All GALAs' are assessed this includes PD GALAs, indicated by the red box surrounding the green box at PD-GALAs.

5.2.2. Upon dopaminergic differentiation of iPSCs, there were no significant differences in reference genome TE colocalisation with GALAs when all genes were considered

To assess TE involvement in GALAs before and after differentiation of iPSCs, Hi-C data were first overlapped with TEs included in the reference human genome (see **Figure 5.3a**). Coordinates for 'reference TEs' were obtained from the RepeatMasker annotation of the human genome hosted on the UCSC genome browser. It was noted that the hg38 RepeatMasker annotation contained many small DNA fragments (<200 bp) annotated as TEs, which were absent in the hg19 version. The investigation undertaken here aimed to access the impact of full-sized TEs rather than fragments on chromatin looping, and so the hg19 RepeatMasker annotation was therefore converted to hg38 for intersection with Hi-C data (**Section 2.2.1.4**). The resulting list of chromatin loop anchors featuring TEs was then overlaid with gene coordinates, and any anchors that did not overlap gene bodies were discarded (**Section 5.2.1**). For each of the 16 available samples (the 8 iPSC lines before and after differentiation) a varying number of Hi-C datapoints were available, such that the intersection of loop anchor, reference TE and gene coordinates produced a total number of gene-associated chromatin loops with a TE in at least one anchor that ranged from 273,908 to 3,120,886 and had a mean of 1,744,922 (**Table 5.1**):

Sample			Total loops		Unique loops		
#	Category	Gender	Day 0	Day 65	Day 0	Day 65	Shared
1	Control	Male	2,287,771	2,630,062	2,175,488	2,517,779	112,283
2	Control	Female	306,440	2,295,179	273,908	2,262,647	32,532
3	Control	Female	2,009,899	3,120,886	1,772,694	2,883,681	237,205
4	Control	Female	1,662,416	3,046,600	1,534,242	2,918,426	128,174
5	Control	Female	406,053	3,005,997	352,677	2,952,621	53,376
6	PD	Male	323,897	653,799	292,931	622,833	30,966
7	PD	Male	1,672,582	701,814	1,640,502	669,734	32,080
8	PD	Male	1,662,747	2,930,718	1,491,262	2,759,233	171,485

Table 5.1 – Numbers of chromatin loops featuring gene loci and TEs in the Hi-C data from FOUNDIN-PD iPSC lines. Arbitrary sample number is listed along with patient PD category and gender. The total number of loops for each sample before (day 0) and after (day 65) a dopaminergic neuronal differentiation protocol, in addition to the number of chromatin loops that were unique to each time point and those that were shared between the two.

Initially, it was assessed whether differentiation of iPSCs was associated with changes in TE colocalisation with chromatin loop anchors at all gene loci in the human genome. Given the wide range of total number of observed TE overlaps with GALAs in the 8 iPSC lines at the 2 timepoints (**Table 5.1**), for each cell line the proportion of GALAs that overlapped a TE was expressed as a percentage of all GALAs that were observed in that line. In doing so, the TE overlap with GALAs was normalised across iPSC lines and their proportional involvement with GALAs could be compared more robustly. All reference TEs were counted together and also as the separate families of *Alu*, HERV, L1 and SVA. A Shapiro-Wilk test indicated that all of these groupings ('All TEs, *Alu*, HERV, L1 and SVA) were non-normally distributed, and so a paired Wilcoxon signed-rank test was used to assess differences in TE-associated looping at day 0 and day 65 of the iPSC differentiation process. It was observed that there

was no change in TE involvement with GALAs upon dopaminergic differentiation of iPSCs, whether all reference TEs were considered together or broken down into individual families (

Figure 5.4a). It had been expected that by day 65 the percentage of TE-associated loops would have increased overall, as it is known that many chromatin domain boundaries and loop anchors become defined upon differentiation [242]. It is likely that while total numbers of chromatin loop contacts at genes increased, the proportion which overlapped TEs was similar to that prior to differentiation.

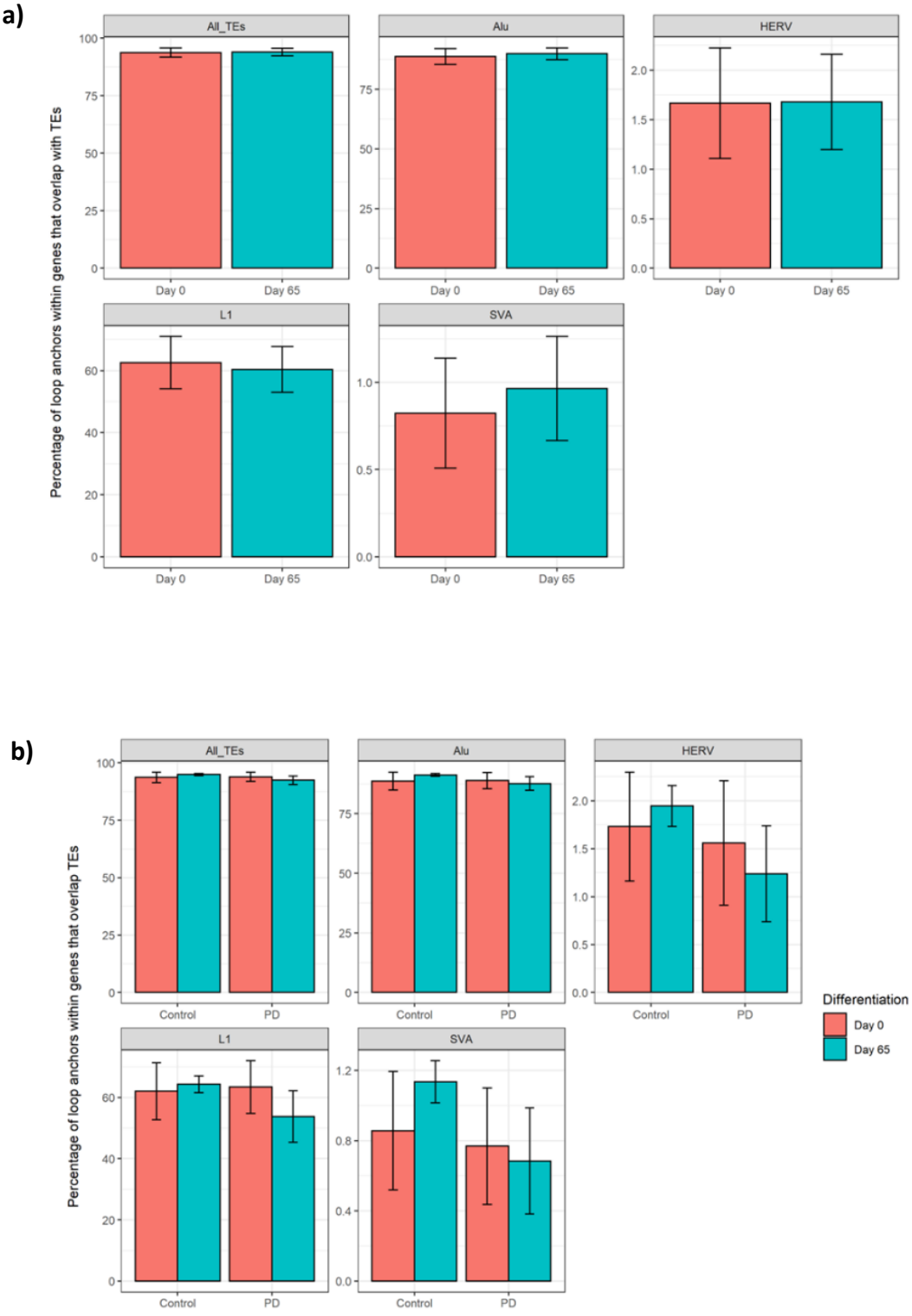


Figure 5.4 – Proportion of all gene-associated chromatin loop anchors that overlap with reference TEs. Chromatin loop anchor coordinates from Hi-C of iPSC lines from FOUNDIN-PD were intersected with reference TE coordinates from RepeatMasker and gene coordinates from the RefSeq hg38 curated

subset. TEs were considered as 'All TEs' and as separate TE families. **a)** TE-associated loops at genes from all iPSC lines were compared at day 0 and day 65 of a dopaminergic neuronal differentiation. Groups underwent paired Wilcoxon signed-rank test. N=8. **b)** Dataset from **(a)** was further broken down into iPSC lines derived from control and PD individuals. Data were divided by diagnosis and differentiation state separately, and each grouping underwent paired Wilcoxon signed-rank test. Control n=5, PD n=3.

Subsequently, the dataset was split to compare iPSCs derived from healthy controls (n=5) and PD patients (n=3). These data were still non-normally distributed (precluding the use of two-way ANOVA for simultaneous examination of effect of diagnosis and differentiation state) so group means were compared using paired Wilcoxon tests; data were split on diagnosis and day 0 of differentiation was compared to day 65, and separately data were split on differentiation state and control samples were compared to PD. It was observed that there were no significant differences in TE colocalisation with GALAs, whether means were compared by differentiation state or diagnosis for any of the TE groupings (

Figure 5.4b). However, a trend was apparent in which overlap with GALAs appeared to increase slightly for all TE groupings upon differentiation of control iPSC lines, while this overlap decreased slightly in PD lines. This hints that the proportion of gene-associated chromatin loop anchors that overlap with TEs might vary in PD.

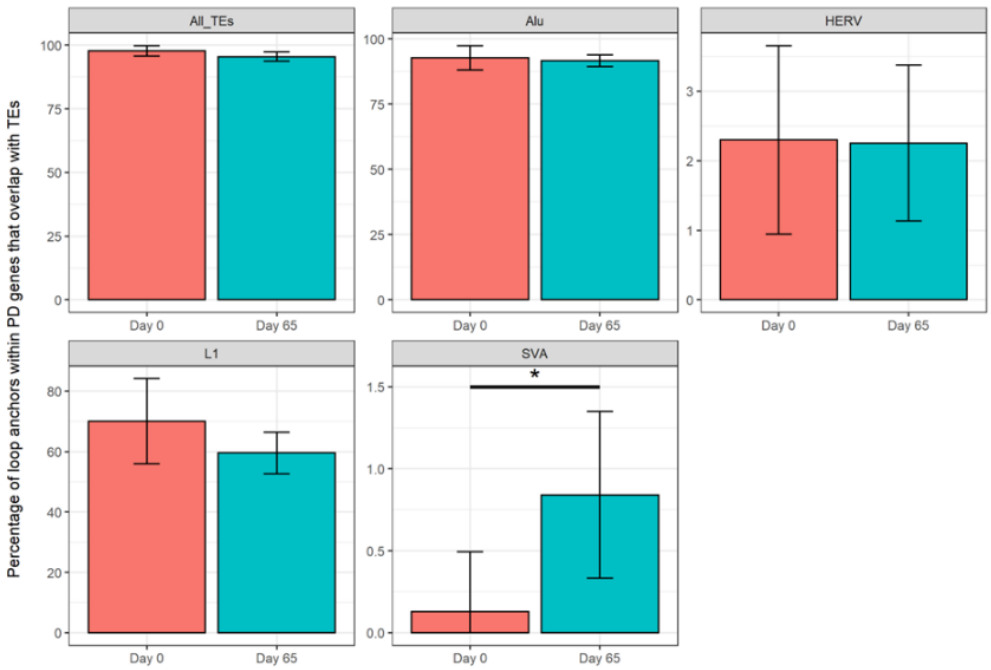
5.2.3. At nominated PD risk genes involvement of reference genome SVAs at GALAs was increased after differentiation in all samples, while involvement of SVAs and HERVs was decreased in PD lines versus controls

To further investigate the small difference in TE colocalisation with gene-associated chromosome loops observed between control and PD iPSC lines, the analysis in

Section 5.2.2 was refined to only consider PD relevant genes (see **Figure 5.3b**). These 'PD relevant' genes were those previously nominated to be functionally associated with PD risk SNPs via QTL analysis in the largest and most recent PD GWAS meta-analysis (Nalls *et al*. 2019, Supplementary Table 2) [23]. For this subset of PD genes it was found that numbers of TE-associated loops for 'All TEs', *Alu*, HERV and L1 groups were normally distributed, while the SVA group was not. Accordingly, when TE colocalisation with 'PD GALAs' was compared between day 0 and day 65 of dopaminergic differentiation for all 8 iPSC lines the SVA group underwent a paired Wilcoxon signed-rank test, while the remainder underwent a paired Student's t-test. It was observed that upon differentiation there was a significant increase in SVA colocalisation with loop anchors at PD genes, and there was no change related to 'All TEs', *Alu*, HERV and L1 groups (**Figure 5.5a**). As before, this dataset was then further split into iPSCs derived from control and PD individuals. Being normally distributed, effects of diagnosis and differentiation state were assessed in 'All TEs', *Alu*, HERV and L1 groups using two-way ANOVA (with Tukey's Honest Significant Difference post hoc test) while the non-normal SVA group underwent multiple Wilcoxon signed-rank tests. Although not achieving statistical significance, it was found that the colocalisation of SVA elements with PD GALAs increased after differentiation in both control and PD iPSC lines (**Figure 5.5b**), consistent with the increase observed when all cell lines were grouped together (**Figure 5.5a**). It was also observed that for HERV and SVA elements that their overlap with PD GALAs was lower in PD lines than for the corresponding timepoint in control iPSCs, finding statistically significant differences for HERVs when making 'Control Day 0 vs PD Day 65' and 'Control Day 65 vs PD Day

0’ comparisons (**Figure 5.5b**). In other words, HERV and SVA colocalisation with PD GALAs was generally lower in PD iPSCs than control lines.

a)



b)

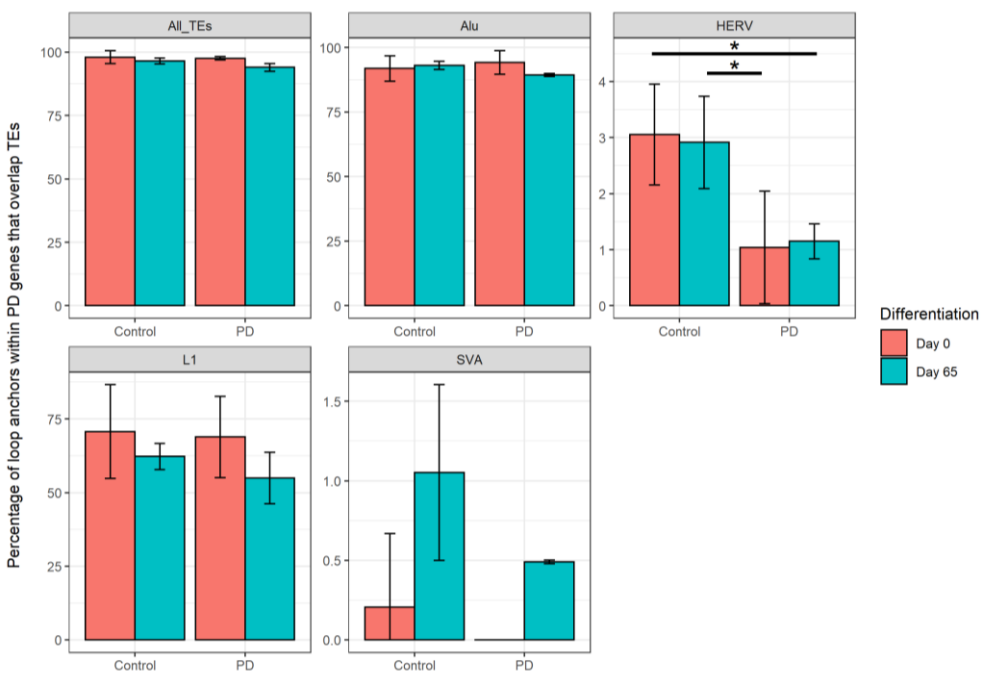


Figure 5.5 – Proportion of PD gene-associated chromatin loop anchors that overlap with reference TEs. Gene coordinates from the RefSeq hg38 curated subset were filtered to only include those previously nominated to be functionally associated with PD risk SNPs [23]. Chromatin loop anchor coordinates from Hi-C of iPSC lines from FOUNDIN-PD were intersected with reference TE coordinates from RepeatMasker and coordinates of the nominated PD genes. TEs were considered as 'All TEs' and as separate TE families. **a)** TE-associated loops at genes from all iPSC lines were compared at day 0 and day 65 of a dopaminergic neuronal differentiation. Groups 'All TEs', *Alu*, HERV and L1 underwent paired Student's t-test, while SVAs underwent paired Wilcoxon signed-rank test. N=8. **b)** Dataset from **(a)** was further broken down into iPSC lines derived from control and PD individuals. Groups 'All TEs', *Alu*, HERV and L1 underwent two-way ANOVA with Tukey's HSD post hoc test, while SVAs were divided by diagnosis and differentiation state separately and underwent multiple Wilcoxon signed-rank tests. Control n=5, PD n=3.

These two observations of significant changes in TE association with PD GALAs – for SVAs generally, and for HERVs when control and PD iPSCs were compared – were investigated further. For individual overlaps the coordinates and subclass of the TE involved were extracted along with coordinates of proximal and distal loop contacts and the names of any genes within these TE-associated loop anchors. Across Day 0 and Day 65 of the dopaminergic differentiation process, 4 different SVAs of the D subfamily were observed to colocalise with the 10 PD GALAs, with 3 of these SVAs appearing in multiple independent loop anchor pairs (**Table 5.2**). Manual inspection of the sequences of these SVA Ds in the reference genome showed that they were all fully intact – possessing a CT element, *Alu*-like domain, central VNTR, SINE region and poly-A signal – and that none were known RIPs listed in dbRIP. It was found that prior

to dopaminergic differentiation there was a single SVA D-associated chromatin loop anchor that overlapped *PRKAR2A* and the nominated PD gene *ARIH2*, which occurred in a single control iPSC line. By contrast, there were 9 SVA-associated loops at PD genes following differentiation, 7 of which involved the nominated PD gene *NEK1* and 2 involved *SCARB2*. These *NEK1*-associated loop anchors colocalised with 3 different SVAs, and the SVAs appear in both the proximal and distal loop anchor to the gene. Notably, at one of these *NEK1*-associated loop anchors featuring an SVA it was observed that the distal loop anchor could be formed from 3 distinct loci, incorporating either *ANXA10*, *CBR4*, or *PALLD*. Additionally, an SVA-associated loop anchor that colocalised with a gene not implicated in PD, *MFAP3L*, was observed to make a looping interaction to *NEK1*. The genes *ANXA10*, *CBR4*, *PALLD* and *MFAP3L* are involved in processes including signal transduction, fatty acid synthesis and cytoskeletal organisation but are not currently implicated in neurodegeneration (www.genecards.org identifiers GC04P168081, GC04M168864, GC04P168497 and GC04M169986, respectively), and therefore a disease-associated coregulatory network with *NEK1* and mediated by SVAs was not supported by this data. Interestingly, there was considerable overlap between distinct chromatin loops that featured the same proximal and distal loci, indicating a degree of redundancy. The proportional incidence of each of these SVA-associated chromatin contacts was slightly increased in PD lines – for the 10 identified loop anchors, they occurred 28% of the time (14 out of 50) in control lines and had 33% occurrence (10 out of 30) in PD lines. It should be noted that this does not directly contradict the data presented in **Figure 5.5b**, which suggests that SVA colocalisation with PD GALAs is decreased in PD iPSCs. The data in **Figure 5.5b** is presented as a percentage of the overall number

PD GALAs, whereas **Table 5.2** presents un-normalised counts of SVA-associated loops. Therefore, while the total number of SVA overlaps with PD GALAs may increase (**Table 5.2**) it is likely that this occurs against a backdrop of increased looping at PD genes that does not involve SVAs, causing their proportional involvement to fall as in **Figure 5.5b**. Altogether, this suggests that upon differentiation there may be some difference between control and PD genomes in formation of chromatin loops at PD gene loci with SVAs, and that the *NEK1* locus in particular may be a prime candidate for further study. Given that long-range chromatin loops have been demonstrated to facilitate enhancer-promoter interactions resulting in gene expression changes and that colocalisation of promoters in 3D space can enable coordinated transcription [241, 242, 246, 247], it is reasonable to speculate that the altered chromatin looping at gene-associated SVA loci observed here might contribute to PD-specific expression patterns of nominated PD genes *NEK1*, *SCARB2*, *ARIH2* or the non-PD genes with which they colocalise.

	TE				Proximal Loop				Distal Loop				Frequency	
Day	Chr	Start	Stop	Name	Chr	Start	Stop	Gene Name	Chr	Start	Stop	Gene Name	Control	PD
0	3	48733458	48735453	SVA_D	3	48725000	48750000	PRKAR2A	3	48950000	48975000	ARIH2	1/5	0/3
65	4	76422691	76424316	SVA_D	4	76420000	76430000	.	4	76170000	76180000	SCARB2	1/5	1/3
65	4	76422691	76424316	SVA_D	4	76420000	76430000	.	4	76180000	76190000	SCARB2	1/5	0/3
65	4	169569425	169571172	SVA_D	4	169550000	169575000	NEK1	4	168175000	168200000	ANXA10	2/5	1/3
65	4	169569425	169571172	SVA_D	4	169560000	169570000	NEK1	4	168180000	168190000	ANXA10	2/5	2/3
65	4	169569425	169571172	SVA_D	4	169560000	169570000	NEK1	4	169010000	169020000	CBR4	0/5	1/3
65	4	169569425	169571172	SVA_D	4	169565000	169570000	NEK1	4	168915000	168920000	PALLD	2/5	0/3
65	4	169569425	169571172	SVA_D	4	169565000	169570000	NEK1	4	168920000	168925000	PALLD	2/5	3/3
65	4	170045721	170047562	SVA_D	4	170025000	170050000	MFAP3L	4	169600000	169625000	NEK1	3/5	0/3
65	4	170045721	170047562	SVA_D	4	170040000	170050000	.	4	169610000	169620000	NEK1	0/5	2/3

Table 5.2 – Breakdown of SVA overlap with chromatin loop anchors at PD genes before and after dopaminergic neuronal differentiation of iPSCs from FOUNDIN-PD. Loops at day 0 are in white, while loops at day 65 are in grey. Nominated PD genes are underlined.

PD GALA overlap with HERVs was also broken down to examine specific TE and gene colocalisation. Likely due to their greater genomic abundance compared to SVA elements, there was a greater number of loop anchors at PD genes that overlapped HERVs than SVAs; a total of 37 chromatin loop contacts were observed before and after dopaminergic differentiation of iPSCs, featuring overlaps with 16 HERV elements (**Table 5.3**). These HERV-associated loops involved a diverse list of nominated PD genes: *RAB29*, *ITPKB*, *MAP4K4*, *KLHDC8B*, *BST1*, *NEK1*, *SH3RF1*, *NDUFAF2*, *ZKSCAN8*, *ZSCAN16*, *ZSCAN16-AS1*, *CTSB*, *SEC23IP*, and *SETD1A*, along with several non-PD genes in the opposite portion of a chromatin loop contact (**Table 5.3**, genes not underlined). Notably, one HERV-associated loop contained the nominated PD genes *SH3RF1* and *NEK1* in opposite anchors of a chromatin loop, suggesting potential for PD-relevant coregulation. Indeed, this interaction was only observed in PD iPSCs, although only in one cell line. The most numerous HERV-associated PD GALAs were those involving the cluster of genes *ZKSCAN8*, *ZSCAN16*, and *ZSCAN16-AS1* on chromosome 6, which made up 15 of 37 observed chromatin loops. It was noted that there was overlap in the coordinates for the multiple distinct HERV-associated loop anchors at each PD gene, which is further suggestive of the loop redundancy postulated previously for SVAs. As expected from **Figure 5.5b**, the occurrence of loops at PD genes featuring HERVs was lower in PD iPSCs than in those derived from controls – for all possible loops in each diagnosis grouping, there was 25% occurrence (46 out of 185) in control lines and 13% occurrence (14 out of 111) in PD lines (**Table 5.3**). Moreover, a given HERV-associated loop anchor at a PD gene occurred in a higher proportion of controls than PD lines for 28 out of 37 loops. In summary, this analysis suggests an altered chromatin landscape around some

genomic HERVs at PD-relevant genes in the PD state versus control individuals, and provides a shortlist of genes where this may be pertinent.

	TE				Proximal Loop				Distal Loop				Frequency	
Day	Chr	Start	Stop	Name	Chr	Start	Stop	Gene(s)	Chr	Start	Stop	Gene(s)	Control	PD
65	1	205777285	205778085	HERVK9-int	1	205775000	205800000	RAB29	1	206100000	206125000	AVPR1B, RHEX	3/5	0/3
0	1	226833944	226834368	HERVH-int	1	226825000	226850000	.	1	226625000	226650000	ITPKB	1/5	0/3
0	1	226833944	226834368	HERVH-int	1	226830000	226835000	.	1	226630000	226635000	ITPKB	1/5	0/3
0	1	226834434	226837969	HERVH-int	1	226825000	226850000	.	1	226625000	226650000	ITPKB	1/5	0/3
0	1	226834434	226837969	HERVH-int	1	226830000	226835000	.	1	226630000	226635000	ITPKB	1/5	0/3
0	2	101992273	101992732	HERVIP10FH-int	2	101975000	102000000	IL1R2, LINC01127	2	101875000	101900000	MAP4K4	1/5	0/3
0	2	101992273	101992732	HERVIP10FH-int	2	101990000	101995000	IL1R2	2	101880000	101885000	MAP4K4	1/5	0/3
65	3	49219649	49220344	HERVK9-int	3	49215000	49220000	CCDC36	3	49170000	49175000	KLHDC8B	0/5	1/3
65	4	15955808	15956133	HERVL40-int	4	15950000	15975000	FGFBP2, PROM1	4	15725000	15750000	BST1	1/5	0/3
0	4	169172167	169172625	HERVL-int	4	169150000	169175000	SH3RF1	4	169475000	169500000	NEK1	0/5	1/3
65	4	169172167	169172625	HERVL-int	4	169150000	169175000	SH3RF1	4	169250000	169275000	SH3RF1	1/5	0/3
65	5	61338974	61339040	HERV16-int	5	61325000	61350000	ZSWIM6	5	61050000	61075000	NDUFAF2	1/5	0/3
0	6	28354638	28354697	HERVL18-int	6	28350000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28140000	ZSCAN16, ZSCAN16-AS1	3/5	0/3
0	6	28354638	28354697	HERVL18-int	6	28350000	28375000	ZKSCAN3, ZSCAN31	6	28125000	28150000	ZKSCAN8, ZSCAN16, ZSCAN16-AS1	2/5	1/3
65	6	28354638	28354697	HERVL18-int	6	28350000	28355000	ZKSCAN3, ZSCAN31	6	28135000	28140000	ZSCAN16-AS1	0/5	1/3
65	6	28354638	28354697	HERVL18-int	6	28350000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28140000	ZSCAN16, ZSCAN16-AS1	1/5	0/3
0	6	28354699	28356361	HERVL18-int	6	28350000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28140000	ZSCAN16, ZSCAN16-AS1	3/5	0/3
0	6	28354699	28356361	HERVL18-int	6	28350000	28375000	ZKSCAN3, ZSCAN31	6	28125000	28150000	ZKSCAN8, ZSCAN16, ZSCAN16-AS1	2/5	1/3
65	6	28354699	28356361	HERVL18-int	6	28350000	28355000	ZKSCAN3, ZSCAN31	6	28135000	28140000	ZSCAN16-AS1	0/5	1/3
65	6	28354699	28356361	HERVL18-int	6	28350000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28140000	ZSCAN16, ZSCAN16-AS1	1/5	0/3
65	6	28354699	28356361	HERVL18-int	6	28355000	28360000	ZKSCAN3, ZSCAN31	6	28120000	28125000	ZSCAN16, ZSCAN16-AS1	1/5	0/3
65	6	28354699	28356361	HERVL18-int	6	28355000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28135000	ZSCAN16, ZSCAN16-AS1	2/5	2/3
0	6	28356368	28356924	HERVL18-int	6	28350000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28140000	ZSCAN16, ZSCAN16-AS1	3/5	0/3
0	6	28356368	28356924	HERVL18-int	6	28350000	28375000	ZKSCAN3, ZSCAN31	6	28125000	28150000	ZKSCAN8, ZSCAN16, ZSCAN16-AS1	2/5	1/3
65	6	28356368	28356924	HERVL18-int	6	28350000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28140000	ZSCAN16, ZSCAN16-AS1	1/5	0/3
65	6	28356368	28356924	HERVL18-int	6	28355000	28360000	ZKSCAN3, ZSCAN31	6	28120000	28125000	ZSCAN16, ZSCAN16-AS1	1/5	0/3
65	6	28356368	28356924	HERVL18-int	6	28355000	28360000	ZKSCAN3, ZSCAN31	6	28130000	28135000	ZSCAN16, ZSCAN16-AS1	2/5	2/3
0	8	11907116	11911625	HERVE-int	8	11900000	11910000	.	8	11840000	11850000	CTSB	1/5	1/3
0	8	11907116	11911625	HERVE-int	8	11900000	11910000	.	8	11850000	11860000	CTSB	1/5	0/3
65	8	11907116	11911625	HERVE-int	8	11900000	11910000	.	8	11840000	11850000	CTSB	0/5	1/3
65	8	11907116	11911625	HERVE-int	8	11900000	11910000	.	8	11850000	11860000	CTSB	0/5	1/3
65	10	120849541	120850126	HERVH-int	10	120850000	120875000	WDR11, WDR11-AS1	10	119900000	119925000	SEC23IP	1/5	0/3
65	10	120849541	120850126	HERVH-int	10	120850000	120875000	WDR11, WDR11-AS1	10	119925000	119950000	SEC23IP	3/5	0/3
65	16	31850017	31850230	HERVIP10FH-int	16	31850000	31875000	ZNF267	16	30950000	30975000	SETD1A	1/5	0/3
65	16	31850238	31850382	HERVIP10F-int	16	31850000	31875000	ZNF267	16	30950000	30975000	SETD1A	1/5	0/3
65	16	31850803	31851322	HERVIP10FH-int	16	31850000	31875000	ZNF267	16	30950000	30975000	SETD1A	1/5	0/3
65	16	31851324	31852021	HERVIP10F-int	16	31850000	31875000	ZNF267	16	30950000	30975000	SETD1A	1/5	0/3

Table 5.3 – Breakdown of HERV overlap with chromatin loop anchors at PD genes before and after dopaminergic neuronal differentiation of iPSCs from FOUNDIN-PD. Loops at day 0 are in white, while loops at day 65 are in grey. Nominated PD genes are underlined.

5.2.4. At sites of known non-reference retrotransposon insertions, differentiation-associated changes in TE overlap with GALAs is similarly divergent for control and PD lines

As has been discussed previously, the recent mobilisation of active retrotransposon classes *Alu*, L1 and SVA can produce RIPs – loci where a given insertion may be present or absent – in the human genome. RIPs may be acutely important in producing interpersonal differences in expression of surrounding genes, yet the reference genome contains a far from complete catalogue of their locations. It is also possible that there are fixed genomic TEs not yet included in the reference genome, which may be similarly underappreciated genetic elements. The Genome Aggregation Database – Structural Variants (gnomAD-SV, https://gnomad.broadinstitute.org/) dataset is a publicly available resource of 445,857 structural variants discovered in WGS from 10,738 unrelated individuals through combinatorial use of four variant detection algorithms (Manta [268], DELLY [269], MELT [218], and cn.MOPS [270]). Essentially, this represents a database of known variants of at least 50 bp not annotated in the reference genome, which includes TEs. By extracting entries for *Alu*, L1 and SVAs from gnomAD-SV a list of 81,350 non-reference retrotransposon coordinates was obtained. HERVs were not found in the gnomAD-SV dataset likely because they are mostly extinct for transposition, making novel annotations of non-reference insertions exceedingly rare [39, 41].

Having obtained a list of known non-reference retrotransposons, their coordinates were intersected with FOUNDIN-PD iPSC line Hi-C chromatin loop anchor and gene coordinates in the same manner as

for reference TEs in **Section 5.2.2** (see **Figure 5.3c**, and **Section 2.2.1.4** for methodology). First, iPSCs derived from controls and PD patients were grouped together and it was found that there was no difference in colocalisation of coordinates of known non-reference TEs with GALAs when cell lines were compared before and after dopaminergic neuronal differentiation (**Figure 5.6a**). This was not unexpected, as non-reference TEs are presumably relatively uncommon and therefore only anticipated to be present at a fraction of possible loci in this small sample size (n=8). However, the same data were then further split into control and PD diagnoses and it was observed that after differentiation GALA overlap with known non-reference TEs was enriched in control iPSC lines but was decreased in PD lines – both when TEs were grouped together or considered as individual *Alu*, HERV, L1 or SVA classes (**Figure 5.6b**). Wilcoxon signed-rank tests indicated that these differences were not significant, and so these data should be interpreted with caution. Nevertheless, this hints that non-reference TEs, which are likely to represent RIPs between individuals, may be differently involved in the establishment of intrachromosomal contacts at gene loci in the PD dopaminergic neuron compared to healthy controls. Interestingly, this trend was also observed for reference TEs at chromatin loops anchors across all genes (compare **Figure 5.6b** with

Figure 5.4b).

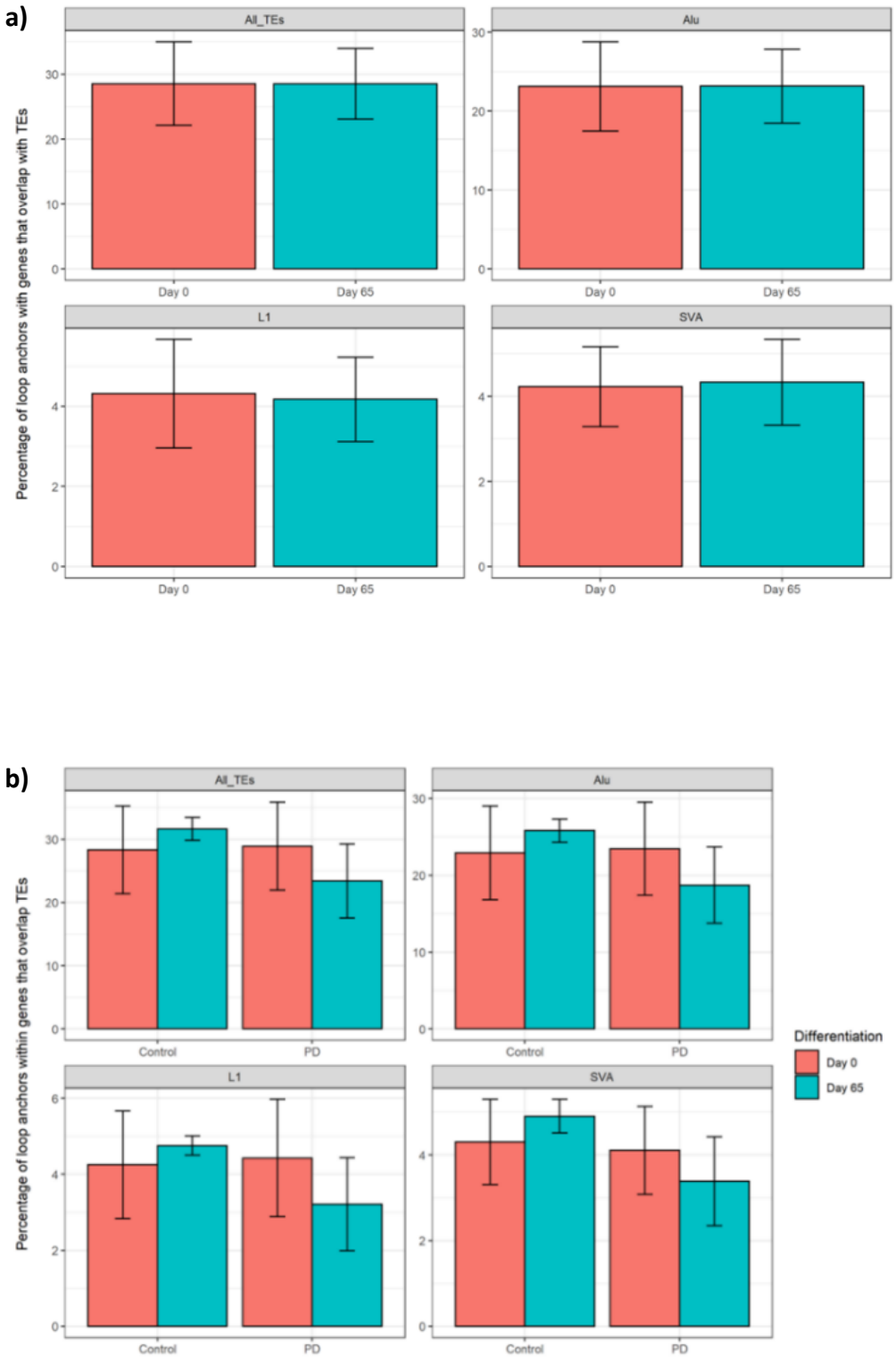


Figure 5.6 – Proportion of all gene-associated chromatin loop anchors that overlap with non-reference TEs. Chromatin loop anchor coordinates from Hi-C of iPSC lines from FOUNDIN-PD were intersected with non-reference retrotransposon coordinates from gnomAD-SV hg38 and gene coordinates from

the RefSeq hg38 curated subset. TEs were considered as 'All TEs' and as separate TE families. **a)** TE-associated loops at genes from all iPSC lines were compared at day 0 and day 65 of a dopaminergic neuronal differentiation. Groups underwent paired Wilcoxon signed-rank test. N=8. **b)** Dataset from **(a)** was further broken down into iPSC lines derived from control and PD individuals. Data were divided by diagnosis and differentiation state separately, and each grouping underwent paired Wilcoxon signed-rank test. Control n=5, PD n=3.

Subsequently, the list of GALAs was again refined to only consider those nominated as potentially functionally associated with PD (Nalls *et al*, 2019 [23]) to examine whether there was differing involvement of non-reference retrotransposons in PD-relevant chromatin loops upon differentiation (see **Figure 5.3d**), and whether this varied between PD-derived iPSCs and control lines. It was observed that dopaminergic differentiation was not associated with significant differences in colocalisation of non-reference TEs with PD GALAs, whether iPSC lines were considered grouped together (**Figure 5.7a**) or as controls versus PD lines (**Figure 5.7b**). However, it was noted that when iPSC lines were studied altogether the overlap of non-reference SVAs with PD GALAs appeared to decrease slightly (**Figure 5.7a**) in contrast to the previously observed increase in reference SVA colocalisation with these PD GALAs (**Figure 5.5a**), suggesting different developmental changes in chromatin architecture at loci of the less common (and likely evolutionarily younger) non-reference SVAs. Additionally, when control and PD iPSCs were compared it was found that non-reference *Alu* and L1 elements followed the same modest trend described previously for genome-wide GALAs in which their overlap with PD GALAs increased with differentiation of control lines and decreased in PD lines (**Figure 5.7b**).

This was contrasted by the behaviour of non-reference SVA retrotransposons – whose overlaps with PD GALAs instead decreased with differentiation of control cell lines and did not change in PD iPSCs (**Figure 5.7b**). This was again markedly different to the response to differentiation of reference SVA elements, which became enriched at PD GALAs in both control and PD lines (**Figure 5.5b**). However, it was interesting to note that overall the proportion of PD GALAs that colocalise with non-reference SVA elements was decreased in the PD lines compared to controls – in line with the trend observed previously for reference SVAs (compare SVA plots in **Figure 5.5b** and **Figure 5.7b**). In other words, in the iPSC lines examined here there was a modest reduction in the overlap of PD GALAs with loci that harbour both putative reference SVAs and non-reference SVA insertions, suggesting a diminishing of their contribution to chromosomal structure in the PD nucleus.

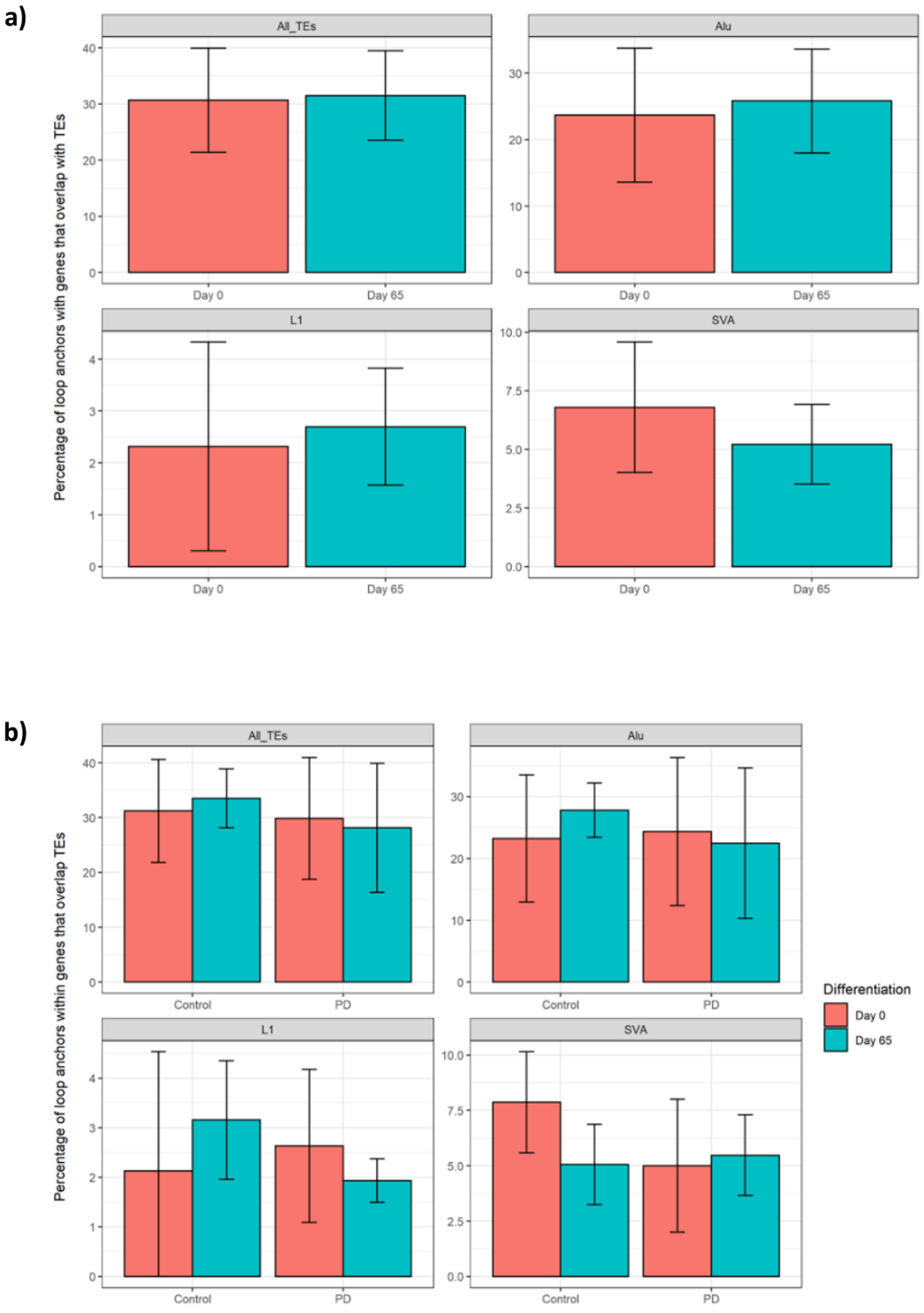


Figure 5.7 – Proportion of PD gene-associated chromatin loop anchors that overlap with non-reference TEs. Gene coordinates from the RefSeq curated subset were filtered to only include those previously

nominated to be functionally associated with PD risk SNPs [23]. Chromatin loop anchor coordinates from Hi-C of iPSC lines from FOUNDIN-PD were intersected with non-reference TE coordinates from gnomAD-SV and coordinates of the nominated PD genes. TEs were considered as 'All TEs' and as separate TE families. **a)** TE-associated loops at genes from all iPSC lines were compared at day 0 and day 65 of a dopaminergic neuronal differentiation. Each TE grouping underwent paired Student's t-test. N=8. **b)** Dataset from **(a)** was further broken down into iPSC lines derived from control and PD individuals. Each TE grouping underwent two-way ANOVA. Control n=5, PD n=3.

5.2.5. *De novo* annotation of non-reference TEs suggests that overall colocalisation with GALAs is reduced in PD, while only *Alu* elements overlapped with PD GALAs

An inherent caveat to the use of non-reference retrotransposon coordinates from the gnomAD-SV database is that this represents an amalgamation of all detected retrotransposons in the dataset, and does not reflect the specific complement of RIPs that may be present in each iPSC line from FOUNDIN-PD. Fortuitously, TE annotations of WGS data made using the Mobile Element Locator Tool (MELT) were available for 2 control and 1 PD iPSC line. In brief, MELT identifies TE insertions in WGS data that are absent in the reference genome, meaning that the presence of specific non-reference insertions could be directly compared to chromatin loop anchor coordinates in these iPSCs. It was noted that this study would be inherently statistically underpowered because MELT data was only available for 3 iPSC lines, but the analysis was nevertheless performed for any preliminary insights it might provide. As described previously for comparisons of Hi-C data with reference and non-reference TE databases, the novel TE annotations from MELT were intersected with coordinates of genes and chromatin loop anchors from iPSC (see **Figure 5.3e**). When

all 3 cell lines were examined before and after dopaminergic neuronal differentiation, paired t-tests indicated that there were no significant differences in colocalisation with GALAs for any of the TE families newly annotated by MELT individually or when grouped together (**Figure 5.8a**). The result of this statistical test was not unexpected at this sample size, and does not preclude the cautionary evaluation of trends in the dataset. Accordingly, it was observed that there was a small increase in colocalisation of novel non-reference *Alu*, HERV, SVA and TEs overall with GALA coordinates, while overlap with L1 elements appeared to decrease slightly.

As before, data were then split to compare the 1 PD and 2 control iPSC lines. This indicated that for MELT-annotated *Alu*, HERV and L1 elements that colocalisation with GALAs was reduced in the PD cell line – indeed, in the PD cell line no L1 overlap was observed after differentiation and no HERV overlap was detected at either timepoint (**Figure 5.8b**). By contrast, colocalisation of MELT-annotated SVA retrotransposons with GALAs appeared to be increased in the PD line versus the control lines. However, this breakdown of the iPSC lines by diagnosis also showed that upon dopaminergic differentiation overlap of SVA elements with GALAs increased in the control lines and decreased slightly in the PD line. This pattern of changes at SVA RIPs (that TE overlap with GALAs was enriched in control iPSC lines and diminished in PD lines after differentiation) had been observed previously for all TE classes when reference and non-reference TE databases were overlaid with loop anchors across all gene coordinates (see

Figure 5.4b and **Figure 5.6b**). Despite a very small sample size this recurring trend, along with the reduced overlap of *de novo Alu*, HERV and L1 with GALAs, further hints at a divergence in TE involvement in chromatin looping at gene bodies between the healthy and PD nuclei that may warrant further investigation.

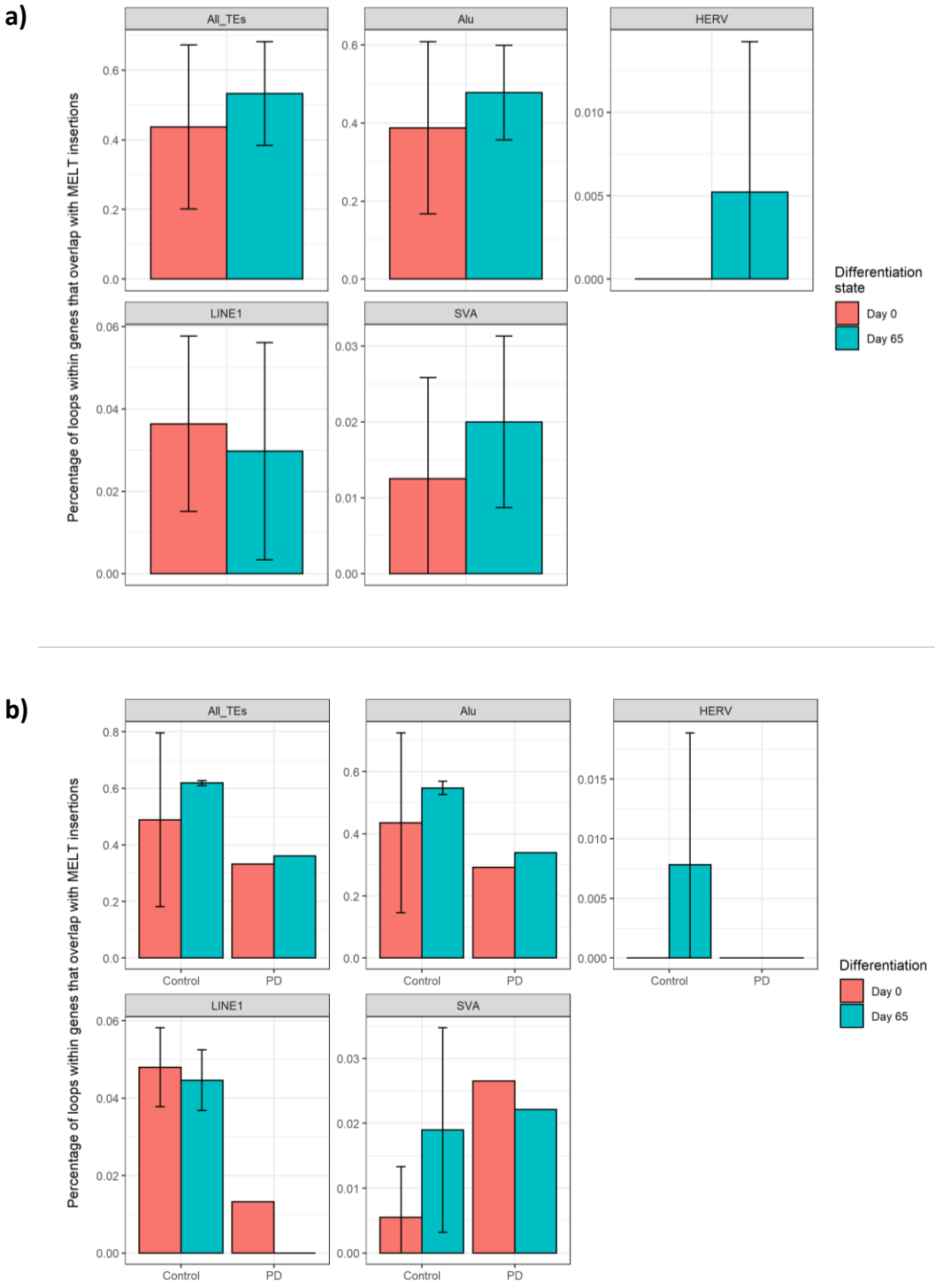


Figure 5.8 – Proportion of all gene-associated chromatin loop anchors that overlap with MELT annotations of novel non-reference TEs. Chromatin loop anchor coordinates from Hi-C of iPSC lines from FOUNDIN-PD were intersected with non-reference retrotransposon coordinates established by MELT in WGS and gene coordinates from the RefSeq hg38 curated subset. TEs were considered as 'All

TEs' and as separate TE families. **a)** TE-associated loops at genes from all iPSC lines were compared at day 0 and day 65 of a dopaminergic neuronal differentiation. Groups underwent paired Student's t-test. N=3. **b)** Dataset from **(a)** was further broken down into iPSC lines derived from control and PD individuals. Data were divided by diagnosis and differentiation state separately, and each grouping underwent two-way ANOVA. Control n=2, PD n=1.

When the overlap of MELT-annotated TEs with chromatin loop anchors and genes was refined to only nominated PD-relevant genes (see **Figure 5.3f**), it was found that the only novel retrotransposon insertions to colocalise with PD GALAs were of the *Alu* family. This is perhaps unsurprising, as this represents a very small portion of human genes (151 out of 20,000-30,000 [28]) and *Alu* are currently the most active family of retrotransposons [79]. *Alu* colocalisation with PD GALAs was found to increase after differentiation in the 3 iPSC lines (**Figure 5.9a**), in line with the previous observation made for MELT-annotated *Alu* elements across all genes (**Figure 5.8a**), but this was not significant in a Student's t-test. Breakdown into healthy control and PD diagnoses for the iPSCs suggested that prior to differentiation all *Alu*-associated loops at PD genes in control iPSCs were absent in the PD line, although after differentiation *Alu* involvement in PD GALAs in the PD line had risen to match that of the control lines (**Figure 5.8b**). Despite arriving at similar proportions after dopaminergic differentiation, this suggests that overall *Alu* overlap with PD GALAs might be lower in the PD line than control lines – an effect that may be born out more clearly at higher sample sizes.

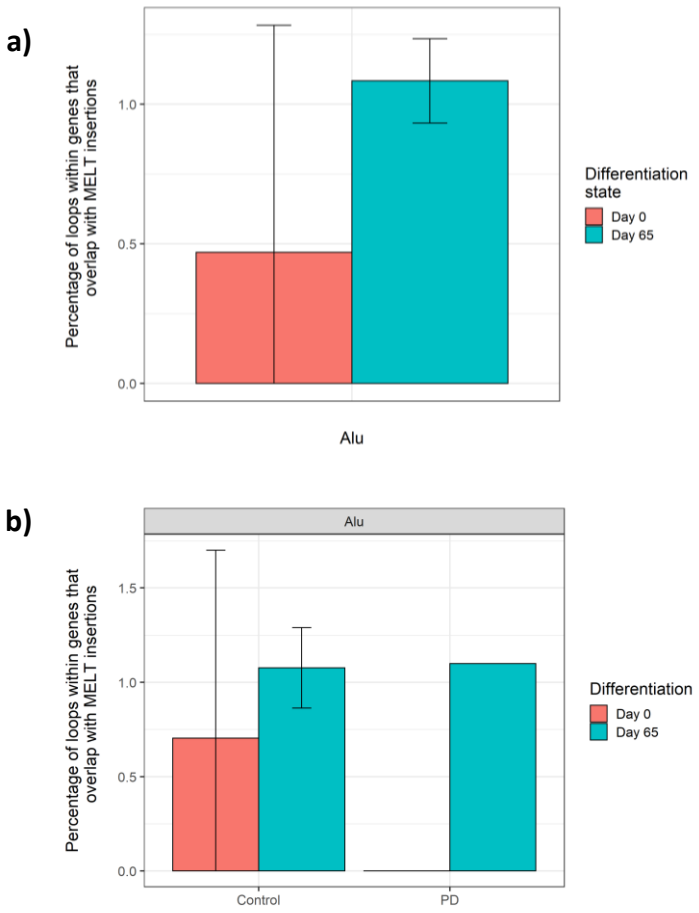


Figure 5.9 – Proportion of PD gene-associated chromatin loop anchors that overlap with MELT annotations of novel non-reference *Alu* retrotransposons. Gene coordinates from the RefSeq curated subset were filtered to only include those previously nominated to be functionally associated with PD risk SNPs [23]. Chromatin loop anchor coordinates from Hi-C of iPSC lines from FOUNDIN-PD were intersected with non-reference retrotransposon coordinates established by MELT in WGS and coordinates of the nominated PD genes. TEs were considered as 'All TEs' and as separate TE families. **a)** TE-associated loops at genes from all iPSC lines were compared at day 0 and day 65 of a dopaminergic neuronal differentiation. Each TE grouping underwent paired Student's t-test. N=3. **b)**

Dataset from **(a)** was further broken down into iPSC lines derived from control and PD individuals. Each TE grouping underwent two-way ANOVA. Control n=2, PD n=1.

5.3. Discussion

In this chapter, Hi-C data from control and PD iPSC lines before and after dopaminergic neuronal differentiation was utilised to assess the involvement of TEs – of which several classes are known the bind the chromatin architectural protein CTCF [161] – in 3D genomic structure in the contexts of development and PD. It has been demonstrated previously that SINE (*Alu*), LINE, LTR (HERV) and DNA transposons have provided CTCFs binding sites in the mouse and human genomes that promote long distance chromatin interactions and formation of loop anchors [258], potentially altering gene expression by bringing promoters and regulatory elements together [241, 242]. To examine how these TE-associated chromatin loops might vary in PD and development, first gene-associated loop anchors, or GALAs, were established by intersecting loop anchor coordinates from Hi-C with gene coordinates obtained from the RefSeq database. By intersecting GALAs with retrotransposon coordinates derived from the RepeatMasker annotation of the reference genome, the non-reference structural variant gnomAD-SV database or *de novo* insertion annotations from the MELT programme, the influence of a variable genomic complement of TEs could be assessed in several ways. Examination of GALAs prior to and after iPSC differentiation provided an assessment of how TEs might be involved in establishing chromatin structure in the mature neuron during development, while comparison of control and PD lines provided insight into how this process might differ in the disease state.

When reference and non-reference TE colocalisation with GALAs were examined genome wide it was observed that there was little difference in TE-GALA overlap before and after iPSC differentiation (

Figure 5.4a and **Figure 5.6a**), indicating that there is not a systemic role for these TEs in establishment of chromatin structure in the mature neuron. This was perhaps unsurprising since locations of TAD boundaries exhibit a degree of evolutionary conservation [248], and therefore TEs - which by contrast display high levels of species-specificity - unlikely to drive changes in GALA formation during differentiation at levels detectable in the sample size available here. However, when the same data were split into iPSC lines derived from PD individuals and controls a pattern emerged in which TE association with GALAs increased in control cell lines but decreased in PD lines after differentiation. Notably, this was true for all TE classes examined in both reference and non-reference data sets (

Figure 5.4b and **Figure 5.6b**, respectively). While these observed changes fell short of statistical significance, the consistent trend across separate TE datasets was encouraging and it was postulated that this trend might persist at higher sample sizes and achieve significance. In this analysis the most pronounced difference between paired groups was a 32.3% change in SVA colocalisation with GALAs after differentiation in control iPSCs (

Figure 5.4b, SVAs from control group); a power analysis was performed (using the R software environment) to determine the number of iPSC lines that would needed for a difference of this magnitude to achieve statistical significance. Setting α significance threshold at 0.05 and test power at 0.95 (1 - Type II error probability of 0.05) it was predicted that statistically significant observations could be made with 12.93, or 13, iPSC lines. Although more than 13 sample pairs (lines before and after differentiation) would be required to demonstrate significance associated with the observed smaller differences in proportional TE involvement with GALAs (

Figure 5.4b, *Alu*, HERV, L1), this nevertheless indicates that confirmation of the trend observed here – namely, that across all GALAs the change in TE involvement upon dopaminergic differentiation is divergent between control and PD conditions – could be achieved with a feasible number of cell lines. While it was assumed that upon differentiation the total number of chromosomal loops in the genome would increase as domain boundaries formed [242, 271, 272], TE overlap with GALAs had been

expected to remain constant if their association with loop anchors was purely coincidental. For colocalisation of all examined classes of TEs with GALAs to be consistently enriched in control lines but reduced in PD lines after differentiation is therefore striking (

Figure 5.4b and **Figure 5.6b**), and suggests a differing 3D genome landscape relative to TEs in the mature dopaminergic neuron in PD. Pertinently, these observations are in line with recent work by Lee *et al* demonstrating that looping of *cis*-regulatory elements containing PD risk SNPs to target genes is down-regulated in cells from the substantia nigra of PD individuals, and that this is functionally associated with expression of these genes [266].

Considering the diverging outcomes for TE association with GALAs after differentiation in control and PD cell lines observed here, it was speculated that differing involvement of TEs might underpin changes in chromatin looping at PD-relevant loci. Therefore, the GALAs under examination were refined to only consider those with genes nominated as functionally associated with PD risk SNPs by the largest PD GWAS meta-analysis to date [23]. Interestingly, when control and PD iPSCs were grouped together it was found that colocalisation of reference genome SVAs with PD GALAs was significantly increased after dopaminergic differentiation (**Figure 5.5a**), suggesting a potential role for this class of retrotransposons in the establishment of chromatin architecture at PD risk loci. When iPSCs were divided into those derived from control and PD individuals it became apparent that this overlap between SVAs and PD GALAs was reduced in PD lines both before and after differentiation (**Figure 5.5b**). Similarly, comparison by PD diagnosis revealed that reference genome HERV association with PD GALAs was significantly diminished in

PD iPSCs versus controls (**Figure 5.5b**). HERV elements were not available in the gnomAD-SV database's list of non-reference TEs as they are almost entirely extinct for transposition, meaning that examples of HERV RIPs have only very recently been identified [41]. Nevertheless, the overlap of the list of non-reference TEs from gnomAD-SV with PD GALAs suggested that overall colocalisation with non-reference SVAs was reduced in PD lines (**Figure 5.7b**), similar to observations for reference genome SVAs (**Figure 5.5b**). Meanwhile, overlap of non-reference *Alu* and L1 elements with PD GALAs appeared to increase in controls and decreased in PD lines after differentiation (**Figure 5.7b**) – mirroring the change that had been observed at all GALAs (**Figure 5.6b**). Altogether, this suggests that PD-associated changes in TE colocalisation with PD GALAs are not only in line with those observed at GALAs genome-wide but are more pronounced for SVA and HERV retrotransposons, further contributing to the hypothesis that TEs may be involved in disease-relevant changes in genome architecture in PD.

Intersecting Hi-C data generated in cell lines with TE coordinates from databases only yields an approximate measure of their association. The RepeatMasker database used to annotate the reference genome contains TEs that are RIPs and may in fact be absent in a given individual, while the non-reference RIPs listed in gnomAD-SV may be present in an individual in any number of permutations. To gain more targeted insight into the influence of specific non-reference TEs on the 3D genome, *de novo* annotations of TE insertions from MELT in a subset of iPSC lines were overlaid with GALAs. Consistent with the behaviour of TEs from reference and non-reference databases, MELT-annotated *Alu*, HERV and L1 element colocalisation with GALAs across the genome was found to be lower in the PD iPSC line (**Figure 5.8b**). MELT-annotated SVA overlap with GALAs was slightly increased in the PD line, in contrast not only with the other MELT-annotated TEs but with the behaviour of SVAs

derived from reference and non-reference databases – for which association with GALAs was reduced in PD (

Figure 5.4b and **Figure 5.6b**). However, after differentiation of the iPSCs overlap of these novel SVA coordinates with GALAs was seen to increase in the control lines and decrease in the PD line (**Figure 5.8b**), a pattern that has recurred throughout this analysis for TE data derived from various sources. When PD GALAs were focussed on, it was found that the only MELT-annotated TEs captured by this analysis were of the *Alu* family. By comparing control and PD iPSCs it was observed that the overall proportion of novel *Alu* elements that overlapped PD GALAs was lower in the PD lines (**Figure 5.9b**). As noted previously, however, the small sample size of 2 control iPSC lines and 1 PD line with available MELT annotations means that these particular interpretations are preliminary. Nevertheless, these data again point towards an altered chromatin landscape at TE insertions in PD.

The most robust observations of differential TE involvement with GALAs in PD, that of reference SVA and HERV elements at candidate PD gene loci (**Figure 5.5a** and **b**, respectively), were explored in greater detail. Notably, it was found that all SVA overlaps with PD GALAs involved members of the SVA D subfamily (**Table 5.2**), the central VNTR of which has been shown to provide binding sites for CTCF and CTCFL [180]. It was observed that a single SVA-associated loop anchor at the nominated PD gene *ARIH2* occurred prior to differentiation and only in one control cell line, while 9 different SVA-associated GALAs were seen after: 2 at gene *SCARB2*, and 7 featuring *NEK1*. There was no clear direction of change across all of these PD GALAs with some

occurring more frequently in controls and others in PD cell lines, and hints that gain and loss of different SVA-associated PD GALAs might be associated with the PD 3D genome. The apparent overrepresentation of SVA-associated chromatin loops at *NEK1* is particularly interesting as this gene has previously been associated with another neurodegenerative disorder, amyotrophic lateral sclerosis (ALS) [273]. Furthermore, recent mouse models have demonstrated that *NEK1* deficiency is associated with disrupted endosomal trafficking and accumulation of α-synuclein, a hallmark of PD [274]. It was noted that an SVA D-associated loop anchor that colocalised with *NEK1* was involved in 3 different chromatin loops to the genes *ANXA10*, *CBR4*, and *PALLD*, while another SVA D-associated anchor overlapping *MFAP3L* featured *NEK1* in the distal loop anchor, but these genes were not immediate candidates for formation of a PD-relevant network of coregulation. Nevertheless, *NEK1* emerges from this analysis as a candidate PD gene that may be subject to altered regulation via changes in SVA-mediated chromatin looping, which was postulated to be consequential for PD aetiology. Similarly, examination of reference HERVs yielded a list of 14 genes that were potentially involved in PD-associated changes in chromatin looping involving these TEs (**Table 5.3**); when compared by loop abundance, the *ZKSCAN8*, *ZSCAN16*, and *ZSCAN16-AS1* gene cluster emerges as the top candidate for differential regulation via HERV-associated chromatin looping, followed by *ITPKB* and *SETD1A*. That HERVs might contribute to an altered chromosomal environment at the zinc nuclease finger genes *ZKSCAN8* and *ZSCAN16*, also known as *ZNF192* and *ZNF392*/*ZNF435*, was noted as it raised the prospect of reciprocal regulation between TEs and their putative repressors – as has been postulated previously for SVA retrotransposons [188]. However, a landmark

paper mapping genomic targets of ZNFs did not observe binding of either of these proteins at TEs (or at promoter regions, as was found for the majority of ZNFs that did not target TEs) [139]. Interestingly, *NEK1* was found to colocalise with a HERV-associated loop anchor in addition to the previous overlap with an SVA element; this HERV-associated anchor at *NEK1* only occurred in one PD line, representing an increase versus control iPSCs – where none were observed. This follows the same trend as SVA-associated loops at *NEK1*, which occurred more frequently in PD lines (9 out of 21 possible occurrences, 42.9%) than in controls (11 out of 35, 31.4%). Altogether, this reinforces the potential for altered *NEK1* regulation by increased TE-mediated chromatin looping in PD. Finally, it was noted that for both SVA- and HERV-associated PD GALAs that the majority of contacts between loci occurred more than once via multiple distinct but overlapping chromatin anchors. This was indicative of a level of redundancy, therefore suggesting that these chromatin anchor sites may be under selective pressure for resistance to genetic perturbation – such as disruption of a CTCF binding site – and may therefore be of functional importance.

In summary, in this chapter a preliminary assessment of TE association with gene-associated chromatin loop anchors was performed to examine putative roles in neuronal development and PD. This was accomplished by leveraging databases of reference and non-reference TE coordinates along with novel annotations for TE insertions, and then comparing the resulting lists of TEs at GALAs prior to and after dopaminergic neuronal differentiation of iPSCs derived from control and PD individuals. Two broad trends became apparent: First, it was noted that at GALAs

across the genome differentiation was associated with divergent changes in colocalisation with TEs – in control iPSCs this overlap generally increased, while in PD lines it was decreased. Secondly, it was observed that overall proportions of TE overlap with GALAs were generally lower in PD lines. Taken together, these suggest an altered chromatin landscape at TEs in gene loci in the PD nucleus, which may amount to changes in gene expression. A significant drawback to the analysis carried out here is that it cannot be ascertained with the available data whether TEs are drivers or simply passengers in the formation of chromatin contacts at genes. Functional studies such as those examining how validated RIP genotypes associate with GALAs in a large number of samples or those involving the specific knock-out (or even knock-in) of GALA-associated TEs may shed valuable insight into their contribution to chromatin architecture, and how this relates to PD. To this end, the candidate PD genes identified here as differentially associated with SVAs and HERVs at loop anchors represent prime candidates for functional study of TE contribution to GALAs in the context of PD. This is especially true for *NEK1*, which featured both a HERV and multiple SVA elements within its numerous chromatin anchors. As described earlier in this chapter, one rationale for the role of TEs in PD is that disease-associated dysregulated methylation leads to inappropriate changes in CTCF recruitment to these elements, subsequently contributing to changes in genome architecture and gene expression. Altogether, the work presented here provides preliminary suggestions that this may indeed be the case, and provides candidates for further study of this hypothesis.

Chapter 6 General Discussion

In this thesis the *cis*-regulatory impacts of genomic retrotransposons, often referred to as transposable elements (TEs), were investigated using a variety of approaches both in the context of their fundamental influences in the human genome and their potential roles in PD, a genetically complex disease. LTR (HERVs) and non-LTR retrotransposons have propagated throughout the human genome and are thought to have influenced gene loci at or near their insertion sites through several mechanisms including direct disruption of protein coding regions [80], alteration of splicing patterns [70, 81, 82], introduction of enhancer or silencer elements [159, 164, 165], or induction of epigenetic changes in the surrounding genome [97, 177]. Additionally, suppression of TE transcription by KRAB-ZFPs via deposition of repressive chromatin marks can result in changes in nearby gene expression [139, 149]. Importantly, the recent and ongoing transpositional activities of the non-LTR retrotransposons L1, *Alu*, and SVA result in retrotransposon insertion polymorphisms (RIPs) throughout the genome. Differences in the complement of these potential regulatory elements among the populace have therefore been speculated to partly underpin interpersonal variation in gene expression patterns in both health and disease. Crucially, short-read sequencing technologies, the current workhorse of WGS approaches, do not in many instances produce reads long enough to unambiguously map TE sequence back to the reference genome, and tools for TE annotation of the resulting incomplete sequence data exhibit varying degrees of accuracy [189-193]. This leads to incomplete capture of TEs (and especially RIPs) in WGS projects such as those examining complex polygenic disorders such as PD, where causal genetic variants remain largely unidentified and small effect alleles like TEs may be important for understanding cumulative genetic risk. Therefore, manual

characterisation of TEs at the lab bench remains essential for fully understanding their roles in gene regulation and disease. SVA retrotransposons were of particular interest as they have been postulated to represent a source of genomic variation important for human-specific gene expression patterns [275, 276], owing to their contemporary transposition [66], hominid specificity [28], potential as CpG islands, and binding of the architectural regulator CTCF [180]. Furthermore, SVA subfamilies E, F and F1 are specific to humans along with more than half of SVA Ds [68], reinforcing their candidacy as contributors to human-specific gene regulation. Indeed, reporter gene studies have shown that SVAs possess regulatory properties [173, 181]. However, to date demonstration of *in situ* modulation of gene expression by SVA retrotransposons is limited to only a few examples including the SVA E insertion into *CASP8* that is protective against prostate cancer [187], while study of highly deleterious disruptive insertions such as the *TAF1*-associated SVA insertion that is causative of XDP are not particularly informative of how SVAs shape normal genome regulation and evolution. Therefore, this area of study became a primary focus of this thesis.

To this end, in **Chapter 3** a common SVA RIP situated upstream of the promoter of the gene *LRIG2* was examined as a model for how presence or absence of an SVA retrotransposon can influence local gene regulation. The 'LRIG2 SVA' was chosen for its high frequency in the general population which allowed it to be interrogated in datasets more easily. This study utilised a wide range of techniques including standard PCR, bioinformatic querying of phenotypic data, CRISPR-Cas9-mediated

deletion, qPCR and pyrosequencing to examine the impact of the LRIG2 SVA on local transcription and DNA methylation in both cohort data and a transgenic cell line model. Although only 96 DNA samples were available from the North American Brain Expression Consortium cohort, by identifying proxy SNPs for the LRIG2 SVA it was possible to infer its genotype in the wider dataset and thereby observe that within this cohort increased allele dosage of the SVA was modestly associated with decreased expression of *LRIG2* and its divergent transcript *LRIG2-DT* (**Figure 3.6**). Importantly, it was also observed that LRIG2 SVA dosage was significantly associated with methylation of the nearest Illumina 450K methylation probe, cg23932873 (**Figure 3.8**), and a weak but significant correlation was found with expression of *LRIG2* (**Figure 3.9**) – suggesting a potential functional relationship. Since these observations could be confounded by other genetic variation present within the NABEC cohort, influence of the LRIG2 SVA RIP in an otherwise genetically identical background was investigated via CRISPR-mediated deletion in SH-SY5Y, an established cell line in which this SVA is homozygous for presence. qPCR and bisulphite sequencing indicated that CRISPR-mediated deletion of the LRIG2 SVA recapitulated the trends in expression and methylation with SVA dosage that were observed in NABEC, albeit at smaller magnitudes (**Figure 3.15** and **Figure 3.16**). Taken together, these data are consistent with the model that the LRIG2 SVA is a subtle modulator of the *LRIG2* promoter region – specifically, the SVA may decrease expression from the locus by increasing local methylation, which may recruit heterochromatin-forming factors such as methyl binding domain proteins and ultimately repress local transcription [176]. The small effect size associated with the LRIG2 SVA is in keeping the proposed role for TEs in shaping the genome in which

novel insertions provide additional enhancer or repressive elements and thereby fine tune gene expression patterns, rather than causing dramatic – and likely deleterious – changes with every insertion. It should also be noted that these model systems only allow a snapshot of expression patterns, and the influence of the SVA could be more dramatic at other points in the human life cycle; for example, disparate intracellular environments such as those associated with early development or ageing could give rise to epigenetic changes resulting in altered TF recruitment or TE activation. Although the modest changes in local expression and methylation associated with the LRIG2 SVA seen here were expected, it would nevertheless be prudent to increase sample sizes to achieve statistical significance for these observations. Two ways forward in this regard were discussed in **Chapter 3**; briefly, the LRIG2 SVA proxy SNPs identified here could readily be used in larger cohort datasets, and suggestions were made for efficiency improvements of the SVA CRISPR deletion pipeline. Should the influences of the LRIG2 SVA be confirmed at statistically significant sample sizes, a logical next step would be to investigate the mechanism by which the SVA affects the surrounding locus. For example, does methylation increase downstream of the SVA as a result of recruitment of methyl binding domain proteins to the SVA's internal methylated CpG-dense regions [176], or does methylation increase as a result of SVA targeting by KRAB-ZFPs and KAP1 [139, 149]? Are other TFs involved? These questions might be answered via chromatin immunoprecipitation (ChIP) approaches. ChIP-seq, in which particular nuclear proteins are extracted and the genomic DNA to which they were crosslinked is sequenced, is unlikely to work due to the previously discussed obstacles of TE sequence mappability using short read sequencing. However, more targeted approaches may bear fruit, such as the PCR-based

interrogation of such DNA pulldowns for the specific LRIG2 SVA sequence. Alternatively, the LRIG2 SVA sequence (amplified by PCR, for instance), could be used as bait for proteins in a nuclear extract and bound proteins could be identified by mass spectroscopy, or even by binding of tagged antibodies if candidate factors are known beforehand. Indeed, such an approach was used to demonstrate CTCF binding at SVAs [180]. It would also be interesting to study whether increased methylation around the SVA LRIG2 insertion site is directly causative of the observed gene expression changes. Recently, it was shown that a 'a programmable epigenetic memory writer' composed of a catalytically dead Cas9 protein fused to a KRAB domain and two DNA methyltransferases was capable of inducible and targeted DNA methylation [277] – such a technique might readily allow the study of consequences of methylation at the *LRIG2* promoter region. In summary, the work presented here demonstrates that the LRIG2 SVA is a tractable model for the influence of SVA RIPs at gene loci, having provided evidence for subtle SVA-mediated modulation of expression and methylation, which might readily be further investigated to elucidate the mechanisms underlying these effects.

In a workflow similar to that of the LRIG2 SVA but with important contextual differences, **Chapter 4** focussed on the study of a novel non-reference genome SVA RIP situated within intron 4 of the gene *KANSL1*, on the premise that the 'KANSL1 SVA' may have been a previously undescribed *cis*-regulatory element that contributed to the observed reduction in PD risk associated with the H2 inverted haplotype at the *MAPT* locus. Primarily, it was postulated that the putatively H2

haplotype-associated 'KANSL1 SVA' was a driver of haplotype-specific *KANSL1* expression patterns, which may then result in changes to KANSL1 regulation of PINK-1-mediated mitophagy and altered PD risk [229]. Here, PCR-based characterisation the KANSL1 SVA was again used in combination with available genotypic and phenotypic data to assess the SVA's association with PD risk and expression changes, and subsequently the molecular biology groundwork was laid for *in vivo* study of the KANSL1 SVA via reporter gene assays. Initially PCR was utilised to validate the existence of the KANSL1 SVA at the genomic coordinates predicted by the Mobile Element Locator Tool (MELT) in NABEC DNA samples (**Figure 4.2**), and these validated genotypes were then used to generate proxy SNPs that confirmed the anticipated association of the SVA insertion with the H2 haplotype (Dr Kimberley Billingsley, NIH, personal correspondence) by assessing the linkage disequilibrium between the two (r^2=0.9855, D'=1). Unexpectedly, this corroboration of KANSL1 SVA association with H2 haplotype was found despite only 35% agreement between MELT annotated and PCR validated SVA genotypes. After extensive troubleshooting it remains unclear how this discrepancy has arisen, but the tight linkage observed between genotypes determined by PCR and the H2-tagging SNP was taken to mean that the PCR-validated genotypes were 'correct', given that close association with the H2 haplotype had been expected. A proxy SNP for the KANSL1 SVA RIP was used to infer its genotype in the wider NABEC cohort and it was observed that increased SVA allele dosage was associated with increased expression of *KANSL1* (**Figure 4.10**), in line with expectations for a H2-associated variant [229, 233]. KANSL1 SVA dosage was also associated with increased methylation at the nearest Illumina 450K methylation probe (as shown in the hg19 build of UCSC), cg18699337 (**Figure 4.11**). Since intronic

hypermethylation has previously been linked to up-regulation of genes [95, 96], this provides a mechanism for the increased *KANSL1* expression observed with the KANSL1 SVA here and with the H2 haplotype in previous studies [229, 233]. When the wider *MAPT* locus was considered, it was found that stratifying gene expression by KANSL1 SVA genotype in NABEC reproduced a list of 10 H1/H2 haplotype-associated genes identified in a previous study (**Table 4.3**) (this prior study's list contained 13 genes, but expression data for the remaining 3 were unavailable in NABEC) [233]. Interestingly, this analysis further indicated that KANSL1 SVA genotype was additionally associated with expression of *WNT3*, a known PD candidate gene that has been functionally nominated as the top candidate for explanation of PD risk at the *MAPT* locus [23], and *ARL17B*, a poorly characterised transcript that has been linked to Alzheimer's disease and progressive supranuclear palsy [236, 237]. Comparison of H1/H2-associated genes to the genotype of the H1/H2 proxy SNP rs8070723 in NABEC data indicated that the KANSL1 SVA was predictive of expression for an almost identical list of genes as the H2 proxy SNP but with slightly weaker associations, suggesting that while the SVA may be a contributor to their regulation it is not the primary driver. By contrast, the H2 proxy SNP was a weaker predictor of *ARL17B* expression and was not associated with *WNT3* at all in NABEC, suggesting that the KANSL1 SVA is a predictor of gene expression that is distinct from H1/H2-tagging SNPs and may therefore represent a previously unstudied eQTL for the potentially important PD gene *WNT3*. The divergence in gene expression patterns associated with the KANSL1 SVA and H2 haplotypes was in line with the tight but imperfect LD observed between a proxy SNP for the SVA and the H2 proxy SNP in NABEC WGS data (r^2=0.9855). Importantly, this incomplete association was

confirmed in samples with PCR-validated KANSL1 SVA genotypes where it was observed that 2 H2-bearing alleles (out of 188 alleles in 94 samples) did not harbour a copy of the KANSL1 SVA (based on *in silico* genotyping data, not shown). The finding that not every instance of the H2 haplotype in the human populace carries a KANSL1 SVA insertion is interesting, and may have divergent functional consequences for transcription associated with this haplotype. Furthermore, the identification of a rare shorter CT variant within the KANSL1 SVA raises the possibility of varying *cis*-regulatory effects driven by the SVA even between individuals with the same RIP genotype, given the established potential for VNTRs (which CT elements constitute) to influence local expression [182]. However, only a single example of this short CT element allele was observed even when expansion via proxy SNPs was attempted, and therefore the data here is not powered to investigate this further. The suggestion that the KANSL1 SVA underpins gene expression associated with the H2 haplotype is not only important for understanding differences between the H1 and H2 inversions, but also for interpreting H2 haplotype contribution to PD risk; based on the observations made here, H2 haplotypes with and without the KANSL1 SVA might exhibit altered gene expression patterns with differing associations with PD risk. Such an effect would represent an invaluable datapoint in multivariate prediction of PD risk, such in an assessment of polygenic risk score (PRS) of PD, and supports the notion of consideration of TEs in polygenic disease risk. Expanding upon these observations, established proxy SNPs for the KANSL1 SVA were utilised to examine RIP associations in a substantially larger dataset – namely, the multi-cohort 'Accelerating Medicines Partnership – Parkinson's Disease' (AMP-PD) database. Genotypes and transcriptomic data were available for 2698 individuals in the 'v1

release' of AMP-PD, and it was found that KANSL1 SVA dosage was associated with increased *KANSL1* expression (**Figure 4.14**), as it had been in NABEC. However, SVA +/+ genotypes inferred by proxy SNP were unexpectedly low in number, totalling 6 instead of the 170 predicted for an allele of 25% frequency. Due to data access limitations it was not possible to determine the cause of this disparity. It was speculated, for example, that overrepresentation of samples from PD patients within the 2698 datapoints could have reduced the number of H2 haplotypes observed, as H2 is associated with reduced PD risk. However, sample diagnoses were unavailable at the time of writing. After extensive bioinformatic interrogation of the KANSL1 SVA and its associations, steps towards functional characterisation of the element were undertaken. It was found that CRISPR-mediated deletion of the SVA was not practical as none of the cell lines available in the lab were of the RIP genotype +/+, and therefore it was decided that the *in vivo* influences of the KANSL1 SVA would instead be investigation via generation of reporter gene constructs. In short, the KANSL1 SVA was successfully inserted upstream of the minimal promoter of the pGL3-Promoter plasmid, but only in the antisense orientation relative to the Firefly luciferase reporter gene (**Figure 4.18**). Unfortunately, due to time constraints it was not possible to also generate a pGL3P construct containing the SVA in the sense orientation or to perform a luciferase assay of the existing antisense SVA-containing construct. However, it was speculated that a minor adjustment of the cloning strategy presented here might yield improvements in efficiency and enable rapid generation of pGL3P carrying a sense-oriented KANSL1 SVA. Since the end of this thesis project the cloning strategy described here has been taken to completion, and preliminary results indicate that the KANSL1 SVA acts as a transcriptional repressor

when upstream of a promoter (Prof John Quinn, personal correspondence). While this finding initially appears at odds with the positive transcriptional associations of the KANSL1 SVA described in this thesis, this likely represents the difference in genetic contexts: in the human genome the KANSL1 SVA is situated within an intron, which may lead to diverse outcomes when compared to SVA retrotransposons upstream of promoters, as is the case in pGL3P. Indeed, transcriptional repression by a promoter-proximal KANSL1 SVA is in line with the repressive effects associated with the LRIG2 SVA observed in **Chapter 3**, and therefore supports a model for context-dependent influences of SVAs. Furthermore, it was proposed here that the KANSL1 SVA could alternatively be inserted into a site within pGL3P that is downstream of the luciferase gene, which may lead to modulation of reporter gene expression that is more representative of the SVA's genomic context. It would be of great interest to see what, if any, transcriptional changes arose from this. It was additionally noted that the divergent transcriptional associations of the KANSL1 SVA observed in NABEC and in the preliminary pGL3P luciferase data could be the result of the disparate intracellular conditions under which these measurements were made, since NABEC transcriptomics are obtained from post-mitotic and putatively 'normal' neurons while luciferase assays are performed in dividing cells in tissue culture. Furthermore, NABEC 'prefrontal cortex' samples are extracted from a mixture of cell types obtained from that brain region, which may convolute measurements of cellular phenotypes. It is also possible that when inserted into the circular DNA topology of a reporter plasmid the KANSL1 SVA does not form the same structures as when it resides in the genome and might not be expected to influence transcription in the same way. Therefore, it can be speculated that the varying SVA influences seen in NABEC RNA-

seq and luciferase assays represent tissue- or stimulus-specific responses driven by the SVA regulatory element, with these model systems each only capturing a snapshot of potentially dynamic gene expression patterns. A better model for the study of the regulatory influences of the KANSL1 SVA might be to undertake a stable transfection in which a reporter cassette with or without the SVA is introduced to a cell line permanently, either through genomic integration or the maintenance of a replicating extra-chromosomal episome. Such an approach would potentially allow the influences of the KANSL1 SVA RIP to be readily examined in a more realistic genomic context over time and in response to various stimuli, enabling its interplay with intracellular changes such as those associated with PD to be explored. Altogether, the work presented here has therefore demonstrated that a novel MELT-annotated SVA F within intron 4 of the *KANSL1* gene may contribute to gene expression patterns at the *MAPT* locus that have previously been attributed to the H2 inverted haplotype as a whole, and furthermore laid the groundwork for *in vivo* validation of this observation – a strategy which is already bearing fruit and may yet be further extended. This not only improves our understanding of the genetically complex *MAPT* locus, at which genomic variants that are causative of PD have not previously been identified, but also supports the wider hypothesis that SVA RIPs might be a hitherto underappreciated source of genetic variation that contribute to disease-associated gene expression changes in PD and other polygenic disorders.

Finally, in light of observed binding of the genomic architectural protein CTCF to numerous classes of TEs [159, 161, 239], the scope of this thesis was expanded in **Chapter 5** to examine a chromatin looping dataset for evidence of enrichment of TEs of all classes at intrachromosomal chromatin contacts in

gene bodies and whether this varied with PD diagnosis or development state. Given established capabilities for CTCF binding at multiple classes of TEs [161], it was postulated that TEs might act as sites for CTCF-mediated chromatin looping that are altered in the perturbed intracellular environment of PD – a hypothesis which, at the time of writing, has not previously been tested. This entirely *in silico* study was performed by using the Bedtools suite of genomic interrogation tools to overlap Hi-C data (pairs of chromosome coordinates that have been found to interact in 3D space) from iPSCs derived from PD individuals and healthy controls, both in the undifferentiated state and then following a dopaminergic differentiation protocol, with coordinates of genes and TEs. These TE coordinates were derived from the RepeatMasker annotation of the reference genome (hg19 annotations converted to hg38 coordinates), the gnomAD-SV database of non-reference genome structural variants (hg38), and specific MELT annotations of the genome sequence of a subset of iPSC lines (annotated relative to hg38), and therefore permit a multifaceted assessment of how gene-associated loop anchors (GALAs) colocalise with TEs (summarised in **Figure 5.3**). Although it was generally expected that differentiation would result in increased chromatin looping as domain boundaries form, consideration of TE-GALAs as a proportion of all GALAs enabled the relative involvement of TEs to be evaluated at both timepoints. When iPSC lines were differentiated it was consistently observed that GALA overlap with reference and non-reference TEs of all classes increased in control cell lines but decreased in PD individual-derived lines (

Figure 5.4b, **Figure 5.6b**, **Figure 5.7b** *Alu* and L1, **Figure 5.8b** SVA), suggesting a diverging chromatin architecture around TEs at genes in PD. In light of this finding, TE colocalisation with loop anchors at meta-analysis-nominated PD genes [23] (PD GALAs) was examined to determine whether TEs were coincident with disease-associated changes at important PD loci. In considering reference TEs it was found that SVAs were significantly enriched at PD GALAs after differentiation when all iPSC lines were grouped together (**Figure 5.5a**), suggesting that SVA retrotransposons may promote to chromatin-chromatin interactions at PD-relevant gene loci upon differentiation – presumably via increased binding of the architectural factor CTCF.

When control and PD lines were separated this trend persisted but was no longer and statistically significant (**Figure 5.5b**), perhaps due to the reduction in sample size (from 8 total iPSC lines to 5 control and 3 PD lines). Interestingly, this comparison of control and PD iPSCs also revealed that the proportional colocalisation of reference SVAs with PD GALAs was reduced in PD. It was notable that when these SVA-associated PD GALAs were examined in detail they all involved SVA Ds (**Table 5.2**), potentially indicating that this subfamily of SVAs is particularly amenable to promotion of chromatin looping. Furthermore, PD GALAs overlapping reference SVAs exhibited an overrepresentation of the gene *NEK1*, which has typically been associated with the neurodegenerative disease ALS but was more recently implicated in PD via its roles in regulation of α-synuclein [273, 274]. Additionally, the comparison of control and PD iPSCs indicated that reference genome HERV associations with PD GALAs were significantly reduced in the PD lines, both before and after dopaminergic neuronal differentiation (**Figure 5.5b**). The diminished formation of PD GALAs colocalised with both reference SVA and HERV retrotransposons in PD was very interesting, as it was suggestive of disease-associated changes in contribution of these elements to genomic structure. Given the established roles for chromatin-chromatin interactions in gene regulation [241, 242, 246, 247], this may have important consequences for regulation of these disease-relevant loci. Breakdown of HERV-associated loop anchors produced a substantially more diverse list of PD gene loci with HERVs than had been observed for SVAs (**Table 5.3**), with the *ZKSCAN8*, *ZSCAN16*, and *ZSCAN16-AS1* gene cluster on chromosome 6 featuring the highest number of distinct HERV-associated loop anchors. It is unclear what the biological consequences of altered chromatin looping at this gene cluster are, as the ZNF

proteins ZKSCAN8 and ZSCAN16 are relatively poorly characterised and were not detected in a comprehensive analysis of genomic ZNF binding [139]. Functional characterisation of these genes and in both normal physiology and PD along with validation of the contribution of HERVs to chromatin structure at their loci might therefore be a candidate area for future study. The examination of constituents at HERV-associated loops again revealed *NEK1* involvement, with looping to the other nominated PD gene *SH3RF1* suggesting a potential mechanism for coregulation of these PD-relevant genes. Notably, while the majority of HERV-associated loops at PD genes occurred proportionally less in PD iPSCs than in control lines (**Table 5.3**), by contrast the loop interaction featuring *NEK1* occurred only in a PD line – in line with the observation of increased overlap of *NEK1* with SVA-associated loop anchors. Therefore, this analysis has identified a shortlist of genes at which the proportional involvement of reference genome SVAs and HERVs at chromatin loop anchors is altered in PD cell lines, which is speculated to contribute to differences in gene expression patterns at these PD-relevant loci. Amongst these genes *NEK1* has emerged as a particularly notable candidate, having exhibited a proportional increase in overlap with both reference SVA- and HERV-associated loop anchors in PD. When PD GALAs were intersected with non-reference TEs it was found that the overall proportion of overlaps with SVAs was decreased in PD lines (**Figure 5.7b**), similar to that of reference SVAs (**Figure 5.5b**), although a clear effect of differentiation was not observed. By contrast, upon differentiation colocalisation of PD GALAs with non-reference *Alu* and L1 elements was increased in control lines and decreased in PD lines (**Figure 5.7b**), recapitulating the divergent response observed for all TE classes at all GALAs. Subsequently, by overlaying the GALAs of 1 PD and 2 control iPSC lines

with MELT annotations of TEs in their WGS it was possibly to gain a more specific view of how non-reference genome TEs colocalised with loop anchors, albeit at a reduced sample size. It was observed that across all GALAs, colocalisation with MELT-annotated *Alu*, L1 and HERV elements was decreased in PD cell lines, while overlap with SVAs increased slightly (**Figure 5.8b**). Interestingly, after dopaminergic differentiation overlap of these novel SVA annotations increased in the control iPSCs and decreased in the PD line (**Figure 5.8b**, SVA) – a pattern which had recurred often for TEs from reference and non-reference genome databases. Furthermore, consideration of MELT-annotated TEs at only PD GALAs – in which *Alu* elements were the only constituents – indicated that overall colocalisation of these novel *Alu* insertions with PD GALAs was reduced in the PD line. Taken together, the analysis presented here uses multiple parallel investigations to demonstrate an altered chromatin landscape around gene-associated TEs in cells from PD patients. Exactly how this relates to gene expression changes in PD remains to be seen, but it is reasonable to speculate that perturbation of the 3D genome is a consequence of disease-related intranuclear changes which then further worsen cellular health in a deleterious positive feedback loop. In this scenario, CTCF-mediated chromatin looping driven by TEs could be inappropriately up- or down-regulated in a manner specific to local context, with interpersonal variation in genomic RIP content resulting in different TE contributions to disease-associated 3D genome architecture.

It was postulated that the Hi-C data from FOUNDIN-PD iPSCs could be leveraged to investigate the architectural roles of the LRIG2 and KANSL1 SVAs, for which local gene

expression and methylation were examined extensively in this thesis. It was found that the coordinates of the KANSL1 SVA insertion did not coincide with any GALAs in the 8 iPSC lines. This does not appear to be due to the presence or absence of H2 haplotype (on which the SVA typically resides) sequence in the WGS of iPSCs because despite their inverted orientations H1 and H2 share considerable sequence homology, such that the predicted target site of the KANSL1 SVA insertion can be found in the H1-containing hg38 reference genome (AAACCAAAAAATC at chr17:46076611). In other words, sequence is sufficiently similar between H1 and H2 haplotypes for the KANSL1 SVA and any H2-associated Hi-C loop anchors to be mapped to the reference genome or *de novo* WGS data. Therefore, lack of colocalisation of loop anchors with the KANSL1 SVA is likely not due to an artifact arising from H1/H2 haplotype sequences. By contrast, the LRIG2 SVA was found to overlap with 8 GALAs which involved interactions between the *LRIG2* promoter region and upstream 4 loci (**Figure 6.1**). These distal loci ranged from 90 kb to 300 kb from the LRIG2-SVA containing loop anchor and were made up of both gene-sparse regions and regions containing predicted and validated genes. It was observed that there was an SVA A element in the loop anchor locus furthest from that containing the LRIG2 SVA (**Figure 6.1**, leftmost blue region), raising the interesting possibility of a long-range chromatin interaction mediated by CTCF binding to an SVA in each anchor. Additionally, it was noted that within the loop anchor locus that contained the LRIG2 SVA there was an SVA C, a subclass that is fixed in the human genome and does not form RIPs, which may contribute to loop anchor formation at the *LRIG2* promoter region. Most significantly, the chromatin loop between the *LRIG2* promoter region and the upstream region featuring the genes *MOV10*, *RHOC*, *PPMIJ* and *TAFA3*

suggests a mechanism by which these genes may be co-regulated, it is possible that this co-regulation would be promoted by the presence of the LRIG2 SVA and its potential for CTCF binding. Although WGS for the 8 iPSC lines was not available, collaborators at the FOUNDIN-PD project extracted genotypes of SNPs that were identified in **Chapter 3** as tagging LRIG2 SVA VNTR-specific genotypes (rs114767321, rs183751190 and rs12744009). From these it was possible to infer the RIP genotype of the LRIG2 SVA and it was found that 2 lines were +/+, 3 were +/- and 3 were -/- for the SVA. When the frequencies of the chromatin loops featuring the LRIG2 SVA were separated into RIP genotypes it was observed that there was a greater number of looping interactions associated with the +/- genotype than -/- but, unexpectedly, the +/+ genotype was associated with the lowest number of loops (**Table 6.1**). The interaction between the LRIG2 SVA locus and the upstream gene *PPMIJ* was observed most frequently as it appeared in every iPSC line after 65 days of dopaminergic differentiation, thereby highlighting *PPMIJ* as the strongest candidate for coregulation with *LRIG2*. Altogether this suggests that there may be varying formation of chromatin loops overlapping the LRIG2 SVA, however the relationship with SVA allele dosage remains unclear from this data and may represent an area of future study. The ΔLRIG2 SVA SH-SY5Y cell lines generated in **Chapter 3** are a model in which the effects of SVA presence on CTCF binding at the *LRIG2* promoter could readily be examined via ChIP-based approaches, while SVA impact on expression of *MOV10*, *RHOC*, *PPMIJ* and *TAFA3* on could be assessed with qPCR. Together, these experiments might quickly provide insight into SVA RIPs can influence distant gene loci by altering 3D genome architecture.

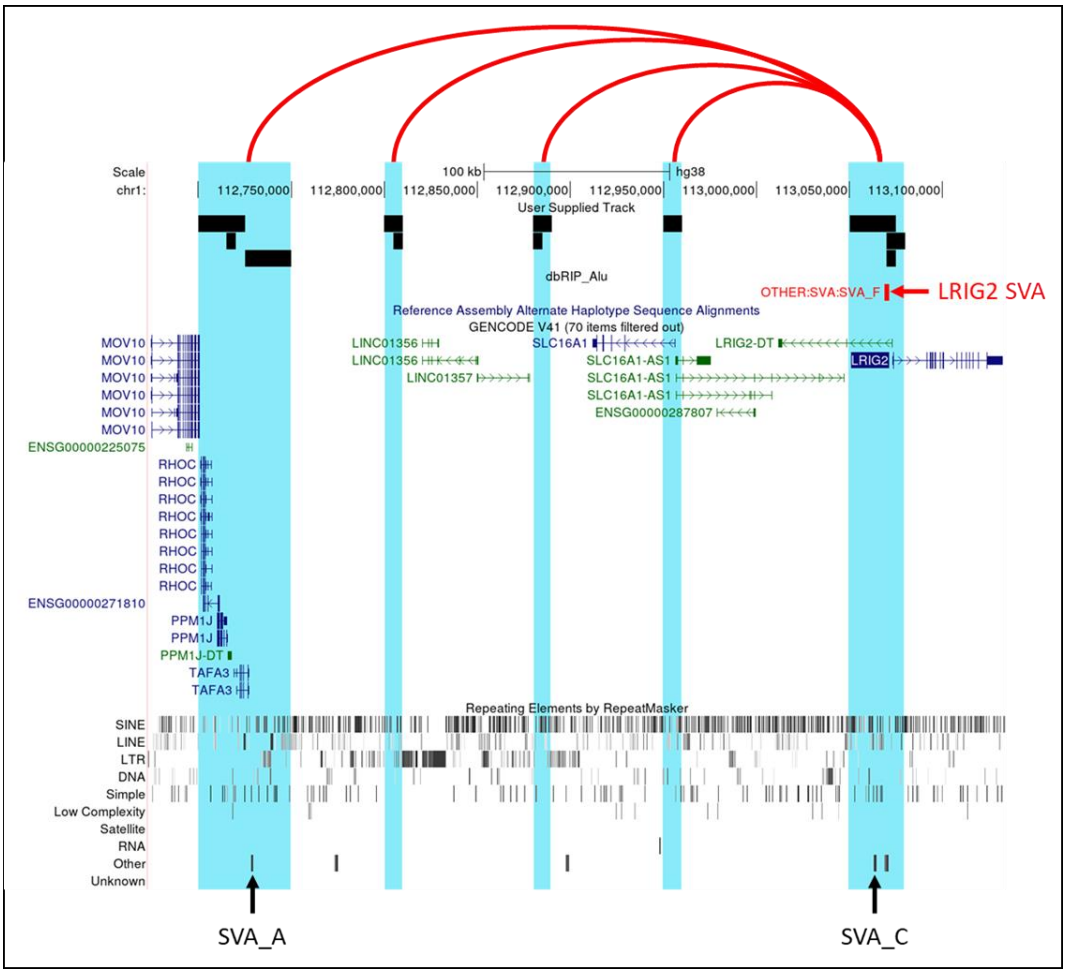


Figure 6.1 – Illustration of chromatin loop anchors from iPSC Hi-C data (FOUNDIN-PD) that overlap with the LRIG2 SVA as shown on the UCSC genome browser, hg38. Loop anchor regions are shown in black block (top) with overlapping genomic features highlighted in blue. Interaction in 3D space between LRIG2 SVA and distal loop anchors is depicted by red lines (top). LRIG2 SVA indicated by red arrow, with other SVAs within loop anchor coordinates indicated by black arrows (bottom).

	LRIG2 SVA RIP genotype:					
	+/+ (n=2)		+/- (n=3)		-/- (n=3)	
Distal Gene	**Day 0**	**Day 65**	**Day 0**	**Day 65**	**Day 0**	**Day 65**
None	0	0	0	1	1	0
MIR11399	0	0	3	0	2	0
MOV10	0	0	0	1	0	0
PPM1J	0	2	0	3	0	2
RHOC	0	0	0	1	0	0
SLC16A1	0	0	0	1	0	0
SLC16A1-AS1	0	0	0	1	0	0
TAFA3	0	0	0	1	1	0

Table 6.1 – Frequencies of chromatin loops featuring the LRIG2 SVA broken down into SVA RIP genotypes. Genes overlapping the distal loop anchor, if any, are listed.

This thesis represents an examination of the roles of TEs in both normal and PD physiology in a variety of genomic contexts, with a particular focus on SVA retrotransposons. This was undertaken using a combination of *in vivo* and *in silico* approaches to leverage the wealth of data currently being produced by high throughput genotyping and phenotyping techniques, and subsequently validate and expand upon these observations at the lab bench. The findings presented here support the hypothesis that SVA retrotransposons possess regulatory capabilities *in situ* and that SVA RIPs may be determinants of small but potentially important differences in gene expression between individuals, which may have consequences for risk of genetically complex disease such as PD. Additionally, this work suggests that association of TEs of all classes with chromatin loop anchors at gene coordinates is altered in the PD nucleus, which was speculated here to underpin widespread disease-related changes in gene expression. In summary, this thesis supports the proposal that TEs are a potent functional source of genomic variation that warrant

greater scrutiny in prediction of gene expression in health and disease, and I anticipate that continuation of the projects described in this thesis will lead to further insight into how SVAs and TEs in general shape the human genome.

References

1. Dorsey, E.R. and B.R. Bloem, *The Parkinson Pandemic-A Call to Action.* JAMA Neurol, 2018. **75**(1): p. 9-10.
2. Yang, W., et al., *Current and projected future economic burden of Parkinson's disease in the U.S.* NPJ Parkinsons Dis, 2020. **6**: p. 15.
3. Poewe, W., et al., *Parkinson disease.* Nat Rev Dis Primers, 2017. **3**: p. 17013.
4. Chaudhuri, K.R. and A.H. Schapira, *Non-motor symptoms of Parkinson's disease: dopaminergic pathophysiology and treatment.* Lancet Neurol, 2009. **8**(5): p. 464-74.
5. Langston, J.W., et al., *Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis.* Science, 1983. **219**(4587): p. 979-80.
6. Cunha, B.A., *Influenza: historical aspects of epidemics and pandemics.* Infect Dis Clin North Am, 2004. **18**(1): p. 141-55.
7. Maurizi, C.P., *Influenza caused epidemic encephalitis (encephalitis lethargica): the circumstantial evidence and a challenge to the nonbelievers.* Med Hypotheses, 2010. **74**(5): p. 798-801.
8. Eldridge, R. and S.E. Ince, *The low concordance rate for Parkinson's disease in twins: a possible explanation.* Neurology, 1984. **34**(10): p. 1354-6.
9. Lesage, S. and A. Brice, *Parkinson's disease: from monogenic forms to genetic susceptibility factors.* Hum Mol Genet, 2009. **18**(R1): p. R48-59.
10. Polymeropoulos, M.H., et al., *Mutation in the alpha-synuclein gene identified in families with Parkinson's disease.* Science, 1997. **276**(5321): p. 2045-7.
11. Valente, E.M., et al., *Localization of a novel locus for autosomal recessive early-onset parkinsonism, PARK6, on human chromosome 1p35-p36.* Am J Hum Genet, 2001. **68**(4): p. 895-900.
12. Bonifati, V., et al., *DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism.* Neurol Sci, 2003. **24**(3): p. 159-60.
13. Kitada, T., et al., *Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism.* Nature, 1998. **392**(6676): p. 605-8.
14. Zimprich, A., et al., *Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology.* Neuron, 2004. **44**(4): p. 601-7.
15. Hernandez, D.G., X. Reed, and A.B. Singleton, *Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance.* J Neurochem, 2016. **139 Suppl 1**: p. 59-74.
16. Lohmueller, K.E., et al., *Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease.* Nat Genet, 2003. **33**(2): p. 177-82.
17. Simón-Sánchez, J., et al., *Genome-wide association study reveals genetic risk underlying Parkinson's disease.* Nat Genet, 2009. **41**(12): p. 1308-12.
18. Kara, E., et al., *Assessment of Parkinson's disease risk loci in Greece.* Neurobiol Aging, 2014. **35**(2): p. 442.e9-442.e16.
19. Satake, W., et al., *Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease.* Nat Genet, 2009. **41**(12): p. 1303-7.

20. Nalls, M.A., et al., *Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies.* Lancet, 2011. **377**(9766): p. 641-9.
21. Nalls, M.A., et al., *Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease.* Nat Genet, 2014. **46**(9): p. 989-93.
22. Chang, D., et al., *A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci.* Nature Genetics, 2017. **49**: p. 1511-1516.
23. Nalls, M.A., et al., *Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies.* Lancet Neurol, 2019. **18**(12): p. 1091-1102.
24. Ibanez, L., et al., *Parkinson disease polygenic risk score is associated with Parkinson disease status and age at onset but not with alpha-synuclein cerebrospinal fluid levels.* BMC Neurol, 2017. **17**(1): p. 198.
25. Escott-Price, V., et al., *Polygenic risk of Parkinson disease is correlated with disease age at onset.* Ann Neurol, 2015. **77**(4): p. 582-91.
26. Paul, K.C., et al., *Association of Polygenic Risk Score With Cognitive Decline and Motor Progression in Parkinson Disease.* JAMA Neurol, 2018. **75**(3): p. 360-366.
27. Nalls, M.A., et al., *Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study.* Lancet Neurol, 2015. **14**(10): p. 1002-9.
28. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**: p. 860-921.
29. McClintock, B., *The origin and behavior of mutable loci in maize.* Proc Natl Acad Sci U S A, 1950. **36**(6): p. 344-55.
30. Erwin, J.A., M.C. Marchetto, and F.H. Gage, *Mobile DNA elements in the generation of diversity and complexity in the brain.* Nature Reviews Neuroscience, 2014. **15**: p. 497-506.
31. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
32. Pace, J.K., 2nd and C. Feschotte, *The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage.* Genome Res, 2007. **17**(4): p. 422-32.
33. Deniz, Ö., J.M. Frost, and M.R. Branco, *Regulation of transposable elements by DNA modifications.* Nat Rev Genet, 2019. **20**(7): p. 417-431.
34. Biechele, S., et al., *Unwind and transcribe: chromatin reprogramming in the early mammalian embryo.* Curr Opin Genet Dev, 2015. **34**: p. 17-23.
35. Richardson, S.R. and G.J. Faulkner, *Heritable L1 Retrotransposition Events During Development: Understanding Their Origins: Examination of heritable, endogenous L1 retrotransposition in mice opens up exciting new questions and research directions.* Bioessays, 2018. **40**(6): p. e1700189.
36. Bannert, N. and R. Kurth, *The evolutionary dynamics of human endogenous retroviral families.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 149-73.

37. Belshaw, R., et al., *Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity.* J Virol, 2005. **79**(19): p. 12507-14.
38. Hughes, J.F. and J.M. Coffin, *Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution.* Proc Natl Acad Sci U S A, 2004. **101**(6): p. 1668-72.
39. Subramanian, R.P., et al., *Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses.* Retrovirology, 2011. **8**: p. 90.
40. Turner, G., et al., *Insertional polymorphisms of full-length endogenous retroviruses in humans.* Curr Biol, 2001. **11**(19): p. 1531-5.
41. Wildschutte, J.H., et al., *Discovery of unfixed endogenous retrovirus insertions in diverse human populations.* Proc Natl Acad Sci U S A, 2016. **113**(16): p. E2326-34.
42. Consortium, G.R.; Available from: https://www.ncbi.nlm.nih.gov/grc/help/faq/#human-reference-genome-individuals.
43. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
44. Hancks, D.C. and H.H. Kazazian, *Roles for retrotransposon insertions in human disease.* Mobile DNA, 2016. **7**: p. 9.
45. Martin, S.L. and F.D. Bushman, *Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon.* Mol Cell Biol, 2001. **21**(2): p. 467-75.
46. Feng, Q., et al., *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell, 1996. **87**(5): p. 905-16.
47. Moran, J.V., et al., *High frequency retrotransposition in cultured mammalian cells.* Cell, 1996. **87**(5): p. 917-27.
48. Swergold, G.D., *Identification, characterization, and cell specificity of a human LINE-1 promoter.* Mol Cell Biol, 1990. **10**(12): p. 6718-29.
49. Alisch, R.S., et al., *Unconventional translation of mammalian LINE-1 retrotransposons.* Genes Dev, 2006. **20**(2): p. 210-24.
50. Kulpa, D.A. and J.V. Moran, *Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles.* Nat Struct Mol Biol, 2006. **13**(7): p. 655-60.
51. Doucet, A.J., et al., *A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition.* Mol Cell, 2015. **60**(5): p. 728-741.
52. Cost, G.J., et al., *Human L1 element target-primed reverse transcription in vitro.* EMBO J, 2002. **21**(21): p. 5899-910.
53. Christensen, S.M. and T.H. Eickbush, *R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA.* Mol Cell Biol, 2005. **25**(15): p. 6617-28.
54. Richardson, S.R., et al., *APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition.* Elife, 2014. **3**: p. e02008.

55. Coufal, N.G., et al., *Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells.* Proc Natl Acad Sci U S A, 2011. **108**(51): p. 20382-7.
56. Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, *Exon shuffling by L1 retrotransposition.* Science, 1999. **283**(5407): p. 1530-4.
57. Symer, D.E., et al., *Human l1 retrotransposition is associated with genetic instability in vivo.* Cell, 2002. **110**(3): p. 327-38.
58. Gilbert, N., S. Lutz-Prigge, and J.V. Moran, *Genomic deletions created upon LINE-1 retrotransposition.* Cell, 2002. **110**(3): p. 315-25.
59. Wei, W., et al., *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol, 2001. **21**(4): p. 1429-39.
60. Batzer, M.A. and P.L. Deininger, *Alu repeats and human genomic diversity.* Nat Rev Genet, 2002. **3**(5): p. 370-9.
61. Paolella, G., et al., *The Alu family repeat promoter has a tRNA-like bipartite structure.* EMBO J, 1983. **2**(5): p. 691-6.
62. Chu, W.M., W.M. Liu, and C.W. Schmid, *RNA polymerase III promoter and terminator elements affect Alu RNA expression.* Nucleic Acids Res, 1995. **23**(10): p. 1750-7.
63. Ahl, V., et al., *Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation.* Mol Cell, 2015. **60**(5): p. 715-727.
64. Bennett, E.A., et al., *Active Alu retrotransposons in the human genome.* Genome Res, 2008. **18**(12): p. 1875-83.
65. Price, A.L., E. Eskin, and P.A. Pevzner, *Whole-genome analysis of Alu repeat elements reveals complex evolutionary history.* Genome Res, 2004. **14**(11): p. 2245-52.
66. Wang, H., et al., *SVA Elements: A Hominid-specific Retroposon Family.* Journal of Molecular Biology, 2005. **354**: p. 994-1007.
67. Bantysh, O.B. and A.A. Buzdin, *Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA.* Biochemistry (Moscow), 2009. **74**: p. 1393-9.
68. Tang, W., et al., *Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase.* DNA Res, 2018. **25**(5): p. 521-533.
69. Damert, A., et al., *5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome.* Genome Res, 2009. **19**(11): p. 1992-2008.
70. Hancks, D.C., et al., *Exon-trapping mediated by the human retrotransposon SVA.* Genome Res, 2009. **19**(11): p. 1983-91.
71. Bantysh, O.B. and A.A. Buzdin, *Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA.* Biochemistry. Biokhimiia, 2009. **74**: p. 1393-9.
72. Bennett, E.A., et al., *Natural Genetic Variation Caused by Transposable Elements in Humans.* Genetics, 2004. **168**: p. 933-951.
73. Garcia-Perez, J.L., et al., *LINE-1 retrotransposition in human embryonic stem cells.* Hum Mol Genet, 2007. **16**(13): p. 1569-77.
74. Wissing, S., et al., *Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility.* Hum Mol Genet, 2012. **21**(1): p. 208-18.

75. Friedli, M., et al., *Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency.* Genome Res, 2014. **24**(8): p. 1251-9.
76. Guenther, M.G., et al., *Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells.* Cell Stem Cell, 2010. **7**(2): p. 249-57.
77. Klawitter, S., et al., *Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells.* Nat Commun, 2016. **7**: p. 10286.
78. Xing, J., et al., *Mobile elements create structural variation: Analysis of a complete human genome.* Genome Research, 2009. **19**: p. 1516-1526.
79. Faulkner, G.J., *Retrotransposons: mobile and mutagenic from conception to death.* FEBS Lett, 2011. **585**(11): p. 1589-94.
80. Kazazian, H.H., et al., *Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.* Nature, 1988. **332**(6160): p. 164-6.
81. Belancio, V.P., A.M. Roy-Engel, and P. Deininger, *The impact of multiple splice sites in human L1 elements.* Gene, 2008. **411**(1-2): p. 38-45.
82. Sorek, R., G. Ast, and D. Graur, *Alu-containing exons are alternatively spliced.* Genome Res, 2002. **12**(7): p. 1060-7.
83. Taniguchi-Ikeda, M., et al., *Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy.* Nature, 2011. **478**(7367): p. 127-31.
84. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-74.
85. Kim, D.S. and Y. Hahn, *Identification of human-specific transcript variants induced by DNA insertions in the human genome.* Bioinformatics, 2011. **27**: p. 14-21.
86. Narita, N., et al., *Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy.* J Clin Invest, 1993. **91**(5): p. 1862-7.
87. Vogt, J., et al., *SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints.* Genome Biol, 2014. **15**(6): p. R80.
88. Deininger, P.L. and M.A. Batzer, *Alu repeats and human disease.* Mol Genet Metab, 1999. **67**(3): p. 183-93.
89. Han, K., et al., *L1 recombination-associated deletions generate human genomic variation.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19366-71.
90. Nazaryan-Petersen, L., et al., *Germline Chromothripsis Driven by L1-Mediated Retrotransposition and Alu/Alu Homologous Recombination.* Hum Mutat, 2016. **37**(4): p. 385-95.
91. Bestor, T.H. and D. Bourc'his, *Transposon silencing and imprint establishment in mammalian germ cells.* Cold Spring Harb Symp Quant Biol, 2004. **69**: p. 381-7.

92. Strichman-Almashanu, L.Z., et al., *A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes.* Genome Res, 2002. **12**(4): p. 543-54.
93. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
94. Lee, J., et al., *High Levels of Sequence Diversity in the 5' UTRs of Human-Specific L1 Elements.* Comp Funct Genomics, 2012. **2012**: p. 129416.
95. Yang, X., et al., *Gene body methylation can alter gene expression and is a therapeutic target in cancer.* Cancer Cell, 2014. **26**(4): p. 577-90.
96. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
97. Garcia-Perez, J.L., et al., *Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells.* Nature, 2010. **466**(7307): p. 769-73.
98. van den Hurk, J.A., et al., *L1 retrotransposition can occur early in human embryonic development.* Hum Mol Genet, 2007. **16**(13): p. 1587-92.
99. Miki, Y., et al., *Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer.* Cancer Res, 1992. **52**(3): p. 643-5.
100. Helman, E., et al., *Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing.* Genome Res, 2014. **24**(7): p. 1053-63.
101. Rodríguez-Martín, C., et al., *Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene.* J Hum Genet, 2016. **61**(5): p. 463-6.
102. Babushok, D.V., et al., *L1 integration in a transgenic mouse model.* Genome Res, 2006. **16**(2): p. 240-50.
103. Kano, H., et al., *L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism.* Genes Dev, 2009. **23**(11): p. 1303-12.
104. An, W., et al., *Active retrotransposition by a synthetic L1 element in mice.* Proc Natl Acad Sci U S A, 2006. **103**(49): p. 18662-7.
105. Muotri, A.R., et al., *Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.* Nature, 2005. **435**(7044): p. 903-10.
106. Coufal, N.G., et al., *L1 retrotransposition in human neural progenitor cells.* Nature, 2009. **460**: p. 1127-1131.
107. Kubo, S., et al., *L1 retrotransposition in nondividing and primary human somatic cells.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 8036-41.
108. Faulkner, G.J. and J.L. Garcia-Perez, *L1 Mosaicism in Mammals: Extent, Effects, and Evolution.* Trends Genet, 2017. **33**(11): p. 802-816.
109. Savage, A.L., et al., *Frequency and methylation status of selected retrotransposition competent L1 loci in amyotrophic lateral sclerosis.* Mol Brain, 2020. **13**(1): p. 154.
110. Sanchez-Luque, F.J., et al., *LINE-1 Evasion of Epigenetic Repression in Humans.* Mol Cell, 2019. **75**(3): p. 590-604.e12.
111. Dubnau, J., *The Retrotransposon storm and the dangers of a Collyer's genome.* Current Opinion in Genetics and Development, 2018. **49**: p. 95-105.
112. Zhao, X. and D.L. Moore, *Neural stem cells: developmental mechanisms and disease modeling.* Cell Tissue Res, 2018. **371**(1): p. 1-6.

113. Mao, Z., et al., *DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells.* Cell Cycle, 2008. **7**(18): p. 2902-6.
114. Rodgers, K. and M. McVey, *Error-prone repair of DNA double-strand breaks.* J Cell Physiol.
115. Lips, J. and B. Kaina, *DNA double-strand breaks trigger apoptosis in p53-deficient fibroblasts.* Carcinogenesis, 2001. **22**(4): p. 579-85.
116. Gasior, S.L., et al., *The human LINE-1 retrotransposon creates DNA double-strand breaks.* J Mol Biol, 2006. **357**(5): p. 1383-93.
117. Erwin, J.A., et al., *L1-associated genomic regions are deleted in somatic cells of the healthy human brain.* Nat Neurosci, 2016. **19**(12): p. 1583-1591.
118. Blaudin de Thé, F.X., et al., *Engrailed homeoprotein blocks degeneration in adult dopaminergic neurons through LINE-1 repression.* EMBO J, 2018. **37**(15).
119. De Cecco, M., et al., *L1 drives IFN in senescent cells and promotes age-associated inflammation.* Nature, 2019. **566**(7742): p. 73-78.
120. Simon, M., et al., *LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation.* Cell Metab, 2019. **29**(4): p. 871-885.e5.
121. Thomas, C.A., et al., *Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation.* Cell Stem Cell, 2017. **21**(3): p. 319-331.e8.
122. Gorbunova, V., et al., *The role of retrotransposable elements in ageing and age-associated diseases.* Nature, 2021. **596**(7870): p. 43-53.
123. Ahmad, S., et al., *Breaching Self-Tolerance to Alu Duplex RNA Underlies MDA5-Mediated Inflammation.* Cell, 2018. **172**(4): p. 797-810.e13.
124. Zhao, K., et al., *LINE1 contributes to autoimmunity through both RIG-I- and MDA5-mediated RNA sensing pathways.* J Autoimmun, 2018. **90**: p. 105-115.
125. Li, W., et al., *Human endogenous retrovirus-K contributes to motor neuron disease.* Sci Transl Med, 2015. **7**(307): p. 307ra153.
126. Steele, A.J., et al., *Detection of serum reverse transcriptase activity in patients with ALS and unaffected blood relatives.* Neurology, 2005. **64**: p. 454-458.
127. Savage, A.L., et al., *Retrotransposons in the development and progression of amyotrophic lateral sclerosis.* Journal of Neurology, Neurosurgery & Psychiatry, 2018: p. jnnp-2018-319210.
128. Czech, B. and G.J. Hannon, *One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing.* Trends Biochem Sci, 2016. **41**(4): p. 324-337.
129. Malone, C.D. and G.J. Hannon, *Small RNAs as guardians of the genome.* Cell, 2009. **136**(4): p. 656-68.
130. Aravin, A.A., et al., *A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice.* Mol Cell, 2008. **31**(6): p. 785-99.
131. Ha, H., et al., *A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements.* BMC Genomics, 2014. **15**: p. 545.
132. Marchetto, M.C.N., et al., *Differential L1 regulation in pluripotent stem cells of humans and apes.* Nature, 2013. **503**(7477): p. 525-529.
133. Ecco, G., M. Imbeault, and D. Trono, *KRAB zinc finger proteins.* Development, 2017. **144**: p. 2719-2729.

134. Robbez-Masson, L., et al., *The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes.* Genome Res, 2018. **28**(6): p. 836-845.
135. Rowe, H.M. and D. Trono, *Dynamic control of endogenous retroviruses during development.* Virology, 2011. **411**(2): p. 273-87.
136. Wolf, G., et al., *The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses.* Genes Dev, 2015. **29**(5): p. 538-54.
137. Kauzlaric, A., et al., *The mouse genome displays highly dynamic populations of KRAB-zinc finger protein genes and related genetic units.* PLoS One, 2017. **12**(3): p. e0173746.
138. Imbeault, M. and D. Trono, *As time goes by: KRABs evolve to KAP endogenous retroelements.* Dev Cell, 2014. **31**(3): p. 257-258.
139. Imbeault, M., P.Y. Helleboid, and D. Trono, *KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks.* Nature, 2017. **543**: p. 550-554.
140. Rowe, H.M., et al., *TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells.* Genome Res, 2013. **23**(3): p. 452-61.
141. Jacobs, F.M., et al., *An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons.* Nature, 2014. **516**(7530): p. 242-5.
142. Haring, N.L., et al., *ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters.* Genome Res, 2021. **31**(4): p. 551-563.
143. Pontis, J., et al., *Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs.* Cell Stem Cell, 2019. **24**(5): p. 724-735.e5.
144. Emerson, R.O. and J.H. Thomas, *Adaptive evolution in zinc finger transcription factors.* PLoS Genet, 2009. **5**(1): p. e1000325.
145. Castro-Diaz, N., et al., *Evolutionally dynamic L1 regulation in embryonic stem cells.* Genes Dev, 2014. **28**(13): p. 1397-409.
146. Thomas, J.H. and S. Schneider, *Coevolution of retroelements and tandem zinc finger genes.* Genome Res, 2011. **21**(11): p. 1800-12.
147. Liu, H., et al., *Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies.* Genome Biol Evol, 2014. **6**(3): p. 510-25.
148. Faulkner, G.J., et al., *The regulated retrotransposon transcriptome of mammalian cells.* Nat Genet, 2009. **41**(5): p. 563-71.
149. Ecco, G., et al., *Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues.* Dev Cell, 2016. **36**(6): p. 611-23.
150. Chuong, E.B., N.C. Elde, and C. Feschotte, *Regulatory activities of transposable elements: from conflicts to benefits.* Nature Reviews Genetics, 2017. **18**: p. 71-86.
151. Wang, J., et al., *Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells.* Nature, 2014. **516**(7531): p. 405-9.
152. Corsinotti, A., et al., *Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells.* PLoS One, 2013. **8**(2): p. e56721.

153. Han, K., et al., *Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages.* Nucleic Acids Res, 2005. **33**(13): p. 4040-52.
154. Callinan, P.A., et al., *Alu retrotransposition-mediated deletion.* J Mol Biol, 2005. **348**(4): p. 791-800.
155. Lee, J., et al., *Human Genomic Deletions Generated by SVA-Associated Events.* Comp Funct Genomics, 2012. **2012**: p. 807270.
156. Xia, B., et al., *The genetic basis of tail-loss evolution in humans and apes.* bioRxiv, 2021: p. 2021.09.14.460388.
157. Young, N.M., G.P. Wagner, and B. Hallgrimsson, *Development and the evolvability of human limbs.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3400-5.
158. Kunarso, G., et al., *Transposable elements have rewired the core regulatory network of human embryonic stem cells.* Nat Genet, 2010. **42**(7): p. 631-4.
159. Schmidt, D., et al., *Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.* Cell, 2012. **148**: p. 335-48.
160. Wang, T., et al., *Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53.* Proc Natl Acad Sci U S A, 2007. **104**(47): p. 18613-8.
161. Sundaram, V., et al., *Widespread contribution of transposable elements to the innovation of gene regulatory networks.* Genome Res, 2014. **24**(12): p. 1963-76.
162. Marnetto, D., et al., *Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion.* Am J Hum Genet, 2018. **102**(2): p. 207-218.
163. Lynch, V.J., et al., *Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.* Nat Genet, 2011. **43**(11): p. 1154-9.
164. Trizzino, M., et al., *Transposable elements are the primary source of novelty in primate gene regulation.* Genome Res, 2017. **27**(10): p. 1623-1633.
165. Prescott, S.L., et al., *Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest.* Cell, 2015. **163**(1): p. 68-83.
166. Cohen, C.J., W.M. Lock, and D.L. Mager, *Endogenous retroviral LTRs as promoters for human genes: a critical assessment.* Gene, 2009. **448**(2): p. 105-14.
167. Conley, A.B., J. Piriyapongsa, and I.K. Jordan, *Retroviral promoters in the human genome.* Bioinformatics, 2008. **24**(14): p. 1563-7.
168. Nigumann, P., et al., *Many human genes are transcribed from the antisense promoter of L1 retrotransposon.* Genomics, 2002. **79**(5): p. 628-34.
169. Babaian, A. and D.L. Mager, *Endogenous retroviral promoter exaptation in human cancer.* Mob DNA, 2016. **7**: p. 24.
170. Kellner, M. and W. Makałowski, *Transposable elements significantly contributed to the core promoters in the human genome.* Sci China Life Sci, 2019. **62**(4): p. 489-497.
171. Jönsson, M.E., et al., *Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors.* Nature Communications, 2019. **10**(1): p. 3182.

172. Mätlik, K., K. Redik, and M. Speek, *L1 antisense promoter drives tissue-specific transcription of human genes.* J Biomed Biotechnol, 2006. **2006**(1): p. 71753.
173. Savage, A.L., et al., *Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns.* BMC Evolutionary Biology, 2013. **13**: p. 101.
174. Szpakowski, S., et al., *Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements.* Gene, 2009. **448**: p. 151-167.
175. Turker, M.S., *Gene silencing in mammalian cells and the spread of DNA methylation.* Oncogene, 2002. **21**(35): p. 5388-93.
176. Kim, J.K., M. Samaranayake, and S. Pradhan, *Epigenetic mechanisms in mammals.* Cell Mol Life Sci, 2009. **66**(4): p. 596-612.
177. Yates, P.A., et al., *Silencing of mouse Aprt is a gradual process in differentiated cells.* Mol Cell Biol, 2003. **23**(13): p. 4461-70.
178. Wang, X., et al., *Spreading of Alu methylation to the promoter of the MLH1 gene in gastrointestinal cancer.* PLoS One, 2011. **6**(10): p. e25913.
179. Zabolotneva, A.A., et al., *Transcriptional regulation of human-specific SVAF1retrotransposons by cis-regulatory MAST2 sequences.* Gene, 2012. **505**: p. 128-136.
180. Pugacheva, E.M., et al., *The cancer-associated CTCFL/BORIS protein targets multiple classes of genomic repeats, with a distinct binding and functional preference for humanoid-specific SVA transposable elements.* Epigenetics & Chromatin, 2016. **9**: p. 35.
181. Savage, A.L., et al., *An Evaluation of a SVA Retrotransposon in the FUS Promoter as a Transcriptional Regulator and Its Association to ALS.* PLoS ONE, 2014. **9**: p. e90833.
182. Vasiliou, S.A., et al., *The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro.* Addict Biol, 2012. **17**(1): p. 156-70.
183. Westenberger, A., et al., *A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism.* Annals of Neurology, 2019. **85**(6): p. 812-822.
184. Aneichyk, T., et al., *Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly.* Cell, 2018. **172**: p. 897-909.e21.
185. Bragg, D.C., et al., *Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1.* Proceedings of the National Academy of Sciences, 2017. **114**: p. E11020-E11028.
186. Petrozziello, T., et al., *SVA insertion in X-linked Dystonia Parkinsonism alters histone H3 acetylation associated with TAF1 gene.* PLoS One, 2020. **15**(12): p. e0243655.
187. Stacey, S.N., et al., *Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer.* Human Molecular Genetics, 2016. **25**: p. 1008-1018.
188. Gianfrancesco, O., et al., *The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome.* Int J Mol Sci, 2019. **20**(23).
189. Goerner-Potvin, P. and G. Bourque, *Computational tools to unmask transposable elements.* Nature Reviews Genetics, 2018. **19**: p. 688-704.

190. Price, A.L., N.C. Jones, and P.A. Pevzner, *De novo identification of repeat families in large genomes.* Bioinformatics, 2005. **21 Suppl 1**: p. i351-8.
191. Smit, A., R. Hubley, and P. Green. *RepeatMasker Open-4.0.* 2013-2015; Available from: http://www.repeatmasker.org.
192. Bao, W., K.K. Kojima, and O. Kohany, *Repbase Update, a database of repetitive elements in eukaryotic genomes.* Mob DNA, 2015. **6**: p. 11.
193. Vendrell-Mir, P., et al., *A benchmark of transposon insertion detection tools using real data.* Mob DNA, 2019. **10**: p. 53.
194. Gibbs, J.R., et al., *Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.* PLoS Genet, 2010. **6**(5): p. e1000952.
195. Hernandez, D.G., et al., *Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain.* Neurobiol Dis, 2012. **47**(1): p. 20-8.
196. Kumar, A., et al., *Age-associated changes in gene expression in human brain and isolated neurons.* Neurobiol Aging, 2013. **34**(4): p. 1199-209.
197. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
198. Bustin, S.A., et al., *The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.* Clin Chem, 2009. **55**(4): p. 611-22.
199. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-2308.
200. Holmlund, C., et al., *Characterization and tissue-specific expression of human LRIG2.* Gene, 2004. **332**: p. 35-43.
201. Stuart, Helen M., et al., *LRIG2 Mutations Cause Urofacial Syndrome.* The American Journal of Human Genetics, 2013. **92**: p. 259-264.
202. Newman, W.G. and A.S. Woolf, *Urofacial Syndrome*, in *GeneReviews((R))*, M.P. Adam, et al., Editors. 1993: Seattle (WA).
203. Roberts, N.A., et al., *Lrig2 and Hpse2, mutated in urofacial syndrome, pattern nerves in the urinary bladder.* Kidney Int, 2019. **95**(5): p. 1138-1152.
204. Seila, A.C., et al., *Divergent transcription from active promoters.* Science, 2008. **322**(5909): p. 1849-51.
205. Haberle, V. and A. Stark, *Eukaryotic core promoters and the functional basis of transcription initiation.* Nat Rev Mol Cell Biol, 2018. **19**(10): p. 621-637.
206. Wang, J., et al., *dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans.* Human Mutation, 2006. **27**: p. 323-329.
207. de Koning, A.P.J., et al., *Repetitive Elements May Comprise Over Two-Thirds of the Human Genome.* PLoS Genetics, 2011. **7**: p. e1002384.
208. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
209. Halldorsson, B.V., S. Istrail, and F.M. De La Vega, *Optimal selection of SNP markers for disease association studies.* Hum Hered, 2004. **58**(3-4): p. 190-202.
210. Moore, L.D., T. Le, and G. Fan, *DNA methylation and its basic function.* Neuropsychopharmacology, 2013. **38**(1): p. 23-38.
211. ATCC. *SH-SY5Y*. Available from: https://www.atcc.org/products/crl-2266#detailed-product-information.

212. Jacob, A.G. and C.W.J. Smith, *Intron retention as a component of regulated gene expression programs.* Hum Genet, 2017. **136**(9): p. 1043-1057.
213. Core, L.J., et al., *Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers.* Nat Genet, 2014. **46**(12): p. 1311-20.
214. Geiman, T.M., et al., *DNMT3B interacts with hSNF2H chromatin remodeling enzyme, HDACs 1 and 2, and components of the histone methylation system.* Biochem Biophys Res Commun, 2004. **318**(2): p. 544-55.
215. Fuks, F., et al., *DNA methyltransferase Dnmt1 associates with histone deacetylase activity.* Nat Genet, 2000. **24**(1): p. 88-91.
216. Fuks, F., et al., *The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase.* Nucleic Acids Res, 2003. **31**(9): p. 2305-12.
217. Pollard, M.O., et al., *Long reads: their purpose and place.* Hum Mol Genet, 2018. **27**(R2): p. R234-R241.
218. Gardner, E.J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Mills, R. E., Pittard, W. S., 1000 Genomes Project Consortium & Devine, S. E., *The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology.* Genome Research, 2017.
219. Strang, K.H., T.E. Golde, and B.I. Giasson, *MAPT mutations, tauopathy, and mechanisms of neurodegeneration.* Lab Invest, 2019. **99**(7): p. 912-928.
220. Zhang, X., et al., *Tau Pathology in Parkinson's Disease.* Front Neurol, 2018. **9**: p. 809.
221. Stefansson, H., et al., *A common inversion under selection in Europeans.* Nat Genet, 2005. **37**(2): p. 129-37.
222. Pittman, A.M., et al., *Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration.* J Med Genet, 2005. **42**(11): p. 837-46.
223. Wall, J.D. and J.K. Pritchard, *Haplotype blocks and linkage disequilibrium in the human genome.* Nat Rev Genet, 2003. **4**(8): p. 587-97.
224. Evans, W., et al., *The tau H2 haplotype is almost exclusively Caucasian in origin.* Neurosci Lett, 2004. **369**(3): p. 183-5.
225. Sirugo, G., S.M. Williams, and S.A. Tishkoff, *The Missing Diversity in Human Genetic Studies.* Cell, 2019. **177**(4): p. 1080.
226. Pastor, P., et al., *Significant association between the tau gene A0/A0 genotype and Parkinson's disease.* Ann Neurol, 2000. **47**(2): p. 242-5.
227. Myers, A.J., et al., *The H1c haplotype at the MAPT locus is associated with Alzheimer's disease.* Hum Mol Genet, 2005. **14**(16): p. 2399-404.
228. Baker, M., et al., *Association of an extended haplotype in the tau gene with progressive supranuclear palsy.* Hum Mol Genet, 1999. **8**(4): p. 711-5.
229. Soutar, M.P.M., et al., *Regulation of mitophagy by the NSL complex underlies genetic risk for Parkinson's disease at Chr16q11.2 and on the MAPT H1 allele.* bioRxiv, 2021: p. 2020.01.06.896241.
230. Plotegher, N. and M.R. Duchen, *Crosstalk between Lysosomes and Mitochondria in Parkinson's Disease.* Front Cell Dev Biol, 2017. **5**: p. 110.
231. Darnell, J.E., et al., *Polyadenylic acid sequences: role in conversion of nuclear RNA into messenger RNA.* Science, 1971. **174**(4008): p. 507-10.

232. Choi, Y.H. and C.H. Hagedorn, *Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3' poly(A) end phenotype.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 7033-8.
233. O'Brien, H.E., et al., *Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders.* Genome Biol, 2018. **19**(1): p. 194.
234. Weber, M., et al., *Cytogenomics of six human trophoblastic cell lines.* Placenta, 2021. **103**: p. 72-75.
235. Marchetti, B., *Wnt/β-Catenin Signaling Pathway Governs a Full Program for Dopaminergic Neuron Survival, Neurorescue and Regeneration in the MPTP Mouse Model of Parkinson's Disease.* Int J Mol Sci, 2018. **19**(12).
236. He, L., et al., *Exome-wide age-of-onset analysis reveals exonic variants in ERN1 and SPPL2C associated with Alzheimer's disease.* Transl Psychiatry, 2021. **11**(1): p. 146.
237. Allen, M., et al., *Gene expression, methylation and neuropathology correlations at progressive supranuclear palsy risk loci.* Acta Neuropathol, 2016. **132**(2): p. 197-211.
238. Sánchez-Juan, P., et al., *The MAPT H1 Haplotype Is a Risk Factor for Alzheimer's Disease in APOE ε4 Non-carriers.* Front Aging Neurosci, 2019. **11**: p. 327.
239. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome Res, 2008. **18**(11): p. 1752-62.
240. Kadauke, S. and G.A. Blobel, *Chromatin loops in gene regulation.* Biochim Biophys Acta, 2009. **1789**(1): p. 17-25.
241. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
242. Bonev, B., et al., *Multiscale 3D Genome Rewiring during Mouse Neural Development.* Cell, 2017. **171**(3): p. 557-572.e24.
243. Robson, M.I., A.R. Ringel, and S. Mundlos, *Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D.* Mol Cell, 2019. **74**(6): p. 1110-1122.
244. de Laat, W. and D. Duboule, *Topology of mammalian developmental enhancers and their regulatory landscapes.* Nature, 2013. **502**(7472): p. 499-506.
245. Lorberbaum, D.S., et al., *An ancient yet flexible cis-regulatory architecture allows localized Hedgehog tuning by patched/Ptch1.* Elife, 2016. **5**.
246. Schoenfelder, S., et al., *Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells.* Nat Genet, 2010. **42**(1): p. 53-61.
247. Fanucchi, S., et al., *Chromosomal contact permits transcription between coregulated genes.* Cell, 2013. **155**(3): p. 606-20.
248. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
249. Osterwalder, M., et al., *Enhancer redundancy provides phenotypic robustness in mammalian development.* Nature, 2018. **554**(7691): p. 239-243.

250. Lupiáñez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-1025.
251. Dixon, J.R., et al., *Integrative detection and analysis of structural variation in cancer genomes.* Nat Genet, 2018. **50**(10): p. 1388-1398.
252. Roy, S.S., A.K. Mukherjee, and S. Chowdhury, *Insights about genome function from spatial organization of the genome.* Hum Genomics, 2018. **12**(1): p. 8.
253. Wang, D.C., et al., *A tour of 3D genome with a focus on CTCF.* Seminars in Cell & Developmental Biology, 2018.
254. Parelho, V., et al., *Cohesins functionally associate with CTCF on mammalian chromosome arms.* Cell, 2008. **132**(3): p. 422-33.
255. Fudenberg, G., et al., *Formation of Chromosomal Domains by Loop Extrusion.* Cell Rep, 2016. **15**(9): p. 2038-49.
256. Nora, E.P., et al., *Molecular basis of CTCF binding polarity in genome folding.* Nat Commun, 2020. **11**(1): p. 5612.
257. Hansen, A.S., et al., *Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF.* Mol Cell, 2019. **76**(3): p. 395-411.e13.
258. Choudhary, M.N., et al., *Co-opted transposons help perpetuate conserved higher-order chromosomal structures.* Genome Biol, 2020. **21**(1): p. 16.
259. Zhang, D., et al., *Alteration of genome folding via contact domain boundary insertion.* Nat Genet, 2020. **52**(10): p. 1076-1087.
260. Zhang, Y., et al., *Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells.* Nat Genet, 2019. **51**(9): p. 1380-1388.
261. Kruse, K., et al., *Transposable elements drive reorganisation of 3D chromatin during early embryogenesis.* bioRxiv, 2019: p. 523712.
262. Wang, H., et al., *Widespread plasticity in CTCF occupancy linked to DNA methylation.* Genome Res, 2012. **22**(9): p. 1680-8.
263. Wüllner, U., et al., *DNA methylation in Parkinson's disease.* J Neurochem, 2016. **139 Suppl 1**: p. 108-120.
264. Corces, M.R., et al., *Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases.* Nature Genetics, 2020. **52**(11): p. 1158-1168.
265. Farrow, S.L., et al., *Establishing gene regulatory networks from Parkinson's disease risk loci.* Brain, 2022. **145**(7): p. 2422-2435.
266. Lee, A.J., et al., *Characterization of altered molecular mechanisms in Parkinson's disease through cell type-resolved multi-omics analyses.* bioRxiv, 2022: p. 2022.02.13.479386.
267. Belton, J.M., et al., *Hi-C: a comprehensive technique to capture the conformation of genomes.* Methods, 2012. **58**(3): p. 268-76.
268. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.* Bioinformatics, 2016. **32**(8): p. 1220-2.
269. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.

270. Klambauer, G., et al., *cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate.* Nucleic Acids Res, 2012. **40**(9): p. e69.
271. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-6.
272. Choi, W.Y., et al., *Chromatin Interaction Changes during the iPSC-NPC Model to Facilitate the Study of Biologically Significant Genes Involved in Differentiation.* Genes (Basel), 2020. **11**(10).
273. Kenna, K.P., et al., *NEK1 variants confer susceptibility to amyotrophic lateral sclerosis.* Nat Genet, 2016. **48**(9): p. 1037-42.
274. Wang, H., et al., *NEK1-mediated retromer trafficking promotes blood-brain barrier integrity by regulating glucose metabolism and RIPK1 activation.* Nat Commun, 2021. **12**(1): p. 4826.
275. Quinn, J.P. and V.J. Bubb, *SVA retrotransposons as modulators of gene expression.* Mobile Genetic Elements, 2014. **4**: p. e32102.
276. Gianfrancesco, O., V.J. Bubb, and J.P. Quinn, *SVA retrotransposons as potential modulators of neuropeptide gene expression.* Neuropeptides, 2017. **64**: p. 3-7.
277. Nuñez, J.K., et al., *Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing.* Cell, 2021. **184**(9): p. 2503-2519.e17.