The role of endogenous retrotransposable elements in Parkinson's disease

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

Ben Middlehurst MRes

May 2021

Acknowledgements

First and foremost, I would like to thank my PhD supervisors, John Quinn and Jill Bubb for their continued support throughout my studies. I have gained so much experience and confidence in my own abilities throughout the course of my PhD which will last a lifetime. The opportunities given to me have allowed me to travel to both Germany and the USA which has broadened my skill set and contacts massively and for that I thank them!

Day to day life as a lab researcher was made all the more enjoyable by the peers that I worked with during my PhD through which I have made new lifelong friendships. I particularly want to thank my laddy, Jack Marshall, who has been one of the best science buddies and close friends throughout both my professional and social life and always knows how to put a smile on my face even in the toughest times! I also want to say the biggest thank you to everyone in the lab during my time with particular mention to Ashley Hall, Emma Price, Anna Illera, Kimberley Billingsley, Olympia Gianfrancesco, Veridiana Pessoa, Maurizio Manca and Abigail Pfaff. They were all extremely helpful in developing me as a scientist at various points during my master's and PhD studies and have given me immeasurable support both inside and outside of the lab!

Last but not least, I wish to thank my parents, who not only have put up with me living at home throughout my studies with all the ups and downs, but also provided me with the ultimate support. I have always been extremely grateful for everything they do for me and I cannot thank them enough for their help finding my feet.

Thank you to the Wellcome Trust for funding my PhD and providing me the opportunity for such a prestigious accolade.

Contents

List of figures and tables

Chapter 1 – General introduction

Chapter 2 – Materials and methods

Chapter 3 - Investigating an SVA retrotransposon within LRRK2 as a novel regulator of genetic function

Chapter 4 – Understanding the SVA retrotransposon architecture of the PD associated locus *INPP5F*/*BAG3*/*TIAL1*

Chapter 5 – Identifying novel retrotransposon insertion polymorphisms in Parkinson's disease using next generation sequencing

Abbreviations

AD	Alzheimer's disease
ALS	Amyotrophic Lateral Sclerosis
BLAST	Basic local alignment search tool
bp	Base pair
Cas9	CRISPR associate protein 9
ChIP-Seq	Chromatin immunoprecipitation sequencing
CMV	Cytomegalovirus
COR	C-terminal of Roc
CRISPR	Clustered regularly interspaced short palindromic repeats
dNTP	Deoxynucleotide triphosphate
FTLD	Frontotemporal lobar degeneration
GWAS	Genome wide association studies
HA	Healthy aged
HERV	Human endogenous retrovirus
iPSCs	Induced pluripotent stem cells
kb	Kilobase
kDa	Kilodalton
KO	Knockout
LD	Linkage disequilibrium
LDSC	LD score regression
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
NABEC	North American brain expression consortium
NT	Non-targeting
ORF	Open reading frame
PCR	Polymerase chain reaction

PD	Parkinson's disease
PRS	Polygenic risk score
QC	Quality control
qPCR	Quantitative PCR
RC-Seq	Retrotransposon capture sequencing
Roc	Ras-of-complex
RPKM	Reads per kilobase of transcript, per million mapped reads
RT-PCR	Reverse transcription PCR
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SV40	Simian virus 40
SVA	SINE-VNTR-Alu
TE	Transposable element
TPI	Triose phosphate isomerase
TSD	Target site duplication
TSS	Transcriptional start site
UCSC	University California Santa Cruz
UTR	Untranslated region
VNTR	Variable number tandem repeat
WGS	Whole genome sequencing
WT	Wild type
XDP	X-linked dystonia Parkinsonism

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder resulting in significant damage to dopaminergic neurons within the basal ganglia of the brain leading to a deficiency of dopamine. PD is classed as a polygenic disease with multiple known mutations which infer risk of development and impact the age of disease onset. The latest genome-wide association studies (GWAS) indicate approximately 90 independent risk loci that are suspected to be involved in the progression of PD which highlights the complexity of the genetic contribution to disease. However, current estimates only attribute approximately 10% of all PD cases to a known genetic mutation, with the majority of GWAS signals being present in non-coding DNA regions. The over-arching theme of the projects presented within this thesis was to explore the potential role of retrotransposable elements in the aetiology of Parkinson's disease as contributing factors to the missing heritability demonstrated by GWAS. Retrotransposable elements are amongst the only remaining active transposable elements in the human genome (primarily LINE-1) which contribute to genomic diversity and evolution through the generation of novel retrotransposition events. These mutations can be advantageous or pathogenic depending on the context of the insertion, for example, a well characterised pathogenic SVA insertion in the *TAF1* locus causes XDP in small populations within the Philippines.

The aims of these studies were to characterise the multi-facets of TE regulation in PD related loci by studying the effects of retrotransposon insertion polymorphisms (RIPs) of the LINE-1, SVA and *Alu* sub-classes on a global scale as well as the effects of fixed SVA retrotransposons in local gene contexts using defined PD related genes. Presented here is the first described example of the use of next generation sequencing (NGS) techniques (RC-Seq and WGS) to identify RIPs within DNA extracted from brain tissue (frontal cortex and cerebellum) of PD and control samples which can act as both somatic *de novo* mutations and predisposition variants and may correlate with disease progression. RC-Seq provided a highly sensitive approach for the detection of LINE-1 RIPs using a capture-sequencing approach with the aim to study somatic *de novo* insertions both in terms of overall numbers and distribution

of the identified insertions. WGS provided a less targeted approach for the detection of LINE-1, SVA and *Alu* RIPs with genome wide coverage. The results of the RC-Seq analysis showed trends of increased polymorphic and somatic LINE-1 elements within the brain of PD patients compared to the control group. The distribution of LINE-1 RIPs between the PD and control groups also differed with more putative somatic insertions being found within PD related haploblocks in the PD group. WGS analysis confirmed the findings of the RC-Seq analysis and suggested that the distribution of SVA RIPs also differ between the PD and control groups. Furthermore, a fundamental analysis of the identified RIPs provided supporting evidence for the previously suggested hypothesis that retrotransposons preferentially insert into brain related genes and loci in a non-pathological context.

In contrast to the global analysis of retrotransposons, two previously described PD risk loci, *LRRK2* and *INPP5F* were studied as they contained three reference SVAs of interest. Using PCR genotyping in case/control matched PD human DNA samples, luciferase-based reporter gene constructs and CRISPR mediated SVA knockouts in the HEK293 cell line, the SVAs in these loci were shown to be functionally active, eliciting repressive effects in multiple reporter gene assays. The PCR genotyping showed that each SVA contained multiple polymorphic domains within the tested populations but did not identify any statistically significant enrichments of genotype or allele frequencies between PD case and control cohorts but does not exclude the possibility for these elements to confer risk in other populations not tested within the scope of this study.

Hypothesis: Retrotransposable elements which are known to be active in the human genome, in particular within the central nervous system, possess the potential to regulate expression of key PD related genes and polymorphisms which confer risk for the development of PD.

Chapter 1 – General introduction

Chapter 1 - General Introduction

1.1 Parkinson's disease

Parkinson's disease (PD) is a chronic, neurodegenerative disorder that affects as many as one in a hundred people over the age of 60, but early onset forms are also common. First described and named by James Parkinson in 1817, the exact mechanisms behind PD development have remained elusive. PD is characterised by a selective loss of dopaminergic neurons in the substantia nigra and the presence of Lewy bodies, leading to a reduction in levels of dopamine and subsequent uncontrolled firing of motor neurons, causing the characteristic resting tremors associated with PD [1]. Other symptoms such as bradykinesia, rapid eye movement sleep behaviour disorder, hyposmia and constipation are amongst the most phenotypic symptoms with a variety of other non-motor symptoms also common, including psychological problems such as cognitive decline, depression and anxiety [2]. The range of symptom presentation indicates the complexity of Parkinson's disease and is likely reflective of the wide range of cellular processes and genetic involvement in the disorder. Lewy body inclusions are defining characteristics of both PD and dementia with Lewy bodies (DLB) that are composed of protein aggregates, primarily α-synuclein [3]. There is a requirement for a deeper knowledge of the molecular mechanisms and cellular pathways underlying PD with a growing need for biomarkers and objective tests for diagnostic purposes. Current methods of diagnosis have severe limitations as they are based on phenotypic assessment of symptoms including ataxia and tremors [4]. At the point phenotypic symptoms present themselves, there has already been significant irreversible damage sustained to the

dopaminergic neurons of the basal ganglia. In an attempt to address this, much research has been invested in identifying genetic influences for the causes and progression of PD in a bid to find novel biomarkers and allow for early diagnosis and treatments.

Considering over 200 years of medical research has been conducted since the first description of Parkinson's disease, 90% of PD cases remain idiopathic, highlighting the complexity of this disease [5]. With the high incidence rate and idiopathic nature of PD, there exists a void in the understanding of the pathological nature of the genetic component present within this disease. Missense mutations in the leucine rich repeat kinase 2 (*LRRK2*) gene contribute the largest known cause of familial PD, accounting for approximately 3% of all PD cases [6]. Although monogenic forms of PD have been described, whereby single mutations of known risk genes, primarily *SNCA*, *LRRK2*, *PRKN*, *PINK1*, *ATP13A2* and *PARK7*, are causative of disease, they collectively only account for approximately 30% of familial and 3-5% of sporadic PD cases [7]. These demonstrate that PD has a significant genetic component that contributes to disease, however the exact proportion of the pathogenicity for genetic and non-genetic factors is highly debated.

Patients with *SCNA* mutations often present with early-onset PD (defined as the onset of symptoms before the age of 50) however the disease in these patients often rapidly progresses with the development of Lewy bodies in the substantia nigra, locus coeruleus, hypothalamus and cerebral cortex [8]. Lewy body inclusions are a common pathological feature of PD and are composed primarily of α-synuclein aggregates (encoded by the *SNCA* gene) [9]. In small numbers of PD patients (<1%)

the aggregation of alpha-synuclein can be attributed to pathogenic single nucleotide variants (SNVs) or copy number variants (CNVs), but this is not true for the majority of PD patients. In fact, the pathology of alpha-synuclein aggregation for most PD cases is much more complex and is influenced by the disruption of many cellular processes such as mitochondrial function, lysosomal pathways, oxidative stress and neuro-inflammation [10]. The process of alpha-synuclein aggregation leading to neuronal death is suggested to act in a prion-like mechanism [11]. This would allow the protein aggregate to propagate throughout the brain causing cell death to multiple neurons and glial cells [12].

A wide array of non-genetic factors have been known for many years to impact the incidence of PD with both risk and protective factors influencing the outcome [13]. The incidence of PD is most strongly correlated with age and gender, with the majority of cases being males over the age of 60 [14]. Other factors include exposure to drugs, pesticides, solvents [15] and heavy metals such as lead [16]. A vast quantity of research explored the validity of chemical induced PD in the early 1980s and 1990s, for example a higher incidence of Parkinson like symptoms were reported with the increased use of the recreational drug, meperidine. This led to the discovery of 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine (MPTP) toxicity in dopaminergic neurons of the substantia nigra and development of PD symptoms [17, 18]. MPTP became widely used in medical research to induce Parkinsonism in animal models including primates and rodents [19, 20]. Independently, such environmental factors have all shown to significantly increase the risk of developing PD, however they do not solely explain the incidence rates. With such a heterogenous cast of both genetic and environmental factors already showcased as important aetiological factors in PD,

there is a growing interest in the role of gene-environment (GxE) interactions in the development of the disease [21].

Neuroinflammation is regarded as a critical feature in the pathogenesis of Parkinson's disease and is defined as the process whereby inflammation occurs within the central nervous system as a response to immune cell mediated release of inflammatory markers such as cytokines [22]. It is unclear if this process is causative or consequence of PD disease pathology, however it is understood that microglia play a role in neuroinflammation. Microglial cells of the CNS are quiescent in the presence of immunosuppressant molecules released from healthy neurons, such as CX3CL1, CD200, CD22, CD47, CD95 and neuronal cell adhesion molecule (NCAM) [23, 24]. Microglial activation has become a popular topic of discussion in PD and can be caused by multiple stimuli, including pathogens (viral and bacterial), aggregate molecules such as α-synuclein and β-amyloid as well as necrotic factors released from dying neurons [25].

1.2 Genetic composition of PD

The examples described previously demonstrate the complexity of PD from a proteomic context, but do not address the genetic composition of PD pathogenesis. The first genes to be identified to have direct causative effects for the development of PD were rare mutations that led to autosomal dominant or recessive PD and included *SNCA*, *PRKN*, *PARK7* and *LRRK2*. These are still regarded to be major contributors in the development of PD; however, they only represent a small portion of a much wider array of genes now implicated in PD pathology by genome wide association studies (GWAS). GWAS has become the primary method of identification

of novel PD risk loci since the use of next generation sequencing techniques. **Figure 1.1** indicates a timeline of the discovery of novel PD risk loci since 1997, highlighting how rapidly the total number of suggested important genes has progressed using GWAS. From the largest GWAS performed to date for PD by Nalls *et al.* 2019, 90 independent risk signals were identified from a meta-analysis of 7.8 million single nucleotide polymorphisms (SNPs) in 37.7K cases, 18.6K proxy-cases and 1.4 million controls (the list of 90 PD nominated rick loci can be found in **table 1.2**) [26]. Using LD score regression (LDSC) and polygenic risk scores (PRS) this study was able to estimate the heritability of the variants identified for PD at 26-36%, indicating that there is a strong genetic contribution for development of PD [26]. However, this also leaves a large unexplained proportion. GWAS has limitations with inferring biological significance from SNPs and is used rather to identify loci of interest (often genomic loci of over 1 mega base). It is possible therefore that the missing heritability not seen in GWAS could be due to other genomic elements, such as repetitive DNA, which is often excluded in bioinformatic studies. These groups of elements are often excluded due to their highly repetitive nature causing large copy numbers which cause mapping problems in next generation sequencing. Furthermore, particular sub-classes tend to have high GC contents (>80%) leading to poor sequencing quality and exclusion at quality control (QC).

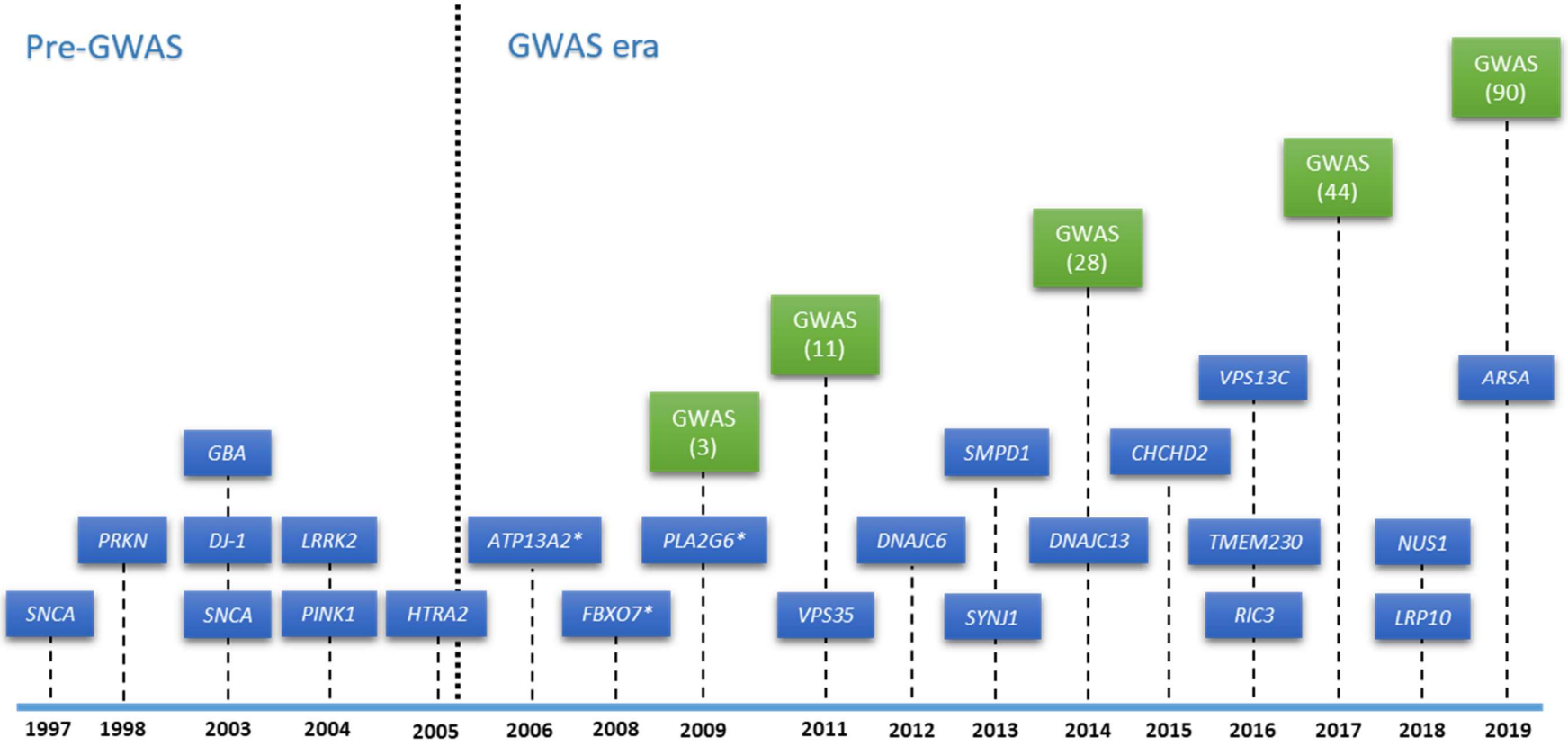


Figure 1.1 – Timeline of nominated risk loci for Parkinson's disease since 1997, adapted from Bandres-Ciga *et al.* 2020 *[27]*. Bracketed numbers indicate total numbers of risk genes identified to date. Current literature nominates 90 total genetic risk loci for Parkinson's disease as identified by GWAS analysis and illustrates the power of GWAS for identification of novel risk loci. Green squares indicate GWAS studies with the cumulative number of discovered risk loci in brackets. Asterisks indicate the genes associated with atypical parkinsonism related disorders.

1.2 Leucine-rich repeat kinase 2 (*LRRK2*)

The Michael J Fox Foundation identified mutations in the leucine-rich repeat kinase-2 (*LRRK2*) gene as the 'greatest known contributor' for Parkinson's disease to date. The *LRRK2* gene encodes the protein LRRK2, which can influence various major cellular processes such as autophagy, neurotransmission, vesicular trafficking and immune responses amongst others (**figure 1.2**) [28]. The LRRK2 protein is a multi-functional enzyme with both kinase and GTPase activity which has been implicated in the mitogen activated protein kinase (MAPK) signalling cascade. It has been proposed to act as a scaffold which may regulate the location of cellular MAPK without altering the level of activity [29]. Other theories suggest that LRRK2 may act through its binding to GTPases, GTPase exchange factors and GTPase-activating proteins such as rac1, cdc42, rab5, rab7L1, endoA, RGS2, ArfGAP1 and ArhGEF7. These interaction partners are all known to be involved in cytoskeletal outgrowth and autophagy. This proposed mechanism of action is thought to be involved with the aggregation of α-synuclein and hence cell toxicity leading to death of dopaminergic neurons and the associated characteristics of PD [30].

1.2.1 LRRK2 protein mutations and involvement in disease

The LRRK2 protein is associated with intracellular membranes, vesicle structures and cytosol including early endosomes, lysosomes, synaptic vesicles, endoplasmic reticulum Golgi apparatus, outer mitochondrial membrane and the plasma membrane [31]. It is also ubiquitously expressed across cell types suggesting an importance in ubiquitous cellular processes [32]. Due to its localisation as a mitochondrial protein, it has been proposed that its role as an indirect calcium

regulator is important. The most common pathogenic mutation in PD is the LRRK2 G2019S mutation, which causes hyperactivity of the kinase domain leading to hyper-auto-phosphorylation and phosphorylation of substrates that causes the enhanced excitatory neurotransmission observed in cortical neurons contributing to the classical dendrite shortening observed in PD models [33]. There are multiple theories to explain the mechanism of the G2019S mutation including stabilisation of the active conformation of LRRK2 and/or introduction of a novel phosphorylation site [34]. The mitochondria, as well as providing the major ATP energy for normal cellular function, also acts as a rapid buffer for cytosolic calcium levels. Expression of mutant LRRK2 is associated with transcriptional upregulation of the mitochondrial calcium uniporter (MCU) and uptake 1 protein (MICU1) mediated by activation of the MAPK3/1 pathway leading to a decrease in cytosolic calcium [35]. In this way, mutant LRRK2 increases susceptibility to mitochondrial calcium imbalance accentuating the shortening of dendritic projections.

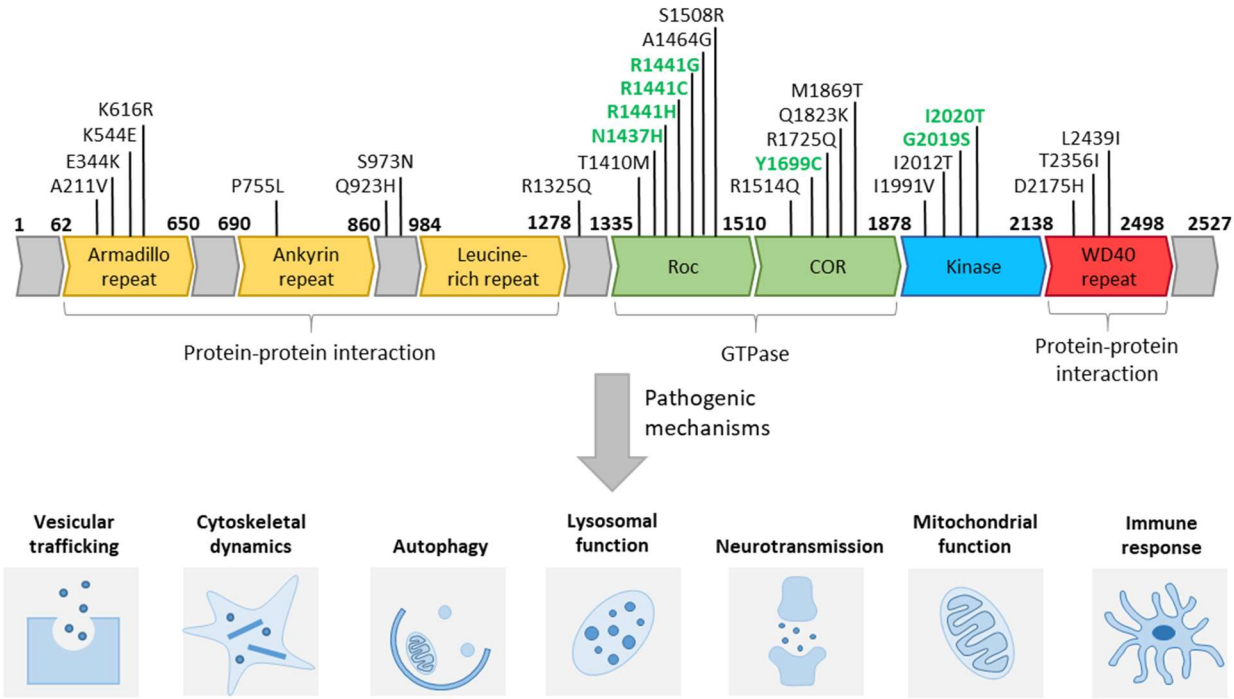


Figure 1.2 – Adapted from Tolosa *et al*. 2020 [28]. **(a)** Schematic domain structure of LRRK2 with associated pathogenic mutations which lead to dysfunctions in multiple pathways **(b)**. Confirmed pathogenic mutations with the most supporting data are highlighted in green with potentially pathogenic mutations having no shading. Domains involved in protein-protein interactions are coloured in yellow and green, the GTPase associated domains in purple and the kinase domain in blue. COR – C-terminal of Roc, Roc – Ras of complex.

However, mutations in LRRK2 extend beyond the commonly reported mutations (primarily G2019S). The LRRK2 protein has also been described in the process of microglial inflammatory responses using murine models (details of microglial activation in PD described in **section 1.1**), whereby the inhibition of kinase activity attenuates tumour necrosis factor alpha (TNFα) secretion and nitric oxide synthase induction, both of which are known triggers for microglial inflammatory responses [36]. This suggests a role for LRRK2 in mediating microglial proinflammatory responses within neuroinflammatory pathways.

The characterisation and implication of LRRK2 in disease has historically been through proteomic analysis with specific attention to the functions of LRRK2 as both a kinase and GTPase with the focus being on the implications of the disruption of these domains. In contrast to this, understanding the regulatory network that govern *LRRK2* gene expression patterns within the CNS is also critically important.

1.3 *INPP5F*/*BAG3*/*TIAL* locus

As a case study for the analysis of the function of retrotransposons within a PD risk locus, the GWAS nominated *INPP5F*/*BAG3*/*TIAL1* locus, located on chromosome 10 (q26.11), has been explored in **chapter 4** as it contained two SVA retrotransposons which were studied in detail. Given this, details of the *INPP5F*/*BAG3*/*TIAL1* locus are provided here with additional details in **section 4.1**.

The inositol polyphosphate-5-phosphatase F (*INPP5F*) gene encodes the protein INPP5F/Sac2, which has phosphatidylinositol (PI) phosphatase enzymatic activity. PI homeostasis is tightly regulated and underlies an extensive amount of cellular physiology, notably vesicular trafficking and lipid signalling. The inositol ring contains hydroxyl groups at the 3, 4 and 5 position, which can be phosphorylated to produce seven PI entities with distinct subcellular localisation and functionality [37]. INPP5F contains a Sac homology domain, primarily thought to function as a 4-phosphatase, and a signature hSac2 domain. INPP5F has been suggested to act as a 5-phosphatase, hydrolysing PIP3 into PI(3,4)P2 to terminate PIP3-Akt signalling [38]. As over activation of the Akt pathway is a common characteristic of many cancers, this suggested that INPP5F could act as a tumour suppressor gene [39]. Additionally, INPP5F is downregulated in gliomas and has been demonstrated to inhibit Signal Transducer and Activator of Transcription 3 (STAT3) to suppress glioma self-renewal and proliferation [40].

However, INPP5F is predominantly regarded as a 4-phosphatase, with a notable specificity for hydrolysis of PI(4)P to PI [24]. PI(4,5)P2 localises at the plasma membrane and is a key recruiter of endocytic factors required for endocytosis. As

INPP5F is primarily localised at early endocytic membranes, it has been suggested that it co-functions alongside a 5-phosphatase such as OCRL or Inositol Polyphosphate-5-Phosphatase B (INPP5B), both of which are close homologues to each other and are localised to endocytic membranes. This would hydrolyse PI(4,5)P2 to PI to terminate endocytosis signalling, followed by action of a PI3 kinase to produce PI(3)P, which is key for endosomal pathway signalling [41].

Furthermore, Synaptojanin 1 (*SYNJ1*) is a dual-function 4- and 5-phosphatase that shares a Sac homology domain with INPP5F. SYNJ1 localises to synaptic membranes and functions in vesicle endocytosis. A missense mutation in the Sac domain of SYNJ1 impairs phosphatase activity and is associated with early-onset autosomal recessive PD [42]. This is not dissimilar to a mutation in the Sac domain of INPP5F that diminishes phosphatase functionality [43]. Mutations in genes that are strongly linked to familial PD, notably α-synuclein (*SNCA*) and Leucine Rich Repeat Kinase 2 (*LRRK2*) are known to disrupt endosomal trafficking. As the endosome-lysosome and ubiquitin-proteasome pathways are both involved in α-synuclein degradation, dysfunction in the endosomal pathway could contribute to α-synuclein aggregation and Lewy body formation and could be important in PD pathology [44].

Bcl-2 associated athanogene (BAG) cochaperone protein 3 (BAG3) is encoded by the *BAG3* gene and is located approximately 48kb upstream of the *INPP5F* transcriptional start site. It was identified as a potential Parkinson's disease risk gene through GWAS analysis and part of the larger *INPP5F*/*BAG3*/*TIAL1* loci and was implicated via linkage disequilibrium analysis using expression quantitative trait loci (eQTL) [45]. However, independent studies had identified BAG3 as having a potential role in PD through

proteomic analysis of neuronal cell types. BAG3 is implicated in the cellular protein quality control (PQC) system which monitors the correct folding of proteins and manages the disposal of misfolded proteins which may be detrimental to normal cell function. BAG3-mediated selective macro-autophagy is a crucial process in the PQC system which is activated during the accumulation of pathological misfolded proteins. BAG3 had been originally shown to be involved in the clearance of hallmark misfolded proteins in several neurodegenerative disease such as Tau aggregation in Alzheimer's disease, huntingtin/polyQ proteins in Huntington's disease and mutated SOD1 in amyotrophic lateral sclerosis (ALS) [46]. Further studies also shown the involvement of BAG3 in PD with the clearance of α-synuclein. Using mutant SCNA transgenic PD mice models, Yu-Lan *et al.* 2017 showed increases in BAG3 expression and correlated the expression of BAG3 with enhanced macro-autophagy protein degradation pathways [47].

The TIA1 cytotoxic granule associated RNA binding protein like 1 (*TIAL1*) gene encodes the TIAR RNA binding protein (RBP). RBPs are proteins that bind double stranded or single stranded RNA to form ribonuclear protein molecules (RNPs) and are involved in a wide array of RNA processes including transcription, splicing, capping, polyadenylation, shuttling of RNA transcripts from the nucleus to cytoplasm, translation and degradation (RNA-binding proteins in human genetic disease). TIAR has been implicated in Parkinson's disease pathophysiology via association with the processing of α-synuclein (*SNCA*) transcripts. Marchese *et al.* 2017, provided evidence that TIAR binds with high affinity to 3' UTRs of *SNCA* alongside a second RBP, ELAVL1, to positively regulate and stabilise *SCNA* transcripts. This function was shown in motor cortex of post-mortem brain tissue and cultured primary fibroblasts

from patients with PD and multiple system atrophy (MSA). Further to this proteomic association, trans-eQTL analysis within the same study also provided evidence to suggest SNPs in the *TIAL1* locus could influence *SCNA* expression in the nucleus accumbens and hippocampus which provided a genetic link to PD in addition to the proteomic studies [48]. The individual associations of *INPP5F*, *BAG3* and *TIAL1* in PD provide evidence to support the role of these genes and encoded proteins in the disease pathology and as such, will be referred to as an entire locus within the analysis performed within this thesis.

1.4 Transposable Elements

1.4.1 Overview

Transposable elements (TEs) can be found in all eukaryotic organisms and constitute approximately 45% of the human genome as a result of many ancient insertion events, however many are no longer active (**figure 1.3**). There are several families of retrotransposons which remain active within the human genome, including members of the long interspersed nuclear elements (LINE), *Alu*, and SINE-VNTR-*Alu* (SVA) families [49-52]. Recent research has also indicated that small numbers of the HERV-K retroviruses also remain transposition capable and could have implications in disease pathology for amyotrophic lateral sclerosis (ALS) [53, 54]. The notion of transposable elements has been known since the 1950s when Barbara McClintock first proposed the idea of “jumping genes” whilst examining colour determination in maize [55]. Since then, TEs have been noted to play important roles in fundamental processes such as chromatin structure remodelling and regulation of gene expression, thus identifying these elements as having regulatory functions [56]. TEs

can transpose throughout the genome via two mechanisms, either by a “copy and paste” (class I elements) or “cut and paste” (class II elements) [57]. It has been hypothesised that TEs drive genetic diversity and evolution of the human genome and therefore do not elicit negative pathophysiological implications. However, there have been many cases reported whereby an insertion into a coding or regulatory region has led to a pathophysiological state. Approximately 0.27%, of human disease mutations have been linked to a direct retro-element insertion [58].

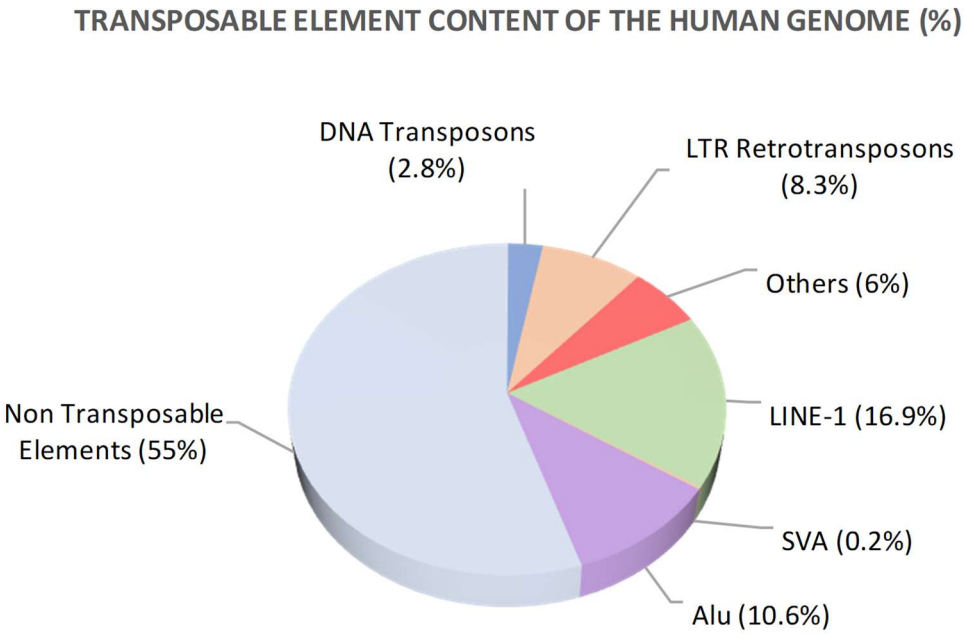


Figure 1.3 – Breakdown of the human genome as percentages contributed by transposable elements (TE’s). TE’s make up a total of 45% of the human genome and are divided into LTR and non-LTR subclasses [59].

Class II elements consist of the DNA transposons which mobilise through the use of transposase enzymes and make up approximately 2% of the human genome. By the nature of their “cut and paste” style mechanism, they do not require an RNA intermediate [60, 61]. The class I retrotransposons differ in that to transpose they

follow a two-step process. Their DNA is transcribed into an RNA intermediate, which must be reverse transcribed and can then insert into a different position in the genome. [62] This class of TEs can be subdivided into two distinct groups, those possessing long terminal repeats (LTRs) and those that do not (non-LTRs). The LTR group consists of human endogenous retroviruses (HERVs) which make up approximately 8% of the human genome and contains around 98,000 elements and fragments [63, 64]. The non-LTR group consists of the LINE, SVA and *Alu* elements and is the most abundant class of retrotransposons, making up approximately 28% of the human genome [62, 65]. The elements focused on within this study are the non-LTR retrotransposable elements (LINE-1, SVA and *Alu*) and will be referred to throughout this thesis.

1.4.2 SINE-VNTR-*Alu* (SVA) elements

Non-truncated SVAs are typically around 2kb in length and are hominid-specific elements [66]. They are mobilised through the LINE-1 protein machinery (ORF1 and ORF2 – details in **section 1.6**) and are derived from several different genomic repeats [67, 68]. These elements represent 0.13% of the total human genome and in accordance with the current reference human genome (Hg38), there are approximately 2700-3000 SVA elements including truncated elements. However, 63% are known to be full length elements containing the five key SVA domains; a 5' hexamer repeat (CCCTCT), an inverted *Alu*-like domain, a GC-rich variable number tandem repeat (VNTR) region, a SINE-R domain derived from the 3' LTR of the retroviral HERV-K10 element and a canonical poly-A signal (AATAAA) at the 3' end which is capped with a poly-A tail (**figure 1.4**) [52, 69]. There are seven SVA sub-

classes, named A-F1, based on evolutionary age and the composition of the SINE-R domain. The SVA-A sub-class is the oldest in evolutionary age at approximately 13.6 million years and SVA-F being the youngest at approximately 3.2 million years [70]. The SVA-F1 sub-class is distinctive from the other classes as it contains a partial sequence of exon 1 of the MAST2 gene which resulted from an alternative splicing event within the MAST2 locus [71].

Primary sequence polymorphisms are known to occur predominantly within the VNTR region and to a lesser extent within the 5' hexamer repeat region [72, 73]. The implications of the effect of SVA orientation with respect to the direction of transcription, i.e. sense versus anti-sense, is not well characterised, however the orientation of LINE retroelements has been studied. It has been reported that LINE elements have a bias against insertion in a sense orientation respective to the host genome, however the implications of this are poorly understood [74].

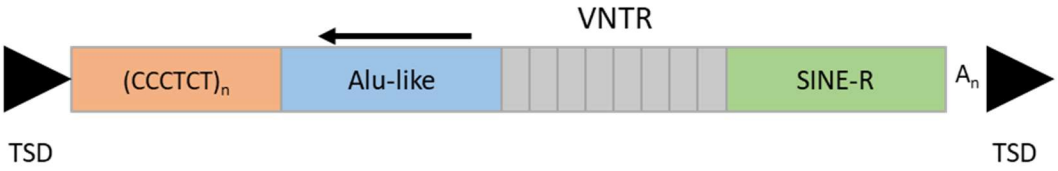


Figure 1.4 – General structure of a full length SVA. A typical non-truncated SVA contains five main domains with flanking regions. Starting from the 5' end is a canonical hexamer repeat (CCCTCT), an inverted *Alu*-like domain (indicated by black arrow), a GC-rich variable number tandem repeat (VNTR), SINE-R, 3' poly-A tail all flanked by target site duplications (TSD). SVAs contain multiple RNA Pol III terminator signals (TTTT) within the *Alu*-like and SINE-R domains [75].

A key function of SVAs has been extensively investigated in which they have been shown to act as modulatory elements under normal conditions [72]. SVAs have been reported to function in at least two ways; either by insertion within single genes,

resulting in pathological conditions, or by acting as regulatory elements to alter gene expression. Regulatory properties of SVA elements have been demonstrated in a Parkinson's disease context within the *PARK7*/*DJ-1* locus, whereby a fixed SVA-D element located approximately 8kb upstream of the *PARK7* transcriptional start site was shown to positively modulate *PARK7* expression using reporter gene assays *in vitro* [76]. A second example by Savage *et al.* 2014 investigated a fixed SVA in a neurodegenerative disease related gene within the *FUS* locus in the context of ALS. In this study, an SVA-D element located 10kb upstream of the *FUS* gene was shown to elicit regulatory properties in a chick embryo model. Multiple component domains of the SVA sequence were tested for regulatory function for driving reporter gene expression with the full-length sequence showing repressive effects and the isolated VNTR domain showing positive driving of reporter gene expression *in vitro*. GFP imaging within the developing chick embryo showed that the *FUS* SVA was expressed in the developing neural tube [73]. To date, at least eight SVA insertions have been directly linked with a variety of disease states (**table 1.1**). This highlights the ability of an SVA insertion to disrupt normal genetic processing within a multitude of diseases including neurodegenerative disorders such as Parkinson's disease.

Table 1.1 – A summary of eight genes with SVA insertions that are causative of disease. Adapted from Hancks *et al.* 2012 and van der Klift *et al.* 2012 [49, 77]. Abbreviations: Chr – Chromosome, S – sense, AS – anti-sense; Disease acronyms: XLA – X-linked agammaglobulinemia, XDP – X-linked dystonia-parkinsonism, ARH – Autosomal recessive hypercholesterolemia, HE and HPP – Hereditary elliptocytosis and hereditary pyropoikilocytosis, FCMD – Fukuyama-type congenital muscular dystrophy, NLSDM – Neutral lipid storage disease with subclinical myopathy, LS – Lynch Syndrome.

Gene	Chr	Disease	Sub-family	Orientation
BTK	X	XLA	N/A	S
TAF1	X	XDP	F	AS
LDRAP1	1	ARH	E	S
SPTA1	1	HE and HPP	E	S
HLA-A	6	Leukemia	F_1	AS
FKTN	9	FCMD	E	S
PNPLA2	11	NLSDM	E	S
PMS2	7	LS	F	-

The table of pathogenic SVA insertions above includes an intronic SVA-F insertion within the *TAF1* gene in an anti-sense orientation which leads to X-linked dystonia-parkinsonism [49]. As a case study to which comparisons will be made to throughout this thesis, the *TAF1* SVA insertion is of particular interest due to the specific mechanisms regarding how the SVA insertion disrupts genomic properties leading to a neuronal disease phenotype. The pathogenic *TAF1* SVA insertion occurs within intron 32 of the *TAF1* gene which causes partial intron retention and lower expression of total *TAF1* mRNA. This effect has been observed in iPSCs, iPSC derived cortical neurons and spiny projection neurons, furthermore the reduction in *TAF1* mRNA was completely reversable by CRISPR mediated excision of the SVA [78, 79]. Further analysis of the pathogenic mechanism demonstrated that the *TAF1* SVA could influence disease onset which correlated with an expansion of the hexameric repeat $(CCCTCT)_n$. An inverse relationship was discovered, namely as the hexamer repeat expanded (ranging between 35 and 52 repeats), the age of onset was reduced [80]. All of these studies highlight the potential influence SVA insertions can have on gene function which can lead to pathogenesis.

The examples described in **table 1.1** demonstrate the pathogenic function novel SVA retrotransposon insertion polymorphisms can cause, but do not reflect the majority of SVA elements within the human genome. Over 60% of SVA elements within the human genome reside within or in close proximity to known gene transcripts (within 10kb of an annotated gene transcript) [76]. The primary sequence of SVAs has high GC content, often exceeding 70% within the central VNTR domain, and this feature has been suggested to have the potential to act as a mobile CpG island [81]. Due to the high GC content of SVAs, the effects on chromatin structure have been of interest

with the prediction that they can participate in the formation of alternative DNA secondary structures such as G-quadruplexes (G4) which can affect transcription [82]. G4 sequences are predicted to form in regions which contain multiple short runs of G nucleotides which have the ability to form planar tetrad structures via Hoogsteen hydrogen boding in stacks which form helices (G-quadruplexes and their regulatory roles in biology). This mechanism of action has been proposed for SVA elements due to their hexameric repeat ($CCCTCT_n$) which provides the multiple short repeats of G necessary for G4 formation. Using bioinformatic analysis, SVAs have been demonstrated to contribute the largest source of putative G4 quadruplexes out of all genetic elements when accounting for their genomic size and provide a potential mechanism of action for SVA sequences [76]. Mutations and stabilisation of G4 quadruplexes have been shown to modify gene expression of genes in close proximity in both *in vivo* and *in vitro* model systems including embryonic models [83-85]. Understanding how SVA elements function in a wider range of contexts is crucial for understanding the effects of novel regulatory domains within complex diseases such as Parkinson's disease and provides a major focus of the work presented in this thesis.

1.4.3 *Alu* elements

Alu elements belong to the short interspersed nuclear elements (SINE) which constitute the largest copy number sub-family of non-LTR retrotransposons consisting of approximately 1.1 million copies (~10.6% of the human genome – **figure 1.3**) [86, 87]. *Alu* elements are comprised of two monomers (left and right) which resulted from a fusion of two ancient 7SL RNA sequences, there are separated by a

short AT rich repeat sequence (A_5TACA_6) and flanked by target site duplications (TSD) (**figure 1.5**) [88, 89]. Mobilisation of *Alu* elements is considered non-autonomous given that they require LINE-1 machinery in order to mobilise *in trans*, whereby the ORF2 protein enables reverse transcription and reintegration into the genome (further L1 mobilisation details in **section 1.5.4**) [90].

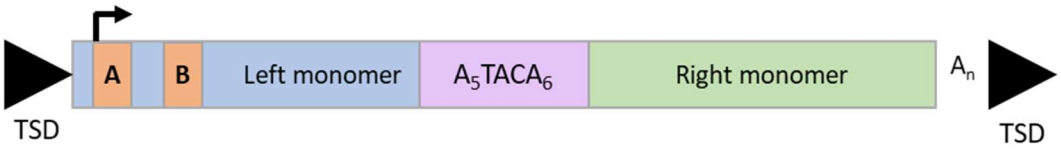


Figure 1.5 – The general structure of a full length *Alu* element is approximately 280bp in length and contains two distinct monomers (left and right) derived from the 7SL RNA gene which are separated by an AT rich linker repeat sequence (A_5TACA_6). The left 7SL derived monomer contains two component sequences (A and B) of the RNA Pol III promoter which are necessary for RNA Pol III mediated transcription. The sequence terminates with a poly-A tail and is flanked by target site duplications (TSDs) [56].

The *Alu* family is divided into 3 sub-families based on the approximated age of the element, with *AluJ* being the oldest at ~65 million years ago, *AluS* at ~30 million years ago and *AluY* being the most recent age. *AluJ* members are widely regarded as inactive in the human genome due to the accumulation of deleterious mutations, the *AluS* family has retained low levels of activity, as measured by the ability to mobilise via L1. The *AluY* family exhibits the highest level of mobilisation activity due, in part, to having the largest number of functionally intact elements [91]. Given *Alu* elements constitute the highest mobilisation rates of the non-LTR retrotransposons, it is estimated that the rate of transposition could be as high as approximately one new *Alu* insertion per ~20 births [92]. This is particularly important when considering *Alu* insertions account for ~60% of the 95 single gene disorders associated with

transposable elements [93]. The ability of *Alu* elements to contribute to disease pathology may stem from their ability to create novel exons within genes via the process of exonisation. That is *Alu* elements, when inserted within an intron in the anti-sense orientation relative to transcription of the host gene, have the ability to generate cryptic splice sites resulting in novel exons [94]. The rate of exonisation of *Alus* is approximately three times higher than all other human transposable elements with ~0.2% of intronic *Alus* being exonised. Exonisation of *Alus* can be considered as a positive driver of novel transcript generation and diversity, but with the potential for negative consequences by causing non-functional protein truncations which could be pathogenic. Early stop codons and frame shift mutations as a result of *Alu* exonisation have been reported in multiple genes including *YY1AP*, *ACAD9* and *AMPK* amongst others with some of these genes being implicated in cancer [95]. *Alu* insertions have also been implicated in neurodegenerative disorders with a notable example involving the accumulation of primate specific *Alu* insertions in the *TOMM40* gene with at least one variant being associated with late-onset Alzheimer's disease [96]. A poly-T variant associated with an anti-sense *Alu* element within the *TOMM40* gene is linked with disrupted processing of *TOMM40* pre-mRNA transcripts by the spliceosome machinery leading to an increase in mRNA degradation [96].

1.4.4 Long Interspersed Nuclear Element 1 (LINE-1/L1)

The LINE family of retrotransposons consist of three sub-classes (LINE-1, 2 and 3) which have co-existed over time with the most distinguishing characteristic between them being the evolution of their 5' UTRs. However, given their sequence similarities, the LINE-1 family (L1) has dominated over classes 2 and 3 and is the only remaining

retrotransposition-competent retrotransposon in humans that can mobilise both autonomously and *in trans* via mobilisation of other non-LTR retrotransposons, primarily *Alu* and SVA elements [97, 98]. Original estimates reported approximately 500,000 copies of L1 existed in the modern human genome which accounted for ~17% of the genome (**figure 1.3**) [99]. Subsequently, the vast majority of these elements were found to be 5' truncated and had lost their capacity to mobilise, with only approximately 80-100 copies currently being regarded as 'active' and retrotransposition competent. Of these active elements, six full-length L1 elements have been classified as 'hot' and account for >80% of all transposition events in the genome [51].

Full length LINE-1 elements are ~6kb in length and encode two open reading frame (ORF) proteins, ORF1p and ORF2p flanked by 5' and 3' UTRs. ORF1 encodes a ~40 kDa RNA binding protein that has nucleic acid chaperone properties whilst ORF2 encodes a ~150 kDa protein which possess endonuclease (EN) and reverse transcriptase (RT) activity (**figure 1.6**) [100, 101]. Whilst ORF1p and ORF2p proteins preferentially bind their own mRNA sequence (cis-mobilisation), they are also co-opted by other retrotransposons *in trans*, namely SINE, *Alu* and SVA elements [102]. Upon insertion of retrotransposons into new genomic loci, small sequences of approximately 7-20bp are often duplicated and are positioned flanking the new insertion, these are termed target site duplications (TSD). The TSDs are considered the boundaries of an L1 insertion and are regarded as hallmarks of LINE-1 mediated retrotransposition [103].

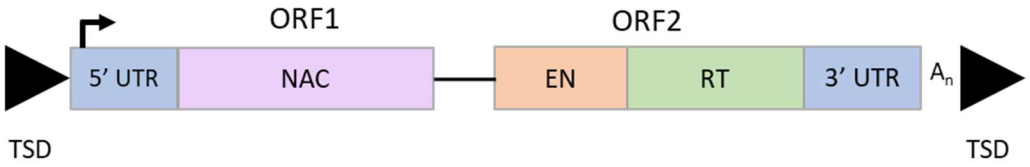


Figure 1.6 – General structure of a full-length LINE-1 element which is approximately 6kb in length and comprised of two distinct well characterised proteins, ORF1p and ORF2p. ORF1 encodes a ~40kDa RNA binding nucleic acid chaperone (NAC) protein whilst ORF2 encodes a larger ~150kDa protein with both endonuclease (EN) and reverse transcriptase (RT) activity. 5' and 3' untranslated regions (UTR) flank the ends of the ORFs with the 5' UTR containing the endogenous L1 promoter. The full length L1 is capped by a poly-A tail at the 3' end and the whole element is flanked by target site duplications (TSD).

The original implication of LINE-1 elements in disease can be traced back to the discovery of a LINE-1 insertion in exon 14 of the factor VIII gene within two haemophilia patients by Kazazian *et al.* 1988 [104]. Since then, multiple examples of pathological LINE-1 insertions have been reported which can cause dysfunctional splicing, exon skipping and double strand DNA breaks with the latter being associated with cancer progression [105-107]. LINE-1 elements have been implicated in a range of neurodegenerative disorders, including Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS) and Parkinson's disease (PD). The presence of Tau neurofibrillary tangles in AD has been shown to increase expression of L1 using RNA-seq from human AD brain tissue and confirmed *in vivo* using human Tau transgenic *Drosophila* models [108]. Within TDP-43 aggregate pathology mediated diseases such as ALS and frontotemporal lobar dementia (FTLD), de-repression of L1 has been linked to the progression of these disorders via human *TDP-43* (*hTDP-43*) transgenic *Drosophila* studies. That is, expression of *hTDP-43* in *Drosophila* neurons and glial cells led to protein aggregation, locomotion impairment and premature death of the

flies. It was shown that *hTDP43* impaired siRNA silencing mechanisms leading to activation of L1 retroelements in these cell types [109]. Specific mechanisms for LINE-1 involvement in the progression of PD have also been reported in which over-expression of LINE-1 triggers oxidative stress in mesencephalic dopaminergic neurons and leads to DNA strand breaks which ultimately leads to cell death. This effect can be rescued either by expression of *Engrailed-1* which directly represses LINE-1 expression, an siRNA knockdown of LINE-1 or administering anti-retroviral drugs to inhibit reverse transcriptase suggesting LINE-1 is the crucial driver in the observed pathological phenotype [110].

1.4.5 The relevance of retrotransposons in neurodegeneration

Retrotransposons that are found in the reference genome are termed 'reference' this is in contrast with retrotransposon insertion polymorphisms (RIPs). These are defined as novel retrotransposon insertions that have resulted from mobilisation and are not found in the reference genome however, they exist across multiple individuals. To demonstrate the relevance of studying the influence of retrotransposons in PD, a simple observational analysis was performed to explore the prevalence of reference SVA retrotransposons in known PD risk loci. Using the 90 identified GWAS nominated PD risk loci (details in section 1.2) by *Nalls et al.* 2019 and a defined list of SVA elements from the RepeatMasker track on UCSC which utilises data from the Repbase database, 21 of the total 90 loci (23%) contained at least one SVA element within 100kb, with a total of 31 SVAs present (**table 1.2**) [111]. Given the pre-existing knowledge that SVA elements have the potential to elicit regulatory function that can be pathological in specific contexts (table 1.1), the presence of SVA elements within almost a quarter of the GWAS nominated PD risk loci is important. Two of the loci identified through this analysis were used to study further within this thesis, the *LRRK2* and *INPP5F*/*BAG3*/*TIAL1* loci as case studies to further understand the potential effects of SVA retrotransposons on gene regulation of key Parkinson's risk genes.

Table 1.2 – Reference SVAs that were present across the 90 PD risk loci identified by GWAS analysis (Nalls et al 2019). The SVA list contained 2676 SVA elements and was taken from the RepeatMasker track, which utilises the Repbase database, via UCSC genome browser using human genome build 19 (hg19) [111]. Of the 90 PD loci in this list, 21 contained at least one SVA (highlighted) with a total of 31 SVAs present due to some loci contained multiple SVAs.

SNP	SNP +/- 100kb (hg19)	SVAs	Nearest Gene	QTL Nominated Gene
rs114138760	chr1:154798185-154998185	0	PMVK	
rs35749011	chr1:155035036-155235036	0	KRTCAP2	
rs76763715	chr1:155105634-155305634	0	GBAP1	GBAP1
rs6658353	**chr1:161369054-161569054**	**1**	**FCGR2A**	**FCGR2A**
rs11578699	**chr1:171619769-171819769**	**1**	**VAMP4**	**VAMP4**
rs823118	chr1:205623572-205823572	0	NUCKS1	NUCKS1
rs11557080	chr1:205637739-205837739	0	RAB29	RAB29
rs4653767	chr1:226816078-227016078	0	ITPKB	ITPKB
rs10797576	chr1:232564611-232764611	0	SIPA1L2	
rs76116224	chr2:18047848-18247848	0	KCNS3	
rs2042477	**chr2:95900943-96100943**	**1**	**KCNIP3**	**KCNIP3**
rs11683001	chr2:102296963-102496963	0	MAP4K4	MAP4K4
rs57891859	chr2:135364616-135564616	0	TMEM163	TMEM163
rs1474055	chr2:169010394-169210394	0	STK39	STK39
rs73038319	chr3:18261759-18461759	0	SATB1	SATB1
rs6808178	chr3:28605690-28805690	0	LINC00693	
rs12497850	**chr3:48648989-48848989**	**1**	**IP6K2**	**IP6K2**
rs55961674	chr3:122096892-122296892	0	KPNA1	KPNA1
rs11707416	chr3:151008965-151208965	0	MED12L	
rs1450522	chr3:160977630-161177630	0	SPTSSB	NMD3
rs10513789	chr3:182660073-182860073	0	MCCC1	MCCC1
rs873786	**chr4:825376-1025376**	**2**	**GAK**	**GAK**
rs34311866	**chr4:851947-1051947**	**1**	**TMEM175**	**TMEM175**
rs4698412	**chr4:15637348-15837348**	**1**	**BST1**	**BST1**
rs34025766	chr4:17868811-18068811	0	LCORL	DCAF16
rs6825004	chr4:77010365-77210365	0	SCARB2	SCARB2
rs4101061	chr4:77047969-77247969	0	FAM47E	FAM47E
rs6854006	chr4:77098054-77298054	0	FAM47E-STBD1	FAM47E
rs356182	chr4:90526111-90726111	0	SNCA	SNCA
rs5019538	chr4:90536630-90736630	0	SNCA	SNCA
rs13117519	chr4:114269065-114469065	0	CAMK2D	CAMK2D
rs62333164	**chr4:170483157-170683157**	**2**	**CLCN3**	**NEK1**
rs1867598	chr5:60037959-60237959	0	ELOVL7	ELOVL7
rs26431	chr5:102265794-102465794	0	PAM	PAM
rs11950533	**chr5:134099105-134299105**	**1**	**C5orf24**	**TXNDC15**
rs4140646	**chr6:27638801-27838801**	**1**	**LOC100131289**	**TXNDC15**
rs9261484	**chr6:30008683-30208683**	**1**	**TRIM40**	**TRIM40**
rs112485576	**chr6:32478772-32678772**	**2**	**HLA-DRB5**	**HLA-DRB5**
rs12528068	chr6:72387762-72587762	0	RIMS1	
rs997368	chr6:112143291-112343291	0	FYN	
rs75859381	chr6:133110361-133310361	0	RPS12	
rs199351	chr7:23200049-23400049	0	GPNMB	GPNMB
rs76949143	**chr7:65909851-66109851**	**1**	**GS1-124K5.11**	**LINC00174**
rs1293298	chr8:11612443-11812443	0	CTSB	CTSB
rs620513	chr8:16597593-16797593	0	FGF20	

SNP	SNP +/- 100kb (hg19)	SVAs	Nearest Gene	QTL Nominated Gene
rs2280104	chr8:22425980-22625980	0	BIN3	BIN3
rs2086641	chr8:130801909-131001909	0	FAM49B	
rs13294100	chr9:17479690-17679690	0	SH3GL2	SH3GL2
rs10756907	chr9:17627065-17827065	0	SH3GL2	SH3GL2
rs6476434	**chr9:33946391-34146391**	**2**	**UBAP2**	**PRSS3**
rs896435	chr10:15457406-15657406	0	ITGA8	ITGA8
rs10748818	**chr10:103915279-104115279**	**1**	**GBF1**	**CUEDC2**
rs72840788	**chr10:121315685-121515685**	**2**	**BAG3**	**BAG3**
rs117896735	**chr10:121436327-121636327**	**2**	**INPP5F**	**SEC23IP**
rs7938782	chr11:10458777-10658777	0	RNF141	RNF141
rs12283611	chr11:83387277-83587277	0	DLG2	
rs3802920	chr11:133687001-133887001	0	IGSF9B	IGSF9B
rs76904798	chr12:40514434-40714434	0	LRRK2	LRRK2
rs34637584	**chr12:40634202-40834202**	**1**	**LRRK2**	**LRRK2**
rs7134559	**chr12:46319086-46519086**	**1**	**SCAF11**	
rs10847864	chr12:123226598-123426598	0	HIP1R	HIP1R
rs11610045	chr12:132963768-133163768	0	FBRSL1	FBRSL1
rs9568188	**chr13:49827732-50027732**	**1**	**CAB39L**	**CAB39L**
rs4771268	chr13:97765021-97965021	0	MBNL2	
rs12147950	**chr14:37889270-38089270**	**1**	**MIPOL1**	
rs11158026	chr14:55248869-55448869	0	GCH1	GCH1
rs3742785	chr14:75273034-75473034	0	RPS6KL1	RPS6KL1
rs979812	chr14:88364264-88564264	0	GALC	GALC
rs2251086	chr15:61897385-62097385	0	VPS13C	
rs6497339	chr16:19177493-19377493	0	SYT17	SYT17
rs2904880	chr16:28844396-29044396	0	CD19	NFATC2IP
rs11150601	chr16:30877799-31077799	0	SETD1A	SETD1A
rs6500328	chr16:50636656-50836656	0	NOD2	NOD2
rs3104783	chr16:52536242-52736242	0	CASC16	CYLD
rs10221156	chr16:52869426-53069426	0	CHD9	CYLD
rs12600861	chr17:7255621-7455621	0	CHRNB1	CHRNB1
rs12951632	chr17:40641013-40841013	0	RETREG3	RETREG3
rs2269906	**chr17:42194337-42394337**	**2**	**UBTF**	**UBTF**
rs850738	chr17:42334630-42534630	0	FAM171A2	GRN
rs62053943	chr17:43644203-43844203	0	CRHR1	WNT3
rs117615688	chr17:43698308-43898308	0	CRHR1	WNT3
rs11658976	chr17:44766805-44966805	0	WNT3	WNT3
rs61169879	chr17:59817366-60017366	0	BRIP1	MED13
rs666463	**chr17:76325480-76525480**	**1**	**DNAH17**	**DNAH17**
rs1941685	chr18:31204318-31404318	0	ASXL3	
rs12456492	chr18:40573380-40773380	0	RIT2	
rs8087969	**chr18:48583589-48783589**	**1**	**MEX3C**	**MEX3C**
rs55818311	chr19:2241047-2441047	0	SPPL2B	
rs77351827	chr20:5906041-6106041	0	CRLS1	
rs2248244	chr21:38752361-38952361	0	DYRK1A	DYRK1A

This preliminary analysis highlights the importance of understanding the fundamental mechanisms of the actions of retrotransposable elements such as SVAs, on gene function when applied in a more global context. Several associations between the increase in retroelement activity, primarily LINE-1 activation, and an increase in the risk of Parkinson's disease have already been described. One such study, by Nielsen *et al.* 2012, linked smoking with the loss of LINE-1 DNA methylation (associated with the de-repression of LINE-1) and suggested the activation of LINE-1 as a potential risk factor for the development of PD [112]. A similar mechanism of LINE-1 activation has also been reported to occur as a result of mitochondria stress leading to loss of LINE-1 methylation and activation of the elements. MPTP (a widely used pharmacological agent in PD model systems) was used to induce reactive oxygen species (ROS) production to destabilise the mitochondrial respiratory chain in human dopaminergic LUHMES cell models, whereby the production of ROS was also found to reduce the levels of LINE-1 methylation. This in turn led to the activation of LINE-1 expression via de-repression and was suggested as a possible mechanism to explain retrotransposon activation of neurological disorders including multiple sclerosis, Alzheimer's disease and Parkinson's disease [113].

Previous studies have further attributed transposable element activation in neurodegenerative disorders, notably amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), whereby the activation of human endogenous retrovirus-K (HERV-K) elements, specifically expression of the envelope (env) domain, led to the retraction and beading of neurite projections which was hypothesised to result from aberrant expression of HERV-K encoded proteins leading to neurotoxicity [54]. Further implication of TEs in ALS has been provided by the study

of TAR DNA binding protein 43 (TDP-43) pathology. Nuclear TDP-43 depletion in conjunction with TDP-43 accumulation in the cytoplasm of neuronal and glial cells is a hallmark of ALS and FTD pathology [114]. TDP-43 was found to directly bind UGUGU pentamer motifs which are regularly found within TE encoded RNAs, primarily from SINE, LINE and ERV family members [115]. Dysfunction of TDP-43 in ALS human neuronal model systems is directly correlated with increased LINE-1 activity and leads to an increase in LINE-1 copy number in neuronal nuclear DNA when nuclear TDP-43 levels are low [116].

1.4.6 Impact of retrotransposons on gene regulation mechanisms

The mechanisms behind gene regulation are complicated and multi-faceted and involve regulating gene expression, chromatin organisation and epigenetic changes amongst others. The process of gene expression starts with an initiation of transcription via binding of general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH to the core promoter of a gene which together form the pre-initiation complex [117-119]. This complex recruit's RNA polymerase II (RNA PolII) and transcription of the gene can occur. However, there are many transcription factors (TFs) that can influence this, these may act to either enhance or repress this process thus increasing or reducing expression. Retrotransposon insertions may impact normal gene regulation in a variety of mechanisms (**figure 1.7**). Both fixed and polymorphic (presence/absence) retrotransposons have the ability to influence gene regulation *in-cis*, including the introduction of novel transcriptional start sites (TSS) (**figure 1.7A**), premature stop codons if the insertion occurs within a coding exon (**figure 1.7B**), generation of novel transcription factor binding sites (TFBSs) which may act as enhancers or repressors (**figure 1.7C**), exonisation of RIPs or novel alternate splice sites (**figure 1.7D**) and epigenomic changes such as introduction of novel DNA methylation sites (**figure 1.7E**).

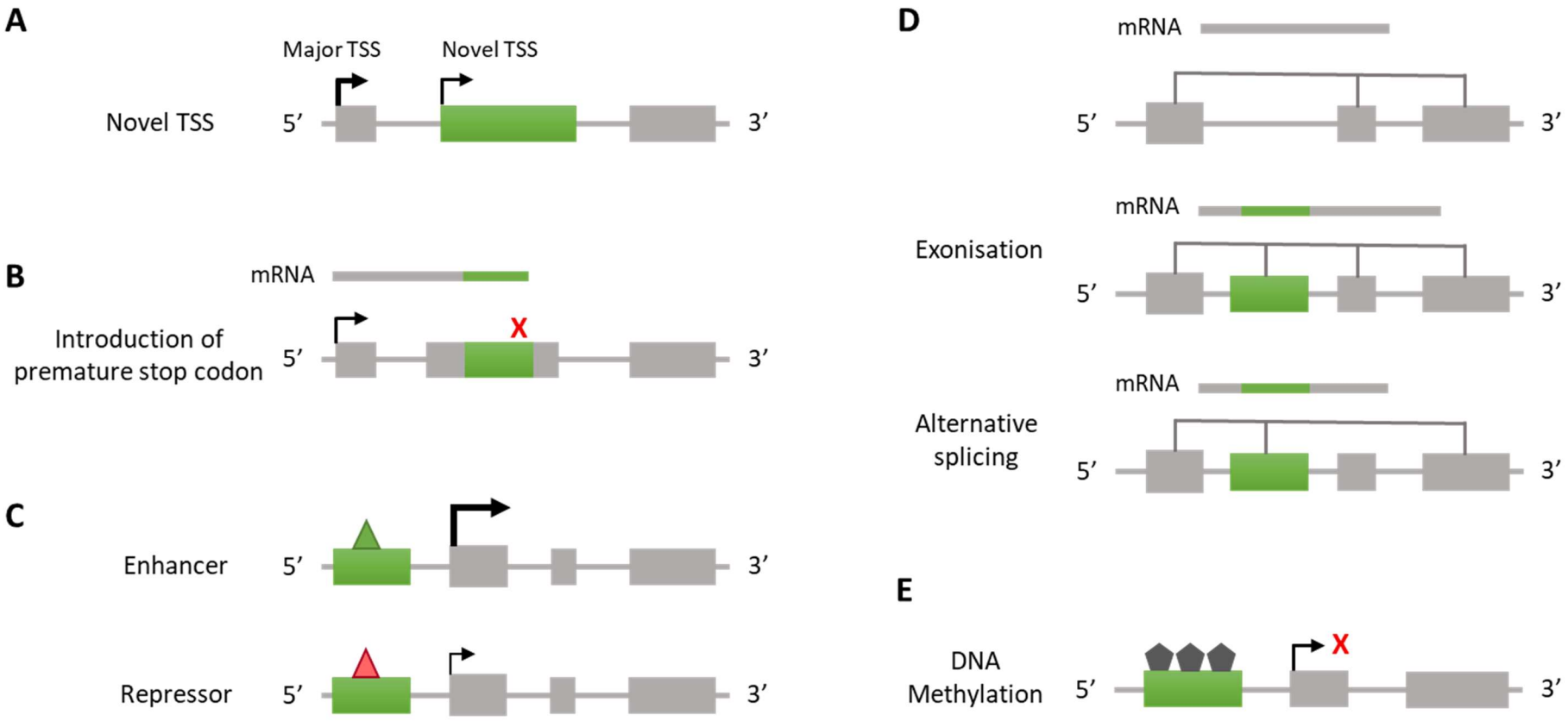


Figure 1.7 – Differential effects of retrotransposons on *cis*-gene regulation. **(A)** Retrotransposon insertion can lead to the creation of novel transcriptional start sites or **(B)** introduction of premature stop codons if the insertion occurs within an exon. **(C)** Retrotransposons can act as both enhancer and repressor elements depending on sub-class and genomic location. **(D)** Intronic insertions of retroelements have the potential to become exonised or cause novel alternate splice sites. **(E)** Retrotransposons can be sites for DNA CpG methylation due to high CpG content which can lead to gene silencing.

An example of the introduction of a TSS as a result of a retrotransposition event was observed in the *NAIP* gene, where an *Alu* insertion upstream of the major *NAIP* promoter led to the production of a TSS from which a novel transcript was expressed resulting in the translation of a novel protein [120]. The generation of novel TFBSs located near gene promoters has been well documented with more than 2 million TFBS being annotated within retrotransposon sequences that are located within 5kb of a promoter [121]. The TFBSs found within all transposable elements account for approximately 17% of all TFBSs located proximal to promoter regions [122]. Furthermore, the production of TFBSs can also be resultant from retrotransposon primary sequence polymorphisms, for example the generation of an enhancer box (E-box) motif due to an expansion polymorphism. This is relevant for the data presented in **chapter 3 – section 3.2.2**, where an expansion polymorphism within the VNTR sequence of the SVA-C element reported in *LRRK2* leads to the generation of an E-box motif (**figure 3.5**). E-box motifs are DNA response elements that have the consensus sequence CANNTG, which bind specific families of transcription factors (TFs) in order to promote transcription [123]. These sequences are associated with, but not limited to, promoters and enhancer sequences that regulate the expression of neuronal, muscle and pancreas-specific genes [123]. The E-box motifs have the highest affinity for helix-loop-helix TFs such as MyoD and the c-Myc/Max complex amongst others, which control major cellular pathways such as cell proliferation and apoptosis [124, 125].

Retrotransposon insertions within intronic regions have the capacity to become exonised or causative of novel splicing variants by influencing the process of alternative splicing. One particular example of pathological alternate splicing is

observed in bilirubin metabolism disorder, Rotor syndrome, where a novel insertion of a full-length LINE-1 element into intron 5 of the *SLCO1B3* gene and is linked with aberrant mRNA splicing resultant in the complete reduction of OATP1B3 protein levels leading to disease onset [126]. Retrotransposon exonisation is particularly prevalent with *Alu* insertions within gene transcripts, with original estimates placing approximately 5% of alternatively spliced exons being derived from *Alu* elements [127]. More recently, discovery of the RNA-binding protein hnRNP C that binds to the cryptic splice sites present within *Alu* sequences, as a mechanism to protect the genome from aberrant exonisation of Alu elements, has led to further understanding how these mechanisms are important for avoiding specific pathologies [94]. There are multiple examples of *Alu* exonisation in disease, with the earliest reports including the discovery of a point mutation within a fixed *Alu* element that introduced a novel 5' splice site within intron 3 of the *OAT* gene which led to ornithine aminotransferase deficiency as a result of partial exonisation of the *Alu* element [128]. Another example includes the constitutive *Alu* exonisation caused by a deletion mutation within an *Alu* element introducing a strong 5' splice site present in the *GUSB* gene which was found in a patient with the autosomal recessive lysosomal storage disease, Sly syndrome [129].

Changes in the epigenetic signature of genomic loci also provide further mechanisms of action for retrotransposons by the introduction of novel sites for DNA methylation into new loci via retrotransposition events or polymorphisms within retrotransposon primary sequence. Epigenetic modification is accepted as a mechanism for the control of transposable elements via either DNA methylation or histone modifications which usually act to silence retrotransposon expression [130]. This has

led to an ‘evolutionary arms race’ between TE activity, in particular the LINE-1 and SVA families, and the epigenetic modifications whose function is to suppress TE activity via the actions of Kruppel associated box (KRAB) zinc finger proteins (ZFPs) which bind to TE primary sequences and recruit TRIM28. TRIM28 acts in a dual function mode to both induce repressive type histone modifications and recruit DNA methyltransferases to methylate the retroelement and ‘silence’ the activity [131]. The generally accepted doctrine regarding methylation status at CpG sites dictates that methylated sites are generally silenced or repressed and unmethylated sites are active or indicate increased expression of a gene. However, in reality, this process is more complicated, with position of the methylated site in relation to the surrounding genes being also critically important. Hypermethylation within a promoter or enhancer region tends to silence the region and block transcription factors from initiating transcription whilst hypermethylation within a gene body between the 5’ TSS and 3’ UTR can lead to increased gene expression [132, 133]. The actions of the epigenetic changes applied to retrotransposable elements is not only implicated in the regulation of the elements themselves but can also have major implications for the surrounding genes within the neighbourhood. This effect was demonstrated in murine neural progenitor cell models, whereby the knockout of TRIM28 resulted in the reduction of TE epigenetic silencing leading to a large increase in gene expression (approximately 3-fold) of genes with endogenous retrovirus elements (ERVs) within close proximity (within 50kb) [134].

1.5 General aims

Within this thesis, the aim is to broaden the understanding of retrotransposable elements, which are often overlooked in computational analysis, as regulatory domains and novel risk factors in the aetiology of Parkinson's disease. To do this, the *LRRK2* and *INPP5F*/*BAG3*/*TIAL1* loci will be used as case studies to explore the effects of SVA retrotransposons on gene function on a local scale. The scope will then be widened to a genome wide scale, using next generation sequencing techniques for the characterisation of LINE-1, SVA and *Alu* RIPs in PD brain extracted DNA. The data presented here aims to further the knowledge of the functions of retrotransposons as sources of regulation within the human genome and highlight their importance for future works.

Chapter 2 – Materials and Methods

Chapter 2 – Materials and methods

2.1 Materials

2.1.1 Commonly used materials

TBE buffer (5X) - 108 g Tris base (Sigma Aldrich), 55 g Boric acid (Sigma Aldrich), 5.84 g EDTA (Sigma Aldrich), made up to 2 L with distilled water. Used at 0.5X for agarose-based gels and running buffers for standard gel electrophoresis.

LB Broth, Miller (L3152) – 25 g/L in distilled water, autoclaved (Sigma Aldrich).

LB Agar, Miller (L3027) – 40g/L in distilled water, autoclaved (Sigma Aldrich)

2.1.2 Primers and oligonucleotides

All DNA and RNA oligonucleotides were obtained from Eurofins Genomics or Sigma Aldrich with purification suitable to the application. For all PCR and synthetic oligos for cloning, a basic de-salt purification was appropriate. Specific details of the primer sequences used for PCR can be found in appendix table 2.

2.1.3 Human PD DNA cohort for genotyping

Purified human genomic DNA (gDNA) extracted from blood, consisted of 192 Parkinson disease case samples with 176 age and gender matched controls of Estonian descent, gifted from Professor Sulev Koks at the University of Tartu, Estonia. Purification of gDNA performed in the University of Tartu with ethical approval. The purity and quantification of gDNA isolation was assessed using Nanodrop light spectroscopy at a 260nm absorbance. 260/280nm and 260/230nm ratios of over 1.8 indicated a pure yield of DNA and passed the quality control check.

2.1.4 NABEC human brain DNA samples

North American Brain Expression Consortium (NABEC) cohort DNA samples obtained from the National Institutes of Health (NIH) were used to genotype several targets of interest. NABEC is a collection of freely available data from healthy brain tissue including Illumina genome-wide genotyping array data, whole genome sequencing (WGS), exome sequencing and RNA-Seq. The genome-wide genotyping arrays from 382 neurologically healthy control samples was leveraged to generate tagging SNPs for the *LRRK2* SVA-C (**section 4.2.2**) and could be correlated with differential gene expression profiles via correlation of RNA-seq data in future works.

2.1.5 Parkinson's disease and control brain DNA for next generation sequencing library preparations; Retrotransposon capture sequencing (RC-Seq) and Whole genome sequencing (WGS)

Genomic DNA isolated from brain tissue gifted from clinical associate Professor, Christos Proukakis in the Queen Square Institute of Neurology at University College London. Approximately 15µg of gDNA was obtained from two brain regions, frontal cortex and cerebellum, from six post-mortem cases of PD and two control healthy patients giving 16 samples total. Sample preparation to isolate gDNA was performed externally at the University College London with full ethical permissions obtained. As a further suitable control group, 11 healthy aged samples (mean age of 88 ranging between 78-94 years old) from the Dyne Steele cohort (based in Manchester) of normal cognition at the time of death were also utilised (further sample details in **chapter 5**). In total, five PD case and one PD control and 11 healthy aged controls were processed for RC-Seq (21 samples total).

For WGS, two PD samples and nine healthy aged samples were processed all of which had also been analysed using RC-Seq. Further details are outlined in **chapter 5**.

2.1.6 Human cell lines and media

Commonly used reagents in cell culture of established cell lines:

Component and stock concentrations	Supplier
PBS pH7.2 (1X)	Gibco
Trypsin-EDTA Solution (0.25%)	Sigma
Foetal bovine serum (FBS), heat inactivated (cat. 10500064)	Gibco
Sodium pyruvate (100mM)	Sigma
L-glutamine (200mM)	Gibco
MEM non-essential amino acids (100X)	Gibco
Penicillin-streptomycin (100X)	Sigma
DMSO (neat)	Sigma

Human cell line SH-SY5Y (ATCC: CRL-2266): The SH-SY5Y cell line was derived from the parental cell line SK-N-SH originally taken from bone marrow tissue extracted from a 4-year-old female with a neuroblastoma phenotype. This cell line exhibits neuronal type characteristics such as catecholaminergic (but not restricted to dopaminergic) neuronal properties making it a suitable model for the experiments performed within this thesis [135].

Growth media for SH-SY5Y: 50:50 mix of Minimal Essential Medium Eagle (Sigma) with Nutrient Mixture F-12 Ham (Sigma), supplemented with 10 % (v/v) FBS, 1% (v/v) penicillin-streptomycin, 1 % (v/v) L-glutamine, and 1 % (v/v) sodium pyruvate.

Human cell line SK-N-AS (ATCC: CRL-2137): Neuroblast cell line originally extracted from bone marrow metastatic site of a brain neuroblastoma from a 6-year-old female of Caucasian ethnicity.

Growth media for SK-N-AS: Dulbecco's minimum essential media containing 4.5g/L D-glucose with 200mM L-glutamine (Gibco - cat no - 11965092) supplemented with 10% (v/v) foetal bovine serum, 1% (v/v) MEM non-essential amino acids and 1% (v/v) penicillin-streptomycin.

Human cell line HEK293 (ATCC: CRL-1573): Embryonic kidney cell line of foetal origin. Although described as a kidney derivative, HEK293 express markers of renal progenitor cells, neuronal cells and adrenal gland tissue leading to much debate of exact cell origin [136]. Ease of transfection and survivability in single cell cultures make these ideal cells for CRISPR protocols.

Growth media for HEK293: Dulbecco's Modified Eagle Medium (DMEM) containing 4.5g/L D-Glucose and L-Glutamine, supplemented with 10% (v/v) FBS, 1% (v/v) sodium pyruvate and 1% (v/v) penicillin-streptomycin.

Human cell line Hap1: Hap1 is a near haploid cell line used for optimisation in the subsequent CRISPR protocols as it contains of one copy of genes of interest in this thesis. Two genes of interest in this thesis are the *LRRK2* and *INPP5F* genes which are located on chromosome 12 and 10 respectively. The Hap1 cell line contains only one copy of these chromosomes making them ideal for CRISPR mediated knockouts. The parental cell line – KBM-7 is a chronic myelogenous leukemia (CML) cell line derived from a 40-year-old male patient.

Growth media for Hap1: Iscove's Modified Dulbecco's Medium (IMDM) with 10% (v/v) foetal bovine serum and 1% (v/v) penicillin-streptomycin.

Episomally derived induced pluripotent stem cells (iPSC) – Commercially available iPS cell line (A18945) is a human line derived from CD34+ cord blood. The protocol to create iSPCs uses a three-plasmid seven-factor Epstein-Barr nuclear antigen (EBNA)-based episomal system to reprogram cells into iPSCs. Factors supplied by plasmids include Sox2, Oct4, Klf4, L-Myc, Nanog and Lin28 for reprogramming. Episomal vectors also contain the SV40 large T antigen. iPS cells are cultured on Matrigel (Corning) coated 6 well plates required for adhesion of cultured cells in Essential 8 medium (Gibco) with Essential 8 supplement (Gibco). For routine culture and splitting of iPSCs, Rho-associated coiled-coil kinase inhibitor (ROCK inhibitor - ROCKi) was used (Tocris (Y-27632 dihydrochloride)), which improves viability of iPSCs during passage. 10 mg of ROCKi was resuspended in 3.12 ml of ultra-pure water to make 10 mM stocks which were aliquoted and stored at -20°C. Matrigel (Corning) was obtained from Fisher Scientific and supplied at 8mg/ml concentration. Matrigel was aliquoted into 2 mg and 0.66 mg aliquots which are appropriate for 6 wells and 2 wells of a 6-well plate respectively using strict manufacturer's guidelines and kept ice-cold at all times to prevent polymerisation. Culture methods of iPSCs are detailed in **section 2.2.9.2.1**.

Forebrain cortical neurons differentiated from iPSCs – Refer to **section 2.2.7.2** for full differentiation protocol. Forebrain neurons were cultured on poly-L ornithine (Sigma), 2µg/ml Fibronectin (Sigma) and 200ng/ml Laminin (Sigma) coated 6 well plates in N4 media (table 2.2). N4 media consists of a DMEM/F12/Neurobasal base media with multiple supplements used as part of the differentiation protocol. Neural induction media constituents include the SMAD and AMPK/BMP inhibitors SB431542

and Dorsomorphin respectively which promote neural differentiation in human pluripotent stem cells.

Table 2.1 – Formulation of N3 media plus the SMAD and AMPK/BMP inhibitors used for neural induction in the differentiation of iPSC into forebrain cortical neurons.

N3 media + SMAD and AMPK inhibitors for neural induction		
Component	**Volume for 500ml total**	**Final concentration**
DMEM/F12 + Glutamax (Gibco)	250ml	-
Neurobasal media (Gibco)	250ml	-
Penicillin/Streptomycin (100X) (Gibco)	5ml	1X
B-27 supplement without Vitamin-A (50X) (Gibco)	5ml	0.5X
N-2 supplement (100X) (Gibco)	2.5ml	0.5X
Glutamax (100X) (Gibco)	2.5ml	1X
Non-essential amino acids (100X) (Gibco)	2.5ml	0.5X
2-Mercaptoethanol (55mM - 1000X) (Gibco)	500µl	55µM
Human recombinant Insulin (10mg/ml) (Gibco)	50µl	1µg/ml
Dorsomorphin (10mM) (Sigma)	75µl	1.5µM
SB431542 (10mM) (Tocris)	500µl	10µM

Table 2.2 – Constituents of neuronal N4 media for culturing differentiated forebrain cortical neurons from iPSC lineage. Post neural differentiation, cells are cultured in this media until they are either used in assays or frozen down and stored long term.

N4 Media		
Component	**Volume for 500ml total**	**Final concentration**
N3 Media (details above)	500ml	-
Retinoic acid (10mM) (Sigma)	2.5µL	0.05µM
Brain-derived neurotrophic factor (BDNF) (25µg/ml) (Sigma)	40µL	2ng/ml
Glial-derived neurotrophic factor (GDNF) (10µg/ml) (Sigma)	100µL	2ng/ml

2.1.7 Plasmid vectors used

Several commercial and non-commercial vectors were used throughout this thesis within reporter gene assays to test SVA function and CRISPR mediated knockouts of established cell lines. The plasmids used for reporter gene assays include the commercially available pCR-Blunt from Invitrogen (**figure 2.1**) and pGL3 based luciferase reporter vectors pGL3b (basic) and pGL3p (promoter) vectors from Promega (**figure 2.2**). The non-commercially available pSHM06 vector was also used in reporter gene assays which was gifted from Professor Gerald Schumann at the Paul Ehrlich Institut (PEI), Langen, Germany. This vector contains an intron of the triose phosphate isomerase (TPI) gene into which SVAs of interest were cloned upstream on a Renilla luciferase reporter cassette (**figure 2.1.7.3**).

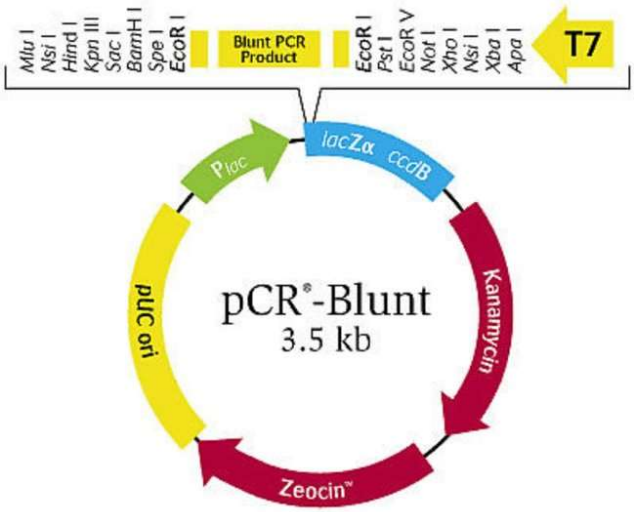


Figure 2.1 – Schematic plasmid map of the pCR-blunt vector from Invitrogen (image supplied from the manufacturer). The multiple cloning site (MCS) contains multiple restriction enzyme cut site to be used for downstream cloning into target expression vectors. The MCS is located in the lacZα and ccdβ cassette which allows for both blue/white X-gal screening and a significantly improved rate of successful cloning. Kanamycin antibiotic selection was used for all cloning of pCR-blunt constructs.

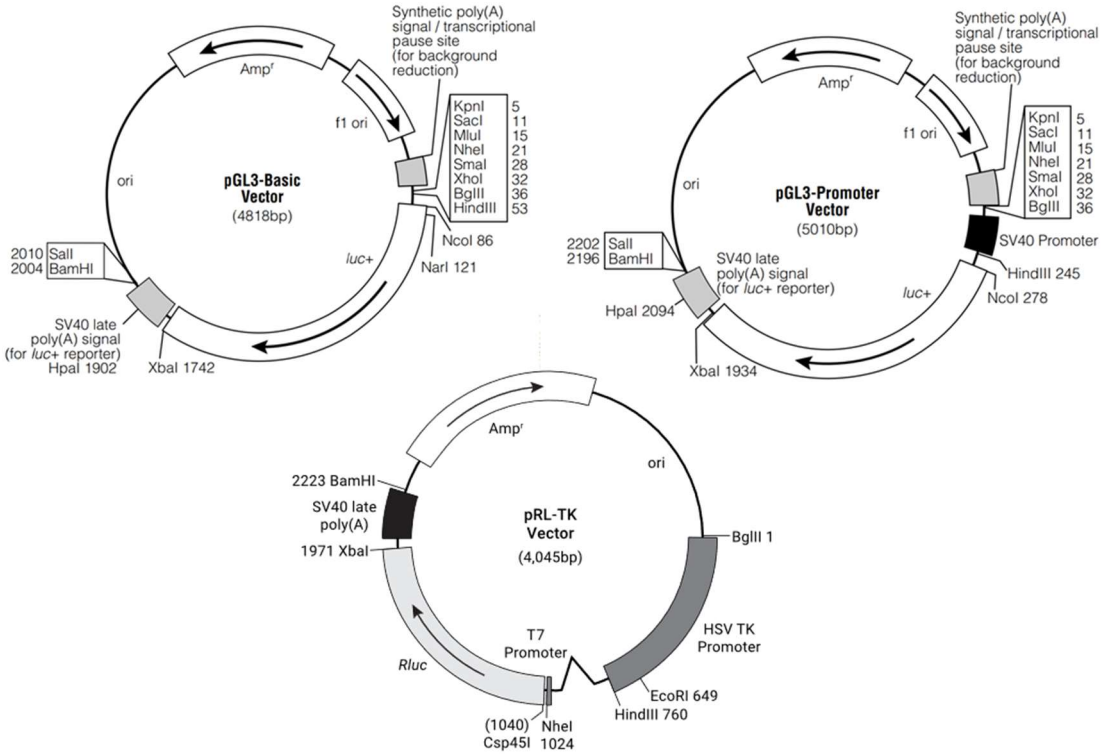


Figure 2.2 – Vector maps of the pGL3-basic (pGL3b), pGL3-promoter (pGL3p) and pRL-TK (Renilla luciferase – thymidine kinase) control vectors (Promega) used in the reporter gene assays to test SVA function. Vector maps supplied from the manufacturer. The pGl3p vector contains an SV40 minimal promoter to drive expression of the Firefly luciferase cassette in contrast to the pGL3b vector which does not contain a promoter. The pRL-TK vector was used in co-transfections as an internal control to normalise against and contains a Renilla luciferase reporter driven by the Herpes simplex virus (HSV) thymidine kinase (TK) promoter for all reporter gene assays.

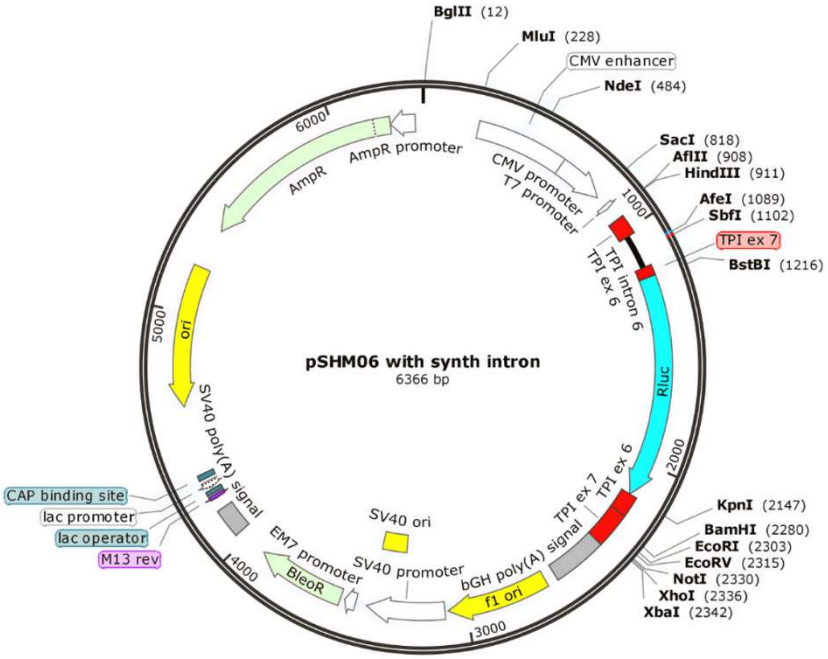


Figure 2.3 – SnapGene vector map of the non-commercially available pSHM06 vector (gifted from Professor Gerald Schumann) with synthetic intron of the triose phosphate isomerase (TPI) gene which contains the AfeI and SbfI restriction sites used for the cloning of SVA elements. The pSHM06 vector contains a Renilla luciferase reporter gene cassette driven by a CMV promoter and enhancer for high expression of luciferase.

The EF1α-pSpCas9(BB)-2A-GFP plasmid was used for the CRISPR mediated knockouts of all SVA targets used throughout this thesis. The EF-1α promoter was chosen to drive Cas9 expression rather than the commercially available CMV driven plasmid due to rapid methylation of CMV promoters that can occur within pluripotent stem cells and derived lines for which the initial CRISPR experiments were intended [137].

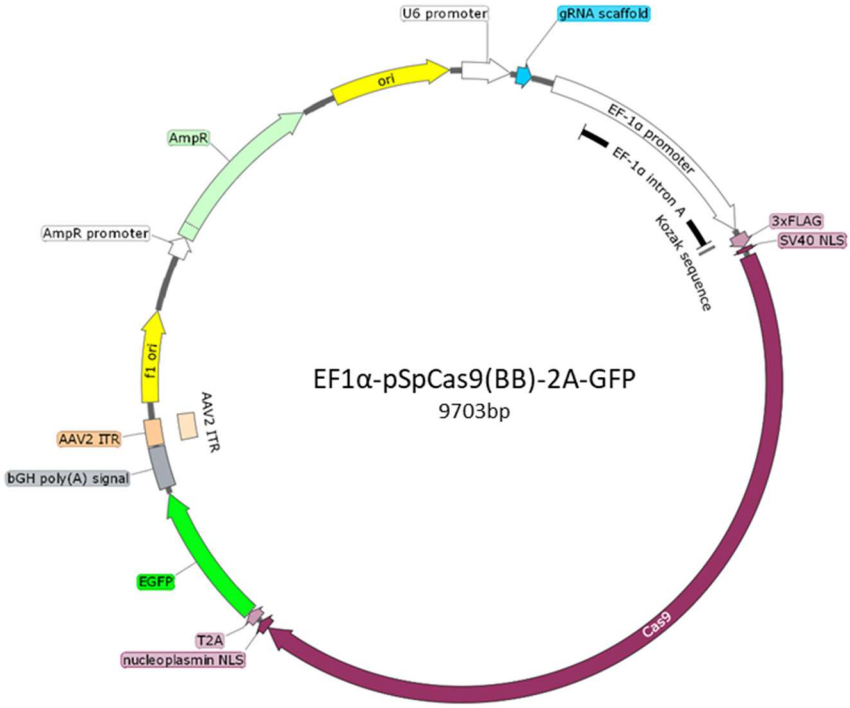


Figure 2.4 – SnapGene generated plasmid map of the EF1α driven pSpCas9(BB)-2A-GFP construct used to modify cell lines of interest using CRISPR. Oligonucleotide sequences that correspond to designed guide RNA (gRNA) were cloned at specific BbsI restriction sites located immediately 5' of the gRNA scaffold (blue). This allows expression of the gRNA sequences with gRNA scaffold and Cas9 together within the same cell to produce the desired CRISPR modification.

2.2 Methods

2.2.1 Polymerase chain reaction (PCR) primer design

Standard three step PCR was employed for all PCR amplifications of genomic SVA elements as a robust assay for the genotyping of case/control PD cohorts. DNA sequences for primer design were obtained from University of California, Santa Cruz genome browser (UCSC), human genome build 38 (hg38) (https://genome.ucsc.edu/). Primer design tools used include:

- Primer3 (http://primer3.ut.ee/)
- OligoAnalyzer tool from Integrated DNA Technologies (IDT) (https://eu.idtdna.com/pages/tools/oligoanalyzer)
- *in-silico* PCR (UCSC) (https://genome.ucsc.edu/cgi-bin/hgPcr)
- Blat (UCSC) https://genome.ucsc.edu/index.html
- Nucleotide blast (Blastn) (NCBI) (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

The most up to date sequences are downloaded from UCSC, hg38 with additional flanking sequences. Primer sequences chosen had to meet pre-assigned parameters including: length of 18—22bp, GC content of 40-60%, Tm 55-65°C, avoidance of repetitive DNA such as retroelements and satellites, minimal homo and hetero dimer delta G v*alue*s between -0kcal/mole and -6kcal/mole and high specificity to reduce non-specific binding. Potential off target binding was assessed using the Blat tool (UCSC) in conjunction with nucleotide Blastn (NCBI) alignment tools to predict where primers would bind to within the genome. Specific primer sequences for targets are described in supplementary table 2.

2.2.2 Standard PCR setup

Standard PCR reactions for all targets used throughout this thesis consisted of amplification with either KOD Hot Start (Merck Millipore-Novagen), KOD xtreme (Merck Millipore-Novagen) or GoTaq Hot Start (Promega) polymerase-based reactions. **Table 2.3** describes a typical PCR reaction setup for a full length SVA. Full descriptions of all PCR reactions described in supplementary tables 1 and 2.

Table 2.3 – An example standard PCR reaction setup for SVA genotyping, each specific amplification reaction may differ. Initial reaction mix based on manufacturer's protocol and optimised with addition of betaine to reduce GC structure. Full thermocycling conditions for genotyping are detailed in supplementary table 2.

Component	Volume per reaction (20ul total)	Final concentration
Nuclease free water	8.8 µL	-
5M Betaine	4 µL	1M
10X KOD Hot Start buffer	2 µL	1X
25mM $MgS0_4$	1.2 µL	1.5mM
dNTPs (2mM each)	2 µL	0.2mM
Fw (5') primer (20µM)	0.3 µL	0.3µM
Rv (3') primer (20µM)	0.3 µL	0.3µM
KOD hot start polymerase (1U/µL)	0.4 µL	0.02U/µL
Template gDNA (5ng/µL)	1 µL	0.25ng/µl

Details of full thermocycling conditions used are listed in supplementary table 2. A typical programme for amplification of a full length SVA using KOD hot start polymerase was as follows: initial denaturation and hot start activation of enzyme 95°C for 2 minutes, followed by 35 cycles of 95°C denaturing for 20 seconds, 57-65°C primer annealing for 10 seconds and an extension of 70°C for 20sec/kb with a final hold temperature of 4°C.

2.2.3 Agarose gel electrophoresis and QIAxcel capillary gel electrophoresis

Agarose gel electrophoresis was used as the primary mode for analysing DNA fragments. Standard gel setup consisted of 1% agarose (Invitrogen – UltraPure cat no. 16500100) made up in 0.5X TBE containing 5µL/100ml ethidium bromide (Sigma - 500µg/ml). The density of the gel being made was dependent on the resolution required. Smaller fragments (<500bp) were analysed using 2-3% agarose to achieve higher resolution in contrast to larger fragments (>1kb) which required lower density gels for adequate resolution, typically 0.8-1%. Prior to loading of DNA samples into the wells of the gel for electrophoresis, 1X (final conc.) Gel loading dye (6X) (NEB) was added to those samples that did not previously contain coloured dyes. This helps visualise the sample being loaded into the wells, increases the density of the sample to ensure it sinks into the wells during loading and also provides visible marker dyes as the sample migrates through the gel to assess how far the sample has migrated. Size markers were added to allow sizing of fragments (100bp and 1Kb DNA ladders – Promega). 0.5X TBE was used as running buffer with sufficient buffer being added to cover the gel during electrophoresis which was typically run at 100V for 1 hour minimum. Gels were visualised using the BioDoc-It imaging system UV transilluminator (UVP) which allows adjustable visualisation (exposure time, UV intensity, magnification and focus) and capturing of digital images.

The QIAxcel advanced system (QIAGEN) was used in multiple genotyping assays as a high throughput, high resolution alternative to gel electrophoresis which allowed for up to 96 samples to be run in one continuous uninterrupted run. With resolution possible of up to 1-2bp within a fragment size of <500bp, the QIAxcel system was

ideal for genotyping amplicons such as CT element and poly-A variation within SVA elements. The selection of appropriate cartridges to use for specific samples required optimisation for each target and was dependent on the resolution required.

2.2.4 Nucleic acid purification

2.2.4.1 Genomic DNA (gDNA) purification

Genomic DNA (gDNA) purification was used in multiple assays including genotyping of SVAs and CRISPR modifications in established cell lines. All gDNA purifications from cell lines were performed using the GenElute Mammalian Genomic DNA miniprep kit (Sigma) using manufacturer's guidelines. Cells were collected by trypsin dissociation and centrifugation at 300 x g for 5 mins to form cell pellets prior to gDNA extraction as per manufacturer's protocols. gDNA was eluted in TE buffer (supplied in kit) and quantified for quality and purity using the nanodrop 8000 (**section 2.2.5**). gDNA was stored at -20°C for long term storage.

2.2.4.2 RNA purification

Messenger RNA (mRNA) extraction was necessary to assess gene expression levels in multiple assays including the CRISPR experiments. All RNA extraction was performed using the Monarch Total RNA Miniprep kit (New England Biolabs - NEB) using manufacturer's guidelines. For maximum yield of high-quality low degradation RNA, the protocol was always performed from freshly harvested cell pellets and processed the same day using RNase free equipment. On-column DNase I treatment (included in kit) was included in all RNA preparations to improve the purity of RNA with little gDNA contamination following manufacturer's instructions. All RNA preparations were eluted in 50µl of nuclease free water and stored at -20°C for short term storage

(up to 1 week) until use or -80°C for longer term storage. The quality and purity of RNA preparations was measured using the nanodrop 8000 spectrophotometer **section 2.2.5**. RNA that was not processed within 3 months was discarded to ensure only high-quality RNA was used within assays. RNA quality was also validated using 1% agarose gel electrophoresis which can identify significant quantities of gDNA contamination and RNA degradation.

2.2.4.3 DNA purification from agarose gel

DNA band excision performed with a disposable scalpel viewed under 302nm UV light and collected in 1.5ml Eppendorf tubes with exposure time kept to a minimal to avoid excessive degradation of the DNA. The DNA purification was performed using the Promega Wizard SV gel and PCR clean up system following manufacturer's guidelines. The purified DNA fraction was eluted in 30-50µl PCR grade nuclease free water.

2.2.5 Quantification of DNA and RNA quality and purity

All nucleic acid quantification and quality control was measured using the Nanodrop 8000 spectrophotometer (Thermo Fisher Scientific) which provides concentration and absorbance ratios for inferring purity. DNA with a 260/280nm ratio above 1.8 and a 260/230 ratio above 1.5 was considered pure and did not require further purification. RNA purity was set at a higher threshold to consider pure with a 260/280nm ratio above 2.0 with a 260/230 ratio above 1.8. RNA that did not meet necessary QC standards was re-purified using the Monarch total RNA miniprep kit. The 260/280nm ratio is generally an indicator of purity with respect to protein contaminants whilst the 260/230nm ratio is indicative of solvent, peptides or phenol contamination amongst others that often come from poor purification processing.

2.2.6 First strand cDNA synthesis

cDNA was synthesised from purified RNA (**section 2.2.4.2**) using the GoScript reverse transcription system (Promega) following manufacturer's guidelines. Typically, 100ng of high QC quality RNA (measured by nanodrop) was used for cDNA synthesis but this was dependent on the RNA yield per sample which ranged from 50ng – 120ng. cDNA synthesis using this system first required an initial denaturing step to the RNA, followed by the reverse transcription reaction. The first step used an equal mix of oligo (dT) and random primers and was setup in a 5µl reaction as follows:

Component	Volume per reaction (5ul total)	Final concentration
Nuclease free water	X µL	-
RNA (~100ng)	Y µL	-
Oligo (dT) primers (500ng/µL)	1 µL	100ng/µL
Random primers (500ng/µL)	1 µL	100ng/µL

The reaction mix was heated to 70°C for 5 minutes to denature the RNA and then cooled rapidly on ice for a further 5 minutes. The reaction tubes were centrifuged and stored on ice until ready to be used in the reverse transcription reaction to generate cDNA. The reverse transcription mix was made up separately to accommodate 15µL per reaction and added directly to the 5µL of RNA/primer mix to give a final reaction volume of 20µL. The reverse transcription reaction was setup as follows:

Component	Volume per reaction (20µL total)	Final concentration
Nuclease free water	5.5 µL	-
GoScript reaction buffer (5X)	4 µL	1X
$MgCl_2$ (25mM)	3 µL	3.75mM
dNTPs (10mM each)	1 µL	0.5mM
RNasin ribonuclease inhibitor (40U/µL)	0.5 µL	1U/µL
GoScript reverse transcriptase (160U/µL)	1 µL	-
RNA/Primer mix from initial step	5 µL	-

The reaction mix was placed in a thermocycler and incubated at 25°C for 5 minutes to anneal, 42°C for 1-hour extension and heat inactivated at 70°C for 15 minutes. Several cDNA dilutions were made up (1:10, 1:100 and 1:1000) with nuclease free water and stored at -20°C until use.

2.2.7 RT-PCR of cDNA to assess endogenous gene expression in cell lines

2.2.7.1 Primer design for targeting isoforms of *INPP5F* and *LRRK2*

The primer design for RT-PCR reactions follows the same basic principles as for gDNA targets described in **section 2.2.1** with some additional points taken into consideration. Amplicon size (<200bp) and specific positioning of primers within unique exons was key to producing targeted isoform specific amplification. Using the UCSC genome browser to extract mRNA sequences for both *INPP5F* and *LRRK2* primers were designed that specifically targeted exons because the target cDNA will contain no introns due to splicing. Where applicable, exons junctions were targeted for primer design because primers that overlap them would specifically bind to the cDNA template only (**figure 2.5**), thereby increasing the specificity if gDNA contamination was present within the sample preparation.

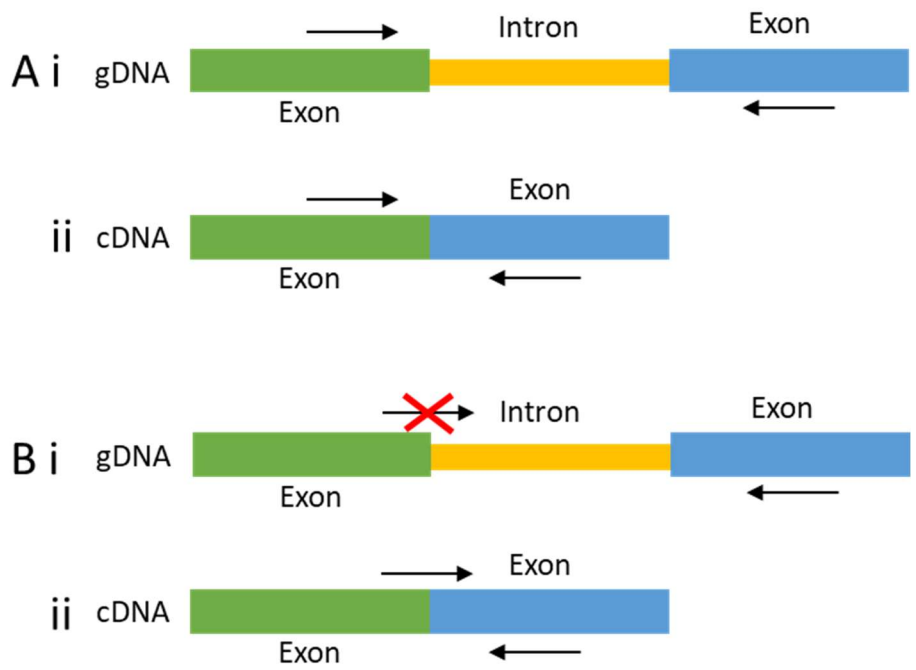


Figure 2.5 – Primer design for targeting cDNA and gDNA templates. (**A**) Represents a standard PCR across exons which would produce PCR amplicons of different lengths from both gDNA and cDNA templates. (**B**) Represents a common primer combination for RT-PCR or qPCR that will uniquely identify cDNA targets only by designing at least one primer across an exon junction.

Individual primer pairs to target the three isoforms of *INPP5F* and three isoforms of *LRRK2* were designed, details in **appendix table 2**, using unique exons within each isoform for specificity.

2.2.7.2 RT-PCR setup

All RT-PCR reactions were performed using the GoTaq Hot Start Polymerase (Promega) enzyme which is suitable for small amplicon sizes (<1kb) using non-GC rich templates (<70% GC content). cDNA generated from first strand synthesis (**section 2.2.6)** was used for all reactions. Full descriptions of each RT-PCR are in **appendix tables 1 and 2**. To identify the optimal quantity of cDNA to use for each target, RT-PCR was first performed using serially diluted cDNA, with 1:10, 1:100 and 1:1000 ratio dilutions in ultrapure water. For all RT-PCR reactions used in the CRISPR analysis of

expression differences of *LRRK2* and *INPP5F*/*BAG3*/*TIAL1* (**sections 3.2.4 and 4.2.3**) in response to SVA knockout, a 1:10 cDNA dilution was used.

A typical RT-PCR reaction was setup as follows:

Component	Volume per reaction (20µL total)	Final concentration
Nuclease free water	12.3 µL	-
Green GoTaq reaction buffer (5X)	4 µL	1X
$MgCl_2$ (25mM)	1.6 µL	1.28mM
dNTPs (10mM each)	0.4 µL	0.2mM
Fw (5') primer (20µM)	0.3 µL	0.3µM
Rv (3') primer (20µM)	0.3 µL	0.3µM
GoTaq hot start polymerase (5U/µL)	0.1 µL	0.025U/µL
cDNA template (dilute)	1µL	-

Cycling conditions included: 95°C for 2minutes, (open cycle) 95°C for 30 seconds, 60°C for 30 seconds, 72°C for 20 seconds (close cycle), for 30-40 cycles and a final extension at 72°C for 2 minutes with an infinite hold at 4°C. The cycle number was dependent on the abundance of the target and was optimised using higher cycles for lower abundance targets.

2.2.8 Methods for cloning

Several clones for testing SVA function have been generated throughout this thesis including luciferase-based reporter gene constructs and cas9 based CRISPR plasmids. The general method for generating these constructs was similar and was based primarily around PCR of the initial sequence of interest, ligation into target vector, transformation using *E. coli* strains, selection of positive colonies followed by purification of the construct.

2.2.8.1 Amplification and ligation into pCR-blunt

For all SVA reporter gene constructs, the target SVA sequence was PCR amplified using KOD hot start polymerase for its high fidelity, proof reading and ability to efficiently amplify high GC sequence, as detailed in **section 2.2.2**. This enzyme leaves blunt ends at both the 5' and 3' ends of the amplicon which can be ligated directly into the pCR-blunt vector (zero blunt PCR cloning kit – Invitrogen cat no. K270040) (**figure 2.1**). This vector is ideal for high efficiency cloning due to the presence of the ccdβ cassette which when intact is lethal to *E. coli* by disrupting DNA gyrase (GyrA) and inhibiting cell division. The cloned sequence inserted will disrupt the ccdβ domain allowing for propagation of the plasmid and survival of the cell. All the cells which do not contain a cloned sequence should not survive.

PCR amplification of the target sequence to be cloned was performed and purified using gel electrophoresis and gel excision purification (**sections 2.2.3 and 2.2.4.3**). The resulting purified DNA was quantified using the nanodrop 8000 (**section 2.2.5**). Prior to ligation, the quantity of PCR product to be used in the reaction was calculated using the following equation:

$$Insert(ng) = \frac{vector\ (ng)\ x\ size\ of\ insert\ (kb)}{size\ of\ vector\ (kb)}\ x\ ratio\ \left(\frac{insert}{vector}\right)$$

The ratio recommended for blunt cloning in the supplied protocol was between 10:1 and 100:1 insert:vector. A 10:1 ratio usually provided high efficiency resulting in large numbers of positive colonies to screen.

The ligation reaction was performed using the zero-blunt PCR cloning kit as follows:

Component	**Volume**
pCR-blunt vector (25ng/µl)	1µl
Insert DNA (PCR product)	1-5µl
5X ExpressLink T4 DNA ligase buffer	2µl
ExpressLink T4 DNA ligase (5U/µL)	1µl
Nuclease free water to a total of	**10µl**

The ligation mix was incubated for a minimum of 4 hours, typically overnight, at 16-20°C (room temperature).

2.2.8.2 Cloning into expression vectors and restriction digest setup

Suitable restriction enzyme-based cloning strategies differed between targets and had to be designed per plasmid. A typical strategy for cloning into the pGL3p (Promega) and pSHM06 (gifted from Prof. G Schumann) expression vectors (**section 2.1.7**) for SVA luciferase reporter gene assays include excising the insert of interest from pCR-blunt using restriction enzymes that cut either side of the sequence and ligating into a complementary restriction site within the target vector. For example, to clone the *INPP5F* SVA-D into the pSHM06 expression vector, the SVA was cut out

of pCR-blunt using *NsiI* which cuts both sides of the SVA in the multiple cloning site (MCS) of pCR-blunt (**figure 2.1**) which generates a fragment that can be efficiently ligated into the *Sbf-I* site of the expression vector due to compatible cohesive ends. The basic protocol for sub-cloning consists of restricting a large quantity (>1µg) of pCR-blunt containing the sequence of interest, separating the insert by migration through an agarose gel (**section 2.2.3)**, purifying the DNA from the gel (**section 2.2.4**), ligation of the insert into an appropriately restricted target vector and transformation of the ligation reaction mix using competent *E. coli*. The setup for standard restriction enzyme reaction setup using Promega or NEB enzymes was set up as follows:

Component	**Quantity**
DNA (pCR-blunt containing insert)	1µg
Restriction enzyme buffer (10X)	2µl
Acetylated BSA (10µg/µl) *	0.2µl
Restriction enzyme (10U/µL)	1µl
Nuclease free water to a total of	**20µl**

*Note, NEB restriction enzyme buffers already contain BSA, so did not require addition.

All restriction enzyme reactions performed used enzymes supplied from Promega or NEB and setup to manufacturer's guidelines. Incubation of the reaction was performed at the appropriate temperature for each enzyme typically for 1-2 hours and was performed using the appropriate buffers for each enzyme for optimal activity.

The ligation of inserts into expression constructs using T4 ligase (Promega) was performed as follows:

Component	**Quantity**
Vector DNA	50ng
T4 ligase buffer (10X)	2µl
Insert DNA	X
T4 ligase (3U/µL)	1µl
Nuclease free water to a total of	**20µl**

For the cloning of SVA constructs into the pSHM06 vectors in which the target vector had been cut with a single cohesive end enzyme (e.g. *Sbf-I*) which would be extremely prone to re-ligation without the insert, de-phosphorylation of the vector was employed. This involved using Antarctic phosphatase (NEB) to remove the 5' and 3' phosphate groups to prevent the vector re-circularising during ligation. The Antarctic phosphatase reactions were setup to contain 5 units of phosphatase enzyme per 1pmol of 5'/3' DNA ends. For reference, 1pmol of DNA ends is approximately 1µg of a 3000bp plasmid. The reaction setup for dephosphorylation of cut pSHM06 vectors to improve ligation efficiency was setup as follows:

Component	**Quantity**
Restricted pSHM06 vector DNA	1µg
Antarctic phosphatase buffer (10X)	2µl
Antarctic phosphatase (5U/µL)	2µl
Nuclease free water to a total of	**20µl**

The reaction mix was incubated at 37°C for 30 minutes followed by heat inactivation at 80°C for 2 minutes. The mix was then used directly in ligation reactions using T4 ligase with no additional purification steps required.

2.2.8.3 Transformation of chemically competent *E. coli*

Ligation mixes were transformed into subcloning efficiency DH5α chemically competent *E. coli* cells (Invitrogen). These cells are suitable for use with the pCR-blunt vector selection and are responsive to the ccdβ selection described in **section 2.2.8.1**. Transformation was performed under manufacturer's guidelines. In brief, 50µl of pre-aliquoted cells were thawed on ice, 1-5µl (1-10ng DNA) of ligation mix was added and incubated for 30 minutes on ice followed by a 20 second heat shock in a 42°C water bath before returning to ice for a further 2 minutes. 950µl of pre-warmed LB-broth was added and the cells were incubated in a shaking incubator for 1 hour at 225rpm. After incubation, 50-200µl of the transformant mix was added and spread using sterilised glass spreaders to pre-warmed antibiotic selection LB-agar plates and incubated overnight at 37°C until individual colonies were seen. Antibiotic selection plates used were dependent on the antibiotic cassettes present in the vectors. For pCR-blunt the antibiotic chosen was kanamycin at a final concentration of 50µg/ml in both LB agar plates and LB broth cultures. For pGL3 based plasmids, pSHM06 and pSp-EF1α-Cas9(BB)-2A-GFP plasmids, ampicillin antibiotic selection was used at a final concentration of 100µg/ml for both LB agar and LB broth cultures.

2.2.8.4 Purification of plasmid DNA from transformed E. coli

2.2.8.4.1 Miniprep purification of plasmid DNA

Individual colonies were picked from LB agar plates on which the transformed *E. coli* had been plated and transferred to 5ml LB broth cultures containing appropriate antibiotic and incubated overnight at 37°C in a shaking incubator at 225rpm. The following day, 1-4 ml of the culture was taken and used for mini prep extraction using the Promega Wizard Plus SV Minipreps DNA purification system following manufacturer's instructions. Purified DNA was eluted in 30µl of nuclease free water, analysed for purification quality using the nanodrop 8000 spectrophotometer (**section 2.2.5**) and stored at -20°C. Restriction enzyme digestion (**section 2.2.8.2**) and agarose gel electrophoresis (**section 2.2.3**) was used to test the presence of correct insert by cutting both flanks of the insert and assessing the presence of a correct size band. If using a non-directional cloning strategy, orientation of insert could also be determined by using enzymes that make a single cut inside the insert and another in the backbone leading to predictable band sizes to infer orientation.

2.2.8.4.2 Maxiprep of plasmid DNA

Maxiprep purification of plasmid DNA was used when large quantities of high purity plasmids were needed for applications such as transfections in cell culture. This includes all the plasmids used for luciferase reporter gene assays and CRISPR cas9 based plasmids. Maxi preps were performed by inoculating 100 ml of LB broth containing appropriate antibiotic with 50-100 µl of bacterial culture from the 5 ml cultures used for minipreps (**section 2.2.8.4.1**). The 100 ml cultures were grown overnight at 37°C in a shaking incubator at 225rpm. Plasmid DNA was extracted from

the bacterial culture using the QIAGEN plasmid maxi kit following the manufacturer's protocols for high copy plasmids. The DNA pellets following isopropanol/ethanol precipitation were resuspended in 200-500µl of nuclease free water depending on the size of pellet obtained. Purity and yield were measured using the nanodrop 8000 before being stored at -20°C until use.

2.2.8.5 Sequence validation of constructs

All sequence validation of the constructs used throughout this thesis was performed externally by Source Bioscience using Sanger sequencing. The requirements for sequencing per reaction include 5µl of 100ng/µl plasmid using 5µl of each sequencing primer. Full list of sequencing primers used for constructs provided in **appendix table 3**. For high GC content (>60%) sequences that are prone to secondary structure such as SVAs, additional dGTP chemistry was including in the reactions.

2.2.8.6 Glycerol stocks

For long term storage of transformed clonal bacteria, glycerol stocks were made. 1-5ml of overnight culture was transferred to a 15ml falcon tube and centrifuged at >4000 x g for 5 minutes and resulting supernatant was discarded. The bacterial pellet was resuspended in sterile filtered (using 0.25µm filter) 15% glycerol (v/v in LB broth) before being transferred to cryovials and immediately frozen at -80°C.

2.2.9 Cell culture methods

2.2.9.1 Culture, passage and freezing down of established cell lines

Multiple established human cell lines have been used throughout this thesis, including SH-SY5Y, SK-N-AS, Hap1 and HEK293. Specific information regarding lineage and media formulations can be found in **section 2.1.6**. Standard culturing techniques

were applied to each cell line used with specific information on recommended split ratios provided by the American Type Culture Collection (ATCC). All established cell lines were typically cultured in T75 flasks at 37°C and 5% CO_2 in a humidified incubator and passaged at 70-80% confluency. Passaging of cells was performed by aspirating the culture media, washing with 10ml pre-warmed cell culture grade PBS, addition of 1ml Trypsin-EDTA (0.25%) (Sigma) and incubation for 3-5 minutes at 37°C until cells detached. The trypsin was inactivated by addition of 10ml of media containing 10% FBS followed by gentle pipetting to triturate the cell clumps into single cells. The cell suspension was transferred to 15ml falcon tube and centrifuged at 130 x g for 5 minutes. The supernatant was aspirated and the remaining cell pellet resuspended in a volume of pre-warmed media appropriate for the desired split ratio, e.g. 10 ml resuspension followed by transferral of 1ml of cell suspension into a new T75 flask would equate to a 1:10 split ratio.

For long term storage of cells, cell aliquots were frozen for later use in freezing media containing 90% FBS and 10% DMSO with a volume appropriate to retain a high density of cells for optimal future recovery (typically 0.5-1ml aliquots). The cell suspension was transferred to cryovial using 1 ml per vial and placed in a Mr Frosty containing isopropanol at stored in a -80°C freezer for 24 hours. After this time, the frozen aliquots were transferred to liquid nitrogen for long term storage.

2.2.9.2 iPSC cell culture and forebrain cortical neuron differentiation protocol

Commercially available human episomal iPSCs (Thermo Fisher – cat no. A18945) were cultured in feeder-free conditions using essential 8 (E8) media containing 1X E8-supplement with media changes being performed daily. Standard incubation

conditions were used, namely 37°C with 5% CO_2. All iPSCs were cultured in Matrigel coated 6-well cell culture treated plates (Corning). Matrigel coated 6-well plates were made fresh just before use by dissolving pre-aliquoted frozen Matrigel in an appropriate volume of cold knockout (KO) Dulbecco's modified eagle medium (DMEM) (Gibco) and adding 2 ml per well (2mg Matrigel resuspended in 12 ml KO DMEM would coat a full 6 well plate). The Matrigel/KO DMEM was incubated for 30 minutes at 37°C and kept at RT until use. Coated plates could be stored at 4°C for up to one week before use but this was not preferred. Matrigel/KO DMEM mix was aspirated immediately prior to plating of iPSCs.

2.2.9.2.1 Splitting iSPC protocol

iPSCs were passaged as clumps of cells rather than single cell suspensions to improve cell viability during routine culture as they were sensitive to heavy trituration. To avoid single cell suspensions, enzyme-free splitting protocols were employed rather than typical trypsin-based passaging. Cell were left to grow as monolayers in 6 well-plates until 80% confluency, at which point the culture media was aspirated and the cells washed with 2 ml Dulbecco's PBS (DPBS – Gibco). DPBS was aspirated and replaced with 1 ml of PBS based enzyme-free cell dissociation buffer (Gibco) and incubated for 5 mins until cells retracted their processes and 'balled up' as observed under a light microscope. At this point, the cell dissociation buffer was aspirated and 3 ml pre-warmed E8 media containing 10 µM ROCK inhibitor (Y-27632 dihydrochloride – Tocris) was added to each well. The cells were then gently pipetted 4-5 times to remove them from the bottom of the wells and 1 ml of the cell suspension for a 1:3 splitting ratio was transferred to a Matrigel pre-coated plate

containing 3 mls of E8 media with 10 µM ROCK inhibitor and evenly distributed using gentle rocking of the plate. The following day, the media was replaced with E8 media without ROCK inhibitor.

2.2.9.2.2 Forebrain cortical neuron differentiation protocol

The differentiation protocol for forebrain cortical neuron differentiation from iPSCs used was supplied by Dr. Mark Cookson at the Laboratory of Neurogenetics, NIH and followed a 24-day SMAD and AMPK/BMP inhibition protocol. Human iPSCs were cultured using the methods described in **section 2.2.9.1** in a 6-well plate format until 95-100% confluency was reached. The constituents of the neural induction media (N3) and induced neuron culture media (N4) are described in **section 2.1.6**. The differentiation protocol is described for one full 6-well plate of iPSCs to be differentiated at 100% confluency and was as follows:

Day 1 – 11 Daily media changes with N3 media + 1.5µM dorsomorphin + 10µM SB431542

Day 12 – 15 Daily media change with N3 media without dorsomorphin and SB431542

Day 16-20 Daily media change with N3 media + 0.5µM retinoic acid (stock 10mM, 1µl in 200ml)

Day 18 Two 6 well plates coated with poly L-ornithine overnight at 37°C in the incubator (1ml per well)

Day 19 Coated wells were washed three times with DPBS then coated with Fibronectin and Laminin in DPBS:

20µl (1mg/ml) fibronectin

2µl (1mg/ml) Laminin

10mL DPBS

1.5ml of coating mix added per well and incubated overnight at 37°C to polymerise.

Day 20 Coated wells were washed 3 times with 2ml DPBS per well with the last wash being left on until cells were ready to be plated to avoid drying of the coating.

Cells were washed with 4ml DPBS, 1 ml of Accutase cell detachment solution (STEMCELL technologies) was added to each well and incubated for 5 minutes at 37°C. 1 ml of pre-warmed N3 media was added and gently pipetted to disrupt the cells. Cells were collected in a single 15ml falcon and centrifuged at 130xg for 5 mins. The supernatant was aspirated and the cell pellet was resuspend in 12mls of N4 media + ROCK inhibitor and distributed evenly (2 ml cell suspension per well) into the 2 pre-coated 6-well plates containing 3 mls of pre-warmed N4 + ROCK inhibitor.

Day 21 – 23 Daily media change with N4 media without ROCKi as cells become more confluent and grow processes.

Day 24 End of differentiation protocol. Cells can be continued to culture with N4 media or frozen down for later use.

At the end of the differentiation protocol (day 24) two wells of differentiated cells were kept for continued culture to be used in luciferase reporter gene assays. The remaining ten wells were frozen down using the Accutase protocol detailed on day 20 of the differentiation protocol. After centrifugation, the supernatant was aspirated, and the cell pellet was resuspended in 10 ml Synthafreeze cryopreservation medium (Gibco). 1 ml of cell suspension was added to each cryovial and stored in a Mr frosty with isopropanol to fill line at -80°C overnight. The following day, the cryovials were transferred to liquid nitrogen for long term storage.

2.2.9.3 Counting cells for experimental assays

A haemocytometer was used for the counting of both established cell lines and iPSCs/differentiated neurons. The methods of cell dissociation differed between these two cell types, the established cell lines were dissociated using a standard

trypsin method (**section 2.2.9.1**) contrasting to the iPSCs and derived neurons, which were dissociated using the Accutase protocol (**section 2.2.9.2.2**). To count dissociated cells, 10µl of cell suspension was applied to one side of the haemocytometer with glass cover slip in place (cleaned with 70% ethanol) to form an even layer of dispersed single cells across the central etched counting square. The counting square was visualised under a light microscope and the cells within the 4 corners of the haemocytometer were counted including those cells touching two of the four boundary lines within each square. The average of four counts was taken and multiplied by 10,000 to give the cell number per ml in suspension. This number was then used to calculate the volume of cell suspension needed for plating in a specific assay.

2.2.9.4 Transient transfections for reporter gene assays

Luciferase reporter gene assays were performed in several established cell lines including HEK293, SH-SY5Y and SK-N-AS as well as the human iPSCs and neuronal derivatives.

2.2.9.4.1 Transfections of established cell lines

The established cell lines HEK293, SH-SY5Y and SK-N-AS were plated at a seeding density of 100,000 cells per well in 24-well plates and cultured for a minimum of 24 hours at 37°C with 5% CO_2 before transfections were initiated. All constructs were delivered into the cells using TurboFect transfection reagent (Thermo Scientific) using 1µg of reporter gene construct, 20ng of internal control plasmid using 2µl of transfection reagent in a 100µl transfection mix made up with serum free media. The internal control plasmid selected was dependent on the reporter gene used. For

pGL3 based constructs the reporter gene expressed was Firefly luciferase thus a thymidine kinase (TK) driven Renilla internal control was selected whilst the pSHM06 (TPI) constructs used a Renilla reporter with a Firefly luciferase internal control. The transfection mix was added to the cells and incubated for 4 hours, thereafter a media change was performed to increase cell viability. Cells were incubated for 48 hours to allow for transgene expression.

2.2.9.4.2 Transfections of iPSCs and differentiated neurons

Luciferase reporter gene assays in iPSCs and derived forebrain cortical neurons used the same constructs as tested in the established lines (HEK293, SH-SY5Y and SK-N-AS) but required a different transfection protocol suitable for pluripotent stem cells and neurons. For this protocol, the iPSCs and forebrain cortical neurons were seeded at different densities to account for the post-mitotic nature of the neurons that would not divide during the experimental procedures. Both the iPSCs and differentiated neurons were plated in 96-well plates with seeding densities of 25,000 and 75,000 cells per well respectively. The plated cells were incubated for 24 hours at 37°C with 5% CO_2 prior to transfection. Lipofectamine STEM reagent (Invitrogen – Thermo Fisher Scientific) was used for the transfections as it was optimised for delivery of plasmid DNA into pluripotent stem cells and neurons which are typically difficult to transfect at high efficiency. The transfection mixes contained 100ng of plasmid construct, 2ng of TK Renilla internal control plasmid, 0.3µl of Lipofectamine STEM made up in 10µl of Opti-MEM-I (Gibco) per well of 96-well plate.

2.2.10 Luciferase reporter gene assays

Luciferase reporter gene assays using established cell lines were performed in Liverpool using the Promega Dual-Luciferase reporter assay system with a Promega Glomax 96 microplate luminometer with injectors following manufacturer's guidelines. Luciferase assays for iPSCs and differentiated neurons were performed in NIH, Bethesda and used the Promega Dual-Glo luciferase assay system with a standard microplate luminometer. All tested samples were performed in quadruplicate (N=4) for each condition within a single assay. Ideally, the assay was then typically repeated 3 times to improve reliability of the results and account for minor changes between cell cultures over a three-week period. Unfortunately, due to time constraints, the iPSC/neuron luciferase reporter gene assays were only performed once.

2.2.10.1 Dual Luciferase assay for established cell lines

Forty-eight hours post-transfection, cell culture media was removed and washed with room temperature cell culture grade PBS and aspirated. For lysis, 100µl 1X passive lysis buffer, diluted from 5X concentrate with water, was added to each well in the 24-well plates. Cells were incubated at room temperature on a rocking plate for 15 minutes to achieve lysis. 20µl of the cell lysates was transferred to white opaque luminometer certified plates and transferred to the Glomax for analysis. Dual-Luciferase reporter assay system reagents were prepared to allow for 100µl of both the luciferase assay reagent II (LARII) and Stop and Glo reagent per sample following manufacturer's guidelines. The Glomax was set up using the pre-set Dual-Luciferase Reporter (DLR) conditions using two injectors with an integration time of

1.5 seconds for both injectors. The LARII reagent was injected first which measures Firefly luciferase signal followed by the Stop and Glo which detected the Renilla luciferase signal.

2.2.10.2 Dual-Glo Luciferase assays for iPSC and differentiated neurons

The Dual-Glo assay system (Promega) was selected for the iPSCs/differentiated neurons as the signal decay from this assay is significantly longer (~2 hours) than the dual luciferase assays (<5minutes). This was an important consideration given that a microplate reader with injectors was not available for these assays which had to be performed manually. The Dual-Glo assay kit negated the need for a separate cell lysis step by combining the lysis and Firefly luciferase signal detection steps using Dual-Glo Luciferase Assay Reagent in one reaction thus streamlining the protocol. At 48-hours post-transfection, the iPSCs and differentiated forebrain neurons were washed once with PBS and 75µl of PBS was added per well to the 96-well plate. To this, 75µl of Dual-Glo Luciferase Assay Reagent was added and incubated at room temperature for 15 minutes before being measured on a microplate reader using the SoftMax (Molecular Devices) software with Dual-Glo pre-set settings using a 500ms integration time. 75µl of Stop and Glo reagent was then added and incubated for a further 15 minutes at room temperature before the luminescent signal was measured on the plate reader using a 500ms integration time.

2.2.10.3 Statistical analysis of luciferase data

To normalise for differences in cell number, pipetting error and transfection efficiencies between wells, the ratios of Firefly to Renilla luciferase were calculated for individual wells and averaged. The averaged ratios were used to calculate fold

changes compared to the appropriate control for each assay. For example, for the pGL3p/SVA constructs the fold changes were normalised to pGL3p as the baseline. Statistical significance was calculated using one-tailed t-tests with a minimum 95% confidence interval and scored at three tiers of significance ($p<0.05$, $p<0.01$, $p<0.001$).

2.2.11 CRISPR of SVAs in established cell lines

The method of CRISPR (clustered regularly interspaced short palindromic repeats) used in this thesis was adapted from Ran *et al.* 2013 and utilised a non-homologous end joining (NHEJ) approach using HEK293 cells [138]. NHEJ is suitable for CRISPR of non-exonic modifications where single base pair fidelity is not critical. For precise gene editing, such as introduction of specific single nucleotide polymorphisms (SNPs), a homology directed repair (HDR) method may be preferable. A plasmid-based method was chosen which used a single plasmid that delivered both the Cas9 machinery and associated gRNA with RNA scaffold necessary for precise editing of a genomic site (**figure 2.1.7.4**). The SpCas9 cassette was derived from the bacteria *S. pyogenes* and required the protospacer adjacent motif (PAM) sequence 5' NGG which is ideal as this sequence appears abundantly in mammalian genomes. The pSpCas9(BB)-2A-GFP also contains an enhanced GFP (EGFP) tag which is expressed as a recombinant protein with Cas9 and subsequently cleaved at the T2A site which allows for an observable indication of the levels of Cas9 present in each cell and is useful to estimate transfection efficiencies.

To produce targeted deletions of SVA elements, two CRISPR constructs were used in a single transfection which contained different gRNAs that would cut either side of

the element to be removed and repaired using NHEJ. The elongation factor 1α promoter (EF-1α) driven plasmid was provided by the Cookson group in the laboratory of neurogenetics at NIH, Bethesda. This promoter was preferred over the commercially available cytomegalovirus (CMV) driven plasmid due to rapid hypermethylation of CMV promoters in pluripotent stem cell models in which the SVA KO models had initially been planned [139].

2.2.11.1 Guide RNA design

The guide RNA design tool used was provided by the Zhang lab at MIT (http://crispr.mit.edu/). This design tool worked best when inputting previously identified unique regions from UCSC that would exclude elements such as simple repeats and retrotransposons as these would produce guides with off-target effects. Suitable guides were generated dependent on the presence of suitable protospacer adjacent motif (PAM) sites and scored, based primarily on the number of predicted off-target sites, with guides that had minimal off-target sites scoring higher on an arbitrary 0-100 scale. All guides selected for the CRISPR experiments met a minimum required score of >70. Information regarding which strand the guide RNA would bind to was also provided, however in this case, was unnecessary to consider with the chosen CRISPR method as the cas9 in the model produced double strand breaks that were repaired by the cell using NHEJ (**figure 2.6**). Identified suitable guide sequences were modified to allow for the golden gate cloning strategy which requires cohesive ends at the 5' and 3' ends of the guide oligo to efficiently ligate into the target EF1α-pSpCas9(BB)-2A-GFP vector (**figure 2.4**). For each guide to be cloned, a double stranded oligo containing the desired gRNA and complementary sequence had to be

designed which contained the necessary modifications. The modifications necessary to generate the cloneable insert sequence for each guide are detailed in the following example:

```
gRNA – 5′ CACCGGCTAAATCGATAATCACAGT      3′
gRNA – 3′      CCGATTTAGCTATTAGTGTCACAAA 5′(Complement)
```

The desired guide sequence is in red where the top strand represents the output from the guide RNA design tool. All additional nucleotides in black represent the modifications necessary to allow the golden gate cloning strategy. The modified oligo sequences selected for each target are detailed in **appendix table 4** and were ordered from Sigma 5′ > 3′.

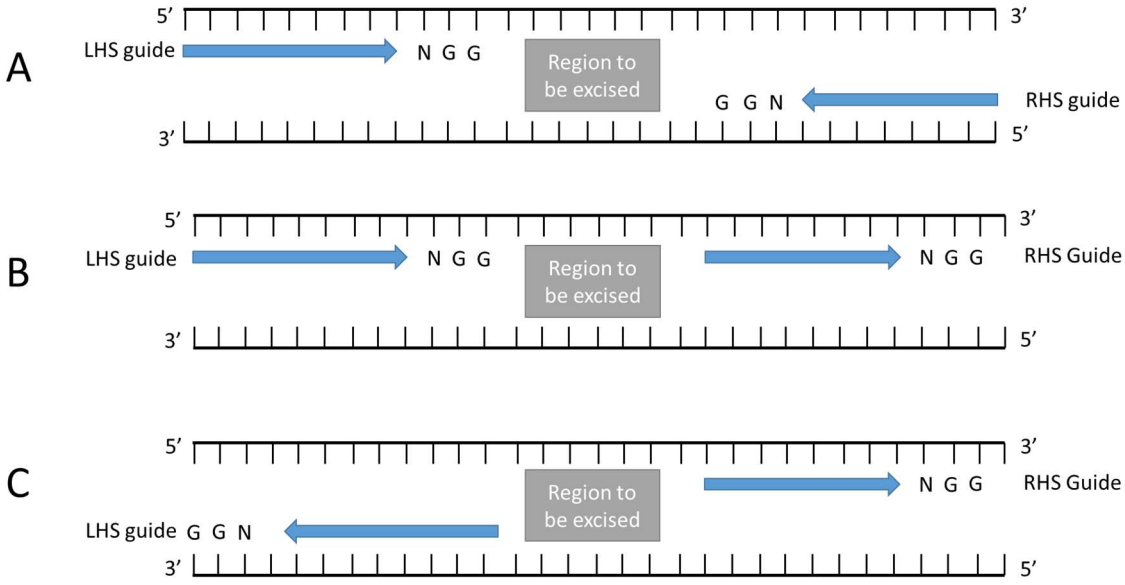


Figure 2.6 – Schematic depiction of the three possible orientations the guide RNAs (gRNA) can be positioned on both sides of the region to be excised. The NGG tri-nucleotide sequence is the protospacer adjacent motif (PAM) site required for the cas9 protein to create the double strand break. Due to the use of unmodified cas9 and repair mechanisms by NHEJ, the positions of guides and orientations did not impact the ability of NHEJ to correctly remove the target region.

In addition to the targeting gRNAs, two non-targeting (NT) gRNA sequences were used as controls throughout the CRISPR experiments within this thesis. These sequences were taken from validated literature sources and constitute sequences which should not bind to any specific genomic locus gRNA sequences and were as follows: NT1 – 5' ACGGAGGCTAAGCGTCGCAA, NT2 – 5' TACTAACGCCGCTCCTACAG. These sequences were checked bioinformatically using the BLAT genome alignment tool (UCSC - https://genome.ucsc.edu/cgi-bin/hgBlat) using the human reference genome build 38 (hg38) to confirm no binding sites existed for each NT guide. This provided a good mimic for the conditions most similar to those in the modified cell clones in an attempt to control for changes in cell characteristics as a result of transfection and Cas9 expression.

2.2.11.2 Cloning of guides into pSpCas9(BB)-2A-GFP

A golden gate approach for the cloning of the designed oligonucleotide sequences was chosen which allowed for simultaneous restriction and ligation of the necessary insert into the EF1α-pSpCas9(BB)-2A-GFP plasmid. The complementary strands of the modified oligonucleotides were first annealed together to create the insert sequence that would be cloned into the EF1α-pSpCas9(BB)-2A-GFP vector.

2.2.11.2.1 Annealing of oligonucleotides

6µl of each the sense and anti-sense single stranded oligonucleotides (100pmol/µl) were added to 83µl of nuclease free water and 5µl T4 DNA ligase buffer, heated to 95°C in a heat block for 5 minutes at which point the heat block was switched off and left to cool for 1 hour.

2.2.11.2.2 Golden gate cloning, transformation and screening of positive clones

Golden gate cloning was ideal for this application as it allowed a simple streamlined protocol when using two type II restriction enzymes sites such as the BbsI restriction enzyme used. Restriction of the EF1α-pSpCas9(BB)-2A-GFP plasmid with BbsI excised a short sequence which competes in the ligation process with the modified gRNA oligonucleotides. Ligation of the modified gRNA sequence destroys the BbsI sites meaning that the sequence cannot be re-excised. However, ligation of the original sequence re-generates the BbsI sites and can be cleaved out in the subsequent cycle of the golden gate reaction.

The golden gate reaction was setup as follows:

Component	Volume per reaction (20ul total)	Final concentration
Nuclease free water	12µL	-
EF1α-pSpCas9(BB)-2A-GFP (100ng/µl)	1.5µL	7.5ng/µl
Annealed gRNA Oligos (50ng/µl)	3µL	7.5ng/µl
T4 DNA ligase buffer (10X)	2µL	1X
BbsI – HF (20U/µL)	0.5µL	0.5U/µL
T4 DNA ligase (3U/µL)	1µL	0.15U/µL

The reaction mix was incubated in a thermocycler for 10 cycles of 37°C for 5 minutes and 16°C for 10 minutes followed by a final 37°C for 30 minutes and a heat inactivation of 80°C for 20 minutes.

For transformation, 2µl of reaction mix was added to 50µl of chemically competent DH5α *E. coli* and the transformation was carried out as detailed in **section 2.2.8.3**. Approximately five transformant colonies were picked and mini-prepped (**section**

2.2.8.4.1) per construct due to the very high efficiency of this cloning method. To screen for positive clones, the BbsI enzyme was utilised since the restriction site in the positive clones should not be present. Restriction digest with BbsI should identify positive clones as the ones which are not digested when run on an 1% agarose gel. Two positive clones for each gRNA/Cas9 construct were taken forward and sequenced externally at Source Bioscience using Sanger sequencing (**section 2.2.8.5**) using the U6 primer (5' GAGGGCCTATTTCCCATGATT) which sequences from the U6 promoter across the gRNA insert site. One correctly cloned construct for each guide was maxi-prepped (**section 2.2.8.4.2**) and taken forward for transfections into established cell lines.

2.2.11.3 Transfection of CRISPR plasmids and clonal cell line isolation

Transfection of the CRISPR plasmids containing the cloned gRNAs used the Turbofect protocol detailed in **section 2.2.9.4** in HEK293 cells. For optimisation, both the Hap1 and SH-SY5Y cells were also tested but did not provide sufficient efficiencies to continue with the clonal isolation of positive CRISPR modified clones. The HEK293 cell line was selected due to significant improvement in transfection and modification efficiencies (results of efficiency optimisation in **section 4.2.3**).

The following protocol is outlined in **figure 2.7** as a schematic workflow. Transfection of the pSpCas9(BB)-2A-GFP plasmid containing gRNA into HEK293 cells was performed using 3 wells of a 24-well plate with 100,000 cells per well (method details in **section 2.2.9.4**). 48 hours post-transfection, the cells were dissociated using trypsin to achieve a single cell suspension, combined in one tube and counted using a haemocytometer (**section 2.2.9.3**). The cells were diluted to 500 cells/ml with

appropriate media and added to 10cm dishes at a seeding density of 1000 cells per dish. The cells were cultured until single isolated colonies were visible without magnification. Individual colonies were picked using a 200µl pipette tip and transferred to 96-well plates and cultured until 70% confluency reached. Duplicate plates were made by splitting the clonal isolates using mechanical dissociation by pipetting. One plate was genotyped using PCR from crude lysates and the other kept for continued culture until positive clones were identified.

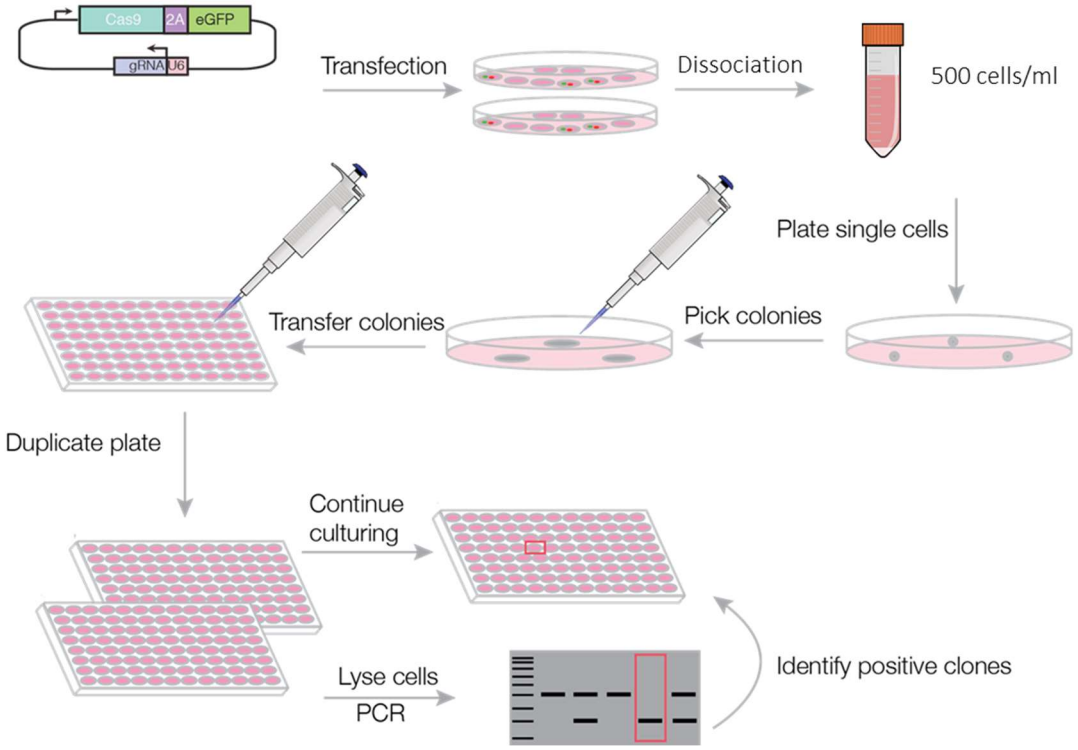


Figure 2.7 - Schematic workflow of the CRISPR clonal isolation process using the HEK293 cell line outlining the method of culturing and identifying positively modified clones. Cells transfected with the EF1α-pSpCas9(BB)-2A-GFP plasmid were dissociated to a single cell suspension. Cells were seeded at low density and cultured until single colonies were visible. Individual colonies were picked using a 200µl pipette tip and transferred to 96-well plates and cultured until 70% confluency reached. Duplicate plates were made so that one plate could be genotyped using PCR and the other kept for continued culture until positive clones were identified.

2.2.11.4 PCR screening of clonal CRISPR isolates

One of the duplicate plates for each target was used for positively modified screening of clonal isolates via crude lysis and PCR. Crude lysis solution was made up using DirectPCR lysis reagent (Cell) (VIAGEN Biotech) containing 1mg/ml proteinase K (Sigma). Cells to be lysed were washed twice with 100µl of pre-warmed PBS before 50µl of lysis solution was added to each well of the 96-well plate. The plates were sealed using ParaFilm M (Sigma) to avoid excessive evaporation and incubated overnight at 55°C in a rotating hybridisation oven (Hybaid) to lyse the cells and preserve the gDNA ready for PCR. The following day 1µl of the crude lysate was used directly in PCR reactions using KOD xtreme hot start polymerase (Novagen) which is suitable for non-purified DNA amplification from crude lysates and difficult to amplify regions. A typical reaction for amplification across the SVA knockout region was set up as follows:

Component	Volume per reaction (20ul total)	Final concentration
Nuclease free water	1.5µL	-
2X xtreme buffer	5µL	1X
dNTPs (2mM)	2µL	0.4mM
Forward Primer (20µM)	0.15µL	0.3µM
Reverse Primer (20µM)	0.15µL	0.3µM
KOD xtreme DNA polymerase (1U/µL)	0.2µL	0.02U/µL

Full details of the reactions used for the *INPP5F* SVA-, SVA-D and *LRRK2* SVA-C knockout PCRs can be found in the **appendix tables 1 and 2**.

2.2.12 qPCR of *INPP5F* isoform expression in ΔSVA KO HEK293 cells

Quantification of the expression of the *INPP5F* isoform 1 using qPCR was performed to assess the effects on expression in response to the knockout of the SVA-F and SVA-D elements. RNA was extracted (**section 2.2.4.2**) from the triplicate clonal isolate (**section 2.2.11.3**) un-transfected control cells (UN), non-targeting gRNA control cells (NT), *INPP5F* SVA-F KO mono and bi-allelic cells and *INPP5F* SVA-D KO mono and bi-allelic cells plated at 100,000 cells per well in 24-well plates under basal conditions (N=3 for each condition). cDNA was generated from the extracted RNA using first strand cDNA synthesis (**section 2.2.6**) using 70ng input RNA for each sample.

All qPCR reactions were performed using the GoTaq qPCR Master Mix (Promega) in accordance with manufacturer's guidelines. The master mix allows for consistent amplification by providing all necessary components including buffer, $MgCl_2$, dNTPs, polymerase enzyme and BRYT Green Dye (Promega's trademark version of SYBR Green) which allows for detection of double stranded DNA through excitation stimulated fluorescence. Carboxy-X-rhodamine (CXR) reference dye, which is identical to the commonly used ROX dye, was also added to as a passive reference dye that allows normalisation between individual samples to account for minor fluctuations in volumes between wells. All reactions were setup in triplicate to increase reliability and reduce the effect of pipetting error. For efficient qPCR, the target amplicon size was limited to <200bp with the average size ~150bp (details of qPCR amplicons and cycling conditions in **appendix table 1 and 2)**. Thermal cycling and qPCR measurements were measured using the Mx3005P Real-time PCR system

(Stratagene). To allow for relative quantification, the housekeeping gene *ACTB* was chosen due to the high primer efficiencies needed for reliable qPCR quantitation.

The qPCR reaction for *INPP5F* isoform 1 was performed using the following reaction setup in 96-well plates in triplicate:

Component	Volume per reaction (10ul total)	Final concentration
Nuclease free water	2.5µL	-
GoTaq qPCR Master Mix (2X)	5µL	1X
Fw (5') *INPP5F* iso 1 primer (10µM)	0.2µL	0.2µM
Rv (3') *INPP5F* iso 1 primer (10µM)	0.2µL	0.2µM
CXR reference dye (30µM)	0.1µL	0.15µM
Template cDNA (1:10 dilute)	2µL	-

The qPCR machine was setup using the following thermocycling conditions: 95°C for 2 minutes, (open cycle), 95°C for 30 seconds, 60°C for 30 seconds, 72°C for 20 seconds, (close cycle) with a max cycle limit of 40. The qPCR machine was setup to detect both SYBR green fluorescence (same wavelength as BRYT green) and ROX (same wavelength as CXR dye) to allow passive reference normalisation.

2.2.12.1 Relative quantification of expression using the $2^{-\Delta\Delta CT}$ method

Ct (cycle threshold) values define the cycle at which the sample signal exceeds background noise within the assay. These values are indicative of the mRNA abundance for each target and so can be used to calculate relative expression of a target compared to a known housekeeping gene that should be stable across all conditions. The housekeeping gene for each assay must be determined as suitable or not prior to the qPCR assay. *ACTB* was chosen as the housekeeping gene for the quantification of relative expression of *INPP5F* isoform 1 as it did not alter

significantly across the SVA knockout conditions for *INPP5F* SVA-F and SVA-D when assessed using RT-PCR and agarose gel electrophoresis (**figure 4.12**). To calculate the $2^{-\Delta\Delta CT}$ values for each condition, the non-targeting gRNA transfected cells (denoted as NT below) were specified as the appropriate control to which relative expression could be generated. Using Microsoft Excel with the raw Ct values, $2^{-\Delta\Delta CT}$ were calculated as follows:

1. Calculate the mean Ct values across the triplicate values within each condition (*INPP5F* isoform 1 and *ACTB*)
2. ΔCt = Mean Ct (*ACTB*) – mean Ct (isoform 1) for each condition
3. Calculate the mean of the ΔCt for NT control group (biological control)
4. ΔΔCt = Mean ΔCt of each sample – mean ΔCt (NT)
5. Convert the ΔΔCt into fold change using $2^{(-\Delta\Delta Ct)}$

2.2.12.2 qPCR primer efficiency calculation and dissociation curves

For reliable results using the $2^{-\Delta\Delta CT}$ method high primer efficiencies within the PCR reactions had to be ensured. A primer efficiency of 100% would infer that the number of copies within the PCR doubles perfectly after every cycle. Ideally, qPCR primer efficiencies should be between 90-100% efficient, with primers between 80-90% being acceptable if corresponding primers being compared where of similar efficiency (within 5-10%). To test this, four qPCRs were setup in triplicate across a ten-fold serial dilution of cDNA including 10^{-1}, 10^{-2}, 10^{-3} and 10^{-4}. By plotting the mean Ct values across the dilution range, the standard curve for each target (*ACTB* and *INPP5F* isoform 1) could be generated. The gradient of the curve can then be used to calculate the primer efficiency using the following equation:

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

To ensure the qPCR is only generating one amplicon, dissociation curves for each target were generated using the built-in program within the qPCR suite. Incrementally increasing the temperature post-PCR to the point of total denaturation of the amplicon allows for measurement of fluorescence decay across specific temperatures in real-time. If the PCR is specific, only one product would be amplified shown as a single peak on the dissociation curve. More than one peak present on the dissociation curve was indicative of non-specific PCR amplification, which would require the primers to be re-designed and tested.

2.2.12.3 Calculating significance using one-way analysis of variance (ANOVA) for *INPP5F* isoform 1 in SVA KO HEK293 cells

The one-way ANOVA statistical test is appropriate to evaluate differences between $2^{-(\Delta\Delta CT)}$ values population means when comparing 6 groups with equal sample sizes within each group. Sum of squares between groups and the sum of squares within groups were calculated using Microsoft Excel. The F ratio was calculated using 5 degrees of freedom (DF) for the sum of squares between groups and 12 DF for the sum of squares within groups (F(5,12)). The critical value needed for significance using these degrees of freedom was 3.11.

2.2.13 Retrotransposon Capture Sequencing (RC-Seq)

The RC-Seq protocol was used as stated in Sanchez-Luque F. J. *et al.* (2016) Retrotransposon Capture Sequencing (RC-Seq): A targeted high-throughput approach to resolve somatic L1 retrotransposition in humans [140]. This work was performed in the Paul-Ehrlich-Institut (PEI) in Langen, Germany in collaboration with Professor Gerald Schumann. The samples used in the analysis (details in section 2.1.5) included cerebellum and frontal cortex derived genomic DNA obtained for 5 PD patients and 1 unaffected control individual giving 12 samples in total that were processed for RC-Seq library preparation. The DNA was extracted and purified in University College London (UCL) with QC data. Specific samples used from UCL included: PD109, PD139, PD184, PD348, PD359 and PDC01 (control). Secondary QC analysis was performed upon receipt of the DNA using the Nanodrop 8000 spectrophotometer (Thermo Fisher Scientific) and Qubit Fluorometer (Thermo Fisher Scientific) using the dsDNA broad range (BR) assay kit. A brief outline of the RC-Seq workflow in the following sections and summarised in **figure 2.8** and includes preparation of Illumina adapter-based DNA libraries, enrichment of LINE-1 elements using LNA-Biotin probes, Illumina sequencing and bioinformatic analysis using the TEBreak pipeline. Sequencing of prepared RC-Seq libraries was performed in the Pfizer/University of Granada and Andalusian Regional Government Center for Genomics and Oncology (GENYO) in Granada, Spain using the Illumina NextSeq 500 platform.

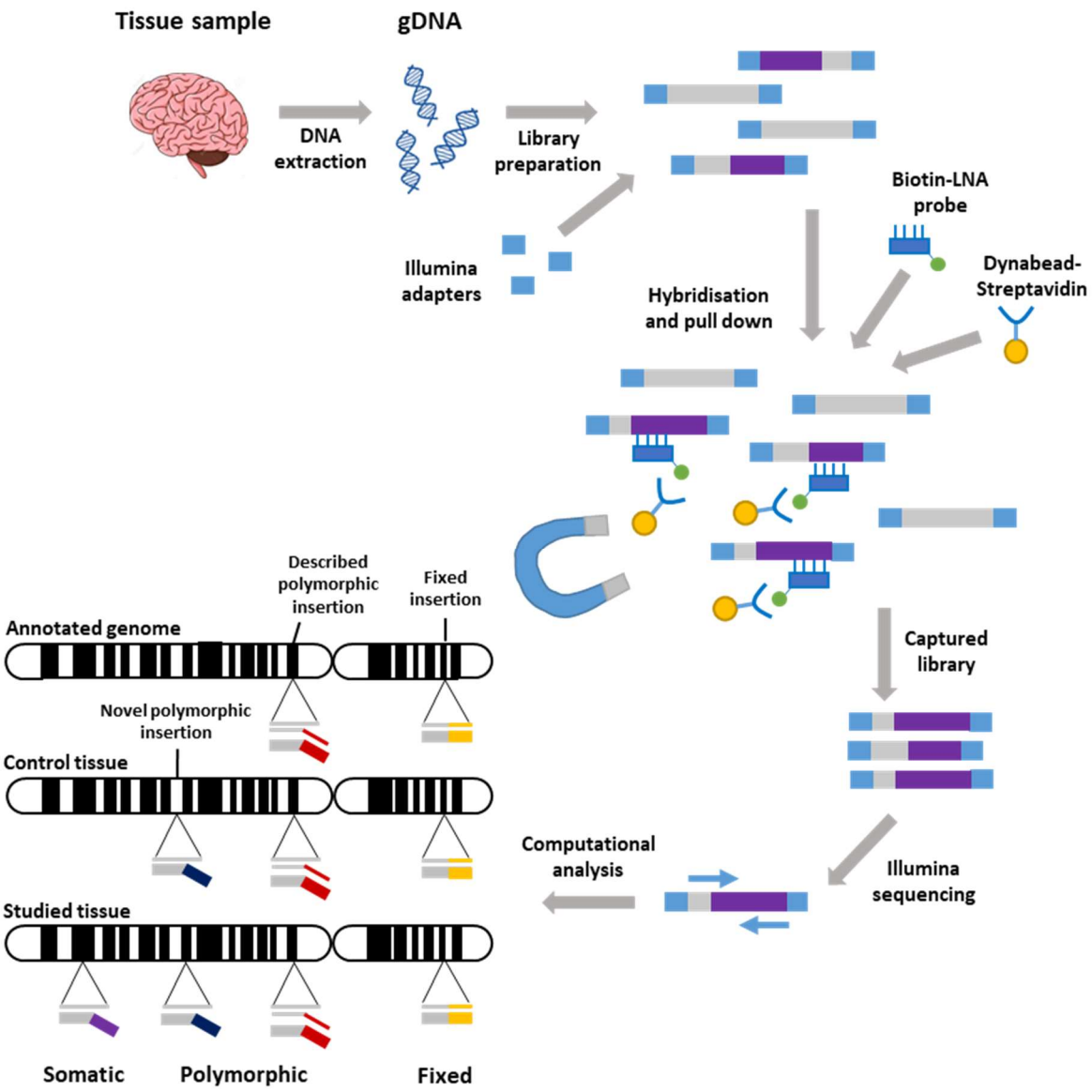


Figure 2.8 - RC-Seq workflow outline adapted from Sanchez-Luque *et al,* 2016 [140]. Illustrates the main steps in library preparation and sequencing to detect novel germline and somatic LINE-1 insertions.

2.2.13.1 DNA shearing by focused ultrasonication

Using the concentrations determined in the Qubit BR assay, 5µg genomic DNA for each tissue (cerebellum and frontal cortex) was diluted in 130µL of TE buffer and sheared using a Covaris M220 focused-ultrasonicator using Sonolab 7 software. DNA was sheared to a target length of 250bp using a peak power of 50, duty factor 20 and 200 cycles per burst for 2 minutes per sample.

2.2.13.2 Illumina LINE-1 library preparation

DNA post-sonication was concentrated using Agencourt AMPure XP beads (Beckman Coulter) at a 1.1:1 ratio of beads to DNA and eluted in 52µL of elution buffer (provided in kit). 1µL of eluate was used for quantification of DNA concentration using the Qubit dsDNA BR assay and 2 µL were used on a Fragment Analyser (Advanced Analytical) for size distribution QC. 1µg of concentrated DNA was used for library preparation using the TruSeq Nano DNA sample prep kit. The sheared DNA was treated with end repair mix at 30°C for 30 minutes, followed by purification using AMPure XP beads. The libraries were then poly-adenylated using an A-tailing mix at 70°C for 5 minutes followed by ligation of Illumina adapters at 30°C for 10 minutes followed by termination of the reaction using stop ligase mix. Two further rounds of purification with AMPure XP beads were performed before agarose size selection.

2.2.13.3 Agarose gel-size selection

A 2% high-resolution agarose (Sigma) gel was prepared using 1X TAE buffer (40mM Tris-HCL, 20mM acetic acid and 1mM EDTA) with the full volume of library preparation being loaded on the gel alongside 1Kb Plus DNA ladder (Thermo Scientific) for sizing. Gel electrophoresis conditions were 120mA for 2.5-3 hours until the separation of DNA was sufficient to comfortably excise gel slices containing fragments differing by ~30-50nt within the 250-400bp range. Four gel slices per lane were excised to include 290-310bp, 310-350bp, 350-380bp and 380-410bp fragment length ranges and placed in separate 1.5ml Eppendorf tubes. The DNA was purified from the gel slices using the MiniElute gel extraction kit (QIAGEN) using manufacturer's guidelines and eluted in 32µL of elution buffer. These eluates were

used in ligation mediated (LM)-PCR using the following conditions: 98°C for 45 secs (open cycle), 98°C for 15 secs, 60°C for 30 secs, 72°C for 30 secs (close cycle), for 6 cycles and a final elongation of 72°C for 5mins with a 4°C infinite hold. For the reaction, 30µL of the eluted DNA from the gel clean-up was added to 50µL of Phusion High Fidelity PCR master mix (2X) (NEB), 18µL of ultrapure water, 1µL of each TS-F primer (5' AATGATACGGCGACCACCGAGA 3' - 100µM) and TS-R primer (5' CAAGCAGAAGACGGCATACGAG 3' - 100µM), mixed and run using the LM-PCR protocol. Resulting DNA/PCR reaction was purified using AMPure XP beads using 32µL of ultrapure water for elution (not resuspension buffer) with 1µL being used for Qubit concentration quantification. A further 2µL was used for sizing the purified fragments using the Fragment Analyser to identify samples with a fragment range of 340-410bp and a median length of 370bp to proceed with.

2.2.13.4 Hybridisation of LNA probes for LINE-1 capture

500ng of both cerebellum and frontal cortex from each individual were pooled together in a 1:1 ratio to be used for hybridisation (1µg total). 10µL of sequence capture developer reagent and 10µL of universal blocking oligo, which binds to the adapters to reduce off-target capture during L1 enrichment, were added per 1µg of pooled DNA. 10µL of index-specific blocker oligo was added per 1µg of library which identifies each pooled library with a unique sequence tag that is defined by the specific adapter added and allows for multiplex sequencing. Samples were air dried using an Eppendorf speed-vac at 70°C for 60 minutes until the samples were completely dry. 7.5µL of 2X hybridisation buffer and 3µL of hybridisation component-A were added to each dried sample and heated at 95°C for 5 minutes. The entire

volume (10.5µL) was transferred to a second tube containing 4.5µL of pre-aliquoted biotin labelled locked nucleic acid (LNA) probes and heated for 3 minutes at 95°C before being transferred to a thermocycler set at 47°C for three days to hybridise. The two LNA probes used bind to the extreme termini (5' and 3' ends) of L1 elements and were as follows: 5' L1 LNA – 5'-Bio/CTCCGGT+C+T+ACAGCTC+C+C+AGC and the 3' L1 LNA – 5'-Bio/AG+A+TGAC+A+C+ATTAGTGGGTGC+A+GCG where 'Bio' denotes the presence of 5' biotin moieties and the '+' represents the locked nucleic acid positions within each probe. They were used because locked nucleic acids do not allow conformational changes in the oligo structure due to an additional bridge connecting the 2' oxygen and 4'carbon atoms within the ribose of oligonucleotides thus increasing hybridisation efficiency to DNA or RNA [141].

2.2.13.5 Capture recovery and amplification

To capture the biotin LNA-hybridised LINE-1 -enriched library, Dynabeads M-270 Streptavidin (Invitrogen - Thermo Fisher) were used. 200µL of Dynabeads of each pair of 5' and 3' captures were prepared with a series of washes using the Roche NimbleGen capture wash kit before being added to the captured pooled library and remaining at 47°C in the thermocycler for 45 minutes with several pipette mixes. Further washes using the Roche NimbleGen capture kit were performed on the captured libraries whilst maintaining 47°C. Each wash step was performed with buffers heated to 47°C and a pre-warmed magnetic rack was used when aspirating each previous wash buffer. At the last wash step the final resuspension was in 50µL of ultrapure water ready for amplification. 100µL of Phusion High Fidelity PCR MasterMix (2X), 46µL of ultrapure water, 2µL of both the TS-F and TS-R (100µM each)

was added directly to the Dynabeads bound to libraries and placed in a thermocycler. LM-PCR was performed with the following conditions: 98°C for 45 secs (open cycle), 98°C for 15 secs, 60°C for30 secs, 72°C for 30 secs (close cycle), for 8 cycles with a final extension of 72°C for 5 mins and a 4°C hold. Post-PCR the amplicon was purified using the MiniElute gel extraction kit according to manufacturer's instructions. 1µL of each library was quantified using the Qubit dsDNA high sensitivity (HS) assay with a target concentration of >1ng/µL for the 3' capture and >7ng/µL needed. If insufficient DNA was present, a second round of LM-PCR was performed with three cycles followed by an AMPure XP bead clean up and quantified again using the Qubit dsDNA HS assay. The size distribution of the PCR amplicon was checked using 2µL of each 5' and 3' capture using the Fragment Analyser. The 5' and 3' capture libraries for each sample were pooled in a 3:7 ratio respectively by molecular mass.

2.2.13.6 Sequencing of RC-Seq libraries

The 12 individual LINE-1 libraries containing captured 5' and 3' ends were pooled in a single tube and multi-plex sequenced on the Illumina NextSeq 500 platform in accordance with the Sanchez-Luque *et al*. protocol. From this, 8 FastQ files per individual (4 forward (R1) and four reverse (R2)) were received and analysed using the TEBreak pipeline.

2.2.14 Whole Genome Sequencing

Whole genome sequencing (WGS) of four brain tissue samples from two individuals (frontal cortex and cerebellum from samples PD109 and PD348) used in the RC-Seq protocols was carried out externally by the Australian Genome Research Facility (AGRF). For comparison, nine healthy aged individuals from the Dyne Steele cohort

(Manchester brain bank) (cohort details in **section 2.1.5**) were also sequenced with temporal cortex and blood samples being sequenced from each individual (18 samples total). Additional sample details can be found in **table 5.1**. One microgram (1µg) of DNA extracted from each tissue was provided in 96-well plate format and sequenced at a depth of 40X on the IlluminaSeq platform. The WGS libraries were sequenced using the IlluminaSeq platform which provided 2 FastQ files per tissue (1 forward (R1) and 1 reverse (R2)) to be analysed using the same TEBreak pipeline as the RC-Seq workflow.

2.2.15 Bioinformatic analysis of next generation sequencing (RC-Seq and WGS)

Downloaded FastQ files were stored locally and uploaded to the Linux based Chadwick server at the University of Liverpool from where all bioinformatic analysis would be performed. To run scripts and manage files on the server, PuTTy terminal and the file directory manager FileZilla were used. The scratch dedicated partition designated for running scripts was used to store and analyse the FastQ files through each step of the analysis pipeline. The TEBreak analysis package created by Adam Ewing was used to interrogate insertion mutations caused by retrotransposons (LINE-1, *Alu* and SVA). This package is able to identify novel insertions in both capture-seq and WGS datasets and provides information regarding positions, tissues the insertion is located, orientations, target site duplications and literature comparisons to ascertain if the insertion has been found before amongst other datasets. The Chadwick server used at the University of Liverpool had a pre-installed TEBreak package that was used to characterise retroelements within the sequencing.

Throughout the following protocol, several common commands and file types were used including:

- **cat** command is the general Linux command for concatenation of files

- **module load** allows a given module e.g. TEBreak to be loaded into the working directory (scratch)

- **qsub** command which submits a specified script to the server tom be run

- **.sh** files denote the scripts that have been written on notepad++

- **.fastq** files are the raw sequencing files that have not been processed

2.2.15.1 Quality control of FastQ files using FastQC software

The 4 FastQ files and 2 FastQ files per tissue from RC-Seq and WGS respectively were first concatenated to proceed with any analysis including quality control (QC). To concatenate FastQ files, the cat command in PuTTy was used by entering the following:

- cat *R1.FastQ.gz > sample.FastQ.gz
- cat *R2.FastQ.gz > sample.FastQ.gz

The wildcard symbol (*) specifies that all files within the directory ending with the suffix after the * symbol will be concatenated.

The free downloadable software FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to analyse the FastQ files for quality. This provides information and interpretation multiple properties including sequence base calling quality, GC content, duplicated

sequences, average read length, overrepresented sequences, and presence of adapters.

2.2.15.2 RC-Seq bioinformatic analysis pipeline

To proceed with the analysis of the LINE-1 elements within the RC-Seq library sequencing files, the FastQ files were first trimmed to remove adapter sequences using the following command:

- qsub trimmomatic_rcseq.sh

The trimmed FastQ files were subjected to a second round of FastQC analysis to ensure no adapter sequences were still present within the reads. The trimmed FastQ files could then be aligned to the reference genome (Hg19) to produce a useable Bam file.

- qsub alignment_rcseq.sh

The output aligned Bam file was then analysed using TEBreak which was available on the Chadwick server. To load the TEBreak module into the scratch workspace the 'module load TEBreak' command was used followed by the following script:

- qsub rcseq_TEBreaknew_2sample.sh

The output file from TEBreak is a PICKLE file which is screened to identify LINE-1 insertions against given consensus sequences for LINE-1 elements using the following command:

- qsub rcseq_picklescreen.sh

Due to the size of the PICKLE file, the resolve step is prone to fail because of running over the maximum allotted run time of 72 hours. In this instance, the PICKLE files would be split into smaller chunks which can be run quicker. To do this, the split pickle script was run using the following command:

- qsub splitpickle.sh

The PICKLE files were then resolved which characterises the screened LINE-1 insertions within the library. Two different scripts were used depending on whether the PICKLE file had been split or not. PICKLE files that had not been split were resolved using the following command:

- qsub rcseq_reslove_new.sh

Split PICKLE files were resolved using:

- qsub rcseq_resolvesplit_new.sh

The output file from the resolve contains all LINE-1 insertions within the sequenced library including background noise and required filtering to identify insertions of interest. To target polymorphic germline insertions the following filters were applied which removes all insertions that do not meet the required characteristics which are: a minimum of 8 split reads per insertion, minimum consensus length of 150bp, a minimum consensus LINE-1 sequence match of 0.9 (90%), a minimum reference genome match of 0.95 (95%) and a maximum number of variants (SNPs) of 2 per insertion. Germline insertions were filtered using the following:

- qsub genfilter_rcseq8_150_0.9_0.95_new.sh

Somatic LINE-1 insertions followed the same properties for the parsing of reads to identify LINE-1 elements as those used for targeting polymorphic germline insertions, except for a lowered split read threshold from 8 to 4. This produced a lower stringency threshold to detect LINE-1 elements which may only be found in smaller numbers of cell populations within the tissue thereby improving the chances of detecting true somatic insertions. The lower threshold also increases the false positive hit rate and had to be considered when choosing the threshold to use. The somatic filtering script was run using the following command:

- qsub genfilter_rcseq4_150_0.9_0.95_new.sh

Details regarding specific scripts used to run the RC-Seq bioinformatic pipeline can be found in appendix 4.

2.2.15.3 Whole genome sequencing (WGS) bioinformatic analysis pipeline

The WGS pipeline used for the identification of novel retroelement insertions followed an almost identical method to that used for RC-Seq for comparable data to be produced in the output. However, where the RC-Seq pipeline detects LINE-1 elements in a more targeted approach to detect somatic insertions, the WGS pipeline identifies LINE-1, *Alu* and SVA elements in addition but may not be as sensitive to low copy number somatic insertions. The additional detection of both *Alu* and SVA as well meant that altered scripts to accommodate these were used. The FastQ files that had been concatenated and quality control checked using FastQC were first trimmed using the following command:

- qsub trimmomatic38_WGS.sh

The trimmed FastQ files were aligned to the reference genome (Hg19) to produce Bam files using the following command:

- qsub alignment_WGS.sh

the Bam files were then analysed using TEBreak by loading the package into the working scratch directory and running the TEBreak scripts:

- module load TEBreak
- qsub TEBreak_new_WGS.sh

The outputted PICKLE files were then screened prior to resolving which identifies LINE-1, Alu and SVA elements using given consensus sequences for these elements using the following command:

- qsub screenpickle_WGS.sh

The screened PICKLE files were then resolved or split depending on file size similar to the RC-Seq pipeline (**section 2.15.2**). To split the PICKLE files the following command was used:

- qsub splitpickle_WGS.sh

Using two scripts to resolve split or unsplit PICKLE files, annotation of the identified LINE-1, *Alu* and SVA insertions was performed using the commands:

For non-split PICKLE files:

- qsub resolve_WGS.sh

For split PICKLE files:

- qsub resolvesplit_WGS.sh

The output from the resolve step is a raw table containing all annotated insertions which requires filtering similar to that used for the RC-Seq pipeline. To target the novel germline insertions the following filters were applied which removes all insertions that do not meet the required characteristics which are: a minimum of 4 split reads per insertion, minimum consensus length of 150bp, a minimum consensus retroelement sequence match of 0.9 (90%) and a minimum reference genome match of 0.95 (95%). Filtering was performed using the following command:

- qsub generalfilter_WGS_new.sh

Details regarding specific scripts used to run the WGS bioinformatic pipeline can be found in appendix 4.

Chapter 3 – Investigating an SVA retrotransposon within *LRRK2* as a novel regulator of genetic function

Chapter 3 - Investigating an SVA retrotransposon within *LRRK2* as a novel regulator of genetic function

3.1 Introduction

Understanding the regulatory landscape surrounding *LRRK2* is important for further characterising the pathogenicity of *LRRK2* and how, not only the coding mutations, but also differences in *LRRK2* expression between individuals could lead to altered risk for the development of Parkinson's disease. There have been multiple pathogenic mechanisms suggested for *LRRK2* across a variety of studies including α-synuclein and tau aggregation models, neuroinflammatory responses, oxidative stress, mitochondrial dysfunction, synaptic dysfunction and autophagy (details for the coding mutations of *LRRK2* in **figure 1.2**) [142-144]. However, there is still a lack of effective treatments to combat the increased *LRRK2* activity observed in both familial and idiopathic PD patients which highlights the need for a better understanding of the regulatory mechanisms of *LRRK2* expression on a genetic level.

Brief analysis of the *LRRK2* locus using the UCSC genome browser (https://genome.ucsc.edu/) identified an SVA retrotransposable element of interest for further investigation within this chapter. There are examples of SVA retrotransposon insertion polymorphisms being implicated in disease via disruption of normal gene processes such as changes in gene expression or splicing [93]. A list of eight known pathogenic SVA insertions in which an SVA insertion is causally linked to the onset of disease are presented in **table 1.1**, **section 1.4.2** which demonstrate the roles SVA retrotransposons can play in disease. However, these examples relate to pathogenic novel insertions of SVAs that disrupt the normal processing of the

genes into which they insert and do not demonstrate the potential functions of 'fixed' SVA elements as sources of genetic regulation for the normal and potentially pathogenic effects on local genes.

Fixed SVA retroelements have been previously demonstrated to possess regulatory properties in neurodegenerative disease contexts and as such, were of interest when evaluating the *LRRK2* locus. Examples of fixed SVA retrotransposons from the literature which have demonstrated the effects these elements can have on nearby genes including influencing changes in gene expression or tissue specificity include the *PARK7*/*DJ-1* and *FUS* loci which have importance within neurodegenerative diseases (details in **section 1.4.2**).

It is therefore important to understand not only the effects of mutations within the coding sequence of *LRRK2*, but also potential novel modes of regulation from SVA retrotransposons within *LRRK2* and potentially other key PD related genomic loci. Disruption of expression of *LRRK2* that could lead to decreased abundance of LRRK2 protein could be equally as important as the known coding mutations within *LRRK2* that lead to enzymatic inactivity of the LRRK2 protein. Given this, it is important to understand the novel regulatory landscape, furthermore potential polymorphic domains within the SVA itself that could alter the transcriptomic profile of *LRRK2*. Both the previously described *PARK7* and *FUS* examples provide evidence that SVA retrotransposons can possess regulatory properties within the central nervous system, and should be considered more carefully when evaluating both novel and characterised loci involved in neurodegenerative disorders such as PD. This chapter focuses on characterisation of an SVA element within one of the most important loci

for PD, the *LRRK2* locus, and explores the potential novel regulatory properties the SVA may possess. To do this, multiple techniques were utilised to study SVA function including bioinformatic analysis with linkage disequilibrium analysis, PCR to discover potential primary sequence polymorphisms and genotyping of a PD case/control cohort, functional inferences using different reporter gene assays to test both expression and splicing effects as well as RT-PCR of *LRRK2* transcripts in CRISPR mediated SVA knockout (KO) cell lines in response to SVA KO. Understanding how SVA elements function in a wider range of contexts is crucial for understanding the effects of novel regulatory domains within complex diseases such as Parkinson's disease and provides a major focus of the work presented in this thesis.

3.1.1 Aims and hypothesis

- To characterise potential primary sequence polymorphisms in the *LRRK2* SVA and evaluate any such polymorphism for heightened risk of PD using genotyping of matched case and control cohorts.
- To test the potential function of the *LRRK2* SVA-C element using luciferase-based reporter gene cell assays and validate any potential effect further using CRISPR technologies to produce SVA-C knockout cell lines and quantifying any changes in *LRRK2* gene expression using RT-PCR and qPCR.

Hypothesis - The SVA-C retrotransposable element within the *LRRK2* locus has the potential to alter the genetic function of *LRRK2* via multiple facets including changes to gene expression and splicing. These changes in gene function may be further influenced by primary sequence polymorphisms that confer risk for the development of Parkinson's disease.

3.2 Results

3.2.1 Bioinformatic analysis of the *LRRK2* locus

Use of the UCSC genome browser (hg38) allowed scrutinisation of the *LRRK2* locus to look for novel potential regulatory domains arising from retrotransposable elements. The presence of an SVA-C element was noted within intron 44 of the full length *LRRK2* isoform 1 (**figure 3.1)**. Given previous literature has highlighted SVA retroelements in the aetiology of disease and provided evidence for roles as regulatory elements that can affect gene function including changes in gene expression and splicing, it was important to explore this SVA further [73, 76].

The subclass of this element (SVA-C) indicated that this insertion occurred approximately 10.88 million years ago meaning that the element is present in the homininae species only (humans, chimpanzees and gorillas) [52]. This is illustrated in **figure 3.1** on the Multiz alignment conservation track, where the SVA-C breaks the homology within all species more distantly related than the gorilla. Highly homologous regions are shown as black blocks whilst gaps refer to sequences which were not conserved between the base genome (human) and the target species.

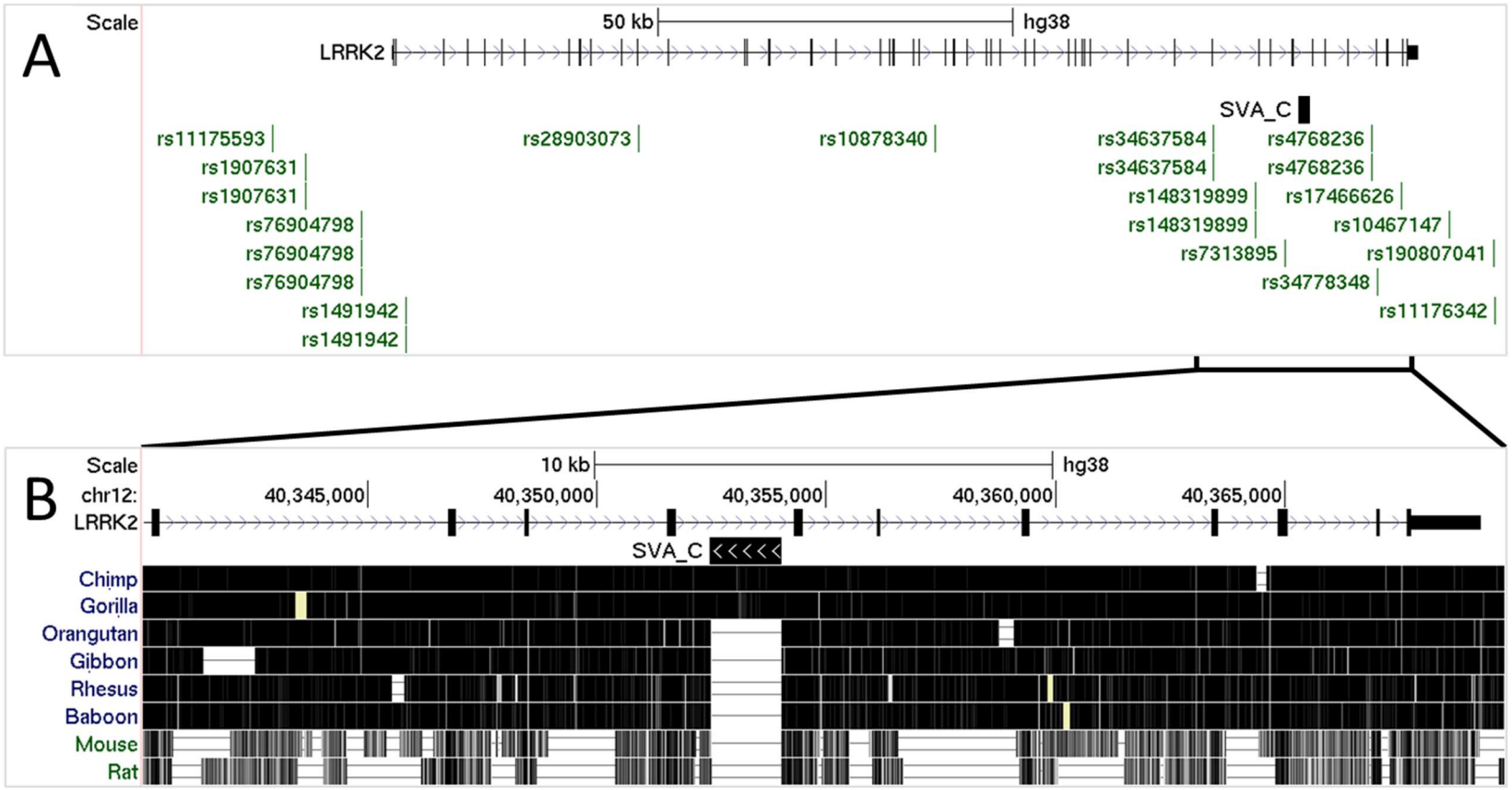


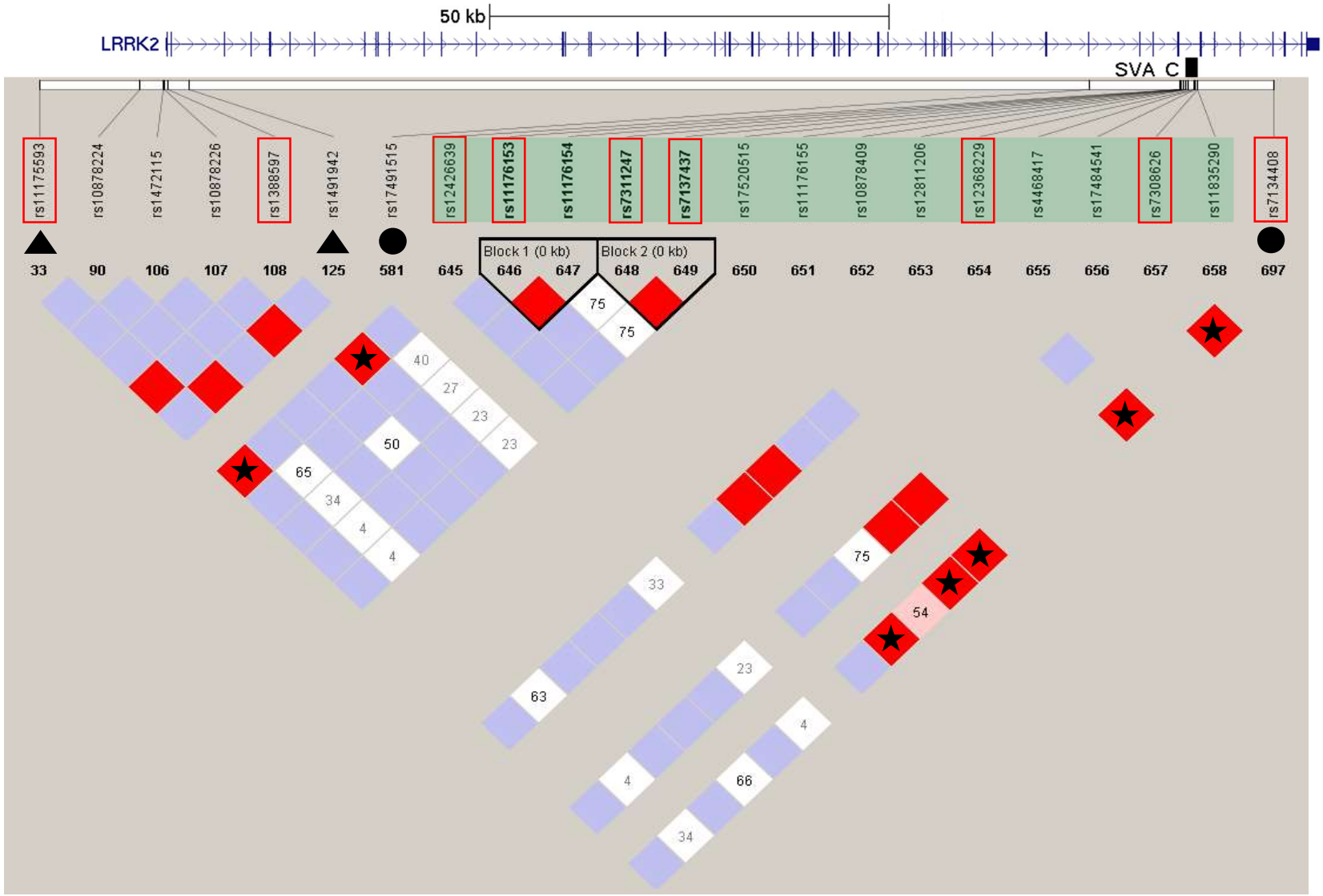
Figure 3.1 – UCSC genome browser bioinformatic analysis of the *LRRK2* locus (hg38). **(A)** The full length *LRRK2* isoform shown from the Gencode V28 database within the gene track which encoded an mRNA transcript of 9,158 bp. An intronic SVA C was noted between exons 44 and 45 with two flanking PD related GWAS SNPs (rs34637584 and rs34778348). **(B)** The SVA in *LRRK2* is homininae specific, only being conserved within humans, chimps and gorillas. The arrows on the gene track and SVA indicate direction of transcription and SVA orientation respectively illustrating the SVA is present in an anti-sense orientation with respect to *LRRK2* transcription. Green SNPs indicate SNPs from GWAS studies from the NHGRI-EBI GWAS catalogue track on UCSC Hg38 with repeated SNPs indicating multiple studies reporting the same SNP.

GWAS SNPs in the *LRRK2* locus highlighted in the GWAS track (provided from the National Human Genome Research Institute and the European Bioinformatics Institute (NHGRI-EBI - https://www.ebi.ac.uk/gwas/docs/about) in **figure 3.1** indicated a range of diseases and traits [146]. The range of GWAS SNPs across the *LRRK2* locus displayed in **figure 3.1** are listed in **table 3.1** and show the associated diseases/traits, reported genes, p-values and the reporting study.

Six of the 15 total GWAS SNPs in the locus were associated with Parkinson's disease (rs76904798, rs1491942, rs28903073, rs34637584, rs34778348 and rs190807041) and have been reported by four independent studies (Nalls MA *et al.* 2019, Lill CM *et al.* 2012, Pickrell JK *et al.* 2016 and Bandres-Ciga S *et al.* 2019) further increasing the confidence that this locus had strong genetic associations with PD [26, 147-149]. Interestingly, other reported diseases included inflammatory disorders such as Crohn's disease, inflammatory bowel disease (IBS) and chronic inflammation which have been previously linked to PD [150, 151].

Table 3.1 – A list of GWAS SNPs from the GWAS catalog (updated - 26/02/2020) in the *LRRK2* locus highlighted in **figure 4.1**. The most highly statistically significant GWAS SNPs are attributed to Parkinson's disease, indicating a strong genetic association for PD across this locus.

GWAS SNP	Disease or trait	Reported genes	Meta p-value	Study
rs11175593	Crohn's disease	*LRRK2*, MUC19	3E-10	Barrett JC et al. 2008
rs1907631	Total cholesterol levels	Not reported	9E-9	Hoffmann TJ et al. 2018
rs76904798	Parkinson's disease	*LRRK2*	2E-28	Nalls MA et al. 2019
rs1491942	Parkinson's disease	*LRRK2*	6E-15	Lill CM et al. 2012
rs28903073	Parkinson's disease	*LRRK2*	1E-39	Pickrell JK et al. 2016
rs10878340	Daytime sleep phenotypes	*LRRK2*, MUC19, SLC2A13, RPL30P13	8E-7	Spada J et al. 2016
rs34637584	Parkinson's disease	*LRRK2*	4E-148	Nalls MA et al. 2019
rs148319899	Inflammatory bowel disease	SLC2A13, *LRRK2*, MUC19	3E-15	de Lange KM et al. 2017
rs7313895	Chronic inflammatory diseases	*LRRK2*	4E-17	Ellinghaus D et al. 2016
rs4768236	Inflammatory bowel disease	Not reported	2E-15	Liu JZ et al. 2015
rs34778348	Parkinson's disease	*LRRK2*	3E-21	Lill CM et al. 2012
rs17466626	Paediatric autoimmune diseases	*LRRK2*	3E-10	Li YR et al. 2015
rs10467147	Obesity-related traits	*LRRK2*	5E-6	Comuzzie AG et al. 2012
rs190807041	Parkinson's disease	*LRRK2*	7E-11	Bandres-Ciga S et al. 2019
rs11176342	Subjective well-being	Not reported	5E-6	Okbay A et al. 2016


50 kb
LRRK2
SVA C
rs11175593
rs10878224
rs1472115
rs10878226
rs1388597
rs1491942
rs17491515
rs12426639
rs11176153
rs11176154
rs7311247
rs7137437
rs17520515
rs11176155
rs10878409
rs12811206
rs12368229
rs4468417
rs17484541
rs7308626
rs11835290
rs7134408
33
90
106
107
108
125
581
645
Block 1 (0 kb)
646
647
Block 2 (0 kb)
648
649
650
651
652
653
654
655
656
657
658
697
75
75
40
27
23
23
50
65
34
4
4
33
63
75
54
23
4
4
66
34

GWAS SNP	Closest HapMap SNP	Distance to GWAS SNP	Disease or trait	Meta p-value
rs11175593	rs11175593	-	Crohn's disease	3E-10
rs76904798	rs10878224	168bp	Parkinson's disease	2E-28
rs1491942	rs1491942	-	Parkinson's disease	6E-15
rs34637584	rs17491515	93bp	Parkinson's disease	4E-148
rs34778348	rs7134408	206bp	Parkinson's disease	3E-21

Figure 3.2 - Linkage disequilibrium (LD) plot of HapMap and GWAS single nucleotide polymorphisms (SNPs) within the SVA-C (+/- 2kb) and promoter loci of *LRRK2*. The plot represents those SNPs that are 'linked' and therefore indicates a genetic association. **Black triangles** indicate disease related GWAS SNPs whilst **black circles** indicate HapMap SNPs that are located closest to a disease related GWAS SNPs and are used as 'proxy' if the GWAS SNP was not available in the HapMap database. Gaps between SNPs show a lack of data available for the associated SNPs. Details of the GWAS and closest HapMap SNPs are described in the table above. Solid red boxes are indicative of perfect LD (LD = 1) between SNPs, violet boxes show no LD (LD = 0) and numbered boxes show a degree of LD with the number representing the percentage of association. Black stars indicate key SNP associations of interest that exhibit high LD between the SNPs related to the SVA-C and the *LRRK2* promoter region or a GWAS related SNP. Missing squares represent the linkage data that is unavailable. Red outlines around SNP codes indicate those SNPs of interest which show high linkage with another SNP. The green highlight across multiple SNPs indicates those SNPs located within or flanking the SVA-C element (+/- 2kb).

Linkage disequilibrium (LD) analysis using the Haploview software package (Broad institute - https://www.broadinstitute.org/haploview/haploview) allowed the viewing of genetic associations between the individual SNPs within the defined set of SNPs from the HapMap project [152]. In the resulting map (**figure 3.2)** black star symbols indicated perfect LD (LD=1) between several SNPs of interest. Key associations included the linkage of rs12426639 and rs1388597 which were located 495bp 3' of the SVA-C and +702bp of the *LRRK2* TSS respectively and could be suggestive of a functional implication of the SVA-C as a regulatory domain influencing the *LRRK2* promoter. Associations with perfect LD were also reported between the GWAS SNP rs11175593 (Crohn's disease) and the HapMap SNP rs12426639 (495bp 3' of the SVA-C). Multiple associations with perfect LD were also noted between rs7134408 (HapMap proxy Parkinson's disease GWAS SNP located 9.7kb upstream of the SVA-C) with five SNPs (rs11176153, rs7311247, rs7137437, rs12368229 and rs7308626) within the SVA locus (+/- 500bp of the SVA). Associations between the Crohn's disease GWAS and PD HapMap GWAS proxy SNPs provided evidence that the SVA-C could be genetically involved in those diseases. The genetic associations identified between the SVA-C locus and the general *LRRK2* promoter region (*LRRK2* TSS +1kb) which could suggest a potential regulatory role of the SVA-C as a regulatory domain for *LRRK2*.

3.2.2 *LRRK2* SVA-C genotyping, sequencing, and tagging SNP generation

Given the large degree of genetic variation within the *LRRK2* locus indicated by highly significant GWAS SNPs for different diseases across the locus (**table 3.1**), it was important to understand potential variation within the SVA-C element also. To do this, primers were designed to target various domains of the SVAs composite structure, the CT element, central VNTR and poly-A tail (schematic in **figure 3.3**). It proved refractory to PCR the VNTR directly by standard methods from genomic DNA (gDNA) templates, therefore a two-step nested PCR strategy was employed, this feature was thought to be due to the repetitive nature of SVA sequences. This involved first amplifying the full length SVA element from gDNA, with a second round of PCR being used to target the VNTR using the product of the first PCR reaction as a template. Details of the nested PCR protocol and primer sets used are detailed in **appendix tables 1 and 2**. All PCR genotyping was performed using human DNA samples from a PD case/control cohort of Estonian origin (case n=192, control n=176, further cohort details in **section 2.1.3**). Using nested and standard PCR reactions to amplify each region using gDNA from human blood samples, primary sequence length polymorphisms within the VNTR and poly-A domains were identified, with two alleles of both the VNTR and poly-A found (**figure 3.3**). No observable length polymorphisms were seen within the CT element in 96 control samples using standard PCR and QIAxcel capillary electrophoresis (methods for QIAxcel in **section 2.2.3**), suggesting no common variants (>1%) were present in the tested population However, this does not exclude the potential for length polymorphisms within the CT element or additional alleles of the VNTR and poly-A within different populations.

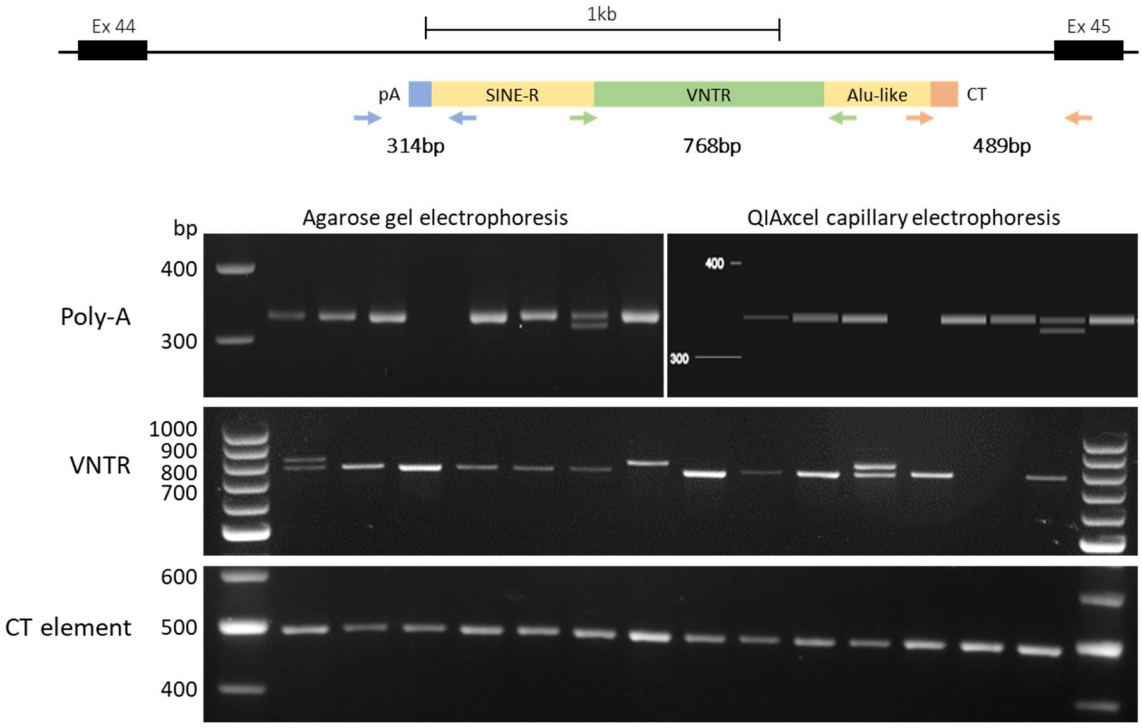
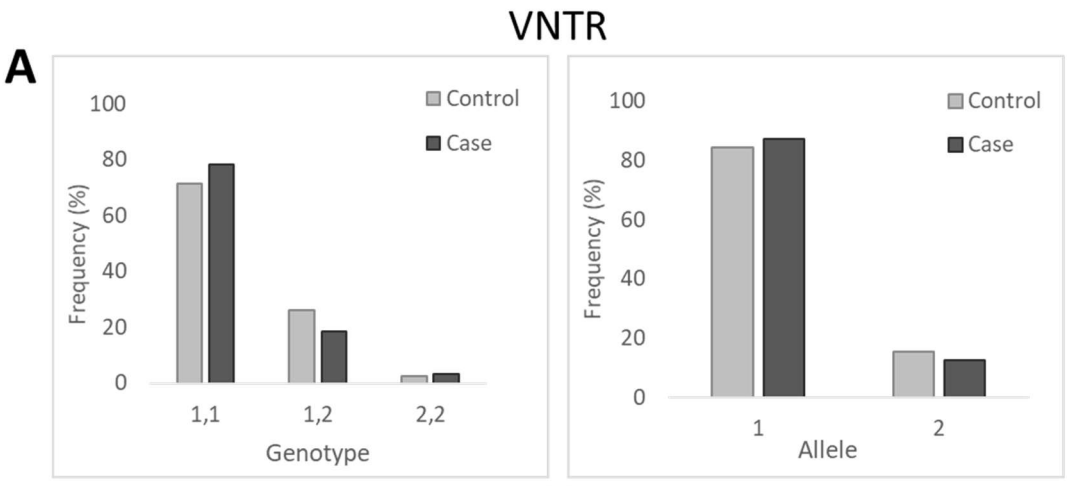


Figure 3.3 – PCR and electrophoresis of the *LRRK2* SVA-C poly-A hexamer CT repeat and VNTR regions highlighting length polymorphisms of both the poly-A and VNTR regions. Schematic depiction of the *LRRK2* SVA-C element within context of the flanking exons 44 and 45 of the *LRRK2* isoform 1. Coloured arrows indicate the placement of primers to amplify the three target regions, poly-A, VNTR and CT element of the SVA-C with predicted PCR amplicon sizes indicated 314bp, 768bp and 489bp respectively. The top panel of the PCR and electrophoresis illustrates comparative amplification of identical human samples of the SVA-C poly-A, with subsequent analysis using conventional agarose gel electrophoresis and the QIAxcel (QIAGEN) advanced capillary electrophoresis system. Validation of the QIAxcel electrophoresis for reliable genotyping of the *LRRK2* SVA-C element allowed for high throughput screening of both the poly-A and VNTR domains of the SVA. Two alleles of both the SVA-C poly-A and VNTR regions were reported with no length polymorphism observed within the hexamer CT element. The two alleles of both the poly-A and VNTR domains were labelled based on their length, with the shorter amplicon labelled as allele 1 and the longer one labelled as allele 2 in both targets. All alleles were subsequently sequenced and compared to the reference genome (**figure 3.5**). All PCRs were performed on healthy (neurodegeneration absent) human control DNA from the Estonian PD case/control cohort (cohort details in **section 2.1.3**).

Having optimised the PCRs to target the three regions of interest in the SVA-C and identifying 2 alleles for both the VNTR and poly-A elements, both polymorphic domains were genotyped in a PD cohort of Estonian ancestry. The central VNTR was genotyped in 147 PD case and 84 matched control samples with no statistical difference between case and control groups identified between genotype or allele frequencies when analysed using Fishers exact tests (confidence interval 0.95) (**figure 3.4 A**). Similar analysis was performed for the poly-A element, in which a total of 182 PD case and 173 matched controls were analysed with no statistical difference in genotype or allele frequency being found using Fisher's exact tests. The results presented only reflect a small sample size (<200 of each case and controls) and may not be adequate to achieve statistical significance. It should also be noted that these analyses were only performed on a single cohort of Estonian ancestry and does not reflect potential differences in other populations.

For further analysis and validation of the identified polymorphisms, Sanger sequencing of each VNTR and poly-A allele was performed. **Figure 3.5** shows the sequences of the reference and alternate alleles identified for each target. The reference allele in each case refers to the allele that appears in the reference genome (hg38), with the shorter allele for each poly-A and VNTR being termed 'allele 1', and the longer allele being labelled as 'allele 2'.

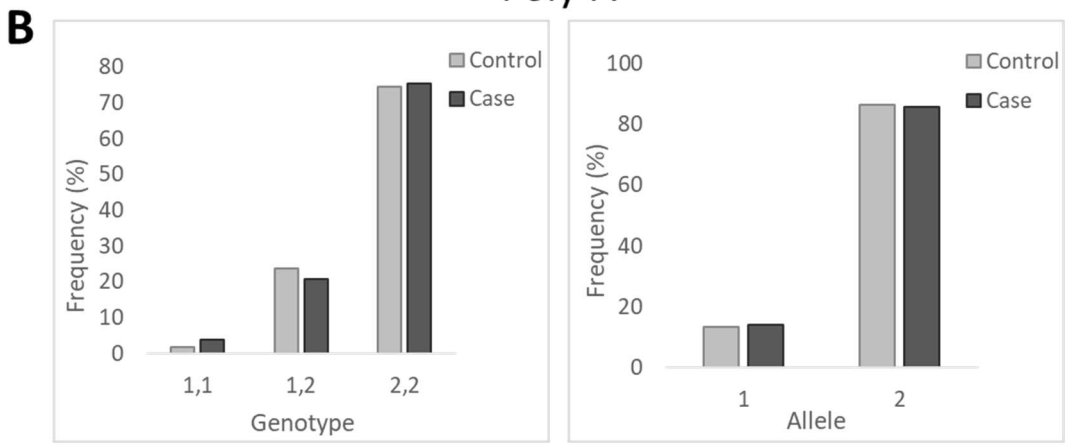


Genotype

Genotype	Control Freq	PD Freq
1,1	60	115
1,2	22	27
2,2	2	5
Total	**84**	**147**

Allele

Allele	Control Freq	PD Freq
1	142	257
2	26	37
Total	**168**	**294**



Genotype

Genotype	Control Freq	PD Freq
1,1	3	7
1,2	41	38
2,2	129	137
Total	**173**	**182**

Allele

Allele	Control Freq	PD Freq
1	57	52
2	299	312
Total	**356**	**364**

Figure 3.4 – Genotyping of the VNTR **(A)** and poly-A **(B)** regions of the *LRRK2* SVA-C element in a matched PD case and control cohort of Estonian ancestry (cohort details in **section 2.1.5**). PCR design and optimisation illustrated in **figure 3.3**. No statistical difference of genotype or allele frequencies were found between case and control samples in the VNTR or poly-A regions when tested using Fishers exact test with a 95% confidence interval.

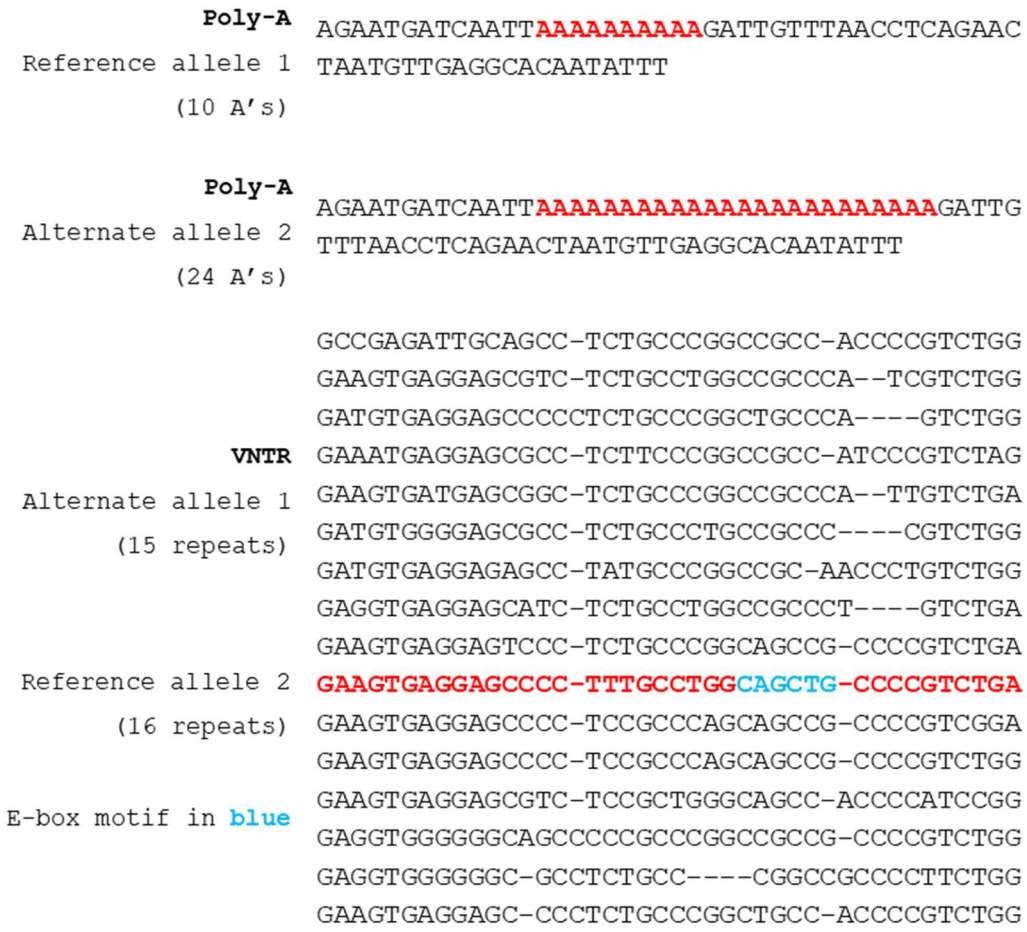


Figure 3.5 – Sequenced alleles of the *LRRK2* SVA-C poly-A and VNTR, showing the two alleles from each region. The poly-A reference allele 1 had a tail of 10 adenine bases with the alternate allele 2 containing an additional 14 bases and thus produced a tail of 24 adenine residues. The VNTR reference allele 2 contained 16 repeats of an approximate 40 base pair motif, with the shorter alternate allele 1 missing one motif (highlighted in red). Interestingly, the missing repeat in the shorter allele is unique due to the presence of an E-box binding motif (CANNTG) highlighted in blue which is not present in the other repeat units.

The two poly-A alleles differed in length by 14 adenine base pairs, whilst the VNTR alleles differed in size by a single 40 base pair repeat. Interestingly, the missing repeat in the VNTR alternate allele 1 was unique when compared with the other repeat units in that it contains an E-box binding motif which could offer a functional difference between the two alleles.

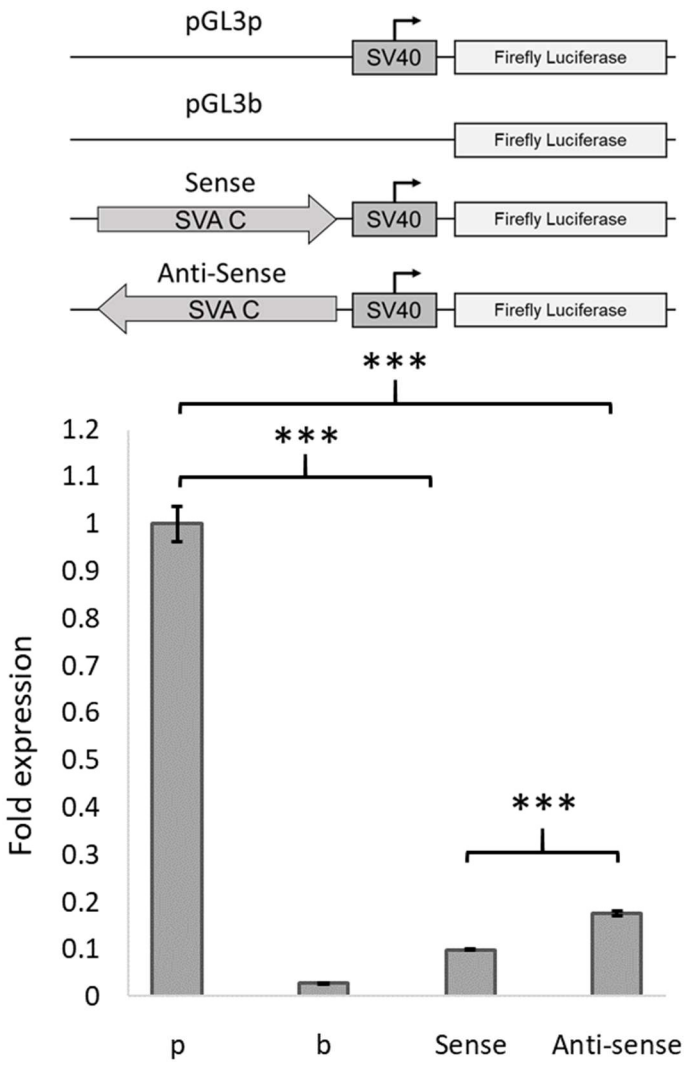
Given the small PD cohort sizes that were available to genotype the *LRRK2* SVA-C element (**figure 3.4**), it was of use to generate bioinformatic tagging SNPs that could be used for future experiments to genotype pre-defined and sequenced, larger PD cohorts such as those available in the International Parkinson Disease Genomics Consortium (IPDGC) (https://pdgenetics.org/) during large scale GWAS analysis [153]. To achieve this a well-defined cohort from the North American Brain Expression Consortium (NABEC) (cohort details in **section 2.1.4**), which consisted of neurologically normal brain DNA samples, which have genome wide genotyping arrays available for use, was used for PCR genotyping of the *LRRK2* SVA-C poly-A on a subset of 96 samples from this cohort. Using the PLINK algorithm command-line program developed by Shaun Purcell (http://pngu.mgh.harvard.edu/purcell/plink/), tagging SNPs were generated for the *LRRK2* SVA-C poly-A with a total of 77 SNPs identified that met the minimum r^2 and D' of >0.8 requirement to be considered a tagging SNP. Tagging SNPs were only generated for the poly-A polymorphisms as a proof on concept and establishment of a pipeline that could be used for future elements including the *LRRK2* SVA-C VNTR polymorphisms.

Table 3.2 – Tagging variants generated for the two *LRRK2* poly-A alleles using the PLINK algorithm pipeline. Seventy-seven tagging variants were generated that satisfied minimum r^2 and D' values of >0.8, indicting high LD between the tag and the SVA genotype that would constitute a suitable tag for reliably bioinformatically genotyping the SVA within large cohorts. Tags generated include both SNPs and short indel sequence variants.

Tags	Co-ords (hg38)	r^2	D'	Tags	Co-ords (hg38)	r^2	D'
rs4767972	chr12:40338321-40338321	1	1	rs7314863	chr12:40350689-40350689	1	1
rs1427271	chr12:40338592-40338592	1	1	rs7311247	chr12:40352112-40352112	1	1
rs1427272	chr12:40338678-40338678	1	1	rs7137437	chr12:40352153-40352153	1	1
rs1427273	chr12:40338928-40338928	1	1	rs11176155	chr12:40352653-40352653	1	1
rs10732751	chr12:40339280-40339280	1	1	rs10878409	chr12:40352707-40352707	1	1
rs12306060	chr12:40339762-40339762	1	1	rs10878410	chr12:40352874-40352874	1	1
rs1365763	chr12:40340781-40340781	1	1	rs10878411	chr12:40353047-40353047	0.91153	1
rs7963697	chr12:40341114-40341114	1	1	rs11176160	chr12:40353065-40353065	1	1
rs7956787	chr12:40341962-40341962	1	1	rs11176161	chr12:40353080-40353080	1	1
rs7956898	chr12:40342002-40342002	1	1	rs10878412	chr12:40353111-40353111	0.954	1
rs7954061	chr12:40342208-40342208	1	1	rs10878413	chr12:40353158-40353158	0.954	1
rs7957151	chr12:40342241-40342241	1	1	rs11610569	chr12:40353217-40353217	0.87222	1
rs919175	chr12:40342603-40342603	1	1	rs142726158	chr12:40353226-40353226	0.87222	1
rs1035812	chr12:40342980-40342980	1	1	rs140722234	chr12:40353344-40353344	0.91153	1
rs1365764	chr12:40343200-40343200	1	1	rs113693842	chr12:40353440-40353440	0.954	1
rs58392855	chr12:40343258-40343265	1	1	rs3930031	chr12:40353732-40353732	1	1
rs17461964	chr12:40343273-40343273	1	1	rs5017705	chr12:40353970-40353970	1	1
rs10784518	chr12:40343313-40343313	1	1	rs7308626	chr12:40354127-40354127	1	1
rs4768231	chr12:40343381-40343381	1	1	rs890575	chr12:40354593-40354593	1	1
rs4768232	chr12:40343438-40343438	1	1	rs10748040	chr12:40354745-40354745	1	1
rs3943893	chr12:40344001-40344001	1	1	rs199872328	chr12:40354836-40354849	0.91153	1
rs1035813	chr12:40344106-40344106	1	1	rs7976724	chr12:40355370-40355370	0.91153	1
rs7296657	chr12:40345925-40345925	1	1	rs7963987	chr12:40355395-40355395	0.91153	1
rs7312497	chr12:40346050-40346050	1	1	rs2162472	chr12:40355416-40355416	0.91153	1
rs10715758	chr12:40346188-40346197	1	1	rs34073451	chr12:40356494-40356501	1	1
rs11289057	chr12:40347027-40347034	1	1	rs7306545	chr12:40356747-40356747	1	1
rs35847030	chr12:40347957-40347967	1	1	rs7306684	chr12:40357018-40357018	1	1
rs1365765	chr12:40348019-40348019	1	1	rs77689380	chr12:40357364-40357373	1	1
rs1365766	chr12:40348271-40348271	1	1	rs145580704	chr12:40359019-40359021	1	1
rs7294958	chr12:40348887-40348887	1	1	rs111612315	chr12:40359243-40359242	1	1
rs7294952	chr12:40349074-40349074	1	1	rs3789330	chr12:40359501-40359501	1	1
rs61579260	chr12:40349510-40349514	1	1	rs10784536	chr12:40360610-40360610	1	1
rs59096461	chr12:40349559-40349576	1	1	rs67472625	chr12:40360878-40360884	1	1
rs4768233	chr12:40349986-40349986	1	1	rs7971919	chr12:40360895-40360895	1	1
rs4768234	chr12:40350028-40350028	1	1	rs7962116	chr12:40360980-40360980	1	1
rs4768235	chr12:40350284-40350284	1	1	rs10548450	chr12:40361243-40361248	0.954	1
rs7314455	chr12:40350343-40350343	1	1	rs2896978	chr12:40361614-40361614	1	1
rs7313525	chr12:40350381-40350381	1	1	rs2404836	chr12:40361674-40361674	1	1
rs7313895	chr12:40350592-40350592	1	1				

3.2.3 Reporter gene assays

Having characterised variation within the *LRRK2* SVA-C at the primary sequence level, it was also important to assess potential function of the SVA-C as a regulatory domain in the context of gene expression. To do this, reporter gene constructs were generated to test the ability of the SVA-C to influence a generic minimal promoter and measure changes in luciferase signal. The SVA-C was cloned into the pGL3p vector (plasmid details in **section 2.1.7**), upstream of a minimal SV40 promoter driving a firefly luciferase reporter (**figure 3.6**). Comparing baseline expression, namely pGL3p, with constructs containing the cloned SVA-C in either the sense or anti-sense orientations, it was possible to infer potential regulatory function of the SVA and measure potential orientation specific effects. **Figure 3.6** shows that the *LRRK2* SVA-C elicited strong repressive effects on firefly luciferase signal when tested in HEK293 cells and that this effect was similar in both the sense and anti-sense orientations, specifically luciferase activity was repressed by 90.2% and 82.2%, respectively. This result was reproducible across the three iterations performed, indicated by the small error bars (**figure 3.6**). The levels of repression were highly significant in both the sense and anti-sense orientations compared to unmodified pGL3p with p values <0.001. The pGL3 based reporter gene assays presented here demonstrated the *LRRK2* SVA-C as possessing repressive characteristics on gene expression, however these models were in context of the minimal SV40 promoter and only represent low expression (**figure 3.6** bottom panel table). Thus, it was also important to demonstrate this effect in a high expression plasmid context such as the pSHM06 vectors driven by a CMV promoter as used in **figure 3.8**.



	Normalised average	Fold change
pGL3p	10.45	1.00 +/- 0.038
pGL3b	0.28	0.03 +/- 0.001
Sense	1.02	0.10 +/-0.002
Anti-sense	1.86	0.18 +/- 0.006

Figure 3.6 – Luciferase reporter gene expression assay performed in HEK293 cells using pGL3p vectors contained the cloned *LRRK2* SVA-C in both the sense and anti-sense orientations shown as schematics in the top panel. Both the sense and anti-sense constructs elicited strong repressive effects with 90.2% and 82.2% reductions in expression, respectively. All data is displayed as a fold change to pGL3p (p) and is normalised to the TK-Renilla internal control. pGL3b (b) plasmids were used as a further control which does not contain a defined promoter and represents background signal within the assay.

Comprehensive pGL3p, pGL3b and pRL-TK control vector maps are presented in **section 2.1.7**. ***p<0.001. Biological replicates N=3 with technical replicates within each assay being performed in quadruplicate (N=4).

For comparison, the same pGL3p constructs were also tested in human induced pluripotent stem cells (iPSCs) (cell line details in **section 2.1.6**) and 28 day differentiated forebrain cortical neuron lineage iPSCs. This allowed for the assessment of potential differential effects within different cell types and further understanding of how the SVA-C may function in a more neuronal cell model than the HEK293 model initially used in **figure 3.6**. Preliminary data from the iPSC models suggested that the *LRRK2* SVA-C may have elicited differential cell specific regulatory effects within the model tested. **Figure 3.7** shows the anti-sense construct eliciting a repressive effect ($p<0.05$) in the differentiated iPSCs, similar to that observed in the HEK293 models (**figure 3.6**), but with weaker effect, with only a 44.9% decrease observed compared to the 82.2% decrease seen in the HEK293 model. The sense construct also demonstrated a repressive effect with a decrease of 34.4% but did not reach statistical significance ($p>0.05$), a possible explanation being due to the lack of biological replicates (N=1) and a higher variability between the technical replicates (N=4). Within the 26-day forebrain cortical neuron differentiated iPSCs the sense construct elicited a significant repressive effect with a 43.6% reduction in luciferase activity (biological replicate N=1, technical replicate N=4) similar to that within the undifferentiated iPSCs where a 34.4% decrease was observed ($p>0.05$) and with a weaker effect compared to the HEK293 model where a 90.2% reduction was seen ($p<0.001$) (**figure 3.6**). In contrast, the anti-sense construct did not present any effect on luciferase signal within the differentiated forebrain neuron iPSCs which could be

indicative of cell dependent effects when compared to the HEK293 model which showed highly significant repression. However, this data is preliminary due to the lack of biological replicates (N=1) and further experiments would be needed to conclude cell dependent effects.

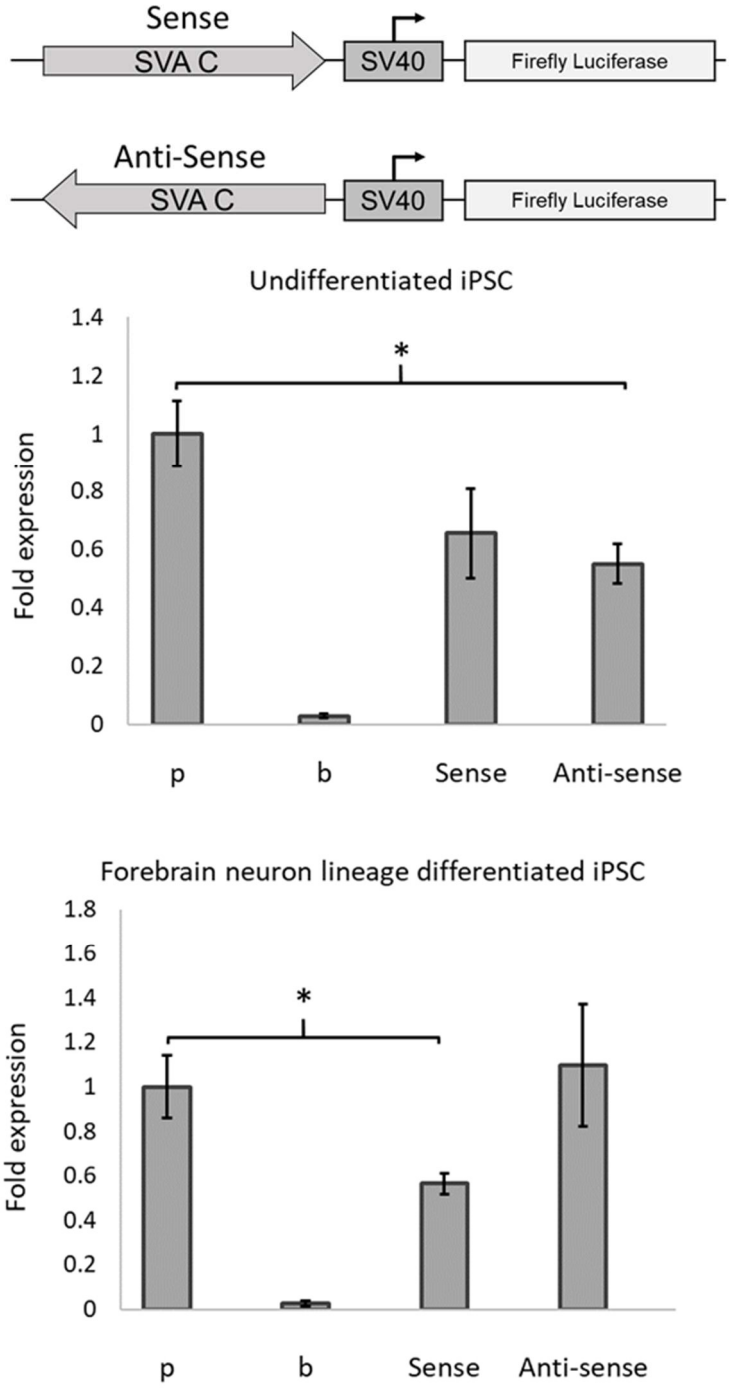
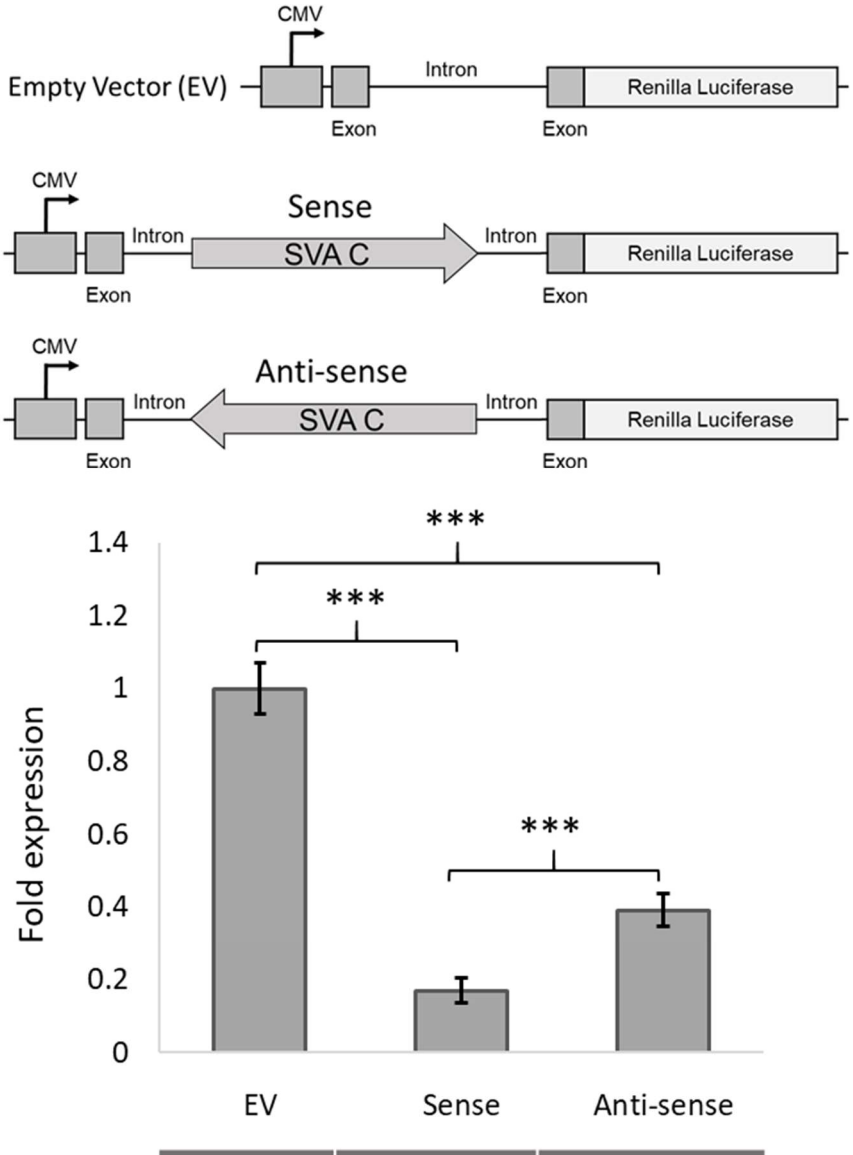


Figure 3.7 – *LRRK2* SVA-C elicited repressive properties in undifferentiated iPSCs and 26-day forebrain cortical neuron differentiated iPSCs with minor orientation specific effects.

Reporter gene constructs containing the *LRRK2* SVA C element in both sense and anti-sense orientations upstream of the minimal SV40 promoter were transfected into undifferentiated human induced pluripotent stem cells (iPSCs) and day 26 forebrain cortical neuron differentiated iPSCs. The sense construct did not elicit any regulatory effects compared to pGL3p (p) within the undifferentiated iPSCs but significantly repressed ($p<0.05$) in the forebrain neuron differentiated iPSCs. A change in trend was observed within the anti-sense construct where there was significant repression ($p<0.05$) within the undifferentiated iPSCs but no difference measured within the forebrain neuron differentiated iPSC model possibly suggesting a tissue specific effect, however without further biological replicates this is only postulation. Fold expression is compared to the SV40 driven pGL3p (p) and normalised to a TK-Renilla internal control (vector details in **section 2.1.7**) to account for transfection efficiency and cell number differences. Biological replicate N=1 with technical replicates performed in quadruplicate (N=4). * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

These data indicated that the *LRRK2* SVA-C possessed regulatory function when tested in the context of the minimal SV40 promoter, but did not provide evidence of how the element may work in context of its genomic position within the *LRRK2* intron 44 of the longer isoform (**figure 3.1**). The pGL3p vector provided basic insights into primary sequence function and identifies elements that may possess enhancers or repressor effects on luciferase gene expression; however, these vectors do not test how the element may impact other regulatory properties such as splicing efficiency. To assess this, a non-commercially available plasmid (gifted from Professor Gerald Schumann, Paul Ehrlich Institute (PEI)) termed pSHM06 (plasmid map in **section 2.1.7**) was used. This vector contains intron 6 of the triose phosphate isomerase (TPI) gene, into which the *LRRK2* SVA-C was cloned, upstream of a Renilla luciferase reporter cassette and driven by a CMV promoter (top panel **figure 3.8**). The CMV promoter is regarded as having high expression compared to the minimal SV40 promoter and provided a more suitable model for measuring the effects of repressive

elements. Sequences cloned within the intron that disrupt the normal splicing process would result in a lower observed luciferase signal (further plasmid details in **section 2.1.7**). The CMV promoter was chosen due to the high expression level compared to the minimal SV40 promoter used in the pGL3 based constructs. Given the initial data (using the pGL3 constructs) demonstrated the SVA-C as a repressive domain, it was useful to test the SVA in the context of a high expression plasmid. Two constructs were generated which contained the *LRRK2* SVA-C cloned in the sense and anti-sense orientations with respect to the direction of transcription within the intron of pSHM06 and transfected in HEK293 cells (**figure 3.8**).



	Normalised average	Fold change
EV	3272	1.00 +/- 0.07
Sense	555	0.17 +/- 0.04
Anti-sense	1284	0.39 +/-0.05

Figure 3.8 – Luciferase reporter gene assay using the pSHM06 constructs indicating the effect of orientation of the *LRRK2* SVA-C element on reporter gene expression in HEK293 cells. Both the sense and anti-sense SVA-C constructs strongly reduce luciferase activity by 83.0% and 60.8% respectively compared to the EV, suggesting a disruption of the normal intron splicing processes. Fold expression is compared to the empty vector (EV) and normalised to a Firefly luciferase internal control. Biological replicates N=3 with each assay containing technical replicates N=4. ***p<0.001. EV – Empty Vector/contained no insert.

The *LRRK2* SVA-C elicited highly significant ($p<0.001$), strong negative effects on splicing of the TPI intron in both the sense and anti-sense orientations within the reporter gene model using the pSHM06 constructs (**figure 3.8**). This effect worked in both orientations with decreases of Renilla luciferase signal of 83.0% and 60.8% being observed in the sense and anti-sense constructs, respectively. The repressive effect had a larger significant effect in the sense orientation over the anti-sense orientation ($p<0.001$) which would suggest an orientation dependent effect. For context, the *LRRK2* SVA-C is found in the anti-sense orientation within the genome with respect to the *LRRK2* direction of transcription and is more similar to the anti-sense construct used in this model. The pSHM06 data presented here shows the repressive effect of the SVA-C on reporter gene expression in a high expression context (driven by CMV promoter) indicated in the bottom table of **figure 3.8**, where the normalised average luciferase signal is approximately 500-fold higher than the normalised average luciferase signal obtained in the pGL3p models in **figure 3.6**. This gives greater confidence that the repressive effects that were observed were comparative and may suggest a more universal effect of SVAs in different contexts. These data along with the pGL3p data in both HEK293 (**figure 3.6**) and iPSC models (**figure 3.7**) provided evidence for the SVA-C as a potential repressive element within different contexts and as a potential modulator of splicing, which is particularly important given the location of the SVA-C within intron 44 of *LRRK2 in situ*.

3.2.4 *In vitro* functional analysis using CRISPR

The collective reporter gene assay data presented here demonstrated that the *LRRK2* SVA-C was a regulatory functional element that could act as a modulator of gene expression (**figures 3.6 and** 3**.7**). The data suggested that the effect may be cell specific and harbours splicing regulatory properties (**figure 3.8**) at the primary sequence. However, the effects observed were out of context of the *LRRK2* locus and so, it would be additionally informative to test the SVA's ability to alter *LRRK2* regulation in an *in vitro* model system. To do this, CRISPR technologies were employed to generate HEK293 knockout (KO) cell lines which were used to measure *LRRK2* expression levels under different conditions in response to SVA KO. Details of the CRISPR protocols employed throughout this thesis are detailed in **section 2.2.11**. The CRISPR protocols used were first optimised on the *INPP5F* SVA elements detailed in **section 4.2.3**. Initial experiments identified the HEK293 cell lines as a suitable model for these experiments due to good transfection and modification efficiencies (**figures 4.7 and 4.8**).

Prior to starting the CRISPR protocols, expression of *LRRK2* was measured to confirm which isoforms were expressed within HEK293 cells under basal conditions. **Figure 3.9-A** shows the three major coding isoforms of *LRRK2* (hg19) and the associated protein domains with the assigned nomenclature to be used throughout this chapter.

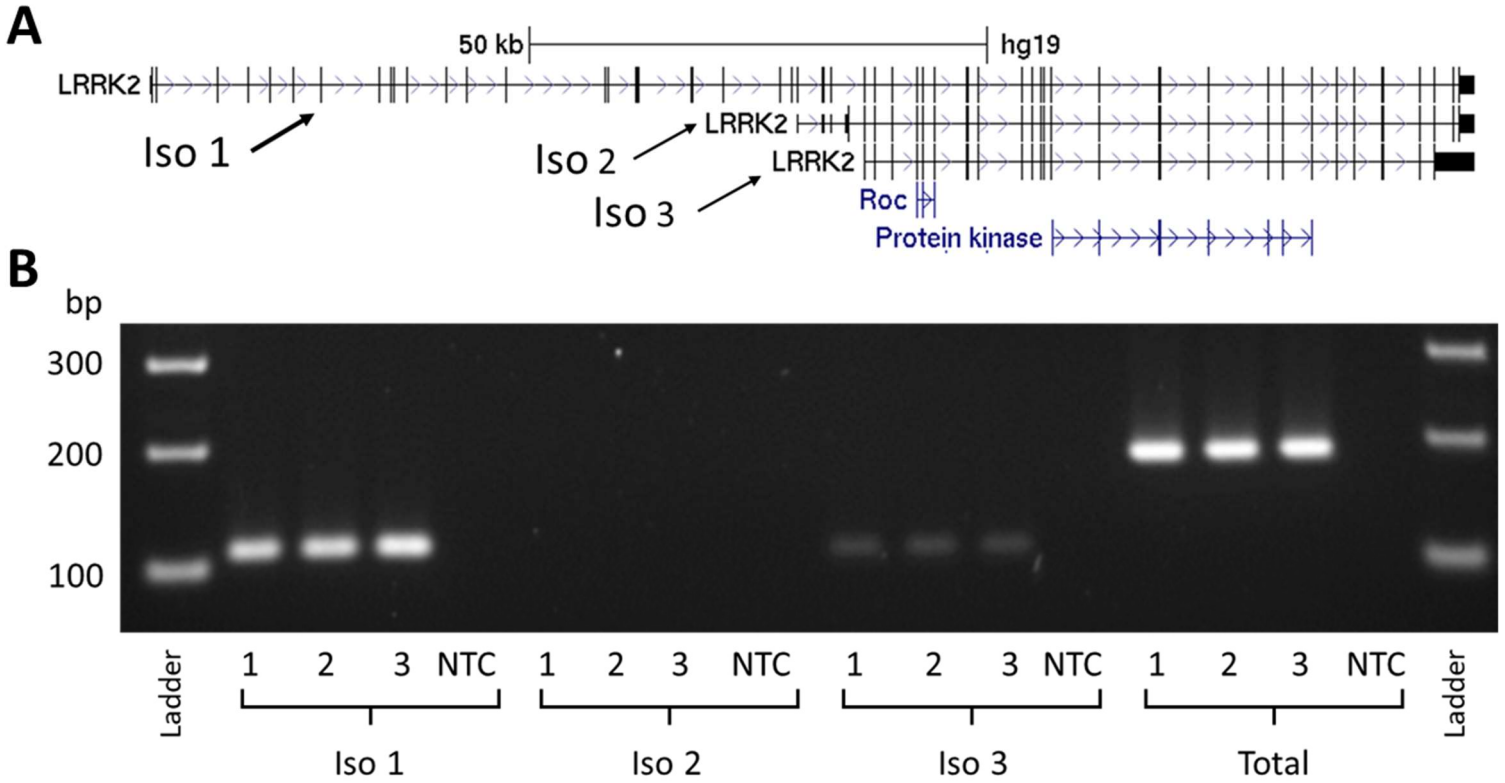
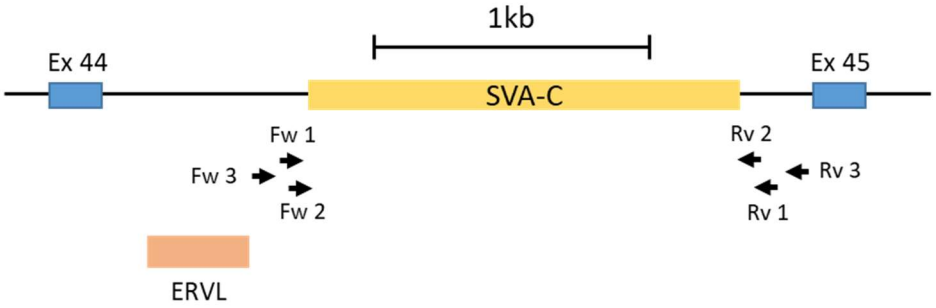


Figure 3.9 - **(A)** UCSC genome browser adaptation indicated three major protein coding isoforms of *LRRK2* to which RT-PCR primers were designed. UniProt protein domain coding exons are indicated in blue and show the roc and kinase domains of *LRRK2*. **(B)** Optimisation of RT-PCR for the three *LRRK2* isoform targets using untreated HEK293 cDNA extracted under basal conditions. Both isoforms 1 and 3 were expressed, with little to no expression of isoform 2. A common primer set was also designed to identify total *LRRK2* expression levels (Total) which used common exons present in all isoforms. Amplicon sizes were limited to <200bp to enable use in qPCR and were designed using the parameters outlined in **section 2.2.7.1**.

The three isoforms identified were termed isoform 1, 2 and 3, to which isoform specific primers were designed to amplify each isoform independently using unique exons available for each target. A fourth primer set was also designed to measure total *LRRK2* expression levels by utilising common exons between the three identified isoforms. RT-PCR and gel electrophoresis were performed on extracted RNA from HEK293 cells under basal conditions prior to initiating the CRISPR protocol which showed both isoforms 1 and 3 were present, with no detectable expression of isoform 2. This was a suitable expression profile given the previous reporter gene data that demonstrated the *LRRK2* SVA-C to act as a repressor and would allow for potential changes in both expressed and not expressed isoforms to be measured in response to the CRISPR mediated excision of the SVA.

Having established HEK293 as a suitable cell model from the transfection and colony selection optimisation using the SVAs of the *INPP5F* locus and confirming appropriate expression of *LRRK2*, guide RNAs (gRNAs) were designed to target the *LRRK2* SVA-C. Three forward and three reverse gRNAs flanking the SVA were tested using all combinations of forward and reverse guides (nine total) to efficiently target the SVA-C for excision. **Figure 3.10** highlights the positions of the gRNAs proximal to the SVA-C whilst avoiding coding exons that could generate *LRRK2* knockouts by excising coding sequences and disrupting protein function. Guides also had to be positioned to avoid overlapping other repeat elements such as the endogenous retrovirus element (ERVL) approximately 140bp away from the SVA-C that may introduce undesirable off target effects by decreasing the specificity of the guides.



Predicted PCR product size (bp)				
gRNA	Fw 1	Fw 2	Fw 3	Control
Rv 1	A – 1033	B – 1039	C – 940	Un - 2690
Rv 2	D –1093	E – 1099	F – 1000	
Rv 3	G – 899	H – 905	I – 806	

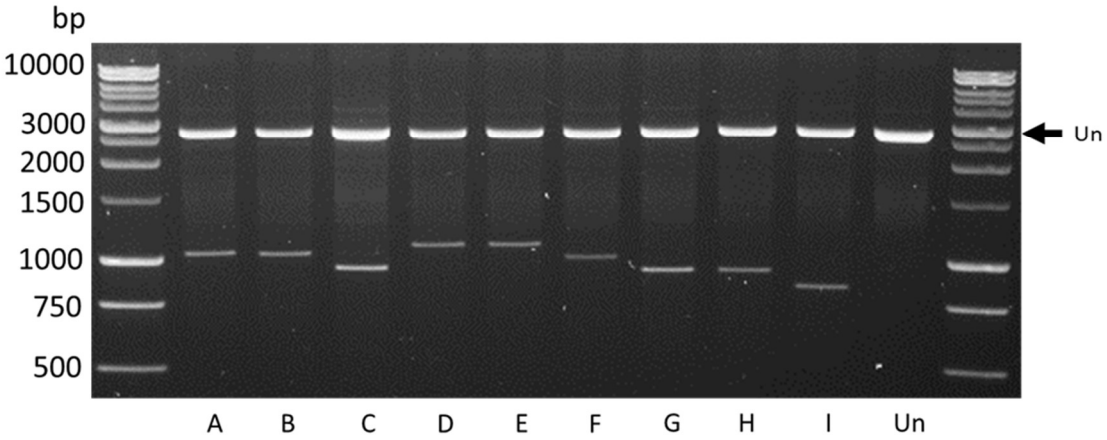


Figure 3.10 – Optimisation of guide RNA (gRNA) sequences to effectively target the SVA-C for CRISPR mediated deletion in HEK293 cells. The gRNAs were designed to precisely target the SVA-C only without editing coding exons of *LRRK2* or other notable sequences such as the endogenous retrovirus (ERVL) sequence. Predicted PCR amplicon sizes are shown in the table upon deletion of the target sequences using the various combinations of gRNAs. PCR amplification and gel electrophoresis indicated an approximate 20% overall editing efficiency across all gRNA combinations (estimated by comparing the ratio of the unmodified band intensity to the modified band) with the highest efficiency observed in combination G (Fw 1, Rv 3). Un – unmodified control HEK293 cells.

Following the CRISPR protocol detailed in **section 2.2.11** using gRNA combination G identified in **figure 3.10**, six putative mono-allelic and three putative bi-allelic SVA-C knockout HEK293 cell lines were generated. Of these cell lines, three of the putative mono-allelic and the three putative bi-allelic lines being chosen for RT-PCR analysis of *LRRK2* isoform expression (highlighted by white arrows in **figure 3.11**). Clonal isolated cell lines were screened using crude lysate PCR (**section 2.2.11.4**). The positive clones were expanded, and DNA was extracted, purified and subsequently PCR amplified across the flanking regions of the SVA-C. The resulting PCRs, when run on an agarose gel, would produce either one or two bands depending on whether the clone contained the correct deletion and could be identified as homozygous (putative bi-allelic) or heterozygous (putative mono-allelic) for the SVA deletion.

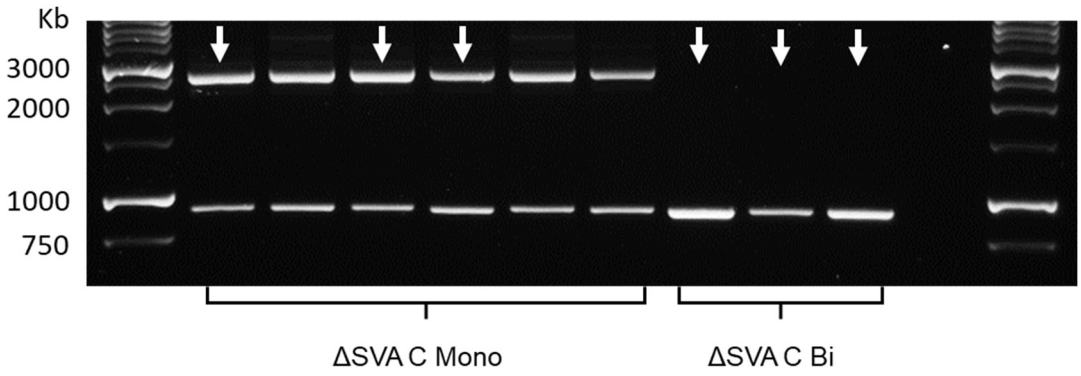


Figure 3.11 - CRISPR modified clonal HEK293 cell lines indicating six putative mono-allelic and three putative bi-allelic deletions of the *LRRK2* SVA-C obtained via PCR screening over the target region. Shift in band size indicates excision of SVA in each respective target. The lanes marked with white arrows indicate the clones that were taken forward for RT-PCR analysis of *LRRK2* expression profiles.

Wild type (WT) and non-targeting (NT) gRNA containing EF1α-pSpCas9(BB)-2A-GFP plasmid transfected control cells were used throughout the entire CRISPR protocol and were treated identically to the SVA knockout (KO) clones across the entire CRISPR

protocol. Both controls underwent the same clonal isolation process throughout the protocol, with three cell lines per condition being chosen to take forward giving a total of twelve clonal HEK293 cell lines used in subsequent analysis.

RT-PCR of the three *LRRK2* isoforms identified (**figure 3.9**) alongside total *LRRK2* expression was performed and analysed using standard agarose gel electrophoresis. **Figure 3.12** indicates the *LRRK2* isoform expression profiles across 12 independent putative clonal cell lines including wild type (WT), non-targeting (NT) gRNA transfected, putative mono allelic and bi allelic *LRRK2* SVA-C knockout HEK293 lines all under basal conditions. When comparing each condition as a collective of the three independent clones within each group, observational analysis indicated little to no consistent change in *LRRK2* expression within any isoform upon excision of the SVA-C element. No changes in both the expression of total *LRRK2* and isoform 1 were noted across the cell lines with inconsistent differences in isoform 2 expression which appeared to be expressed in two of the WT and one putative bi-allelic KO line with no expression seen in any other cell lines. *LRRK2* isoform 3 expression did not change across the cell lines, however the band intensity appeared to be slightly less consistent with minor observable increases in the three putative mono-allelic KO lines and one control NT cell line. To quantify changes in *LRRK2* isoform expression qPCR would be required, however due to time constraints that was not possible. The housekeeping genes *GAPDH* and *ACT-B* were chosen due to the consistent expression profile observed across all samples which indicated that these genes were expressed across all cell lines.

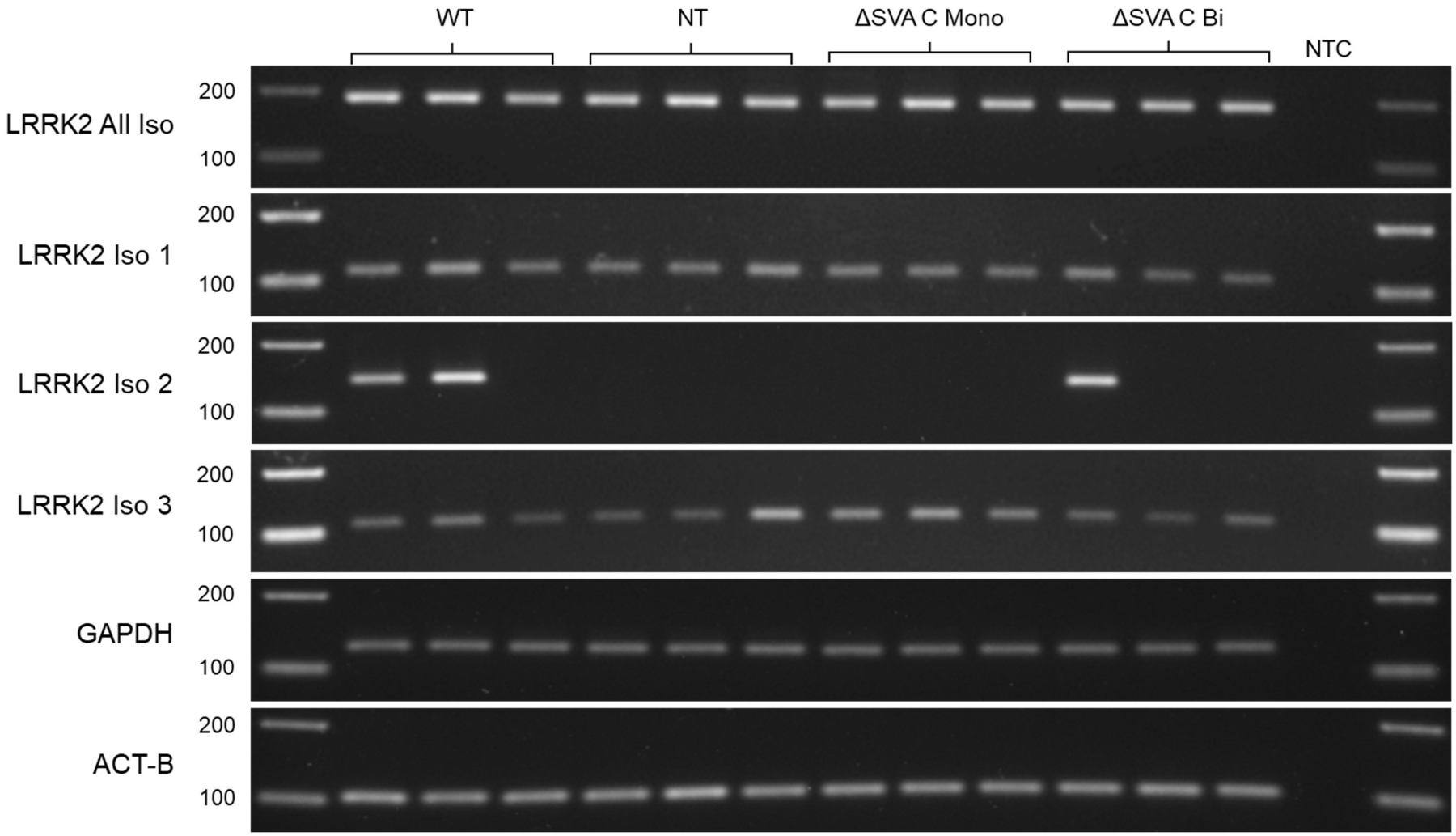

WT
NT
ΔSVA C Mono
ΔSVA C Bi
NTC
LRRK2 All Iso
LRRK2 Iso 1
LRRK2 Iso 2
LRRK2 Iso 3
GAPDH
ACT-B
200
100

Figure 3.12 – RT-PCR and agarose gel electrophoresis of *LRRK2* SVA-C CRISPR knockout HEK293 clonal cell lines targeting various isoforms of *LRRK2* with housekeeping genes *GAPDH* and *ACT-B*. Three independent clones tested for each condition. Abbreviations: WT – wild type, NT – non-target guide control. Predicted amplicon sizes: *LRRK2* all iso – 186bp, iso 1 – 110bp, iso 2 – 139bp, iso 3 – 107bp, *GAPDH* – 130bp, *ACT-B* – 110bp

Under basal conditions, there was little to no effect of SVA-C knockout on the *LRRK2* isoform expression profile (**figure 3.12**). To explore further potential effects of SVA KO, appropriate challenges that could be applied were explored. A serum starvation with reintroduction of serum after a 24-hour starvation period was selected as an appropriate challenge based on bioinformatic analysis of *LRRK2* promoter ChIP-seq data from the ENCODE project data available through the UCSC genome browser (hg38). **Figure 3.13** identified multiple FOS, JUNB and JUND binding sites within the major *LRRK2* promoter, indicated by the CpG island (green) and epigenetic histone marks H3K4 tri-methylation (Me3), and multiple binding sites within potential regulatory regions indicated by the H3K27 mono-methylation marks (Me1) and H3K27Ac (acetyl) marks. FOS and JUN form the AP-1 complex which is activated by the serum response factor upon stimulation by changes in serum levels.

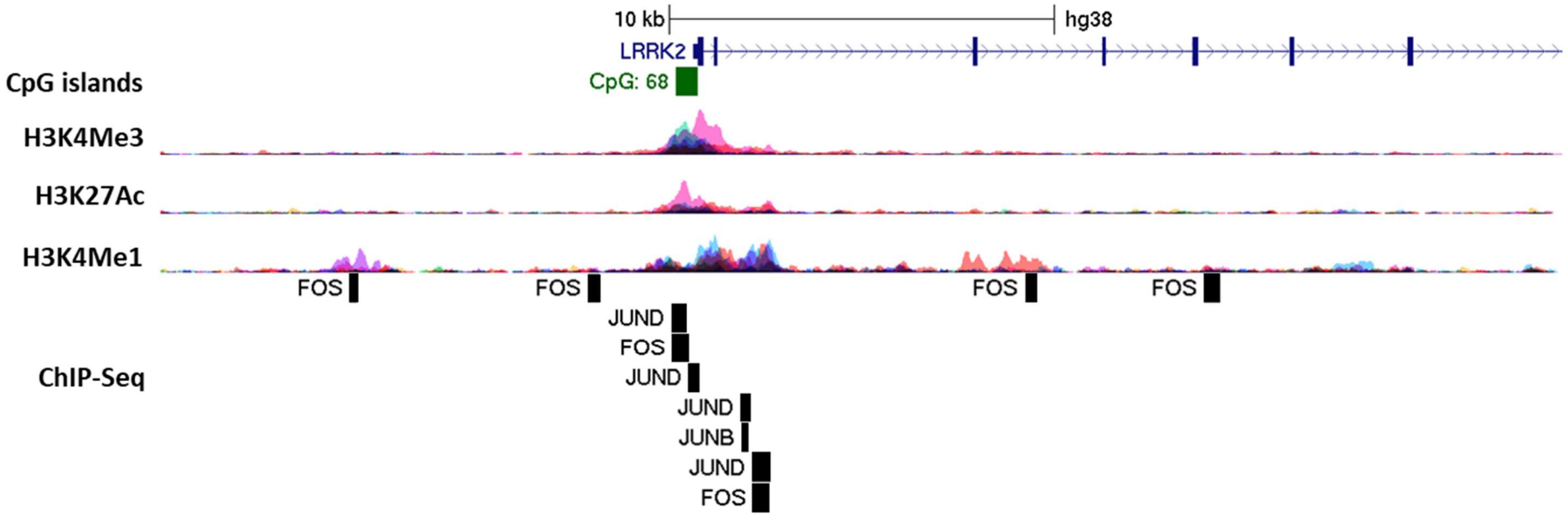


Figure 3.13 – FOS, JUNB and JUND binding sites across the *LRRK2* promoter locus (chr12:40211170-40247566) adapted from the UCSC genome browser (hg38) identified using ChIP-Seq data from the ENCODE project. Multiple binding sites were identified within the *LRRK2* promoter, indicated by the CpG island and H3K4Me3 (tri-methyl) histone marks, and regulatory regions indicated by the H3K27Ac (acetyl) and H3K4Me1 (mono-methyl) marks.

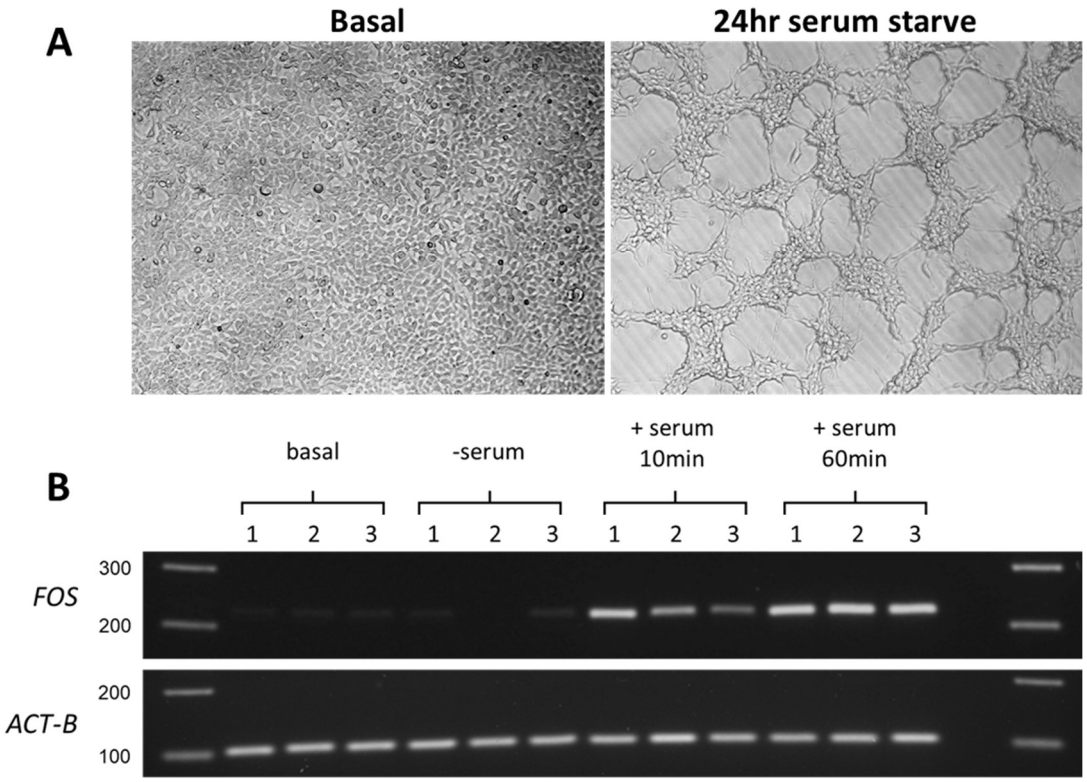


Figure 3.14 – **(A)** Phase contrast imaging of wild type HEK293 cells in culture under basal conditions compared to a cell sample from the identical culture 24 hours post-serum starvation (0% serum). The cells, when serum starved, group together and send out projections. **(B)** Induction of *FOS* after addition of serum to the 24-hour serum starved cells at 10-minute and 60-minute post re-introduction (10% serum reintroduced) time points with a more robust induction achieved with the 60-minute time point. RT-PCRs performed in three independent clonal un-transfected wild type (WT) HEK293 cell lines generated within the CRISPR assays (labelled 1, 2 and 3). Beta-actin (*ACT-B*) was used as a housekeeping gene within this assay which presented stable expression patterns across the various serum conditions used.

Wild type HEK293 control cells were used to test the efficiency of the induction of a serum response to activate the *FOS* pathway as a response to changes in culture media serum concentrations. **Figure 3.14-A** shows phase contrast imaging of WT cells under basal conditions (10% serum) compared with the same cell culture; 24-hours post serum starvation (0% serum). The individual cells group together and appeared to send out projections as a response to the lack of serum. In **figure 3.14 -B**, a robust induction of *FOS* was demonstrated as a response to re-introduction of foetal bovine serum (10% serum) after a 24-hour serum starvation period. Two time points were tested for the incubation period with serum present post-starvation, 10-minute, and 60-minute with the 60-minute re-introduction time point producing the most effective induction of *FOS* in wild type (WT) HEK293 cells.

Using the optimised serum response factor (*FOS*) induction protocol with a 60-minute serum incubation time point, a preliminary experiment to test potential effects of *FOS/JUN* induction on *LRRK2* expression within the KO SVA-C cells compared to control cell lines (**figure 3.15**) was tested. The identical WT, putative mono and bi-allelic SVA-C KO cell lines were used as in the basal measurements of *LRRK2* expression (**figure 3.12**). The NT lines were not included as they exhibited no difference in *LRRK2* expression patterns compared to WT cells prior to the serum starvation assay (**figure 3.12**).

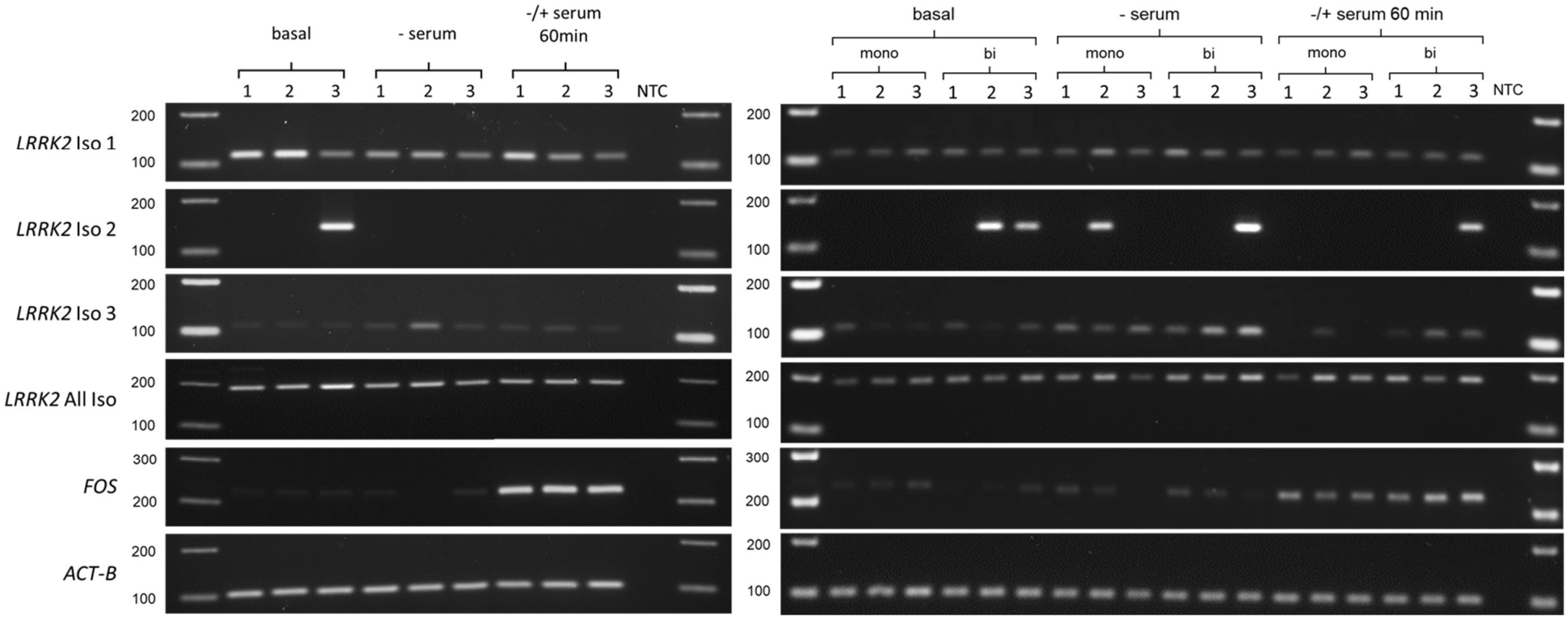


Figure 3.15 – RT-PCR and gel electrophoresis of *LRRK2* SVA-C CRISPR KO HEK293 clonal cell lines, targeting the three identified isoforms of *LRRK2,* total *LRRK2* expression, *FOS* and *ACT-B* under induction of serum response factor (*FOS*) with three independent clones tested for each condition. The left panel shows the untransfected (Un) control cell lines (1, 2 and 3) in response to a 24-hour serum starve followed by a 60-minute serum re-introduction. The right panel indicated the response to the same serum starve challenge in *LRRK2* SVA-C KO HEK293 cell lines using the putative bi-allelic and putative mono-allelic knockout cell lines **(figure 3.11)**. Predicted amplicon sizes: *LRRK2* iso 1 – 110bp, iso 2 – 139bp, iso 3 – 107bp, all iso – 186bp, *GAPDH* – 130bp, *ACT-B* – 110bp

The 60-minute serum re-introduction time point gave the greatest induction of *FOS* out of the time-points tested and was therefore chosen to take forward (**figure 3.14**). **Figure 3.15** represents the preliminary results from the serum starvation challenge showing RT-PCR and agarose gel electrophoresis of three *LRRK2* isoforms (identified in **figure 3.9**), total *LRRK2* expression, *FOS* and *ACT-B* expression levels in control wild type (WT) cell lines (left panel) compared to the *LRRK2* SVA-C knockout (KO) lines (right panel). Observational results indicated no difference in *LRRK2* isoform 1 expression between WT and KO cell lines in the response to *FOS* induction. Isoform 1 expression was expressed across the treated and non-treated WT cell lines with a similar expression profile observed across the KO lines. Minor band intensity differences of isoform 1 expression in response to serum starvation were noted between the WT and KO cells, where the expression appeared to decrease within the WT cells compared to basal but increased marginally within the KO cells. The inconsistency of *LRRK2* isoform 2 expression observed across all samples could not be attributed to any changed condition within the model and must have been under the influence of factors that were not considered within the scope of this assay. The *LRRK2* isoform 3 appeared to increase within the WT1 and bi-allelic 2 and 3 cell lines in the serum starved condition (-serum) compared to the basal condition. *LRRK2* isoform 3 expression observably returned to an expression level similar to that seen in the basal condition after reintroduction of serum (-/+ serum 60 min). The results discussed here are observational and preliminary only. Quantification of relative expression changes in response to changing serum levels using qPCR would be necessary for the comparison of response to changing serum levels, however this was not performed and would be considered for future experiments.

Beta-actin (*ACT-B*) was chosen as a suitable housekeeping gene within this model due to the stability of expression levels across all tested conditions. The induction of *FOS* as a result in changing serum levels within the culture medium exhibited a clearer and stronger response in the WT cells (left panel) compared to the KO cell lines (right panel). This was unexpected given that all cells were treated identically and in tandem. Ideally, this assay would have been repeated to replicate the results described here i.e. the level of serum induced *FOS* expression observed in both WT and KO cell lines, but due to time constraints this was not possible.

3.3 Discussion

Initial bioinformatic analysis of the *LRRK2* locus provided insight into genetic involvement of *LRRK2* within multiple diseases and traits. GWAS signals across the *LRRK2* locus (**figure 3.1 and table 3.1**) indicate multiple assocations to various diseases including PD, Crohn's disease and inflammatory bowel disease (IBD). There are several studies which suggest the gut microbiome and inflammation is involved in the onset on Parkinson's disease as well as an association between patients with IBD having increased risk for developing Parkinson's disease [151]. Interestingly, *LRRK2* has been reported to have pleiotropic effects confering risk for both Parkinson's disease and Chrohn's disease which was reflected within the analysis presented here [150]. Using linkage disequilibrium (LD) analysis between characterised HapMap SNPs selected around the *LRRK2* promoter, SVA +/- 2kb, known GWAS SNPs and HapMap GWAS proxy SNPs (the nearest HapMap SNP to a known GWAS SNP if no HapMap SNP I.D. was available) multiple association were identified. LD is useful for making basic inferences of potential links between

otherwise seemingly unassociated regions and is defined as the non-random assocation of alleles at different loci [154]. Perfect LD (LD = 1.00) was found between rs12426639 and rs1388597 which were located 495bp 3' of the SVA-C and +702bp of the *LRRK2* TSS respectively, in addition to perfect LD between the GWAS SNP rs11175593 (for Crohn's disease) and rs12426639 (located 495bp 3' of the SVA-C). These links were suggestive of genetic associations between the SVA-C locus and *LRRK2* promoter as well as a potential implication of the SVA-C in Crohn's disease. Genotyping of the SVA in a Crohn's disease cohort could be of interest for future studies given this association by LD. Other notable associations with an LD of 1.00 were also found between rs7134408 (proxy GWAS SNP for PD withing *LRRK2*) with five HapMap SNPs (rs11176153, rs7311247, rs7137437, rs12368229 and rs7308626) which were all located within or proximal to the SVA (+/- 500bp). The linkage between the SNPs within the SVA locus and GWAS, GWAS proxy and SNPs within the promoter, all provide evidence to suggest a potential regulatory function for the SVA with respect to *LRRK2*.

To explore potential novel risk variation within the *LRRK2* SVA-C, PCR genotyping was performed to identify primary sequence length polymorphisms and correlate the allele frequencies with disease in a human PD cohort of Estonian origin. Length polymorphisms were found within the central VNTR and poly-A domains with two alleles of each identified within the tested cohort (**figure 3.3**). There were no significantly over-represented alleles or genotype frequencies detected within the case or control populations (**figure 3.4**). However, this analysis only reflects a small sample size (<200 of each case and controls) from a single ethnic background and does not exclude the possibility for additional polymorphisms within other

populations. An additional consideration when approaching the genotyping of retrotransposable elements is to determine if the element in question is fixed or polymorphic for presence/absence, this is because many elements present within the reference genome are assumed to be fixed elements, however this assumption may be erroneous. This assumption is flawed because of the lack of diversity within the reference genome, being that it is only compiled from 20 individuals, with 70% of the build coming from a single sample (donor RPC-11) [155]. This means that annotated retrotransposons within the reference genome have the potential to be polymorphic for presence/absence within different populations. The genotyping of the *LRRK2* SVA-C presented in **section 3.2.2** provided no evidence to suggest the SVA was absent in any of the tested samples. This was determined from the lack of absent amplicons present within the genotyping data (i.e. all samples tested produced amplicons consistent with the presence of the SVA). There was no detection of smaller amplicons during the full-length PCRs that would have been indicative of an empty site PCR product. This result was not unexpected given that the SVA element is classed as an SVA-C, indicating it is likely conserved within other primate genomes. This was confirmed in the initial bioinformatic analysis of the *LRRK2* locus (**figure 3.1**) where the conservation analysis indicated that the SVA-C was present in human, chimpanzee and gorilla genomes indicating that the SVA-C insertion in the LRRK2 locus occurred in a common ancestor and, as such, is extremely unlikely to be polymorphic for presence/absence across human populations.

To bioinformatically explore the potential differences between the two VNTR alleles identified by PCR genotyping, Sanger sequencing was performed which revealed that the alternate allele had a deletion of one 40mer repeat compared to the reference

allele (**figure 3.5**). Interestingly, the repeat that was deleted in the alternate allele is unique compared to the other repeats within the VNTR and contains a unique E-box motif which may suggest a differential functional implication between the alleles. Basic helix-loop-helix (bHLH) transcription factors (TFs) are amongst the highest affinity TFs that bind E-box motifs with the consensus CANNTG sequence [156]. More than 240 bHLH TFs are known to exist in eukaryotes with the specific binding of TFs dependent on the nature of the NN dinucleotide within the consensus sequence as well as flanking sequences [157]. This gives E-box motif binding proteins a wide range of implications in both normal cellular function and pathological dysfunction with many characterised examples of E-box binding proteins in cancers and neurogenesis such as Myc, Mad family members (involved with cancer progression) and Nhlh, NeuroD family members (involved in neurogenesis) [158, 159].

To effectively increase the sample size and range of ethnic backgrounds when genotyping, the use of tagging SNPs to computationally infer SVA genotypes using larger pre-defined cohorts, can be used. The tagging SNPs generated for the *LRRK2* SVA-C (**table 3.2**) provide an invaluable resource for future genotyping of this element within larger cohorts such as those collected by the International Parkinson Disease Genomics Consortium (IPDGC) in which approximately 40,000 PD cases and 1,400,000 control samples have been genotyped [26]. By using sub-sets of these samples, it would be possible to generate *in silico* data to identify novel risk variants of the *LRRK2* SVA-C within specific populations of different disease phenotypes and ethnicities. Using a PCR approach to do this in large sample sets would be an inefficient use of time and resources to obtain the necessary sample sizes from a variety of ethnicities and populations to the *in silico*-based approach using the

tagging SNPs generated within this thesis. The PLINK algorithm command-line program developed by Shaun Purcell at the Broad Institute (http://pngu.mgh.harvard.edu/purcell/plink/) was used to generate tagging SNPs for the *LRRK2* SVA-C poly-A with a total of 88 SNPs being identified. However due to time constraints, further analysis of the identified SNPs using other data resources was not performed. The *LRRK2* SVA-C tagging SNPs can also be used to correlate with existing data within the NABEC databases such as exome sequencing, cap analysis gene expression (CAGE) sequencing, mRNA sequencing and NeuroChip genotyping data of known neuropathological variants (dbGaP study accession: phs001300.v1.p1). Using these data sets it is possible to explore the potential impact of the SVA variants on *LRRK2* gene function in a variety of contexts including differential isoform expression and inheritance with known PD pathological variants.

The functional analysis performed for the *LRRK2* SVA included the use of reporter gene assays and CRISPR mediated SVA knockout (KO) HEK293 cell lines. Reporter gene constructs were generated using two different vector contexts; pGL3 based SV40 driven luciferase reporter vectors and a non-commercially available vector for testing splicing effects of cloned inserts (pSHM06) (vector details in **section 2.1.7**). The pGL3p constructs provided evidence that the SVA-C elicits repressive effects on luciferase reporter expression in both the sense and anti-sense orientations when transfected in HEK293 cells (**figure 3.6**). This repressive effect was also observed in the anti-sense and sense pGL3p constructs when transfected in undifferentiated iPSCs and forebrain neuron lineage differentiated iPSCs, respectively (**figure 3.7**). Savage *et al.* 2013, have previously demonstrated that SVAs can elicit regulatory function using *in vitro* models with both cell specific and orientation specific effects.

In their study, an SVA in the *PARK7* locus cloned into the pGL3p vector (the same as used in this study) was transfected in both SK-N-AS and MCF-7 cell lines, where it was shown that the sense SVA construct had no effect on the expression of luciferase in SK-N-AS which contrasted to a significant increase in expression observed in the MCF-7 cell line. However, the anti-sense construct induced significant repression of luciferase activity in both the SK-N-AS and MCF-7 cell lines indicating that SVA elements can have cell specific and orientation specific effects [76].

Reporter gene assays provide some insight into potential functions of cloned sequences, but do not provide contextual data given that the cloned sequence has been taken out of the context of its *in-situ* state within the genome whereby flanking sequences and chromatin structure have both been removed. To address this, CRISPR technologies were employed in HEK293 cells to generate *LRRK2* SVA-C KO cell lines with the aim to use both RT-PCR and qPCR to measure changes in *LRRK2* isoform expression in response to SVA KO. One consideration in the use of HEK293 cells was the difficulty to specify the exact genotype of the SVA knockout produced without performing karyotyping prior to the CRISPR mediated genetic manipulation. The karyotype of HEK293 is largely disputed in the literature, with different studies suggesting different chromosomal copy numbers [136]. *LRRK2* SVA knockout HEK293 cell lines were successfully generated with a minimum of six putative mono-allelic KO and 3 bi-allelic KO cell lines being confirmed using PCR (**figure 3.11**). The protocols that were optimised for the generation of reference SVA KO HEK293 cells are the first detailed example, to date, of CRISPR technologies being utilised in this way. The results using RT-PCR analysis indicated that under basal conditions, there was little to no effect of SVA-C knockout on the three *LRRK2* isoform expression profiles in

HEK293 cells (**figure 3.12**). This lack of effect could have been due to cell recovery and stabilisation of *LRRK2* expression over the course of the clonal isolation and expansion immediately post-transfection of the Cas9/gRNA delivery plasmids as a result of feedback mechanisms. The results presented here were based solely on observational analysis of RT-PCR reactions, analysed using agarose gel electrophoresis, by comparing band intensity differences between KO cell lines and both the wild type (WT) and non-targeting (NT) controls (**figure 3.12**). There were minor fluctuations in band intensity observed in the *LRRK2* isoform 1 and 3 between the WT, NT, putative mono and bi allelic SVA KO cell lines with the plan to analyse these using qPCR to obtain quantifiable relative expression values between the cell lines (following the same protocol used for the *INPP5F* SVA CRISPR qPCR experiments detailed in **section 4.2.3**). However, due to time constraints it was not possible to perform qPCR for the *LRRK2* SVA KO lines.

To explore the possibility that the lack of effects seen as a result of SVA KO, a serum starvation challenge was chosen to employ to each HEK293 cell line. This challenge was chosen due to the presence of FOS, JUNB and JUND binding sites present within the *LRRK2* promoter and flanking regulatory regions (**figure 3.13**) which were hypothesised to respond to changes in serum levels through activation of the serum response element (SRE) [160]. c-Fos, alongside c-Jun, create a heterodimer that form the AP-1 complex for transcriptional regulation across a vast array of gene targets including *LRRK2*. The serum starvation was performed for 24 hours (0% FBS) before serum was reintroduced into the culture medium (10% FBS) to activate the serum response factor which increases transcription of downstream targets including c-Fos and c-Jun, as demonstrated by the increase in *FOS* expression observed by RT-PCR of

the *FOS* mRNA (**figure 3.14**) [161]. It was predicted that this induction would have promoted an increase in the activation of the AP-1 complex and consequent changes in *LRRK2* expression levels as a response to serum concentration changes from 0% to 10% FBS. If the hypothesis that the SVA-C element influenced the regulation of *LRRK2* were correct, there could have been a change in the expression profile of *LRRK2* in response to the serum starve challenge in SVA-C knockout HEK293 cells compared to the WT or NT controls. The results of the serum starve challenge indicated mixed effects when using observational analysis of RT-PCR band intensities as viewed by agarose gel electrophoresis. Minor potential variances of expression of the three *LRRK2* isoforms were noted within and between the tested conditions particularly across the putative mono and bi-allelic KO cell lines, most notably, the total *LRRK2* isoform (all iso) and isoform 3 RT-PCRs (**figure 3.15**). The RT-PCR observational analysis was highly convoluted, with a qPCR approach being necessary for concise comparison of relative expression differences between the different groups, however due to time constraints this was not possible and so the RT-PCR approach was presented as preliminary data only. The CRIPSR data presented here demonstrated cutting edge genetic manipulation techniques for studying the functions of reference SVA retrotransposons which is supplemented by reporter gene assay data to provide further evidence of the functional implications of SVA elements.

Chapter 4 – Understanding the SVA retrotransposon architecture of the PD associated locus *INPP5F/BAG3/TIAL1*

Chapter 4 – Understanding the SVA retrotransposon architecture of the PD associated locus *INPP5F/BAG3/TIAL1*

4.1 Introduction

The data presented in chapter 3 demonstrated how SVA elements may possess functional properties in a key PD gene (*LRRK2*). To broaden the understanding of the role of SVA retrotransposons in PD, a second GWAS nominated PD locus was targeted, the *INPP5F*/*BAG3*/*TIAL1* locus, which contained two SVA elements. These SVAs were located in different genomic positions within the locus, one adjacent to the *INPP5F* promoter and a second within intron 6 of *INPP5F*, which allowed for multiple facets of potential gene regulatory functions to be assessed in different contexts.

The *INPP5F*/*BAG3*/*TIAL1* locus on chromosome 10 (q26.11) was first identified as a novel PD related locus via GWAS analysis by Nalls *et al* 2014 [45]. The most significant GWAS SNP (rs117896735) in the region identified had a meta p-value of 1.21×10^{-11} and was considered as a risk locus for PD. Subsequently, a second larger GWAS analysis has been carried out by the same group which updated the meta p-value for this SNP to 2.36×10^{-28} suggesting that this locus had a strong genetic association with PD [26]. The Nalls *et al.* 2019 meta-analysis also performed a gene nomination process using expression quantitative trait loci (eQTL) data and Mendelian randomisation which correlated the GWAS SNPs with expression of neighbouring genes, to predict which genes, if any, were likely causative of the increased risk detected by the GWAS signals within any given loci. Within the *INPP5F*/*BAG3*/*TIAL1* locus, both *INPP5F* and *BAG3* were reported as potential risk genes with the locus

extending to include *TIAL1*. As is often the case with GWAS analysis, identification of causative risk variation was not identified within the *INPP5F*/*BAG*/*TIAL1* locus in the Nalls *et al.* 2019 meta-analysis and therefore requires further investigation.

To understand the *INPP5F/BAG3/TIAL1* locus in the terms of regulatory domains, multiple features were evaluated including evolutionary conserved regions (ECR) and epigenetic marks. ECRs allow identification of key domains considered to be important due to the lack of divergence between species. Conservation of non-coding sequences has been a powerful technique for the identification of regulatory elements since the first sequencing data became available across multiple species [162]. In contrast with this, novel retroelement sequences produced from retro transposition-events throughout evolution has led to primate and human specific sequences that may hold novel regulatory potential for human specific signatures of gene regulation. Using conservation and key epigenetic marks, including histone modifications and CpG DNA methylation, potential sites for novel regulatory domains can be identified and explored further using molecular biology techniques such as reporter gene assays and CRISPR mediated deletions for *in vitro* analysis of function.

Within this chapter, the characterisation of two SVA retrotransposons within the *INPP5F*/*BAG3*/*TIAL1* locus was undertaken in order to provide further understanding of the roles of SVA retrotransposons as novel sources of regulation influencing PD related genes. To do this, various approaches were taken including bioinformatic analysis of the locus, reporter gene functional assays, generation of CRISPR mediated SVA knockout HEK293 cell lines and genotyping of potential polymorphisms in a PD case/control cohort. The data reported in this chapter aimed to provide evidence to

support the importance of studying retrotransposable elements within neurodegenerative disorder related genes. In **section 1.5.5**, **table 1.2**, the proportion of the current 90 GWAS nominated PD risk genes which contained SVAs was found to be 23% with a total of 31 SVA elements being located within 100kb of the reported GWAS SNP. Following the results within the analysis presented within this chapter, SVA elements should be considered as potential novel regulators within the other GWAS nominated PD risk loci which also contain SVA elements.

4.1.1 Aims and hypothesis

- To test functional relevance of both the SVA-F and SVA-D elements within the context of *INPP5F* expression and their regulatory properties *in vitro* using reporter gene assays and CRISPR technologies.
- To identify and assess polymorphic domains within the SVA elements and correlate potential differences in genotype and allele frequencies in a case/control Parkinson's disease cohort.
- To critically evaluate novel PD GWAS signals within SVA elements in the *INPP5F* loci.

Hypothesis – SVA elements within the *INPP5F/BAG3/TIAL1* locus have regulatory properties and can influence gene function of key PD genes. Furthermore, polymorphisms within primary sequence confer risk for Parkinson's disease and should be considered as novel PD risk factors.

4.2 Results

4.2.1 Bioinformatic analysis of *INPP5F/BAG3/TIAL1* locus

The human *INPP5F/BAG3/TIAL1* locus covers a genomic span of approximately 250kb on chromosome 10 band 10q26.11 (hg38 – chr10: 119570493-119829293). Using both the UCSC genome browser and ECR browser it was possible to identify potential regulatory domains in the locus. **Figure 4.1** shows the histone marks (H3K27Ac – acetylation – active enhancers, H3K4Me1 – mono-methylation – active or primed enhancers, H3K4Me3 – tri-methylation – promoters) and GWAS SNPs from UCSC aligned with ECR data. There were strong histone modification peaks across all tested cell lines surrounding the promoters of *INPP5F*, *BAG3* and *TIAL1* primarily tri-methylation and acetylation marks directly over the promoter with flanking mono-methylation marks which suggested expression of these genes across multiple tissue types. Cell lines and derivations included: GM12878 – B-lymphocytes, H1-hESC – embryonic stem cells, HSMM – skeletal muscle myoblasts, HUVEC – umbilical vein endothelial cells, K562 – bone marrow from CML patient, NHEK – epidermal keratinocytes and NHLF –primary lung fibroblasts. The histone 3 lysine 4 tri-methylation mark (H3K4Me3) tracks show that *INPP5F*, *BAG3* and *TIAL 1* all expressed strong epigenetic marks at the 5’ termini in all of the aforementioned cell lines, indicating these promoters were active and were likely to be expressed across this wide variety of tissue types.

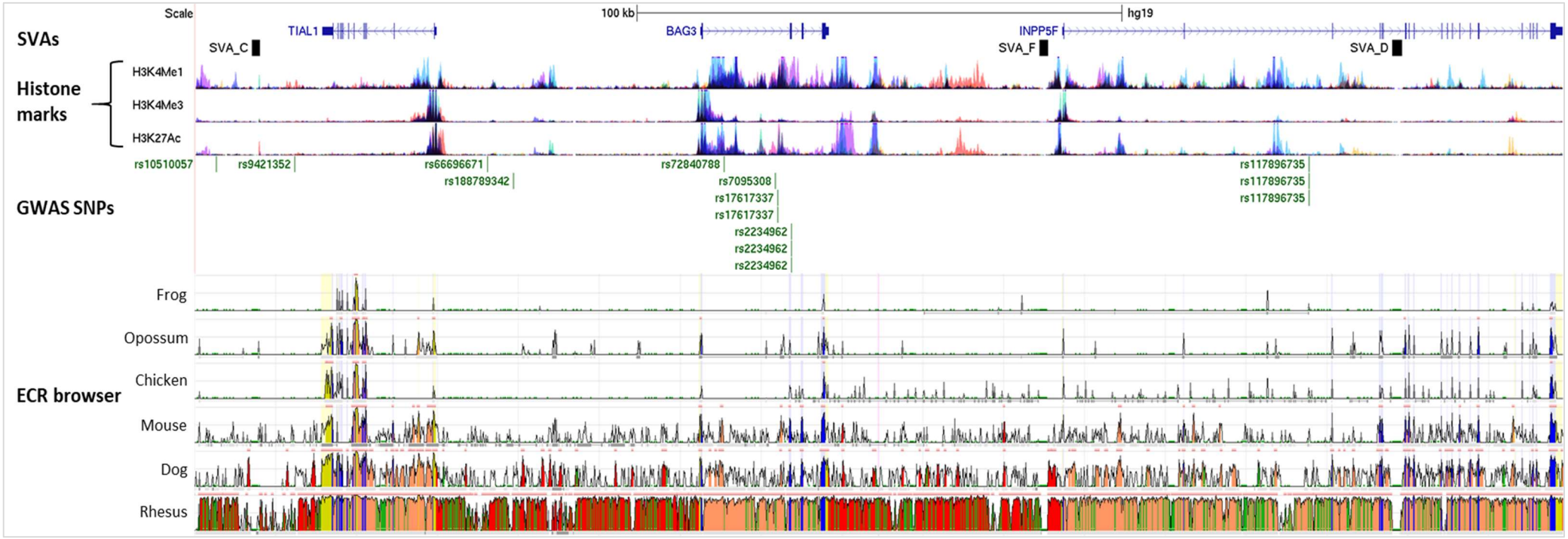

Scale
100 kb
hg19
SVAs
TIAL1
BAG3
INPP5F
SVA_C
SVA_F
SVA_D
Histone marks
H3K4Me1
H3K4Me3
H3K27Ac
GWAS SNPs
rs10510057
rs9421352
rs66696671
rs188789342
rs72840788
rs7095308
rs17617337
rs17617337
rs2234962
rs2234962
rs2234962
rs117896735
rs117896735
rs117896735
ECR browser
Frog
Opossum
Chicken
Mouse
Dog
Rhesus

Figure 4.1 – Genome and ECR browser view of the *INPP5F/BAG3/TIAL1* loci indicating the SVA architecture. Three SVAs were identified in this locus including an SVA-C ~13kb downstream of *TIAL1* 3' UTR, an SVA-F ~3kb from the 5' UTR of *INPP5F* and an SVA-D within intron 6 of *INPP5F*. Histone epigenetic marks from several cell lines indicated sites of putative regulatory domains or promoters (H3K27Ac – active enhancers, H3K4Me1 – active or primed enhancers, H3K4Me3 – promoters) and the green GWAS SNP tracks show genome wide significant hits for a variety of diseases/traits (listed in figure 5.2.2). The ECR browser panel highlighted core ECR's defined as a sequence with at least 77% homology between species and a minimum length of 350bp. Peaks with no colour did not meet this requirement but shared some homology. The height of the peaks indicates the percentage of homology between the species and the human genome. Colour scheme for histone marks indicate different cell lines and were as follows: GM12878 – salmon, H1-hESC – yellow, HSMM – light green, HUVEC – light blue, K562 – dark blue, NHEK – purple and NHLF – pink. The colour scheme within the ECR track indicates different genomic features and are as follows: Blue – exons, yellow – UTR, salmon – intronic regions, red – intragenic regions, green – transposons and simple repeats.

Using ECR browser to identify ECRs across multiple species, (human, Rhesus macaque, dog, mouse, chicken, opossum and frog) potential regions of interest were identified (**figure 4.1**). More highly conserved regions appeared as higher peaks across a wider range of species beyond primate species. There was a high degree of conservation within primate species as demonstrated in the comparison between human (base genome not displayed within the graphic) and Rhesus macaque track and confirmed using the Multiz conservation alignment track on UCSC shown in **figure 4.2.** There were no major core ECR's, defined as sequences that share a minimum of 77% homology and are at least 350bp in length, identified across the *INPP5F/BAG3/TIAL1* region which the exception of exons and un-translated regions (UTR's), suggesting that this region has a high rate of divergence of non-coding sequences. Given the lack of non-coding ECR's in the region, focus was given to

unique human specific sequences including SVA retrotransposons. **Figure 4.1** indicates three SVA elements in this locus, an SVA-C downstream of *TIAL1*, an SVA-F adjacent to the *INPP5F* promoter and an SVA-D within an intron of *INPP5F*. SVAs inherently do not share conservation with any other species other than primates and so do not contain ECRs so they appear as a flat line on the ECR track.

GWAS SNPs in the locus indicate genetic variants that are commonly found within populations afflicted with specific diseases. The GWAS track in **figure 4.1** highlights nine SNPs which have been found across various diseases/traits including Parkinson's disease, heart disease/function, depressive symptoms and height. The exhaustive list to date is described in **table 4.1**. The most statistically significant GWAS SNP in the loci is rs117896735 ($p - 2x10^{-28}$) which is strongly associated with PD and found within intron 2 of *INPP5F*. Four independent studies (Nalls *et al.* 2014, Nalls *et al.* 2019, Pickrell *et al.* 2016 and Chang *et al.* 2017) identified three genome wide significant PD SNPs across this locus which improved the confidence that the locus was genetically associated with PD [26, 45, 147, 163]. It is also important to note, that the GWAS signals identified did not identify causative risk alleles and may be tagging other regions that would be important such as regulatory domains.

Table 4.1 – List of GWAS SNPs with associated disease/trait and reported gene from the *INPP5F*/*BAG3*/*TIAL1* loci highlighted in **figure 5.2.1**. Highlighted in green is the most significant SNP for Parkinson's disease and is the primary reason for the loci's association with PD as of the time of writing. The SNPs highlighted in red indicate those that fall underneath the standard genome wide significance threshold p value – $5x10^{-8}$ and should not be considered as true GWAS signals. Reported genes are proposed by either QTL analysis or defined by the nearest gene and are supplied from the study if such analysis had been performed.

GWAS SNP	Disease or trait	Reported genes	Meta p-value	Study
rs10510057	Depressive symptoms	RGS10, TIAL1	$4x10^{-8}$	Otowa T et al. 2016
rs9421352	Breast cancer	N/A	$1x10^{-7}$	Michailidou K et al. 2017
rs66696671	Response to carboplatin and paclitaxel in ovarian cancer	TIAL1	$7x10^{-7}$	Fridley BL et al. 2016
rs188789342	Parkinson's disease	TIAL1, BAG3	$3x10^{-11}$	Pickrell JK et al. 2016
rs72840788	Parkinson's disease	BAG3	$2x10^{-11}$	Nalls MA et al. 2019
rs7095308	Systolic blood pressure	N/A	$6x10^{-9}$	Kichaev G et al. 2018
rs17617337	Heart failure	BAG3	$4x10^{-9}$	Shah S et al. 2020
rs17617337	Diastolic blood pressure	BAG3	$3x10^{-9}$	Hoffmann TJ et al. 2016
rs2234962	Lung function (FEV1/FVC)	N/A	$3x10^{-9}$	Kichaev G et al. 2018
rs2234962	Height	N/A	$3x10^{-27}$	Kichaev G et al. 2018
rs2234962	Idiopathic dilated cardiomyopathy	BAG3	$4x10^{-12}$	Villard E et al. 2011
rs117896735	Parkinson's disease	INPP5F	$2x10^{-28}$	Nalls MA et al. 2019
rs117896735	Parkinson's disease	BAG3	$2x10^{-19}$	Chang D et al. 2017
rs117896735	Parkinson's disease	INPP5F	$1x10^{-11}$	Nalls MA et al. 2014

Given the absence of non-coding ECRs and the proximity of two of the SVAs, the SVA-F and SVA-D elements, identified to the major PD GWAS SNP (rs117896735) it was decided to assess the potential regulatory role of the SVAs in the *INPP5F*/*BAG3*/*TIAL1* locus. SVAs can be categorised into sub-families based on the point in evolution in which these sequences inserted into the genome with SVA-D being evolutionarily 'older' (~9.6 Myrs) and the F being 'younger' (~3.2 Myrs) [70]. These two elements pose different potential mechanisms of action given their difference in genomic location with respect to *INPP5F*. The SVA-F's position proximal to the *INPP5F* promoter gives it the potential to bind transcription factors (TF's) and directly

influence the properties of *INPP5F* expression. It could also be important in chromatin structure changes between hetero/euchromatin upon stimulation which could increase or reduce expression. The SVA-D could have an impact on expression levels via binding of TF's but could also be involved with regulation of splicing given its position within an intron. As highlighted in **table 1.1**, SVA elements have previously been shown to alter multiple functions regarding gene regulation.

Figure 4.2 shows the *INPP5F* gene in more detail highlighting three isoforms of *INPP5F* and the two SVAs that were focused on for further study for potential function. Both the SVA-F and SVA-D elements illustrated in **figure 4.2** are human specific, indicated by the clear break in homology within the Multiz alignment conservation track which shows multiple species of closely related primates and more distantly related rodents. A high level of conservation is illustrated by black lines/blocks, with the majority of the *INPP5F* loci being strongly conserved across the closely related primate species shown (chimp, gorilla, orangutan and gibbon) with the only exceptions to this being the SVA elements themselves. Conservation across non-coding *INPP5F* sequences was significantly reduced in rodents (mouse and rat) which was also reflected in the ECR data in **figure 4.1**.

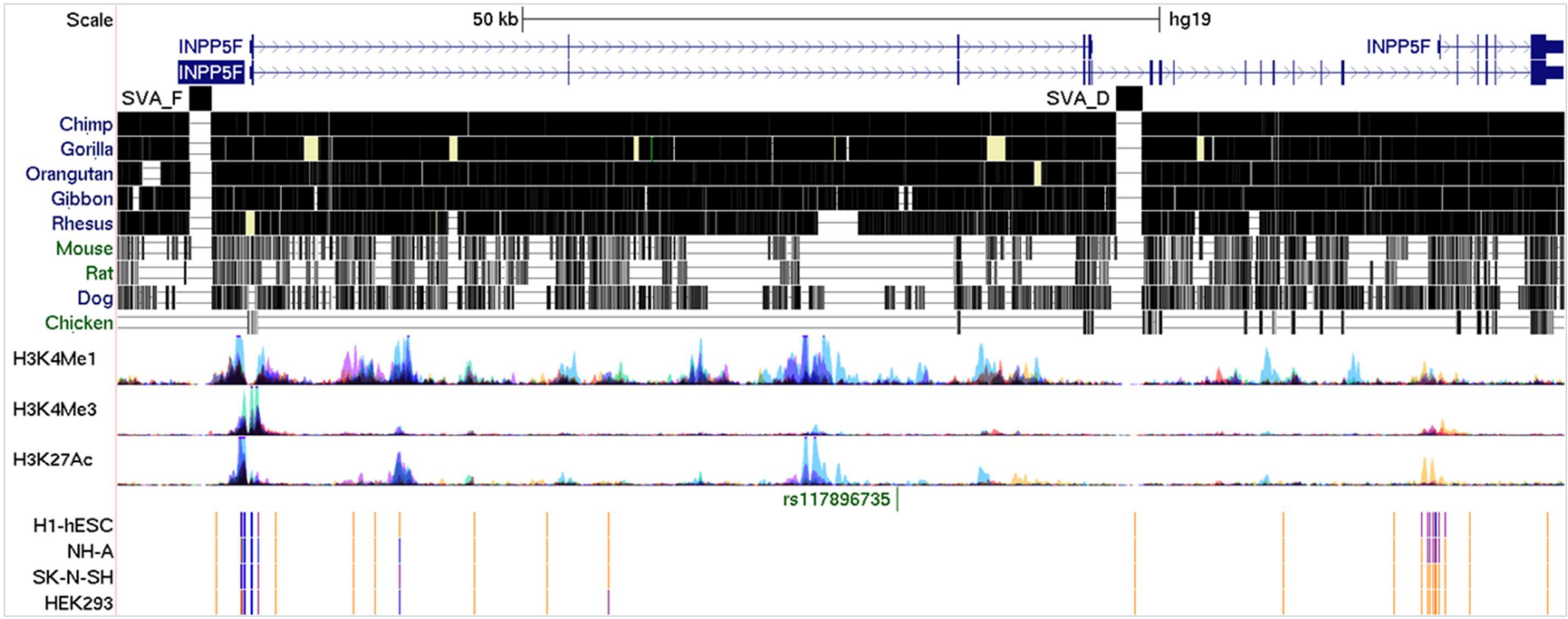


Figure 4.2 – *INPP5F* loci adapted from UCSC genome browser (Hg19) - chr10:121474988-121588682. The two SVA elements (F and D) are human specific indicated by the break in the Multiz alignment conservation track. Layered histone epigenetic marks indicate sites of regulatory domains or promoters (H3K27Ac – active enhancers, H3K4Me1 – active or primed enhancers, H3K4Me3 – promoters) and CpG 450K DNA methylation probe array indicated across four cell lines (orange – methylated, purple – partial methylation, blue – unmethylated). There were seven cell lines used in the histone mark tracks from Encode indicated by different colours: GM12878 – salmon, H1-hESC – yellow, HSMM – light green, HUVEC – light blue, K562 – dark blue, NHEK – purple and NHLF – pink.

The ENCODE histone modification track shows, in greater detail than **figure 4.2,** the regions within the *INPP5F* loci with strong epigenetic marks. The separate colours represent different cell lines, with the darker colours inferring overlap of cell lines. There were strong dark peaks around the promoter of the main isoform (the longest isoform) indicating the major transcriptional start site (TSS) and promoter related enhancer elements with a more open chromatin structure. The SVA-F element is situated 3kb upstream of the major TSS for *INPP5F* which would allow for recruitment of transcription factors to enhance or reduce *INPP5F* expression or alter chromatin structure to similar effect. The proximal promoter here was defined as the sequence immediately prior to the TSS and covered a ~250bp span.

The ENCODE CpG methylation 450K probe array in **figure 4.2** indicated CpG sites that were either methylated (orange), partially methylated (purple) or unmethylated (blue). The data shown includes four different cell lines of human origin from different tissue types: human embryonic stem cells (H1-hESC), astrocyte cells (NH-A), neuroblastoma cells (SK-N-SH) and embryonic kidney cells (HEK293). These cell lines were chosen for the illustration because they approximate most closely to the cell lines used for functional assays to test SVA function in this thesis including iPSC's (similar to hESCs), SH-SY5Y (subline of the SK-N-SH parental cell line) and HEK293 cells. NH-A are a primary normal human astrocyte cell line and were included as they are the only cell line within the CpG array that were taken directly from brain tissue. Major sites of hypomethylation (blue) lie around the *INPP5F* 5' promoter indicating that in these cell lines, *INPP5F* is likely expressed. H1-hESC and NH-A cell lines show a mostly unmethylated state for the shorter *INPP5F* isoform downstream of the SVA-

D whilst SK-N-SH and HEK293 show methylated states suggesting the short *INPP5F* isoform is likely expressed in brain tissue/astrocytes and undifferentiated cells.

Genotype-tissue expression project (GTEx) data was used in conjunction with CpG methylation data to assess which isoforms of *INPP5F* were differentially expressed between tissue types. The GTEx data is described as reads per kilobase of transcript, per million mapped reads (RPKM) and is used to normalise differences in RNA transcript length to account for more reads coming from longer transcripts. For all further analysis, the *INPP5F* isoforms have been labelled as illustrated in **figure 4.3 – A**, with the full-length isoform (iso 1), a shorter 3' transcript (iso 2) and an isoform with a truncated carboxy-terminus (iso 3). The full-length *INPP5F* isoform 1 codes for a protein which contains both the catalytic SAC and hSac2 domains named Sac2. Isoforms 2 and 3 encode proteins that contain partial SAC and hSac2 domains respectively with physiological relevance of these truncated proteins being unknown.

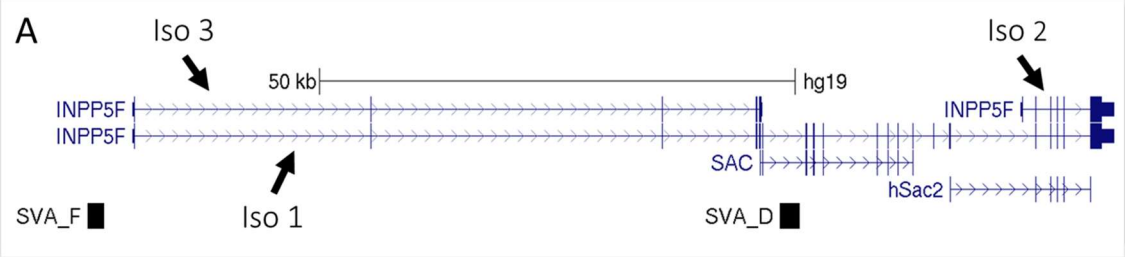


Figure 4.3 – A – Illustration of the three coding transcript variants and associated protein coding exon domains of *INPP5F* labelled iso 1, 2 and 3 to be referred to in further analysis. *INPP5F* contains a catalytic phosphatase protein domain (SAC) and a Sac2 homology domain (hSac2). Accession codes: isoform 1 - NM_014937, isoform 2 - NM_001243194, isoform 3 - NM_001243195.

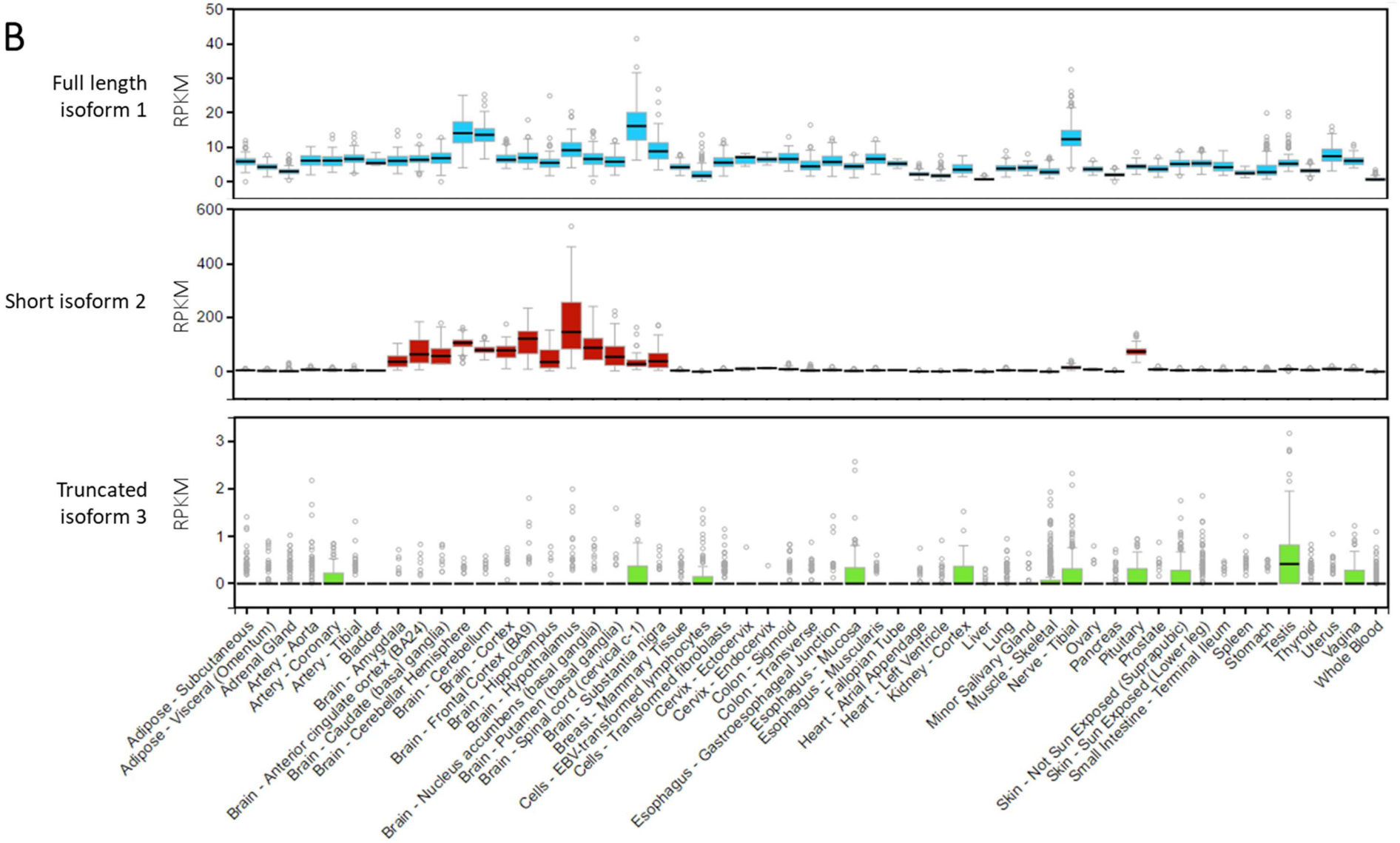

B
Full length isoform 1
Short isoform 2
Truncated isoform 3
RPKM
50
40
30
20
10
0
600
400
200
0
3
2
1
0
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

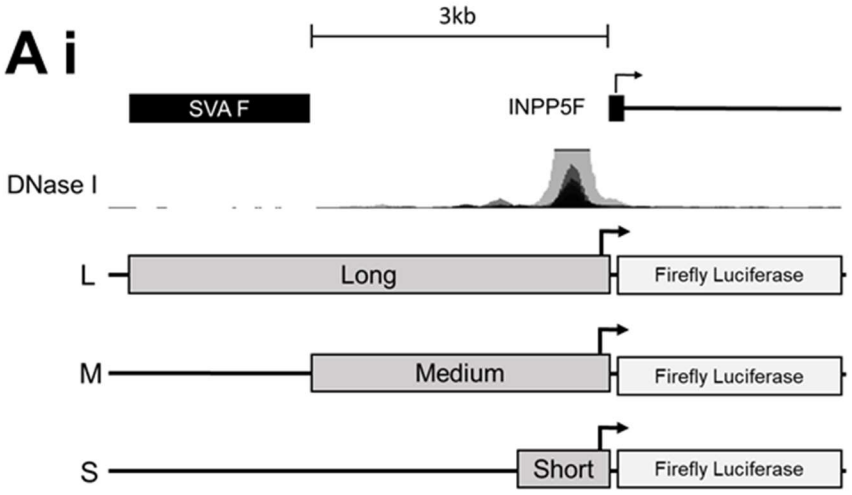
Figure 4.3 – B – Isoform expression data for *INPP5F* taken from the GTEx database which uses Illumina RNA-seq and Affymetrix arrays to sample across a wide range of human tissue types under basal conditions. Isoform 1 of *INPP5F* is broadly ubiquitously expressed. Isoform 2 had negligible expression in many of the analysed tissues, however it had the highest expression levels within all analysed brain regions, pituitary and tibial nerve of the three isoforms studied. Isoform 3 was reported as having little no expression with average RPKM values <1 across all tissue types.

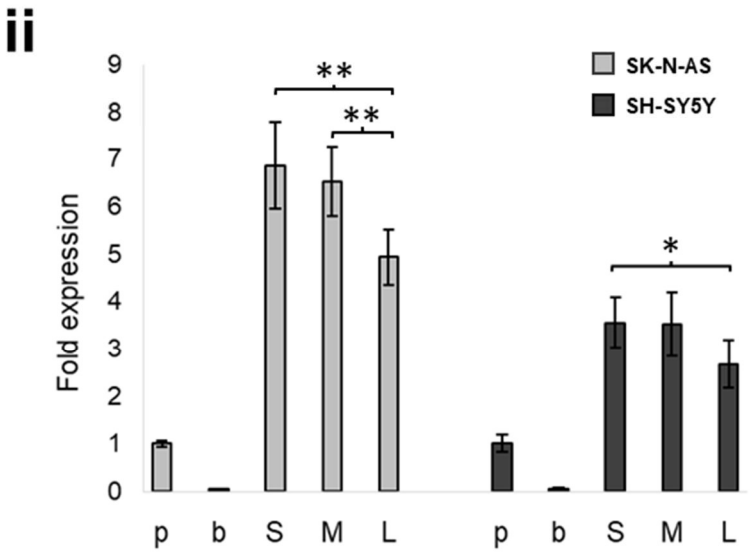
The GTEx data in **figure 4.3 – B** indicated that the full-length isoform of *INPP5F* is ubiquitously expressed at relatively low levels with the median expression across all tissues <20 RPKM. The highest median expression for isoform 1 was observed in the cerebellar hemispheres (15.2 - RPKM), cerebellum (14.2 - RPKM), spinal cord – cervical C-1 (10.5 - RPKM) and the tibial nerve (11.7 - RPKM). This suggested an important universal function for Sac2 across most cell types which was consistent with the published literature regarding its role in major signalling pathways such as AKT and STAT signalling. Interestingly, the shorter isoform 2 was more highly expressed within various regions of the brain including the substantia nigra (19.8 RPKM) giving it potential importance for Parkinson's disease. The highest median expression levels for isoform 2 were observed in the cerebellar hemispheres (87.0 RPKM), hypothalamus (80.6 RPKM) and the frontal cortex (79.1 RPKM), approximately 5.7, 9.0 and 11.5 times higher than that of isoform 1, respectively. The truncated isoform 3 appears to have little to no expression across all tissue types with all median RPKM values <1. This isoform has been reported as a coding, validated transcript within Encode, and so has been included in these analyses but appears to not be expressed under basal conditions. It is possible that this isoform is only expressed under specific conditions which would explain why it is not present in GTEx data which does not include tissues in response to a stimulus.

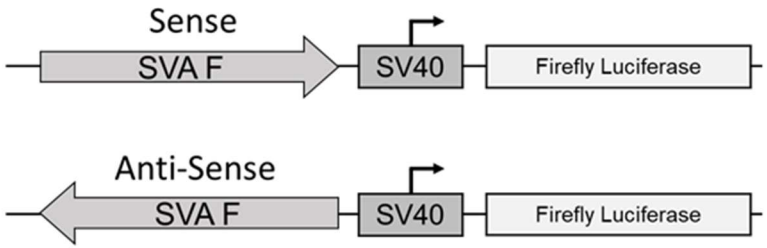
Given the close proximity of the SVA-F element to the *INPP5F* proximal promoter, it was hypothesised that this element has regulatory potential for the expression of *INPP5F* most likely via TF binding or changes to chromatin structure. Due to its location, it was hypothesised that the intronic SVA-D may have a role in alternative splicing and expression of differential isoforms, however it could also have a role in promoter related activity. This is due to the dynamic nature of DNA coiling within chromatin structure, thus the intronic SVA-D could be located proximal to the *INPP5F* promoter where it could recruit TFs to influence gene expression even though when viewed on a linear genome browser it appears to be located approximately 68kb away from the major promoter. This property of chromatin structure gives regulatory domains the ability to recruit TFs that can impact gene expression over far-reaching genomic distances in excess of 100,000 kb and even between chromosomes using mechanisms such as CCCTC-binding factor (CTCF) driven DNA looping and transcription factories [164, 165].

4.2.2 *INPP5F* SVA luciferase reporter assays

In order to assess the regulatory roles of the above SVA elements on *INPP5F* gene expression *in vitro*, several luciferase-based reporter gene constructs were designed and generated. Given the difference in genomic location of the SVA-F (promoter) and SVA-D (intronic) elements, different models were established to test distinct functions. Standard commercially available pGL3 based vectors (Promega) were used to evaluate the SVA-F element in two different contexts: within the context of the endogenous *INPP5F* promoter and an exogenous SV40 minimal promoter. pGL3b (basic) vectors were utilised to assess the SVA-F element within the context of the *INPP5F* promoter by cloning three different sections of the promoter region to both include and exclude the SVA (**figure 4.4 - Ai**). Construct S (short) contained an 860bp fragment from the *INPP5F* isoform 1 (accession code: NM_014937) 5' UTR (hg38 - chr10:119726050 -698/+162bp) and encompassed the major DNase I hypersensitivity peak from ENCODE which contained the core promoter necessary for *INPP5F* expression. DNase I hypersensitivity sites indicate regions of chromatin that have lost their condensed structure and are more accessible to TFs which also allow these areas to be cleaved by DNase I. These sites are generally associated with cis-regulatory elements and often highlight key areas around promoters which are crucial for expression. Construct M (medium) contained a 3097bp insert (hg38 - chr10:119726050 -2935/+162bp) which includes the short construct sequence with additional 5' flanking sequence up to, but not including the SVA-F element. Construct L (long) contained a 4894bp insert spanning the *INPP5F* promoter and adjacent SVA-F element (hg38 - chr10:11972605 -4732/+162bp).


A i
3kb
SVA F
INPP5F
DNase I
L
Long
Firefly Luciferase
M
Medium
Firefly Luciferase
S
Short
Firefly Luciferase


ii
SK-N-AS
SH-SY5Y
Fold expression
0
1
2
3
4
5
6
7
8
9
**
**
*
p
b
S
M
L
p
b
S
M
L


B i
Sense
SVA F
SV40
Firefly Luciferase
Anti-Sense
SVA F
SV40
Firefly Luciferase

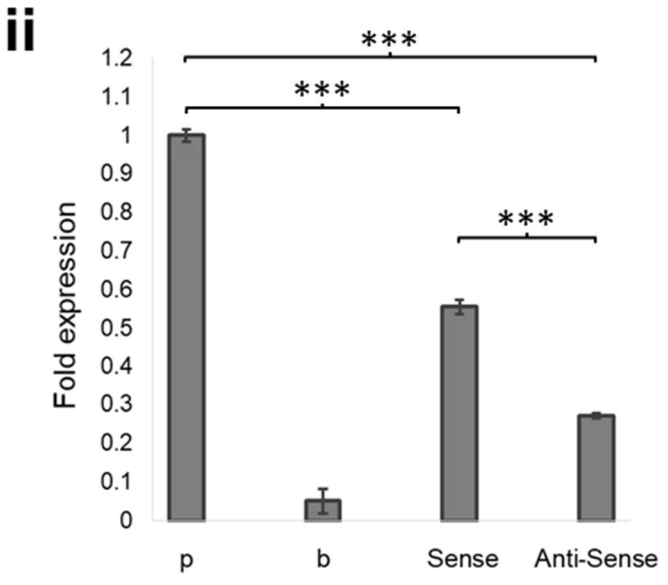

ii
Fold expression
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
***
***
***
p
b
Sense
Anti-Sense

Figure 4.4 – *INPP5F* SVA F elicits repressive effects on firefly luciferase expression within reporter gene constructs. (Ai) Reporter gene constructs were generated which contained composite parts of the *INPP5F* promoter with and without the SVA present. Construct L (long) contained a 4894bp insert encompassing the *INPP5F* isoform 1 (NM_014937) 5' transcriptional start site (TSS) and the SVA F element (hg38 chr10:11972605 -4732/+162bp), Construct M (medium) contained a 3097bp insert spanning the *INPP5F* promoter from the 5' TSS up to but not including the SVA-F element (hg38 chr10:119726050 -2935/+162bp) and construct S (short) contained a 860bp insert from the 5' *INPP5F* TSS with the core promoter (hg38 chr10:119726050 -698/+162bp) which included the major DNase I hypersensitivity peak from ENCODE. **(ii)** These constructs were expressed in SK-N-AS (light grey) N=4, and SH-SY5Y (dark grey), N=3 neuroblastoma cell lines and demonstrate significant repression of luciferase signal in the constructs containing the intact SVA in both cell lines. **(Bi)** Reporter gene constructs containing the *INPP5F* SVA F element in both sense and anti-sense orientations cloned upstream of the SV40 minimal promoter were generated. **(ii)** These constructs were tested in the SH-SY5Y neuroblastoma cell line, N=3. Fold expression is compared to pGL3p (SV40 minimal promoter) and normalised to a TK-Renilla internal control to account for transfection efficiency differences. A similar repressive effect was observed in both the sense and anti-sense orientations with a more pronounced effect in the anti-sense construct indicating a bi-directional response. *** P<0.001, ** p<0.01, *p<0.05. All technical replicates performed in quadruplicate (N=4) within each biological assay.

By comparing the regulatory effects of the three deletion constructs, (**figure 4.4 – A**) across two neuroblastoma cell lines (SK-N-AS and SH-SY5Y), inclusion of the SVA-F element elicited clear repressive characteristics, significantly repressing luciferase activity in both cell lines. A highly significant ($p<0.001$) 28% decrease in luciferase activity from the construct containing the SVA-F (long) was observed, compared to the short construct in the SK-N-AS cell line and a similar significant decrease ($p<0.05$) of 25% in luciferase activity within the SH-SY5Y cell line. This may suggest the primary sequence has the ability to recruit repressive TFs to reduce expression or the SVA has an effect on DNA structure to alter binding of TFs. This effect was confirmed by the use of SV40 driven pGL3p (promoter) vectors (**figure 4.4 – B**) which contained the cloned SVA-F element only, absent of any flanking *INPP5F* promoter related sequences in both the sense and anti-sense orientations with respect to the SV40 promoter. These constructs indicated that the repressive effect worked in a bi-directional way with strong repression ($p<0.001$) seen in both the sense and anti-sense directions with decreases of ~45% and ~73% respectively compared to the pGL3p backbone. The sense orientation of the SVA is defined as (5' > 3') the CT element at the 5' and poly-A tail at the 3' end [166]. For context, the SVA-F in the genome is present in the sense orientation with respect to the *INPP5F* promoter. This data showed that the SVA-F element elicited significant repressive effects on *INPP5F* promoter activity (**figure 4.4 – A**) and suggested this effect would work in both orientations (**figure 4.4 – B**).

To test the regulatory capabilities of the intronic SVA-D element a different strategy was employed to measure the effect of an SVA within an intron and infer potential effects on splicing (**figure 4.5 – A**). A non-commercially available plasmid (gifted from

Professor Gerald Schumann, Paul Ehrlich Institute) termed pSHM06 was used which contains an intron into which the SVA was cloned between two exons of the triose phosphate isomerase gene (TPI). The pSHM06 vector contains a Renilla luciferase reporter gene which requires the correct splicing of the intron in order to generate the Renilla protein and produce a detectable signal within this assay. Sequences that disrupt the splicing of the intron would result in less Renilla protein and a lower luciferase signal. These vectors have been utilised previously to measure the effects of LINE-1 insertions within various introns and the effects on mammalian gene expression within human induced pluripotent stem cells (hiPSCs) [167]. To allow for cloning of SVA elements into the intronic sequence of pSHM06, suitable restriction enzyme sites were required. To produce these, modification of the pSHM06 plasmid was performed in-house by replacing the existing intronic sequence with a synthetic one which contained appropriate restriction sites for downstream cloning strategies (plasmid details in **section 2.1.7**).

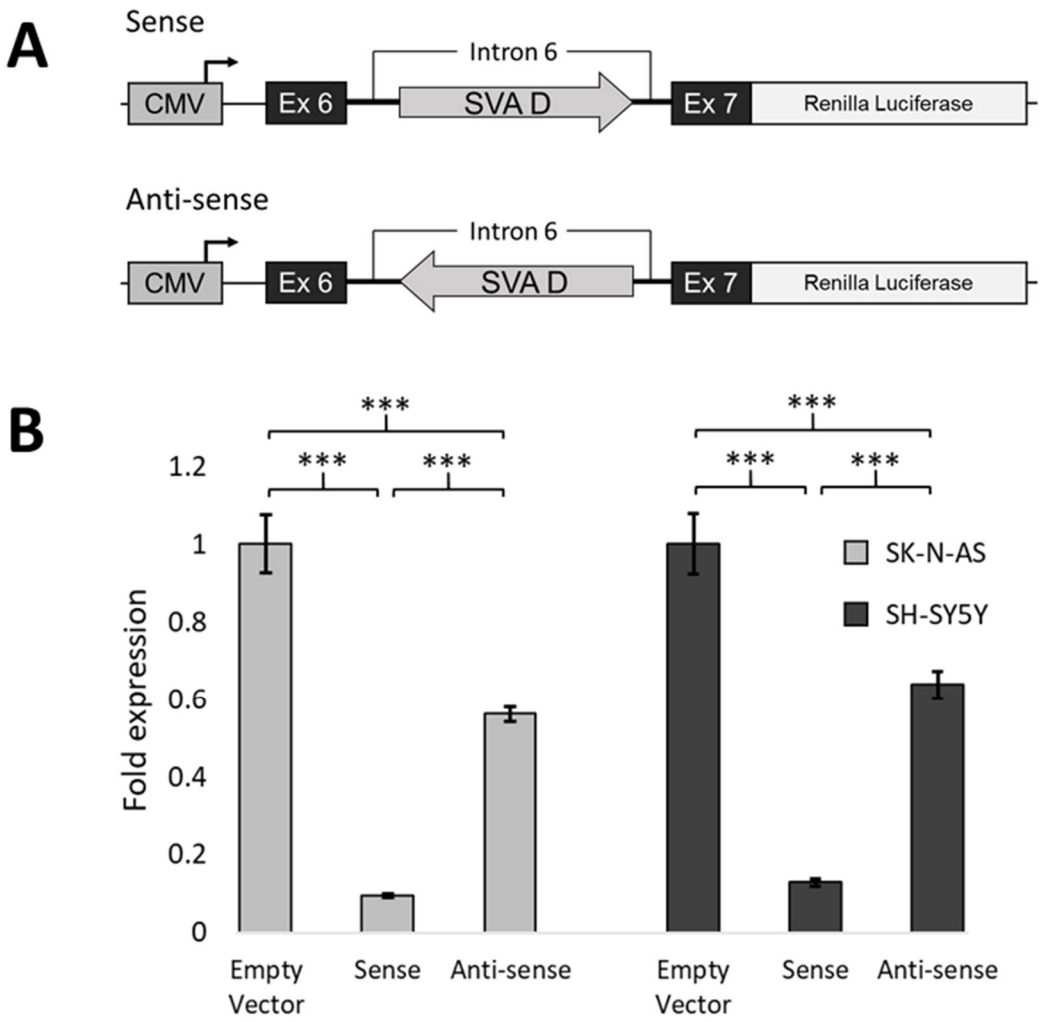


Figure 4.5 – Luciferase reporter gene expression assay indicating the effect of orientation of the *INPP5F* SVA D element on reporter gene expression in SHSY5Y (dark grey) and SK-N-AS (light grey) cell lines. (A) Schematic representation of the generated pSHM06 constructs indicating the SVA-D cloned within the triose phosphate isomerase (TPI) intron in two orientations termed sense and anti-sense with respect to the CMV promoter. **(B)** Both constructs were transfected in SK-N-AS and SHSY5Y neuroblastoma cells to test the effect of the SVA on Renilla luciferase expression. In the sense orientation the SVA strongly repressed luciferase expression by 87% in SK-N-AS and 90% in SH-SY5Y with a lesser effect observed in the anti-sense orientation with repression of luciferase signal of 36% in SK-N-AS and 44% in SH-SY5Y cells. SK-N-AS assay biological replicates N=2 with technical replicates N=4, SH-SY5Y assay biological replicates N=3 with technical replicates N=4. ***p<0.001.

The SVA-D in this model elicited statistically significant ($p<0.001$) repressive characteristics on luciferase activity in both the sense and anti-sense orientations **(figure 4.5)**. Strong repression was observed in the sense orientation, with 87% and 90% reductions of luciferase signal measured in the SK-N-AS and SH-SY5Y cell lines respectively. A lesser, but still highly significant repression effect was also observed in the anti-sense constructs with 36% and 44% reductions of luciferase observed in the SK-N-AS and SH-SY5Y cell lines respectively suggesting this effect works in a bi-directional manner.

Having shown robust and repeatable repression of SVAs in established cell lines, the *INPP5F* SVA F pGL3p constructs were tested in induced pluripotent stem cells (iPSCs) derived from CD34+ cord blood and cells differentiated into a cortical neuron lineage to assess if SVA function is universal or harbours differential effects in different cell types. The use of iPSCs and neuronal lineage differentiated cells was chosen as an appropriate model to test SVA elements given previous published findings that showed activation of retrotransposable elements including LINE-1, *Alu* and SVAs during the reprogramming of iPSCs [167]. This model allowed further exploration of the potential cell-specific differences that may exist between established cell line cultures and neuronal cell models. Commercially available iPSC's (A18945) were used to culture and differentiate into a cortical forebrain neuron lineage to test the generated SVA/pGL3p constructs (methods outlined in **section 2.2.9**). Post day 25 of the cortical neuron differentiation, induced neurons and iPSCs under basal conditions were transfected with the *INPP5F* SVA-F/pGL3p constructs containing the sense and anti-sense SVA-F elements (**figure 4.6)**.

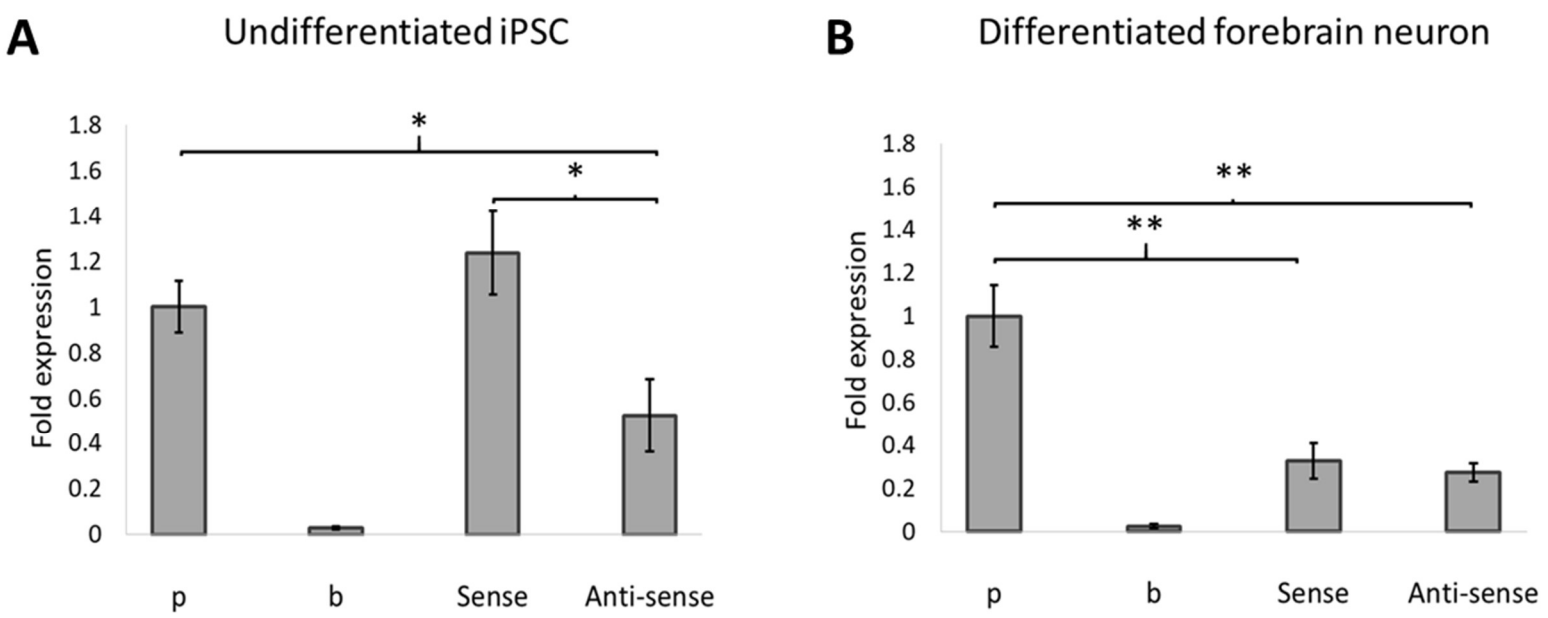


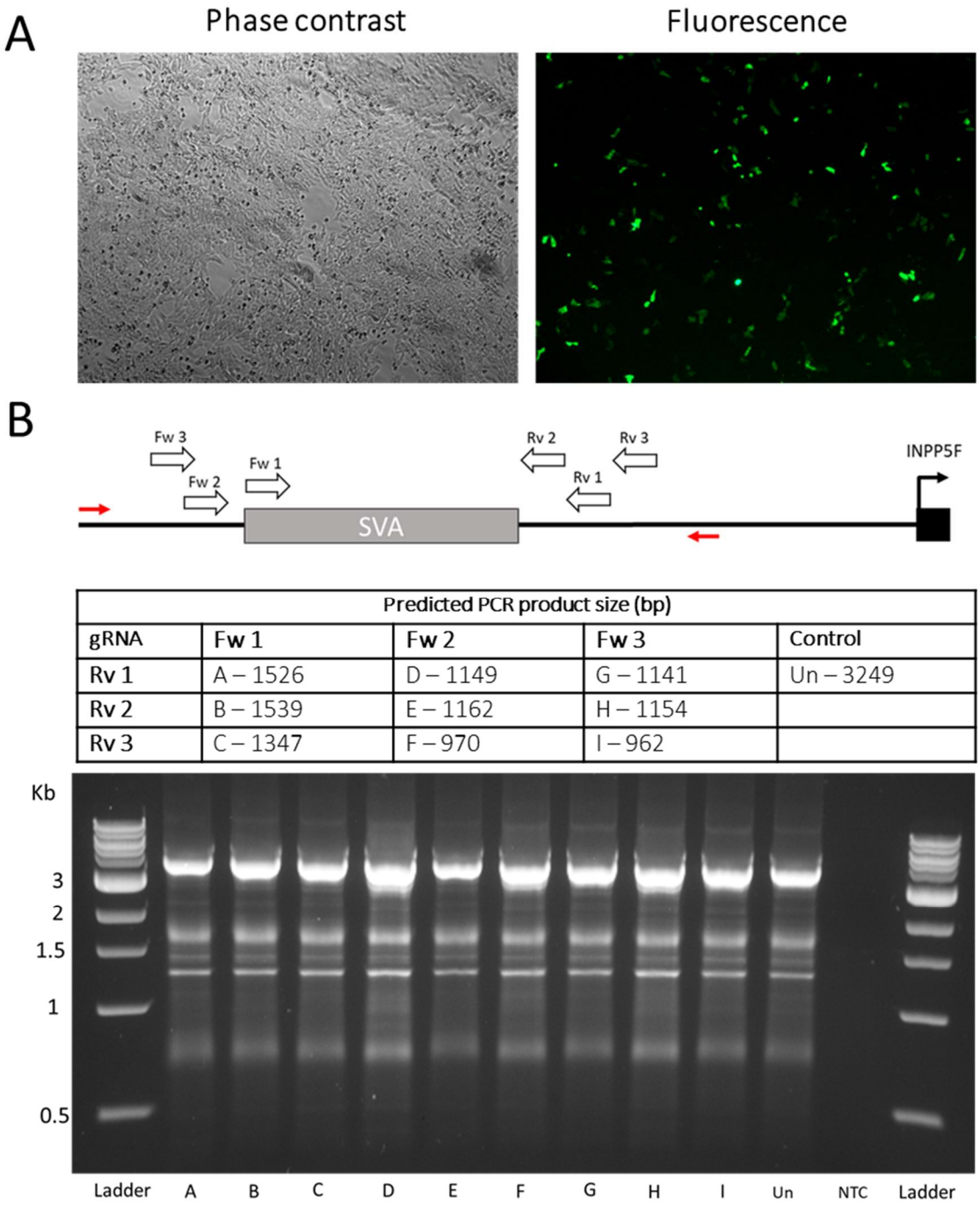
Figure 4.6 - *INPP5F* SVA-F harbours differential regulatory properties when tested in undifferentiated human iPSCs and 26 day differentiated iPSCs (forebrain cortical neuronal lineage). (A) Reporter gene constructs containing the *INPP5F* SVA F element in both sense and anti-sense orientations upstream of the minimal promoter SV40, expressed in undifferentiated human induced pluripotent stem cells (iPSCs). The anti-sense construct elicits repressive effects upon luciferase signal. **(B)** The same constructs transfected in differentiated forebrain neurons show that the SVA F in the sense orientation significantly represses luciferase in contrast with iPSCs. Fold expression is compared to the SV40 driven pGL3p (p) and normalised to a TK-Renilla internal control to account for transfection efficiency differences. Biological replicates N=1 with technical replicates N=4. ** $p<0.01$, *$p<0.05$

The SVA-F anti-sense construct within the iPSC model significantly repressed luciferase activity ($p<0.05$) by a 48% decrease of luciferase signal when compared to the SV40 minimum protomer driven pGL3p (**figure 4.6**). This was similar to, although to a lesser extent, the 73% decrease observed in the SH-SY5Y cell line model using the anti-sense construct (**figure 4.4 - B**). However, the SVA-F sense construct did not alter luciferase activity compared to pGL3p in the iPSC model. This differs from both the cortical neuron response in **figure 4.6-B** and the SH-SY5Y response in **figure 4.4-B** where the SVA-F sense construct significantly repressed activity in both models. This would suggest a difference in transcription factors present between the different cell models that would influence expression levels. In contrast, the same effect was observed in the differentiated cortical neuron model as seen in the established cell lines with strong levels of repression of 67% and 73% from the sense and anti-sense SVA-F constructs, respectively. Interestingly in this example, the difference in effect between the sense and anti-sense constructs was minimal, with no statistical difference between the constructs suggesting the same degree of repression in a neuronal context. However, this model was preliminary and only represents one biological replicate which would require further validation.

4.2.3 Analysing SVA function *in vitro* using CRISPR mediated knockouts

Having demonstrated that the SVA elements of *INPP5F* had function in luciferase-based cell assays, the function of the SVA elements *in vitro* was addressed by employing CRISPR technology to generate SVA-F and SVA-D deletion clonal cell lines and correlated with expression profiles (CRISPR outline detailed in **section-2.2.11**). All guides were designed which conformed to the necessary parameters outlined in **section 2.2.11.1**, which included minimal off target effects and presence of the appropriate PAM motif (NGG) to facilitate Cas9 cleavage. To test the feasibility of this approach, initial experiments were conducted in the near haploid cell line Hap1. Hap1 cells were chosen with the hypothesis that it would be simpler to detect a change in gene function upon deletion of the SVAs in a cell line which had only one copy of chromosome 10 in which *INPP5F* is located. This also allowed for an increased efficiency of CRISPR editing as only one allele would require knockout (KO) rather than the two in a diploid cell line.

Hap1 cell line transfection protocols were optimised by transfection of the EF1α-pSpCas9(BB)-2A-GFP, which contained a GFP cassette to indicate expression of the Cas9-GFP post-transfection, using varying concentrations of transfection reagent within the Turbofect transfection protocol (**section 2.2.9.4**) and assessed using UV light microscopy to observe the number of GFP positive cells and thus estimate the transfection efficiencies (**figure 4.7**). The optimal transfection ratio for these assays was determined to be 1µg plasmid DNA with 2µl Turbofect with a media change 4 hours post-transfection to improve cell viability. The cells were imaged 48 hours post transfection. Using these ratios, the transfection efficiency was estimated to be ~20%.



Predicted PCR product size (bp)				
gRNA	Fw 1	Fw 2	Fw 3	Control
Rv 1	A – 1526	D – 1149	G – 1141	Un – 3249
Rv 2	B – 1539	E – 1162	H – 1154	
Rv 3	C – 1347	F – 970	I – 962	

Figure 4.7 – Optimisation of the Hap1 CRISPR transfections to assess transfection and CRISPR editing efficiencies of the *INPP5F* SVA-F. **(A)** Indicates the expression of EGFP under UV light microscopy compared with total cell number observed with phase contrast imaging. Transfection efficiencies were estimated to be <20%. **(B)** Schematic representation of gRNA sequences (Fw – forward and Rv – reverse guides) used to excise the SVA-F element. Red arrows illustrate the approximate positions of the PCR primers used to analyse modification efficiency which were designed flanking the most extreme gRNAs. PCR of the *INPP5F* SVA-F region indicated approximate modification efficiency within Hap1 cell populations that had

been transfected with combinations of forward (Fw) and reverse (Rv) gRNAs. Clear bands at the unmodified target size of 3249bp can be seen in all conditions with no observable modification within any of the tested gRNA combinations suggesting an extremely poor editing efficiency. Predicted PCR amplicon sizes for correct modifications for each gRNA combination are listed in the table. PCR primers used are detail in **appendix 1**. Un – unmodified HEK293 cells.

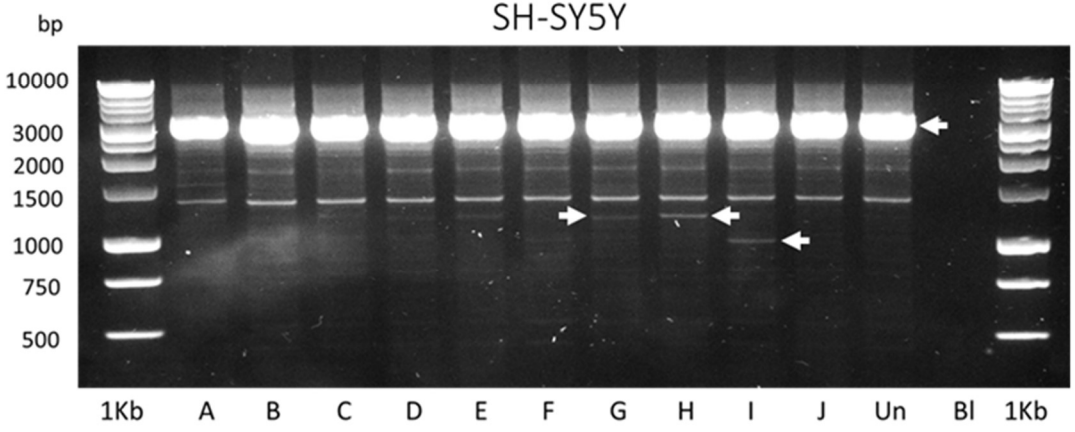
To assess CRISPR modification efficiency, Hap1 cells were transfected with combinations of three forward and three reverse guides located on the 5' and 3' flanks of the SVA-F element, respectively. Using PCR over the modified region, there was no observable modifications of the predicted amplicon sizes that would indicate correct a CRISPR modification (**figure 4.7 - B**). To improve the transfection and editing efficiencies, two alternative cell lines were investigated: the neuroblastoma cell line SH-SY5Y and embryonic kidney derived line HEK293. The SH-SY5Y cell line was selected to test SVA function *in vitro*, given the data generated from the luciferase-based cell assays. Furthermore, the SH-SY5Y cell line has been demonstrated as a suitable model for Parkinson's disease in previous literature [135]. The HEK293 cell line was also selected to test given the high transfection rates consistently observed within this cell line, with high survivability upon transfection and stable expression of recombinant proteins, making them an ideal candidate for these experiments. Furthermore, the HEK293 cell line exhibits markers of neuronal cells making them suitable as a neuronal model system [168].

Following an identical approach to the optimisation used in the Hap1 model (**figure 4.7**), SH-SY5Y cells were transfected with the EF1α-pSpCas9(BB)-2A-GFP plasmid and transfection efficiency was approximated to be ~50%, using UV imaging 48-hour post transfection. Transfection of the SH-SY5Y cells with the same guide combinations as

used in the Hap1 model (three forward and three reverse with nine total combinations), allowed identification of the optimal gRNAs to be taken forward. Guide RNA combinations H and I were identified that efficiently targeted the SVA-F element and produced a detectable PCR amplicon indicating the correct modification size. Both these pairs shared a common 5' gRNA (Fw 3) meaning that the guide effectively targeted the SVA-F element. However unfortunately, the overall modification efficiency in the SH-SY5Y cells was lower (estimated ~1%) than would be necessary to permit efficient clonal isolation of positive clones using gRNA combinations H and I (**figure 4.8 – A)**. The estimated modification efficiency was based on the ratio of unmodified to modified band intensity. Poor overall modification efficiency was likely due to the combination of both low transfection and editing efficiencies given that the transfection efficiency of 50% resulted in only ~1% modification efficiency as observed by PCR. Given this result observed within SH-SY5Y cells, the two highest efficiency gRNA combinations, H and I, were tested in HEK293 cells for comparison, which showed an improved modification efficiency of approximately 20-30%. The gRNA combination H (Fw 3, Rv 2) was selected to take forward as this combination targeted a more accurate removal of the SVA-F with less flanking sequence being excised compared to combination I. Note, in **figure 4.8**, the PCR amplicon sizes and off-target amplification differed between the cell lines tested due to the use of two different PCR primer sets because both the Hap1 and SH-SY5Y PCRs produced smears and non-specific amplification which was partially improved with re-designed PCR primers for the HEK293 line (primer details for each cell line in **appendix 1**).

A

Predicted PCR product size (bp)				
gRNA	Fw 1	Fw 2	Fw 3	Controls
Rv 1	A – 1526	D – 1149	G – 1141	J – 3249
Rv 2	B – 1539	E – 1162	H – 1154	Un – 3249
Rv 3	C – 1347	F – 970	I – 962	



B

Predicted PCR product size (bp)		
gRNA	Fw 3	Control
Rv 2	H – 1828	Un – 3926
Rv 3	I – 1640	

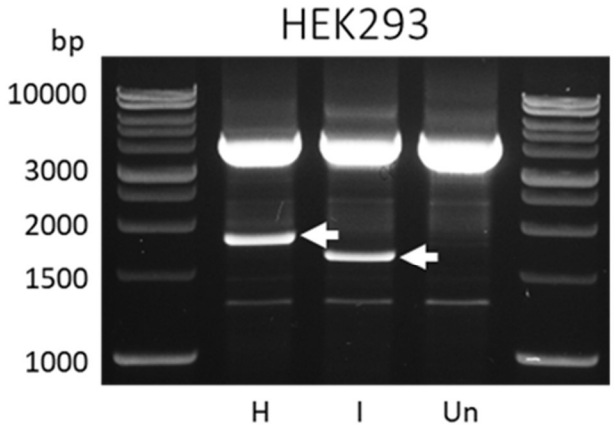


Figure 4.8 – Optimisation for guide RNA (gRNA) and cell lines to be used for subsequent CRISPR knockouts (KO) of *INPP5F* SVA-F. (**A**) Predicted PCR amplicon sizes for detection of CRISPR modified SH-SY5Y cells and associated PCR for each gRNA combination. White arrows indicate the correct modified amplicon sizes observed in samples G (1141bp), H (1154bp) and I (962bp) along with the unmodified amplicon size of 3249bp. These guide combinations were taken forward for testing in the HEK293 cell line. (**B**) Shows the predicted PCR amplicon sizes for the HEK293 deletion lines with associated PCR gel which indicated a modification efficiency of ~20-30%. The PCR amplicon sizes differed between the cell lines due to a different PCR protocol being employed which produced a cleaner PCR with less non-specific amplification within the HEK293 protocol. Guide RNA combinations H and I were taken forward from the initial SH-SY5Y optimisation assay and tested in the HEK293 cell line. Greatly increased modification efficiencies were observed in the HEK293 model, indicated by the greater band intensity (white arrows) compared to the SH-SY5Y cells. The gRNA combination H was selected to take forward due to the more specific modification region which excised a smaller portion of flanking sequence resulting in a more specific removal of the SVA-F.

The efficiency of modification in HEK293 cells was significantly higher than in SH-SY5Y cells as observed via PCR analysis indicated by the brighter modified bands (white arrows in **figure 4.8 – Bii**) with an approximated modification efficiency of 20-30%. Guide combination H was chosen for further experiments because it yielded a smaller deletion of the region but still removed the whole SVA and thus would produce smaller confounding influences on the assessment of SVA activity as more of the flanking sequences remained intact. The same optimisation methods and premises were performed to identify optimal guide RNAs for *INPP5F* SVA-D using HEK293 cells (**figure 4.10**). To generate clonal KO HEK293 cell lines, the optimal gRNA/Cas9 plasmids for both SVA-F and SVA-D were transfected, alongside suitable controls which included un-transfected wild type (WT) cells and a co-transfection of non-targeting (NT) gRNA containing Cas9 plasmid (NT gRNA description and sequences

provided in **section 2.2.11.1**). The non-targeting guide sequences are gRNAs that should not target a specific genomic location. This provided a good mimic for the culture conditions most similar to the modified clones in an attempt to control for changes in cell characteristics as a result of transfection and Cas9 expression. All conditions (WT, NT and CRISPR modified cells) underwent identical clonal isolation protocols and expansion to account for differences in handling procedure (**section 2.2.11.3**).

Before selecting which cell line to proceed with (SH-SY5Y or HEK293), it was important to ascertain the expression profile of *INPP5F* within both the SH-SY5Y and HEK293 lines using RT-PCR of cDNA synthesised from RNA extracted from both cell lines under basal culture conditions. The SH-SY5Y cell line expressed all three isoforms of *INPP5F* (**figure 4.9**) in a pattern that resembled the GTEx data for brain tissue shown in **figure 4.3 – B**. Isoform 2 expression appeared stronger than isoform 1, whilst isoform 3 had little expression. The HEK293 cells presented a different pattern with isoform 2 not being expressed and isoforms 1 and 3 being expressed. The observed band at approximately 600bp (predicted gDNA amplicon 599bp) in the β-actin (*ACT-B*) lane within the HEK293 preparations (**figure 4.9 – B**) matched to the expected size for amplification of contaminating gDNA present within the RNA preparation. The expression patterns observed in both cell lines validated the assumption that these two cell lines would constitute an appropriate model to investigate the effect of SVA deletion on the expression profiles of *INPP5F*. The reporter gene assay data presented in **figures 4.4 and 4.5** demonstrated SVAs acting as repressive elements so the hypothesis was made that *INPP5F* expression would increase upon excision of either the SVA-F or SVA-D elements.

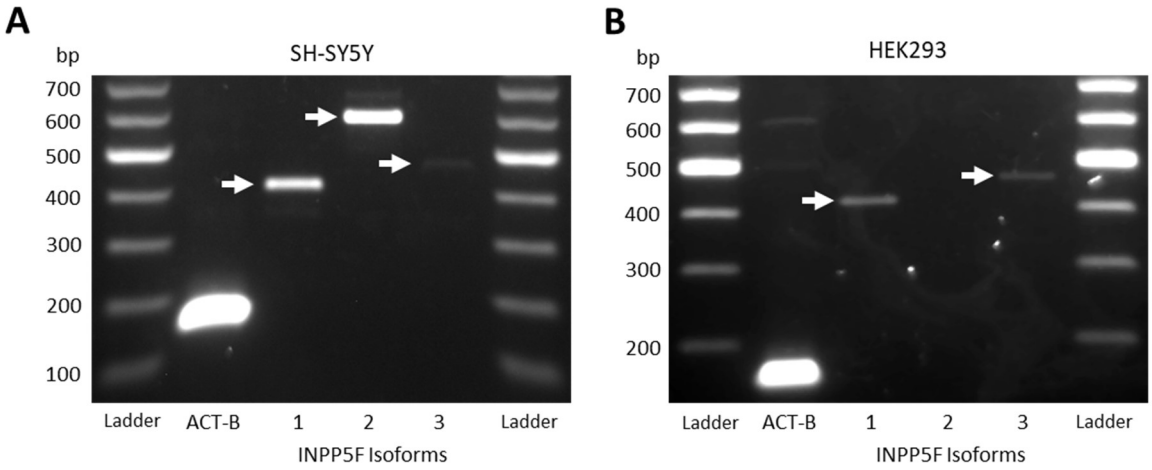
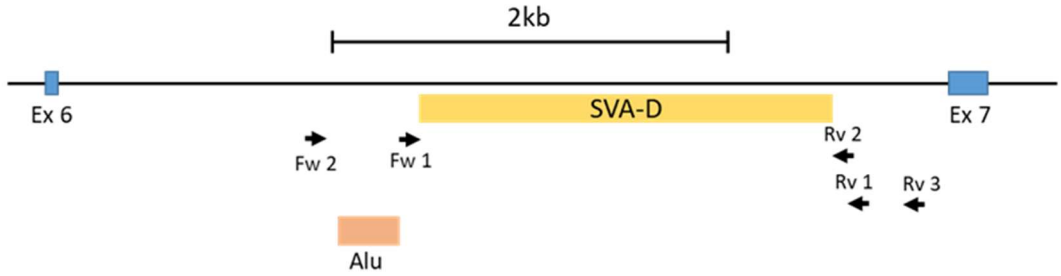


Figure 4.9 – Endogenous expression of the three *INPP5F* isoforms (1, 2 and 3) in SH-SY5Y and HEK293 cell lines under basal conditions using RT-PCR and agarose gel electrophoresis. White arrows indicate the correct band sizes for each corresponding isoform (PCR details in appendix 1) iso 1 – 374bp, iso 2 - 595bp and iso 3 - 459bp. Isoform 2 in HEK293 was not expressed in the model, indicated by the absence of a band at 595bp.

The HEK293 cell line was chosen for all subsequent CRISPR experiments because the modification efficiency was significantly higher than in the SH-SY5Y model for the excision of the *INPP5F* SVA-F (**figure 4.8**) and it demonstrated endogenous expression of *INPP5F* under basal conditions (**figure 4.9**). Using the HEK293 line, gRNAs to target the *INPP5F* SVA-D were optimised following the same process as used for the SVA-F element in which three forward (5') and three reverse (3') gRNAs flanking the SVA-D were designed. One of the forward guides originally designed (Fw 3) was discarded due to issues with plasmid propagation within the cloning process leaving two forward gRNAs and three reverse guides that were tested within HEK293 cells (**figure 4.10**).



Predicted PCR product size (bp)			
gRNA	Fw 1	Fw 2	Control
Rv 1	A – 1186	B – 712	Un – 3249
Rv 2	C – 1253	D – 779	
Rv 3	E – 927	F – 453	

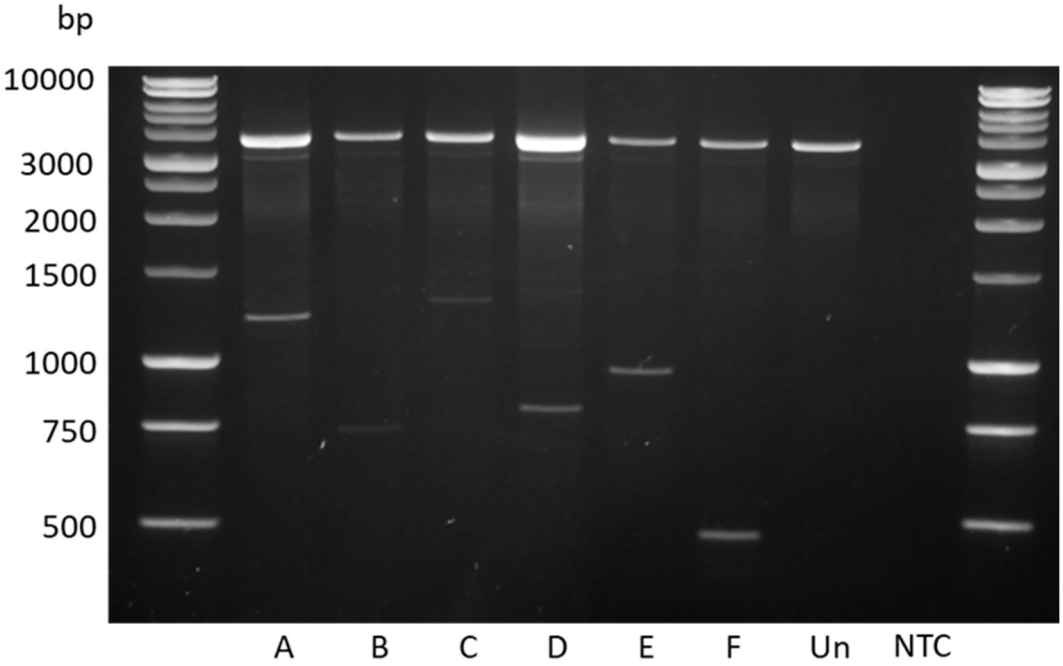


Figure 4.10 – Guide RNA (gRNA) optimisation to target the *INPP5F* SVA-D for CRISPR mediated deletion. Two 5' forward (Fw) guides and three reverse (Rv) 3' guides were cloned into the EF1α-pSpCas9(BB)-2A-GFP plasmid and transfected into HEK293 cells using various combinations of forward and reverse guides. The combinations used with predicted PCR amplicon sizes upon deletion of the SVA element are detailed in the table. PCR amplification and gel electrophoresis indicated that all combinations produced editing efficiencies that ranged ~10-40% with combination E (Fw1, Rv3) being the highest at ~40% (observed using the ratio of modified to unmodified bands intensity). Guide combination Un – unmodified HEK293 cells, NTC – non-template control PCR.

By comparing the band intensity ratios of modified and unmodified PCR amplicons, the modification efficiency was estimated. Guide combination E (Fw 1, Rv 3) was selected to be taken forward for the targeting of *INPP5F* SVA-D, because it presented the highest ratio of modified to unmodified band intensity. Furthermore, the Fw 1 guide RNA would leave the adjacent *Alu* element intact which is depicted in the schematic (top panel – **figure 4.10**) and thus would help to reduce confounding effects that would be introduced by deletion of the *Alu*.

Using the optimised guide RNAs for both *INPP5F* SVA-F and SVA-D (combinations H and E respectively), HEK293 cells were transfected, seeded at low densities (1000 cells per dish in 10cm plates) and cultured until single colonies were formed. Clonal cell populations were isolated, cultured until sufficient numbers were obtained to split into duplicate plates and screened using PCR from crude lysates (full method protocols detailed in **sections 2.2.11.3** and **2.2.11.4**). Three putative mono allelic deletions and three bi-allelic deletions for both SVA-F and SVA-D were selected to take forward (**figure 4.11**). Methods for the PCR screening from crude lysates of clonal isolates are detailed in **section 2.2.11.4**. The putative mono-allelic clones selected were chosen as they presented a more equal 50:50 split in band intensity between the modified and un-modified alleles to avoid selecting colonies that may contain contamination with unmodified cells and would not be truly clonal.

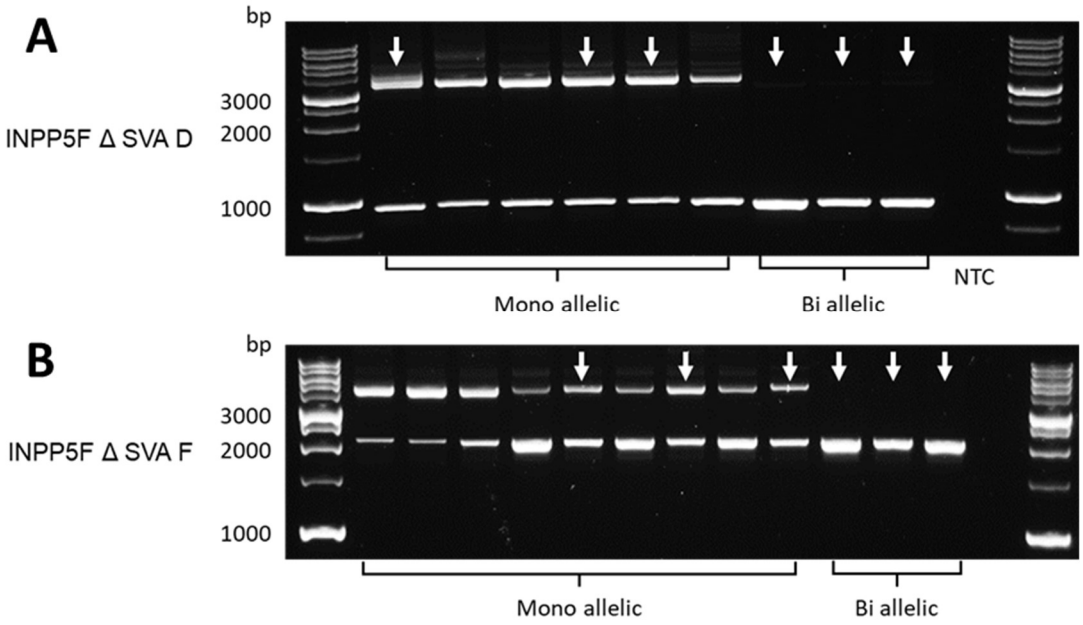


Figure 4.11 – CRISPR deletion PCRs from isolated clonal cell lines of HEK293 cells with deletions of either the *INPP5F* SVA-D **(A)** or SVA-F **(B)**. Three mono and three bi-allelic deletion clones were selected for each target to take forward for *INPP5F* isoform expression analysis (indicated by white arrows). Predicted amplicon sizes: *INPP5F* SVA-D unmodified – 3249bp, modified – 927bp, *INPP5F* SVA-F unmodified – 3926bp and modified – 1828bp.

The selected modified clonal lines and associated controls (WT and NT) were plated and cultured under basal conditions in 24-well plates until 70-80% confluency was reached at which point the cells were dissociated and pelleted. RNA was extracted (**section 2.2.4.2**) and cDNA was generated using first strand synthesis (**section 2.2.6**). The three isoforms of *INPP5F* (**figure 4.3-A**) as well as total *TIAL1* and *BAG3* expression was measured using RT-PCR and agarose gel electrophoresis (**figure 4.12**). Total expression for *TIAL1* and *BAG3* was measured due to the difficulty in generating primer pairs that would target unique exons. The housekeeping genes beta actin (ACT-B) and GAPDH were chosen as these presented stable expression levels across all conditions tested.

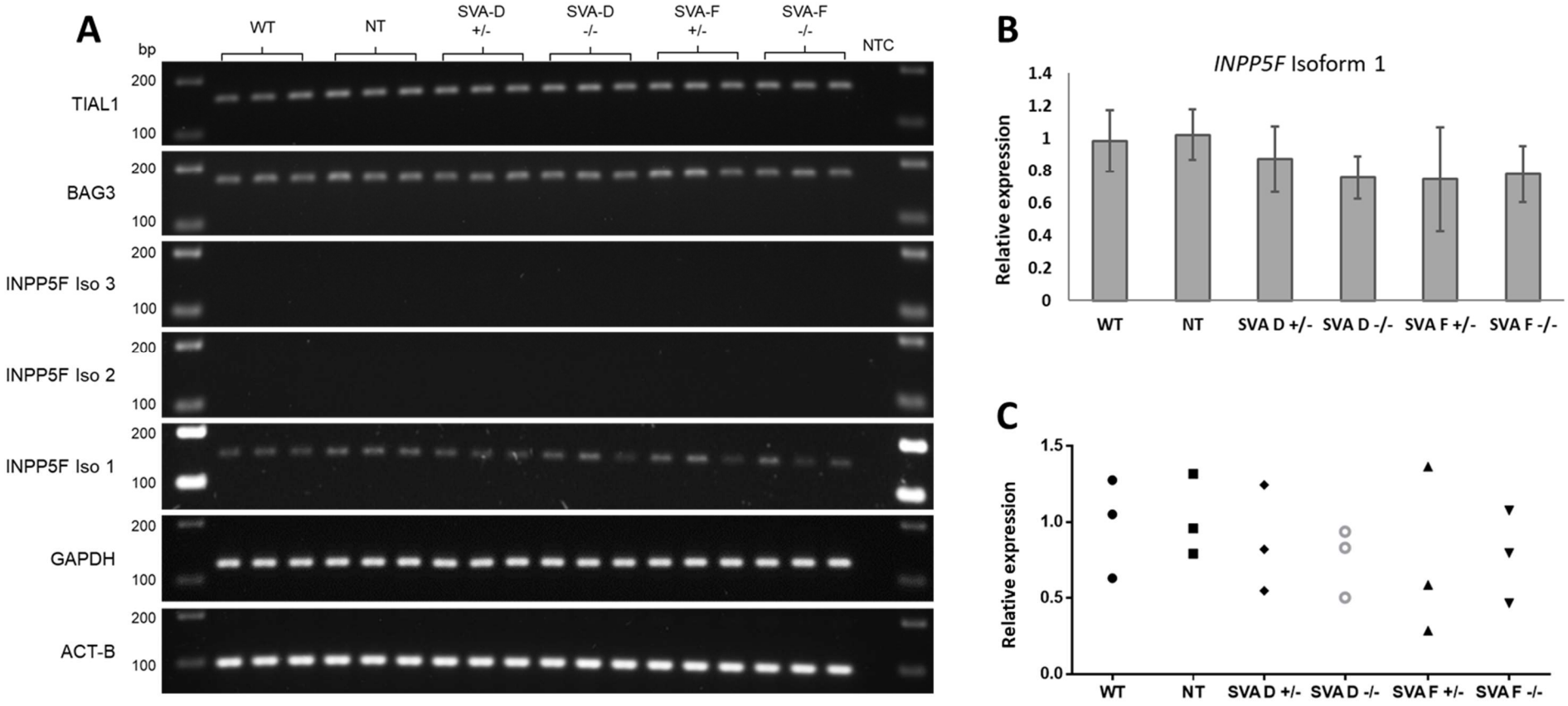

A
WT
NT
SVA-D +/-
SVA-D -/-
SVA-F +/-
SVA-F -/-
NTC
bp
TIAL1
BAG3
INPP5F Iso 3
INPP5F Iso 2
INPP5F Iso 1
GAPDH
ACT-B
200
100
B
INPP5F Isoform 1
Relative expression
1.4
1.2
1
0.8
0.6
0.4
0.2
0
WT
NT
SVA D +/-
SVA D -/-
SVA F +/-
SVA F -/-
C
Relative expression
1.5
1.0
0.5
0.0
WT
NT
SVA D +/-
SVA D -/-
SVA F +/-
SVA F -/-

Figure 4.12 – **(A)** RT-PCR of three isoforms of *INPP5F* along with neighbouring genes *BAG3* and *TIAL1* in response to CRISPR deletion of either SVA F or SVA D in HEK293 cells under basal conditions. *GAPDH* and *ACT-B* were used as housekeeping genes. No change in expression Small potential fluxes in gene expression was observed for *INPP5F* isoform 1 across the different knockout lines indicated by the minor changes in band intensity. **(B)** Quantitative PCR (qPCR) of *INPP5F* isoform 1 expression patterns normalised to *ACTB* with fold changes indicated as relative expression to non-target gRNA (NT) control. No statistical difference observed between CRISPR modified lines using one-way analysis of variance (ANOVA) ($F(5,12) = 0.345897$). **(C)** Breakdown of the individual expression values for each CRISPR modified cell line. Predicted amplicon sizes: *TIAL1* – 160bp, *BAG3* – 176bp, *INPP5F* isoform 1 – 150bp, isoform 2 - 177bp, isoform 3 – 131bp, ACT-B – 110bp, GAPDH – 130bp.

Prior to analysis of the *INPP5F* isoform 1 expression levels, qPCR conditions were analysed using standard, dissociation, and amplification curves in order to calculate primer efficiencies, specificity of amplification and quantity of cDNA required for the qPCR reactions (**figure 4.13**). Primer efficiencies for *ACT-B* and *INPP5F* isoform 1 were calculated at 83.73% and 110.3% respectively. Calculated primer efficiencies above 100% are possible due to the presence of inhibitors within the PCR reaction which become diluted across the serial dilution leading to an increase in amplification between cycles. The difference of approximately 20% efficiency between *ACT-B* and *INPP5F* isoform 1 is not ideal and primers would have been redesigned and tested, however due to time constraints this could not be performed. The dissociation curves showed specific amplification with no off-target amplification, indicated by a single peak for each target. Given the relatively low expression of *INPP5F*, a cDNA dilution of 1:10 (10^{-1}) was deemed optimal for sufficient amplification (blue curve in the amplification plots).

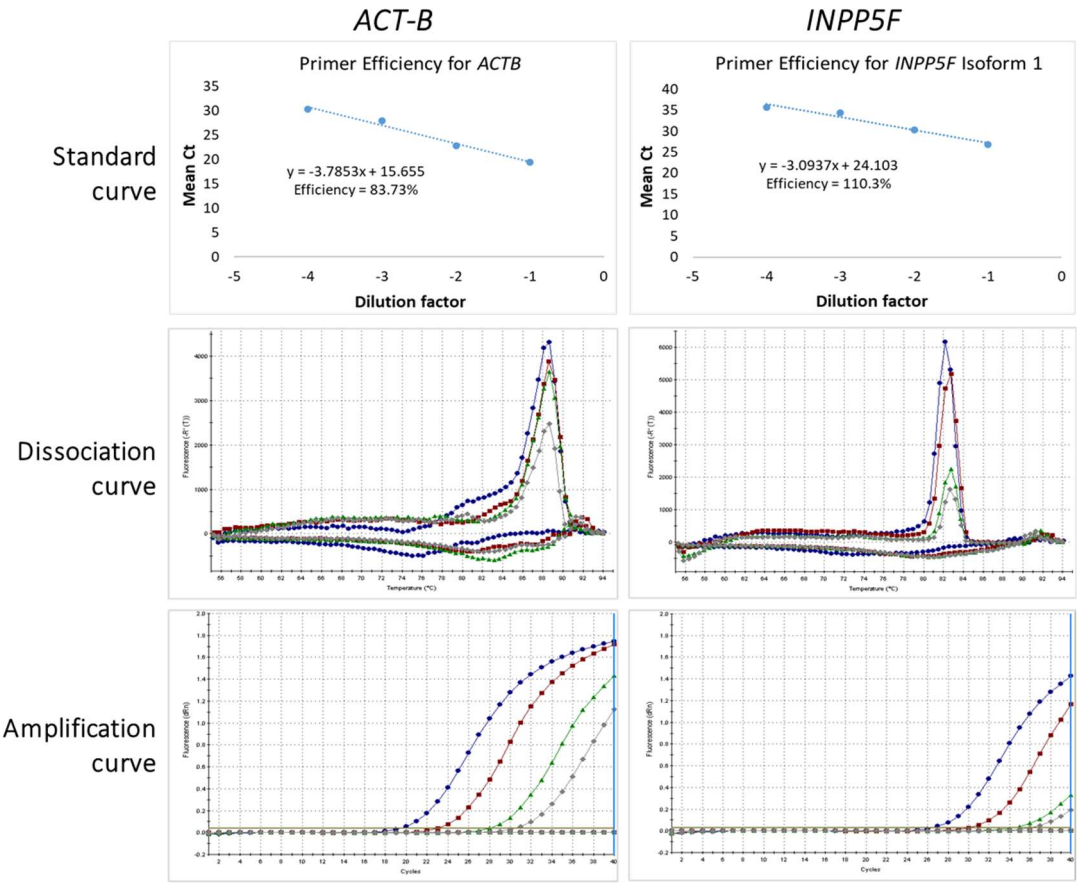


Figure 4.13 – Standard curves, dissociation curves and amplification curves for the qPCR of *INPP5F* isoform 1 and beta-actin (*ACT-B)* using basal wild type HEK293 cDNA. Standard, amplification and dissociation curves for both targets used serial dilutions of cDNA from 10^{-1} to 10^{-4} in order to calculate primary efficiencies and ascertain necessary concentration of cDNA required for amplification. Dissociation and amplification plots are coloured based on dilution factor of input cDNA, blue-10^{-1}, red-10^{-2}, green-10^{-3} and grey 10^{-4}.

Interrogation of the RT-PCR agarose gels showed no consistent change between control and knockout lines for either *TIAL1* or *BAG3* total expression (**figure 4.12 – A**). *INPP5F* isoforms 2 and 3 appeared to have no expression across all conditions within the sensitivity of this assay. This indicates the cell populations may have altered the expression profile of *INPP5F* isoform 3 across the course of the CRISPR protocol given that prior to the generation of knockouts, the HEK293 cells expressed isoform 3 at low levels (**figure 4.9**). Different PCR protocols and primer sets were

employed within this analysis compared to those used in **figure 4.9** so as to keep the amplicon size small (<200bp) to be suitable for potential qPCR analysis. To confirm that the redesigned primer pairs and PCR protocols correctly amplified the correct *INPP5F* isoforms, a second PCR using the same cDNA as in **figure 4.9** was performed using the new PCR conditions. **Figure 4.14** indicates the PCR conditions were suitable for detection of isoforms 2 and 3 within alternative cell line cDNA with isoform 2 being detected in basal Hap1 cDNA and SH-SY5Y cells and isoform 3 being detected in basal SH-SY5Y cells. This negated the possibility of an un-optimised PCR protocol being employed within the CRISPR modified HEK293 cell lines (**figure 4.12**) and suggested that all the HEK293 lines had lost the expression of isoform 3 over the course of the CRISPR protocol.

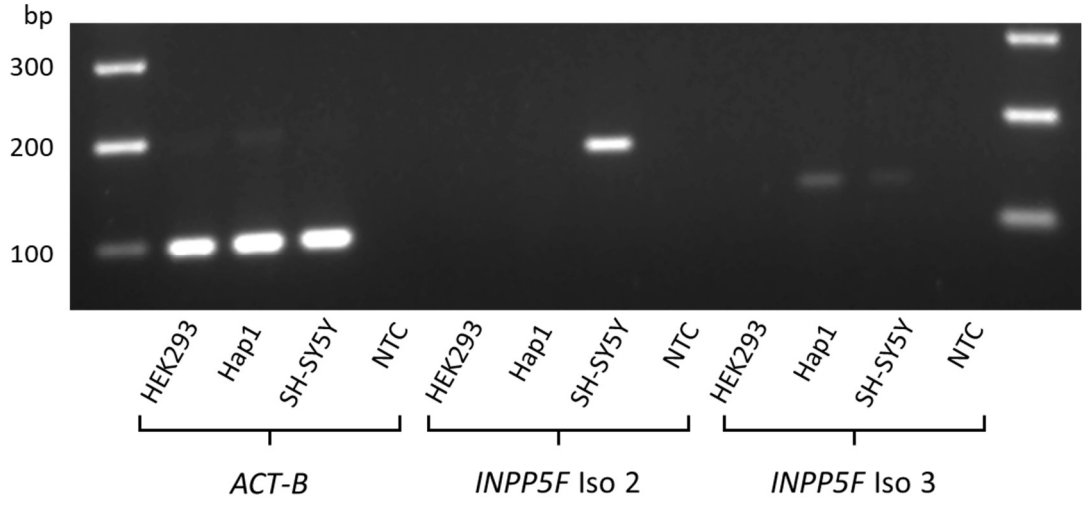


Figure 4.14 – RT-PCR for *INPP5F* isoforms 2 and 3 in un-transfected control HEK293 cells, Hap1 and SH-SY5Y cell lines. Isoform 2 is expressed in SH-SY5Y cells only, whilst isoform 3 is expressed at low levels in both Hap1 and SH-SY5Y but not HEK293. Beta-actin (ACT-B) was used as a housekeeping gene to show amplification of cDNA. Predicted amplicon sizes: *ACT-B* – 110bp, *INPP5F* isoform 2 – 177bp, *INPP5F* isoform 3 – 131bp.

Although there was no consistent change between the triplicate conditions for expression of isoform 1, there were minor variations between individual clonal cell lines (**figure 4.12 - A**). Within the putative mono-allelic SVA-D (SVA-D +/-), mono-allelic SVA-F (+/-) and bi-allelic SVA-F (SVA-F -/-) deletions there were fluctuations in band intensity of *INPP5F* isoform 1 which were explored further using qPCR. **Figure 4.12 – B** indicates the relative expression of each condition (averaged over the three cell lines for each condition) for *INPP5F* isoform 1 normalised to *ACTB*. There was a reduction of average *INPP5F* isoform 1 expression in response to SVA F and SVA D knockout with the largest most consistent decrease observed in the bi-allelic SVA D KO (SVA D -/-) which showed a 35% decrease in average expression (expressed as mean) compared to the NT control (**figure 4.12 – B**). Observational interrogation appeared to show variation of expression within the SVA-F mono-allelic putative knockouts (**figure 4.12 - C**) which could suggest the effect of a mono-allelic SVA knockout was causing individual cell populations to mis-regulate *INPP5F* expression under the tested conditions. To statistically interrogate this potential effect, a one-way analysis of variance (ANOVA) was employed to test statistical significance by calculating the F ratio using 5 degrees of freedom for the sum of squares between groups and 12 DF for the sum of squares within groups (F(5,12)). The critical value needed for significance was 3.11 which the calculated F ratio of 0.345897 did not satisfy and so concluded there was no difference between any of the control or SVA knockout cell lines. The lack of significance could be consistent with an insufficient N numbers within each condition set to account for inherent variability between clones. To overcome this, larger numbers of clones within each condition would be required.

4.2.4 Genetic variation of *INPP5F* SVA-F and SVA-D

Having demonstrated SVA function in reporter gene constructs within transient assays, with failure to replicate the result using *in vitro* knockout of the *INPP5F* SVAs using CRISPR under basal conditions, it was decided to address potential polymorphisms within SVA primary sequence that could infer risk for PD. The two latest PD GWAS meta-analysis from Nalls *et al.* 2014 and Nalls *et al.* 2019 which constitute the largest GWAS studies performed for PD to date, were used to evaluate GWAS SNPs within or flanking (+/- 5Kb) of the SVA elements in the *INPP5F* loci. The presence of three GWAS signals that resided within the SVA elements were noted which strengthened the hypothesis that these elements have genetic association with PD (**figure 4.15)**. The SVA-F element contains a genome wide significance SNP-B (p-value = 3.38×10^{-8}) located within the SINE-R domain whilst the SVA-D contains two statistically significant but not genome wide significant GWAS SNPs, F and G, with p-values of 2.85×10^{-5} and 5.09×10^{-5} respectively.

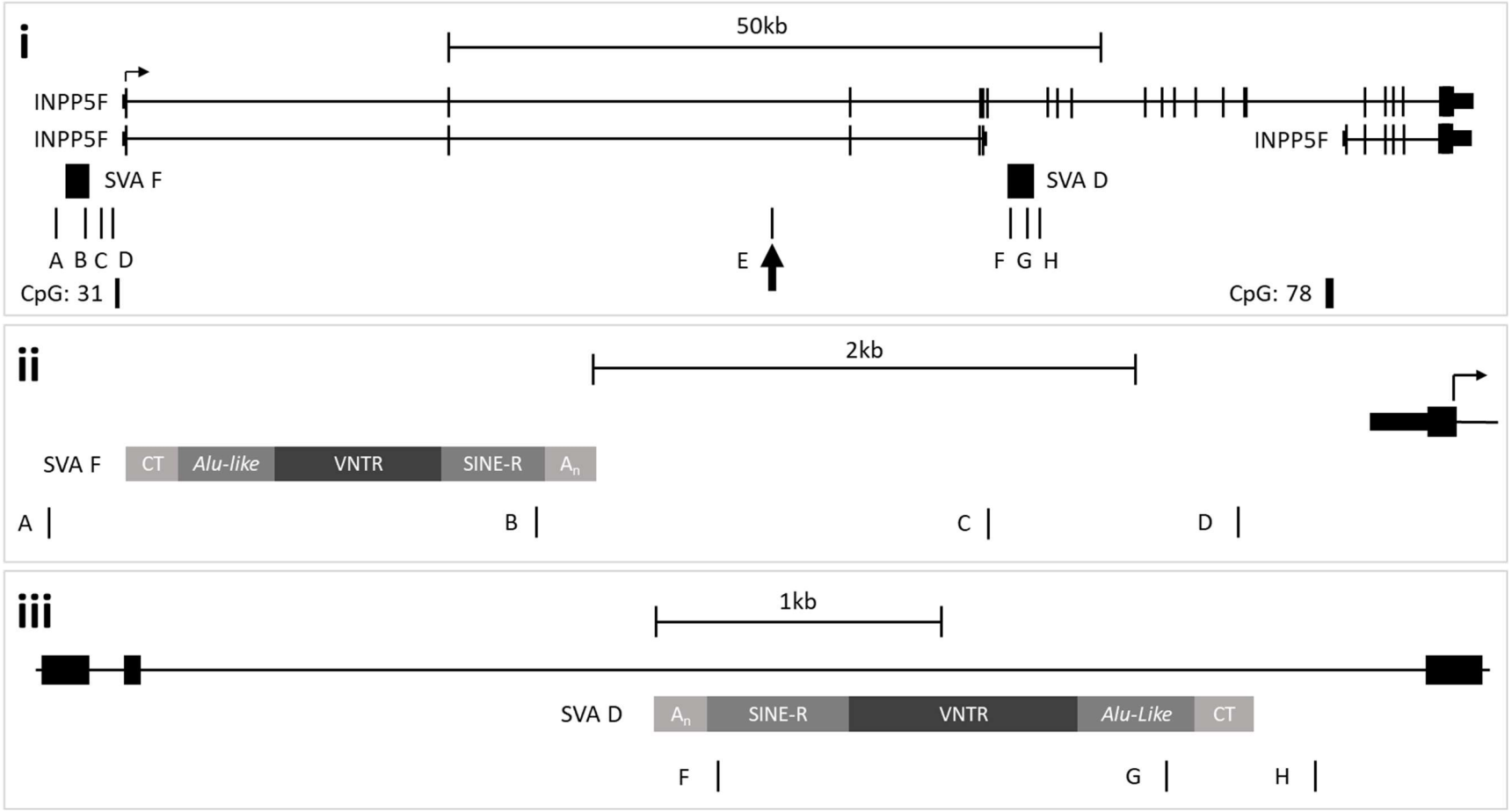

i
50kb
INPP5F
INPP5F
INPP5F
SVA F
SVA D
A B C D
E
F G H
CpG: 31
CpG: 78
ii
2kb
SVA F
CT
Alu-like
VNTR
SINE-R
An
A
B
C
D
iii
1kb
SVA D
An
SINE-R
VNTR
Alu-Like
CT
F
G
H

Figure 4.15 – **(i)** Schematic depiction of the GWAS identified *INPP5F* locus as a novel potential risk for PD via the association with the SNP, rs117896735 (indicated by black arrow at SNP E) found within intron 2 of the larger *INPP5F* isoform 1. CpG islands indicate the two major transcriptional start sites for the two isoforms of *INPP5F* shown. Bioinformatic analysis of this region provided evidence for additional GWAS SNPs within or near (+/- 2.5kb) reference SVA elements denoted as A-H with the exception of SNP E as this is the major GWAS hit for this locus. **(ii)** GWAS SNPs located within and around SVA F and the promoter region of *INPP5F*. SNP B located within the SINE-R domain of SVA F. **(iii)** SNPs F and G are located within the SINE and *Alu*-like domains of SVA D respectively.

GWAS SNPs and associated meta p-values: A - rs196255 - $1.23x10^{-5}$, B – rs12779721 - $3.38x10^{-8}$, C - Rs196212 - $1.35x10^{-5}$, D - rs196213 - $1.28x10^{-5}$, E - rs117896735 - $2.36e^{-28}$, F - Rs12784999 - $2.85x10^{-5}$, G - Rs61867939 - $5.09x10^{-5}$, H - rs12774619 - $2.75x10^{-5}$. p-values for SNPs A, C, D, F, G and H are taken from the Nalls *et al.* 2014 meta-analysis whilst SNPs B and E are taken from the Nalls *et al.* 2019 meta-analysis.

The GWAS SNPs present within the SVA-F and SVA-D primary sequences (SNPs B, F and G) may suggest an important role of specific sequence polymorphisms within the SVA that could confer risk for development of PD. It is important to note that GWAS analysis seldom identifies specific risk SNPs that contribute to disease progression, but rather identify general loci of interest.

4.2.4.1 GWAS SNPs altering transcription factor binding sites within SVAs

As an initial analysis to infer potential function of the GWAS SNPs within the SVA-F and SVA-D elements, the impact the GWAS SNPs on transcription factor (TF) binding sites was analysed. To do this, the TF binding to the wild type (WT) SNP with flanking sequence (+/- ~50bp) was compared to an identical sequence containing the three identified GWAS risk SNPs within both the SVA-F and SVA-D elements using the rVista 2.0 TF analysis package (http:// rvista.dcode.org) [169]. The rVISTA package uses the curated TRANSFAC database as a reference to identify human TF binding motifs within the target sequence and can compare multiple sequences simultaneously.

Figure 4.16 shows six sequences that were inputted to rVista which represent the three identified GWAS SNPs (shown in red) within the SVA-F and SVA-D elements, highlighted in **figure 4.15** (SNPs B, F and G), with the flanking sequence. The sequences in blue represent a binding motif region that was altered by the GWAS SNP compared to the wild type sequence. Rs12779721 (SNP-B) within the SVA-F altered the binding of three human TFs, PPARG, SPZ1 and EGR, which were predicted to bind to the WT sequence but not the sequence containing the GWAS SNP. Rs12784999 (SNP-F) within SVA-D introduced a novel SMAD binding site whilst the rs61867939 (SNP-G) altered the SMAD binding motif. It is important to consider both the introduction or disruption of binding sites and also any changes in binding motifs, as these can alter the affinity of the TF binding to the site and alter the regulatory properties of the TF.

Rs12779721 - SVA-F (SNP B)

Wild type

AGGAAAACCAGAGACCTTTGTTCACTTGTTTATCTGCTGACCTTCCCTCCACTATTGTCCCATG
ACCCTGCCAAATCCCCCTCTGTGAGAAACACCCAA

GWAS Risk

AGGAAAACCAGAGACCTTTGTTCACTTGTTTATCTGCTGACCTTCCCTGCACTATTGTCCCATG
ACCCTGCCAAATCCCCCTCTGTGAGAAACACCCAA

TGACCTTCCCTCCACTA = PPARG

GACCTTCCCTCCACT = SPZ1

CTTCCCTCCAC = EGR

Rs12784999 – SVA D (SNP F)

Wild type

TGGGTACTTGAGATTAGGGAGTGGTGATGACTCTTAAAGAGCATGCTGCCTTCAAGCATCTGTT
TAACAAAGCACATCTTGCACCGCCCTTAATCCATT

GWAS Risk

TGGGTACTTGAGATTAGGGAGTGGTGATGACTCTTAAAGAGCATGCTGTCTTCAAGCATCTGTT
TAACAAAGCACATCTTGCACCGCCCTTAATCCATT

ATGCTGTCT = SMAD

Rs61867939 – SVA D (SNP G)

Wild type

CACTGCACTCCAGCCTGGGCACCATTGAGCACTGAGTGAACGAGACTCCGTCTGCAATCCTGGC
ACCTCGGGAGGCCAAGGCTGGCGGATCACTCGCGGTT

GWAS Risk

CACTGCACTCCAGCCTGGGCACCATTGAGCACTGAGTGAACGAGACTCTGTCTGCAATCCTGGC
ACCTCGGGAGGCCAAGGCTGGCGGATCACTCGCGGTT

AGACTCCGT = SMAD

CTGTCTGCAAT = SMAD

Figure 4.16 – Analysis of the *INPP5F* SVA-F and SVA-D GWAS SNPs impact on transcription factor binding sites. The three GWAS SNPs within SVA-F and SVA-D, identified in **figure 4.11** (SNPs B, F and G), were taken with approximately +/- 50bp of flanking sequence and analysed using the rVista 2.0 transcription factor (TF) binding site analysis software [169]. Using two identical sequences for each GWAS SNP, one containing the wild type allele and the other the risk allele, the binding site alterations could be observed. Rs12779721 (SNP-B) within SVA-F disrupts three predicted human TF binding sites: PPARG, SPZ1 and EGR. SNP rs12784999 (SNP-F) within SVA-D, introduces a novel SMAD binding site and SNP rs61867939 (SNP-G), also within SVA-D, alters the specific SMAD binding motif. PPARG – Peroxisome proliferator activated receptor gamma, SPZ1 – spermatogenic leucine zipper protein 1, EGR – Early growth response.

The presence of GWAS SNPs within SVA-F and SVA-D that alter SMAD binding could be important for Parkinson's disease given that SMAD proteins are involved in the BMP signalling cascade which has been associated with the development of dopaminergic neurons within the substantia nigra [170]. Studies have shown that the BMP/SMAD signalling pathway is critical for the neurogenesis of these neurons, with SMAD inhibition being used for the generation of induced dopaminergic neurons via pluripotent stem cells differentiations [171]. EGR has also been demonstrated as crucial in memory formation, brain plasticity and neuropsychiatric disorders and provides further evidence for the potential implication of the SVA-F in brain related functions [172].

4.2.4.2 *INPP5F* SVA-F genotyping

Using DNA sequences extracted from the UCSC genome browser human genome build 38 (hg38) (chr10:119721360-119723094) it was possible to split the SVA-F sequence into its composite parts, highlighting the CT repeat, *Alu*-like, tandem repeat (TR), VNTR, SINE-R and poly-A regions (**figure 4.17**).

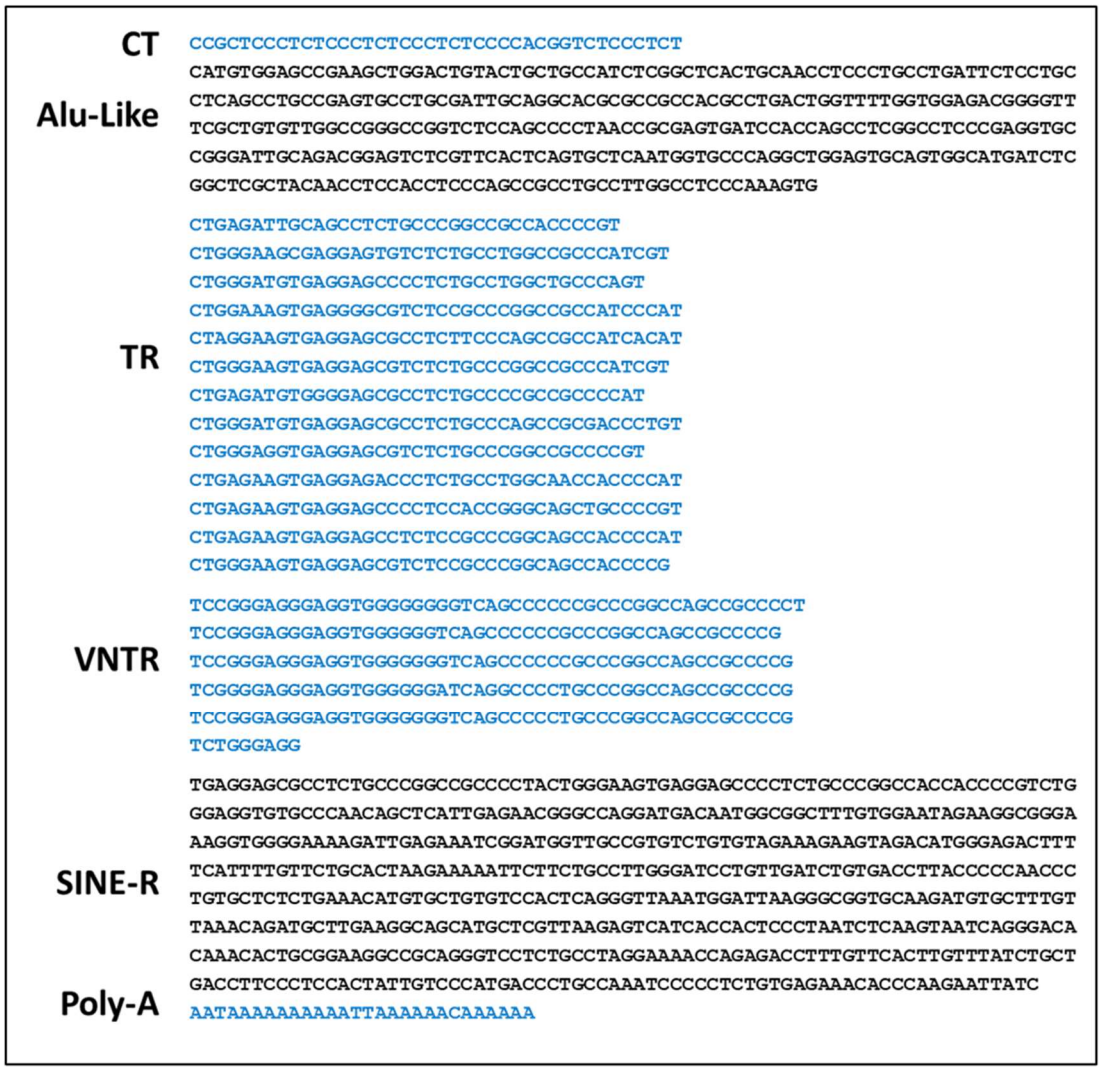
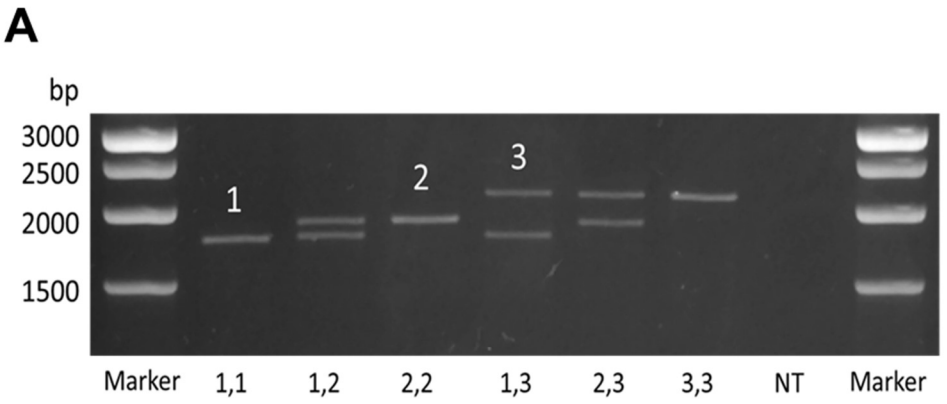
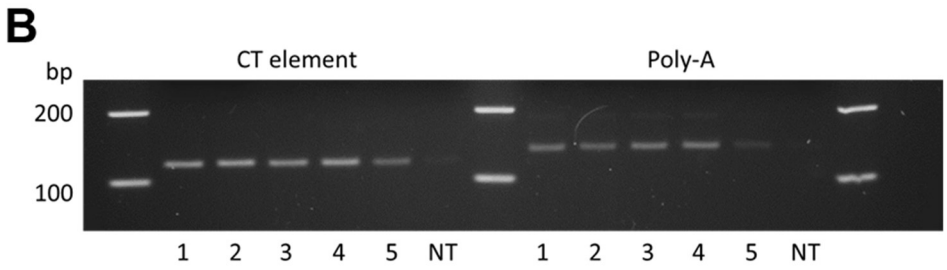
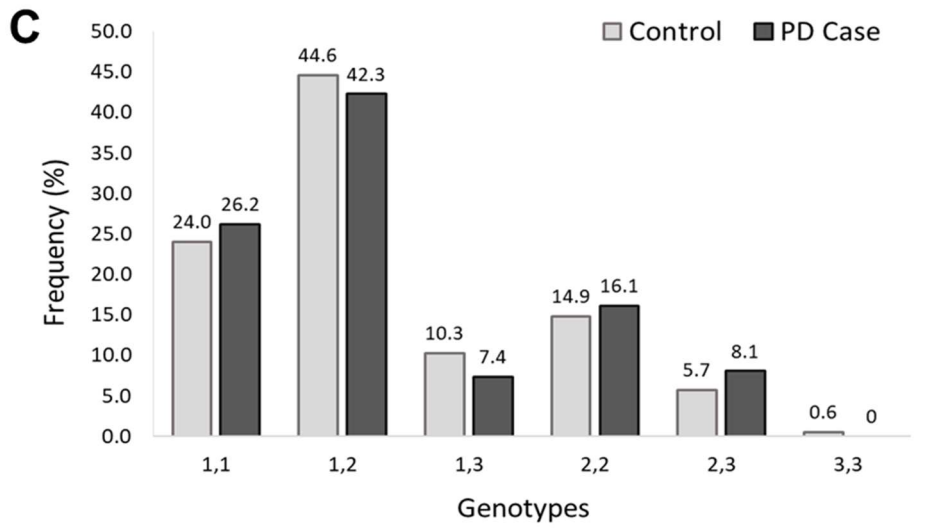


Figure 4.17 – Break down of the reference *INPP5F* SVA-F sequence (downloaded from UCSC genome browser - hg38 - chr10:119721360-119723094) indicating composite parts. The sequence is displayed in the 5' to 3' orientation from CT hexamer repeat to terminal poly-A. Sequences in blue represent the regions which were targeted to analyse for polymorphisms.

The sequence shown in **figure 4.17** demonstrates the composite nature of the SVA-F element and includes two tandem repeats (TR) as expected with subclass F SVAs. The second distinct TR is generally considered as the potential polymorphic TR and has been labelled as a variable number TR (VNTR). Primers were designed to target the CT element, poly-A tail, and the full length SVA and were used to genotype an Estonian human PD case/control matched cohort using PCR (**cohort details in section 2.1.3**). PCR of the full length SVA-F showed three distinct alleles with a large variation (~100bp) between alleles 1 and 2, and approximately 300bp difference between alleles 2 and 3 when observed on an agarose gel. It was hypothesised that a polymorphism of this size was likely to be due to VNTR repeat expansions which was confirmed by Sanger sequencing (**figure 4.15**). Subsequent genotyping in 149 PD case and 175 control samples showed no statistical difference in genotype or allele frequency using a two tailed Fishers exact test (**figure 4.18**). No polymorphism was observed within the CT repeat region when genotyped in 182 PD case and 173 control human samples. Similarly, no polymorphism was observed in the poly-A tail when genotyped in 96 control samples. The lack of significance in this analysis does not rule out the potential importance of the SVA-F as a genetic risk factor given the small sample size and use of only one ethnic cohort. There could be disease specific alleles for this element found in specific ethnic groups or at lower frequencies outside of the sensitivity range within this assay.







Genotype

Genotype	Control Freq	PD Freq
1,1	42	39
1,2	78	63
1,3	18	11
2,2	26	24
2,3	10	12
3,3	1	0
Total	**175**	**149**

Allele

	Control		PD	
Allele	Freq	%	Freq	%
1	180	51.4	152	51.0
2	140	40.0	123	41.3
3	30	8.6	23	7.7
Total	**350**		**298**	

Figure 4.18 – PCR amplification and agarose gel electrophoresis of the full length *INPP5F* SVA F, CT element and poly-A with subsequent genotyping of the full length SVA-F in PD human samples with matched healthy controls. **(A)** Three alleles for SVA F identified and termed **1** - short**, 2** - medium and **3** – long giving rise to six potential genotypes. **(B)** PCR amplification of the CT repeat and poly-A domains of the SVA in control (neurodegeneration free) human samples showing no polymorphism. **(C)** Genotyping of the full length SVA-F in a PD cohort of Estonian ethnic origin shows no statistical significance between different genotypes using two tailed Fishers exact test. Control N=175, PD case N=149. NT = no template control.

The three alleles of the *INPP5F* SVA-F found within this cohort were sequence verified using Sanger sequencing (performed externally at Source Bioscience). This indicated that the second tandem repeat within the SVA contained the polymorphic region. **Figure 4.19** highlights the polymorphic region and shows the difference between the three alleles and have been split according to the corresponding VNTR repeat with one repeat per line. Dashes within a single repeat indicate missing bases compared to different repeat within the same allele. Allele 2 is the sequence found within the reference genome (hg38) and contains 2 additional repeats of a 49bp sequence and a 14bp partial repeat indicated in blue in compared to allele 1. Allele 3 contains a more complicated imperfect repeat expansion with an additional 3 repeats of the 49mer sequence, and 2 repeats which contain partial deletions (indicated by dashes).

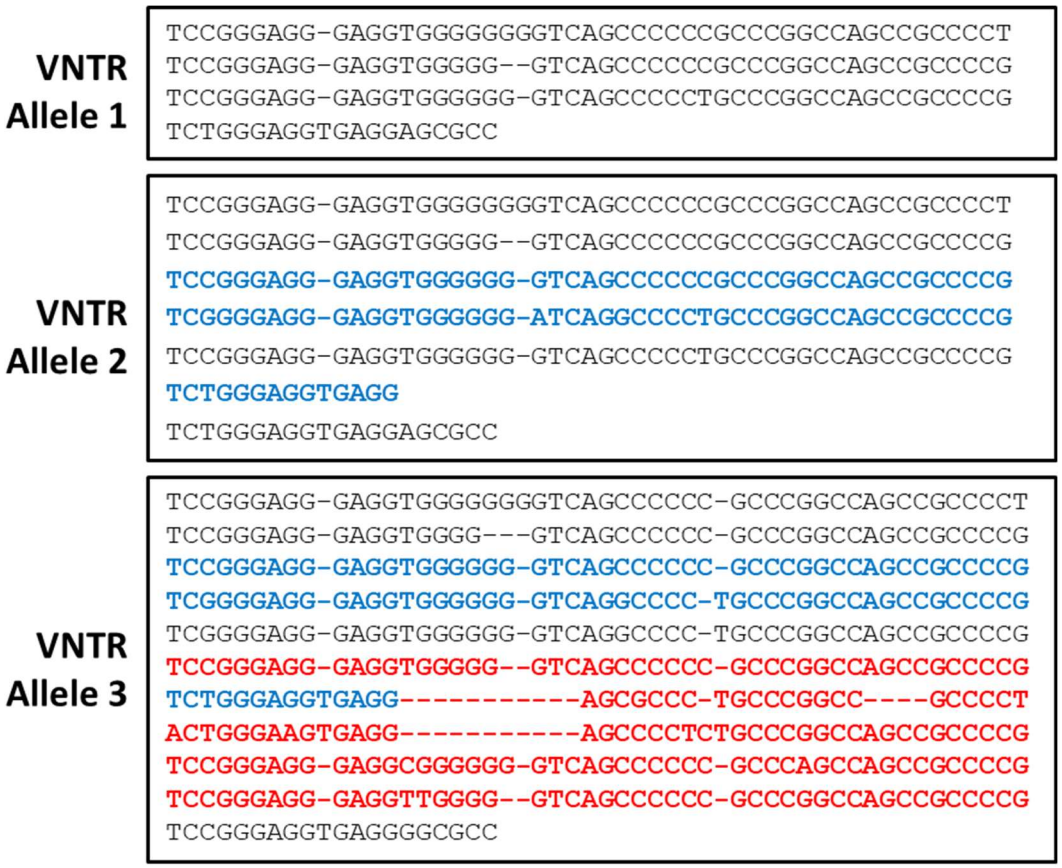


Figure 4.19 – Sequenced polymorphisms of the three VNTR alleles VNTR of the *INPP5F* SVA-F. The blue and red coloured sequences in alleles 2 and 3 indicate the additional bases compared to allele 1 (shortest). Dashes in the sequences indicate missing bases compared to a different repeat within the same allele.

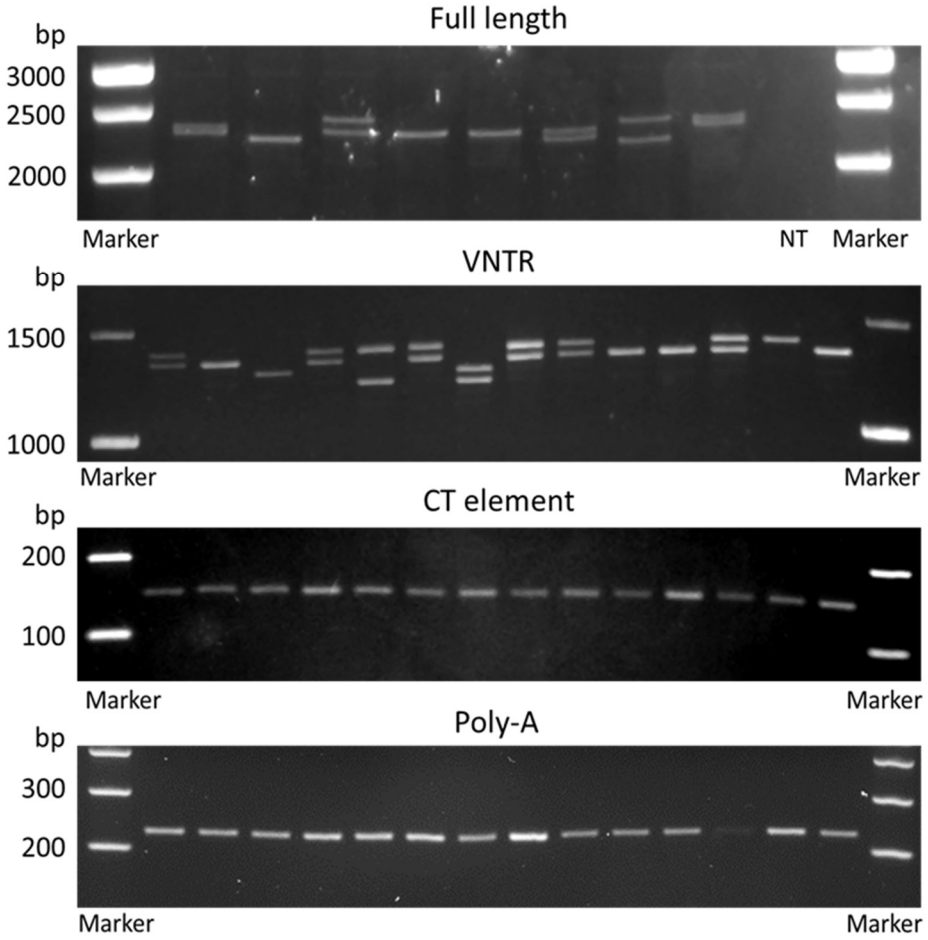
4.2.4.3 *INPP5F* SVA-D genotyping

Following a similar process for the SVA-F analysis, the SVA-D was genotyped using the same PD case/control cohort to assess potential novel risk factors for PD. Using the SVA-D sequences taken from the UCSC genome browser (hg38) RepeatMasker database (chr10:119794026-119796112), the SVA was split into its composite domains (**figure 4.20**).

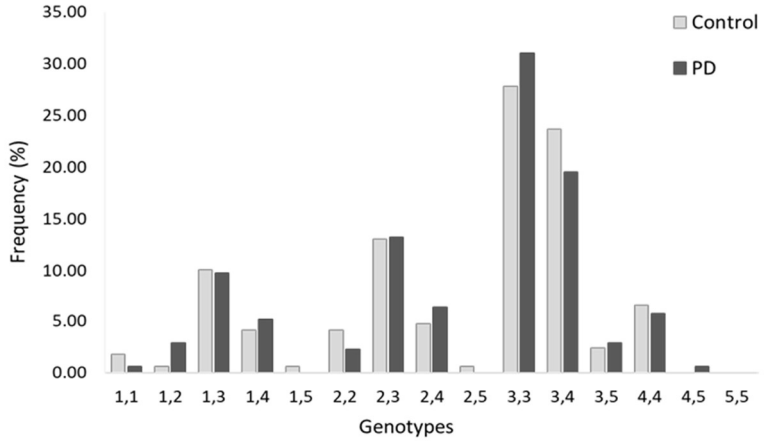
Figure 4.20 – Breakdown of the *INPP5F* SVA-D into composite domains using sequences downloaded from UCSC genome browser RepeatMasker database. The sequence is displayed in the 5'>3' orientation, CT element to poly-A and is located on the negative (-) strand within the reference genome.

Full length PCR amplification of the *INPP5F* SVA-D indicated multiple large polymorphisms (>50bp) using agarose gel electrophoresis (**figure 4.21 panel A**). However, the polymorphisms between individual samples were difficult to genotype due to the size of the SVA-D limiting the resolution on an agarose gel. Employing a nested PCR approach allowed the amplification of the isolated VNTR domain, producing a smaller amplicon size and allowed for greater resolution between polymorphisms. Five alleles of the VNTR were identified and subsequently genotyped in an Estonian PD case/control cohort of 169 controls and 174 case samples. No statistically significant differences were found for both genotype and allele frequencies between case and control when analysed using a two-tailed Fishers exact test. The CT element and poly-A regions were also amplified to identify potential polymorphisms. A total of 72 control samples for the CT element and 93 control samples for the poly-A regions were genotyped with no observable polymorphisms observed in either target. This does not exclude the possibility for rare variants (<1%) or variation within different populations given the sample cohort were of Estonian ancestry.

A



B



Genotype	Control Freq	PD Freq
1,1	3	1
1,2	1	5
1,3	17	17
1,4	7	9
1,5	1	0
2,2	7	4
2,3	22	23
2,4	8	11
2,5	1	0
3,3	47	54
3,4	40	34
3,5	4	5
4,4	11	10
4,5	0	1
5,5	0	0
Total	169	174

Allele	Control Freq	Control %	PD Freq	PD %
1	32	9.5	33	9.5
2	46	13.6	47	13.5
3	177	52.4	187	53.7
4	77	22.8	75	21.6
5	6	1.8	6	1.7
Total	338		348	

Figure 4.21 – **(A)** PCR amplification and agarose gel electrophoresis of the full length, nested VNTR, CT element and poly-A tail of the *INPP5F* SVA-D in control (neurodegeneration free) human samples. NT – No template control. **(B)** Genotyping of the *INPP5F* SVA-D VNTR using nested PCR and analysed by agarose gel electrophoresis. No statistical difference in genotype or allele frequencies between case and control groups using two tailed Fishers exact tests. Control N=169, PD case N=174. NT = no template control.

4.3 Discussion

Currently in large scale short read genetic analyses such as GWAS for Parkinson's disease, SVA retrotransposable elements are often excluded from analysis due to computational difficulties. However, it was demonstrated that such elements within key PD GWAS nominated loci such as *INPP5F/BAG3/TIAL1* are important to consider as potential sources of novel regulatory elements that can impact gene expression. The data presented here shows SVA retroelements as regulators of gene expression using reporter gene assays, with two different SVA subclasses (F and D) eliciting strong inhibitory effects on luciferase expression. Furthermore, these elements may impact various properties of gene regulation, not only expression levels, but also alterations to splicing. This was demonstrated within the intron splicing model using the pSHM06 vectors containing the cloned SVA-D element within an intron, which significantly reduced expression of Renilla luciferase (**figure 4.5**). The model of intron retention produced from SVA insertions within genes which would not normally contain an SVA element, has been explored before in the case of X-linked dystonia Parkinsonism (XDP) where an SVA insertion within the *TAF1* locus leads to intron retention and subsequent reduction of *TAF1* mRNA levels [78].

A difference was observed in the SVA-F sense constructs that were transfected into the undifferentiated iPSCs compared to both the forebrain cortical neuron lineage

differentiated iPSCs and the mammalian cell line models (tested in **figure 4.4**), whereby there was no repression within the undifferentiated cells (**figure 4.6**). The absence of repression observed within this model has been reflected in previous literature where retroelements, including LINE-1, SVA and HERVK elements, become de-repressed during reprogramming of different cell types [173]. Interestingly, in the reprogramming events of CD34+ cord blood cells (of the same cell type used in the data presented within this thesis), this study reported global up-regulation of SVA elements over the course of reprogramming and would be consistent with the absence of repression observed within the models tested here. The changes in the expression of retroelements within previous studies appeared to correlate with changes in KRAB zinc finger proteins (ZFPs) which, together with the co-factor TRIM28, control expression of retroelements [174, 175]. Further to this, the transcription factor BORIS (Brother of the regulator of imprinted site) which is known to bind and repress SVA expression *in vivo* via interaction with the SVA VNTR domain, is expressed in various cancers, testis and pluripotent stem cells but not within other differentiated somatic tissues including mature neurons [176-179]. Within the iPSC models tested in **section 4.2.2**, expression of luciferase from the *INPP5F* SVA-F containing pGL3p constructs was observed within pluripotent iPSCs with significant repression reported within the cortical neuronal lineage differentiated iPSCs which could correlate with expression of BORIS within these models (**figure 4.6**) and could be of interest to study in future works. The presence of BORIS may act to de-repress the activity of the SVA-F element, allowing for expression of luciferase within the undifferentiated iPSCs under basal conditions. To test this hypothesis, measurements

of BORIS expression and protein levels within both the undifferentiated iPSCs and cortical neuron differentiated iPSCs would be necessary.

Optimisations for future works to test the effects of BORIS on SVA function were attempted using over-expression constructs containing a CMV driven BORIS cassette containing the cDNA sequence for the major isoform of BORIS in HEK293 cells. Achieving stable high expression of BORIS in HEK293 cells was difficult for unknown reasons, as measured using western blotting which showed no expression of BORIS in basal HEK293 cells as expected and insufficient expression of BORIS using the over-expression constructs. For future studies, other cell types including the neuroblastoma cell line SH-SY5Y and iPSCs with neuronal derivatives would provide more suitable model systems for SVA function in a more neuronal model than HEK293. Using the BORIS over-expression constructs in a co-expression assay with the *INPP5F* SVA-F and SVA-D containing reporter gene constructs as well as the CRISPR SVA KO HEK293 cell lines generated within this chapter greater insight could be provided into potential novel SVA function with respect to the effects of BORIS.

The pSHM06 SVA constructs utilised within **figure 4.5** provided a different model for the potential mechanisms of action for the tested SVA elements compared to the pGL3 based vectors as they were driven by high expression CMV promoters in contrast with the SV40 driven pGL3 based constructs presented in **figures 4.4 and 4.6**. The position of the cloned SVA elements within the pSHM06 constructs was downstream of the CMV promoter within an intronic sequence in comparison to upstream of the SV40 promoter within the pGL3 vectors. The SVA in this model could be reducing the level of luciferase observed by three possible modes. The SVA

primary sequence could recruit TF's or possess distinct 3D genomic structure which could block or reduce the action of RNA polymerase to reduce luciferase expression, or the SVA could be reducing the efficiency of splicing leading to a reduction in functional luciferase due to the SVA and associated intron being retained in the mature mRNA.

To study the potential effects of the SVA-F and SVA-D elements on the modulation of *INPP5F*, *BAG3* and *TIAL1* expression, CRISPR mediated SVA knockout HEK293 cell lines were successfully generated (**figure 4.11**). Using a combination of RT-PCR and qPCR, expression of *INPP5F*, *BAG3* and *TIAL1* could be analysed in response to SVA KO. Inconsistent fluctuations were observed within the *INPP5F* isoform 1 expression profile as observed by RT-PCR with little to no effect seen within the *INPP5F* isoforms 2, 3, *BAG3* or *TIAL1* expression patterns when comparing the WT, NT, SVA-D and SVA-F knockout cell lines (**figure 4.12**). To quantify the potential changes in *INPP5F* isoform 1 expression patterns across the various cell lines, qPCR was performed which indicated no significant changes in gene expression in response to SVA knockout (**figure 4.12 b and c**). However, average decreases in *INPP5F* isoform 1 expression were observed across all SVA knockout cell lines (SVA-D +/-, SVA-D -/-, SVA-F +/- and SVA-F -/-) with the largest effect seen in the homozygous SVA-D KO (SVA-D -/-) and heterozygous SVA-F KO (SVA-F +/-) cell lines, with an average 35% decrease in expression reported in both conditions. Larger variability was also observed within the SVA-F heterozygous KO cell lines which could suggest an involvement of the SVA-F in *INPP5F* regulation which was impacted by confounding variables not controlled for within the scope of this study. The lack of a statistically significant change in the level of *INPP5F* isoform 1 observed within the CRISPR qPCR

is potentially due to insufficient N numbers within each condition to account for inter-cell line variability between the clones. Rakovic *et al.* 2018 demonstrated a significant increase in relative *TAF1* expression as a result of SVA KO (unedited n=12, edited n=10) and reported relatively small changes in relative expression of approximately 0.2 from ~0.35 to ~0.55 (median) p<0.0001 [79]. The *INPP5F* SVA-D CRISPR data presented in **figure 4.12** showed a change in relative expression compared to the non-targeting controls (NT) where a decrease in *INPP5F* isoform 1 expression of 0.25 from ~1.02 (NT) to ~0.75 (SVA-D -/-) was observed, comparable to that observed in Rakovic *et al.* 2018. However, the changes in *INPP5F* isoform 1 expression within the *INPP5F* SVA-D data did not reach statistical significance between these groups.

The functional data from the reporter gene and CRISPR assays provided insight into the potential implications of SVA elements with respect to gene regulation but did not provide information on primary sequence variation which could be important in PD pathology. Detailed analysis of the GWAS variants described in Nalls *et al.* 2019 revealed three GWAS SNPs present within the SVA-F and SVA-D elements collectively, with one of these (rs12779721, p-value 3.38×10^{-8}), located in the SVA-F SINE-R domain, being statistically genome wide significant [26]. Analysis of the transcription factor binding sites (TFBSs) which were altered by the presence of the GWAS SNP showed the loss of three TFBSs, namely PPARG, SPZ1 and EGR when the GWAS risk variant was present. Interestingly, PPARG has been suggested as a potential therapeutic target in Parkinson's disease whereby agonists of PPARG have been suggested to promote anti-inflammatory pathways and attenuation of microgliosis to provide neuroprotection against the development of PD [180]. The

exact mechanisms of the protective pathways stimulated by PPARG remain elusive, which could be due in part to the lack of understanding of potential novel TFBSs including those found within retrotransposable elements which are often over-looked. Early growth response (EGR) factors have also been implicated in brain plasticity and neuropsychiatric disorders including stress related mood disorders, Schizophrenia and also pathways responsible for drug reward, withdrawal and relapse mechanisms [172, 181-183]. Both PPARG and EGR represent example mechanisms to associate the *INPP5F* SVA-F in brain related functions, whereby the PD GWAS SNP identified within the SVA disrupts the binding sites of these TFs which could have influence over the expression patterns of *INPP5F* within the brain.

The SNP variation within the *INPP5F* SVA-F and SVA-D elements described in **section 4.2.4.1** does not represent the largest source of genetic variation within these elements, as larger primary sequence polymorphisms were also described (**sections 4.2.4.2 and 4.2.4.3**). Both the SVA-F and SVA-D elements were found to have VNTR polymorphisms with multiple alleles present within the tested human DNA samples (**figures 4.18 and 4.21**). None of the alleles found in either SVA-F or SVA-D were reported to have increased prevalence within the PD cohort tested compared to controls which may have indicated a novel potential risk variant (SVA-F - control N=175 PD case N=149, SVA-D - control N=169, PD case N=174). However, the samples tested only represent one ethnicity as the cohort was or primarily Estonian origin and utilises small sample sizes (<200). VNTR copy number polymorphisms have been previously associated with genetic predispositions for both the serotonin and dopamine transporters which may correlate with disease [184-186]. It would be of interest to further explore the identified SVA-F and SVA-D VNTR polymorphisms

further in larger more ethnically diverse cohorts for the potential of identifying novel variants that could be implicated in disease.

The data presented within this chapter alongside that reported for the *LRRK2* SVA-C in chapter 3 represent two examples of the effects of SVA retrotransposons within documented PD risk loci and provide evidence for the necessity for further understanding of the novel mechanisms that underlie retrotransposable element function in neurodegeneration. However, these examples only represent the effects of single isolated 'fixed' elements within the context of neighbouring genes and do not reflect the variation that may exist in PD as a result of retrotransposition events on a genome wide scale.

Chapter 5 – Identifying novel retrotransposon insertion polymorphisms in Parkinson's disease using next generation sequencing

Chapter 5 - Identifying novel retrotransposon insertion polymorphisms in Parkinson's disease using next generation sequencing

Thus far within this thesis, non-LTR retrotransposons, primarily the SVA sub-class, have been explored in context of known PD risk loci within the examples of both *LRRK2* and *INPP5F*/*BAG3*/*TIAL1*. The methods employed allowed the examination of retrotransposon effects within the context of specific characteristics such as expression changes and splicing with respect to neighbouring genes. However, this approach did not allow for global examination of retroelements within the context of PD in an unbiased approach. The focus within this chapter was the utilisation of next generation sequencing techniques, primarily retrotransposon capture sequencing (RC-Seq) and whole genome sequencing (WGS), to study the genomic localisation of novel retrotransposon insertion polymorphisms (RIPs) within the context of Parkinson's disease. Whole genome sequencing provided blanket coverage of the genome in an unbiased approach that was used to identify LINE-1, SVA and *Alu* elements, but can be limited in its detection sensitivity by the depth of sequencing used. This contrasted to the nature of RC-Seq enrichment which used capture probes to pull down LINE-1 sequences and improved sensitivity for low abundance LINE-1 elements such as somatic insertions that may only be found in small cell populations. This type of analysis was used as a proof of principle to assess the effectiveness of using these techniques for studying the roles of retrotransposons in a genome wide scale within the context of disease and provides only preliminary results due to a lack of large sample sizes and resources. The techniques employed here represent the first recorded use of both RC-Seq and WGS for the studying of

retroelements to identify both novel somatic and germline retro-transposition events within the context of PD.

It has been estimated that there exist approximately 1 million LINE-1 copies within the human genome, with only 80-100 of these being considered active and retrotransposition competent [187]. Previous literature had suggested that in some neurodegenerative disorders such as amyotrophic lateral sclerosis (ALS), an increase in retrotransposon activity which would lead to a “tsunami” of LINE-1 insertions could be pathogenic [109]. This argument poses an attractive idea given the strong evidence that already exists which suggests many forms of complex disorders such as PD are multi-faceted and are likely caused by a cumulation of mutations rather than mono-genic, although putative mono-genic forms of PD do exist, they only constitute a minority of total cases, with approximately 30% of familial and 3-5% of sporadic cases being mono-genic [7]. As part of the basis for performing these experiments, addressing any potential dramatic increases in retrotransposition events similar to that described in previous literature, was a key reason for performing these analyses. Both the RC-Seq and WGS pipelines allowed for assessment of both overall numbers of RIPs with the potential for studying of individual insertions of interest selected from the larger datasets.

The analysis of both the RC-Seq and WGS datasets utilised the TEBreak pipeline, which characterises retrotransposable elements by comparing the sequencing data to the reference genome to search for novel insertions. Using defined consensus sequences for each retrotransposable element sub-class, it classifies each insertion and provides information on, but not limited to: genomic location of insertion, TE

sub-class, percentage match to consensus, length of insertion (including details on truncations), orientations, target site duplications, basic annotations of the insertion (whether the insertion is located within a gene or known regulatory domain) and whether the insertion has been previously reported.

RC-Seq and WGS was performed on human PD patient derived DNA samples gifted from clinical associate Professor, Christos Proukakis in the Queen Square Institute of Neurology at University College London (sample details in **section 2.1.5**). A total of 16 samples were obtained which included two brain region DNA samples (frontal cortex and cerebellum) extracted from six PD patients and two healthy controls (neurodegeneration free). Of these samples, five PD and one control were processed for RC-Seq due to insufficient quantities of DNA to proceed with further processing with the other PD and control samples. Two of the same PD samples used for RC-Seq were also processed for WGS analysis (samples PD109 and PD348) details of which are provided in **table 5.1**. With respect to the context of PD, it would appear to have been more appropriate to use DNA extracted from the substantia nigra as this region is the most heavily implicated in the pathology of PD. However, in the latter stages of PD progression, the vast majority of the neurons and supporting cells within the substantia nigra suffer attrition and cell death resulting in inappropriate tissue for DNA extraction and analysis. Any retrotransposition events that may have occurred at the point of tissue acquisition may have been due to cellular death processes rather than PD making this tissue unsuitable for these analyses. In addition to these, samples from the Dyne Steele cohort were processed alongside the PD samples as controls to compare to by a colleague, Ana Illera. A total of nine healthy aged samples were used with a mean age of 88 (ranged between 78-94 years old). **Table 5.1**

summarises the PD and Dyne Steele sample information used for both RC-Seq and WGS analyses. Mitochondrial DNA sequencing (mtDNA-seq) was also performed for the majority of samples to be used for comparisons to the RC-Seq and WGS data, however this data was not processed in time to be reported within this thesis.

Table 5.1 – Breakdown of the sample I.Ds used for RC-Seq and WGS analyses. A total of five PD cases, one PD control and eleven healthy aged (HA) controls were used for RC-Seq with two of the PD samples also being used for WGS. Subsequent analysis using mitochondrial DNA sequencing (mtDNA-seq) was also performed on a subset of samples, the data of which is not presented within this thesis. Green – samples have been processed, pink – samples have not been processed for the corresponding analysis.

Sample I.D	Status	Age	Sex	RC-Seq	WGS	mtDNA-Seq
PDC01	Control	-	-			
PD109	PD	-	-			
PD139	PD	-	-			
PD184	PD	-	-			
PD348	PD	-	-			
PD359	PD	-	-			
09/24	HA	78	M			
09/26	HA	84	M			
09/31	HA	94	F			
11/06	HA	91	F			
11/07	HA	80	F			
11/22	HA	89	F			
11/29	HA	89	M			
14/04	HA	89	F			
14/46	HA	94	F			
15/01	HA	90	M			
15/28	HA	91	F			

Figure 5.1 represents the flow through of the bioinformatic pipeline that was used for both RC-Seq and WGS and included the general quality control (QC) of raw FastQ sequencing files, alignment to the reference genome (hg19), inputting to TEBreak, resolving and filtering.

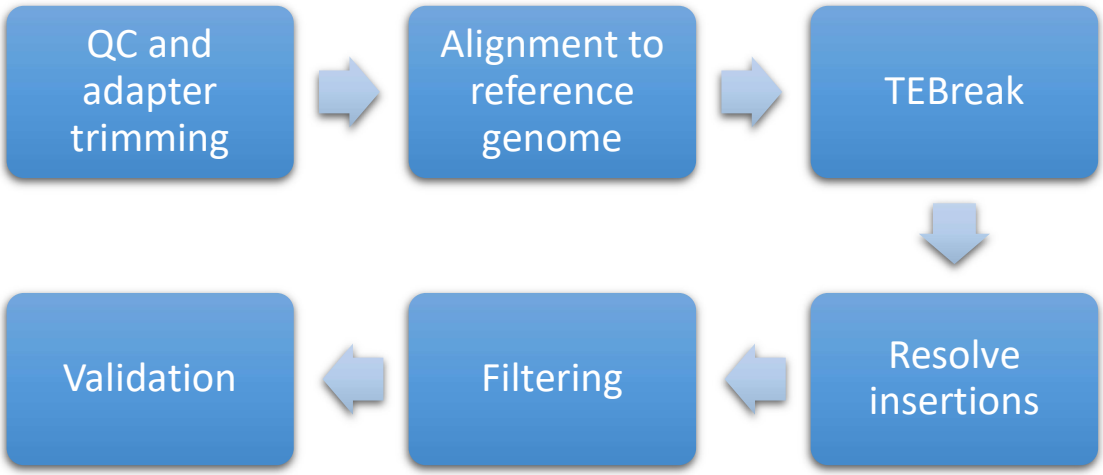


Figure 5.1 – RC-Seq and WGS pipeline for the bioinformatic analysis of retrotransposon insertion polymorphisms. Raw FastQ files from sequencing were quality controlled (QC) using FastQC software (https://www.bioinformatics.babraham.ac.uk/projects/FastQc/), aligned against the reference genome (hg19), inputted to TEBreak and resolved and then filtered using manually defined parameters. The resulting identified insertions could then be validated using PCR validation.

The filtering parameters in TEBreak were user defined and could be altered to increase or decrease the stringency. The filters used are defined as follows:

- Split read – Specifies the minimum number of supporting split read mappings for insertions. Higher stringencies are often desired for the confident calling of somatic insertions to decrease the number of false positives. Split reads are defined as reads which have partial alignments to both the reference genome and another sequence and usually defines the junction of a non-reference transposition event in which a novel sequence has inserted into the genome.
- Disc read - the minimum number of discordant read pairs. This refers to the read within a pair that is not matched to the reference genome in paired ended sequencing.
- Conslen – specifies the minimum total consensus sequence length required for a read to match the consensus TE sequences.
- Eltmatch – this defines the minimum element match (%) of each read to the consensus TE sequence.
- Refmatch – this defines the minimum reference match (%) of each read to the reference genome.
- Maxvar – maximum number of variants (SNPs) allowed within a read.

5.1 Using retrotransposon capture sequencing (RC-Seq) to identify LINE-1 insertion polymorphisms and somatic variation within the context of Parkinson's disease

5.1.1 Introduction

Rapidly advancing genetic techniques such as next generation sequencing allows the genome to be characterised in greater detail than ever before. Large scale genetic studies including GWAS and exome sequencing are currently being expanded in Parkinson's disease in an attempt to unveil novel risk factors with a large focus on SNPs [188]. These studies generally identify novel coding variants or allude to novel risk loci that could be important. However, a large limitation exists within these studies, whereby variation that stems from repetitive elements is often excluded from analysis due to computational difficulties including difficulties mapping unique reads due to the repetitive nature of these elements. It is widely accepted that non-LTR retrotransposons have helped shape the human genome by driving genetic diversity via mutation, which is often useful, but can also be regarded as potentially pathogenic, for example in the case of an SVA insertion within the *TAF1* locus leading to X-linked dystonia Parkinsonism (XDP) [78, 189]. Retrotransposons play fundamental roles in both the structure and function of the human genome, with multiple studies suggesting their putative role in neuronal specific contexts such as altering the expression of neuronal genes [190, 191]. The ability of retrotransposons to target specific genes in a tissue specific context highlights the importance of genomic location of both static (reference) elements and mobile elements which lead to the generation of retrotransposon insertion polymorphisms.

Parkinson's disease is a complex disorder with multiple likely causative factors rather than one specific cause although monogenic forms of the disease do exist [192]. Given this, it is interesting to consider the potential role of complex mutational drivers such as retrotransposons in the aetiology of complex diseases. Retrotransposon insertion polymorphisms (RIPs) are mediated by the LINE-1 machinery and are capable of causing multiple insertion mutations across a wide range of genes giving the potential to impact a variety of cellular processes. It is therefore important to understand the RIP landscape in PD and how these polymorphisms could potentially impact key genes and haploblocks involved in the disease pathogenesis. LINE-1 elements have been implicated in Parkinson's disease via two mechanisms whereby the activation of LINE-1 resultant from distressed mitochondria and the reduction of LINE-1 methylation as a result of smoking led to an increase in activity [112, 113]. The activity of LINE-1 is associated with LINE-1 propagation and the increase in novel retrotransposition events which could be linked with the progression of PD.

Within this chapter, a technique termed retrotransposon capture sequencing (RC-Seq) was utilised which allows a high sensitivity approach for the detection of both novel L1 somatic insertion polymorphisms which may only be present within single cells within a bulk tissue as well as polymorphic L1 insertions which could be found across multiple tissue types within an individual and could be thought of as germline variants. It would be possible for a novel insertion within an individual to be found across multiple tissue types depending on when the insertion occurred, for example, insertions during embryogenesis occur resulting in somatic mosaicism [193]. For an identified L1 insertion to be confidently classed as a germline variant, multiple tissue

developmental lineages would be necessary (e.g. central nervous system tissue, blood or liver). Due to sample acquisition limitations, only PD brain tissues were utilised (frontal cortex and cerebellum for each individual) in this protocol, and as such, potential germline insertions that were identified (present within both tested tissues) were classed as polymorphic and were to be distinguished from somatic polymorphisms (only present in one tested tissue). Within this chapter, only LINE-1 polymorphisms were analysed due to the use of specific probes within the RC-Seq protocol that isolates LINE-1 elements from other retroelements (full RC-Seq methods outlined in **sections 2.2.13 and 5.1.3.1**).

5.1.2 Aims and hypothesis

To address if dramatic increases in retrotransposition event number and/or localisation of insertions could be linked to neurodegeneration in the context of Parkinson's disease by comparing brain extracted DNA from PD patients and healthy aged individuals. In order to assess this, RC-Seq was employed which enriches for L1 insertions prior to sequencing to identify both:

- Increased or novel L1 retro-transposition events occurring in the CNS of individuals with PD which would correlate with neurodegeneration of these neurons.
- Specific germline/polymorphic L1 insertions predicted to be a predisposing factor for PD.

Hypothesis: Endogenous non-LTR retrotransposons of the LINE-1 family result in novel insertions that can act as both germline predisposition variants or *de novo* mutations that affects the progression of PD in a cumulative way.

5.1.3 Methods

5.1.3.1 RC-Seq

The RC-Seq approach used was adapted from the protocols outlined in Sanchez-Luque *et al.* 2016, with full methods described in **section 2.2.13** [140]. This approach is a next generation sequencing method which enriches for LINE-1 insertions using specific hybridisation probes which bind to the most extreme 5' and 3' termini or L1 elements (**figure 5.2**). In this way, it can distinguish between full length or partial insertions and polymorphic elements with respect to the reference genome build.

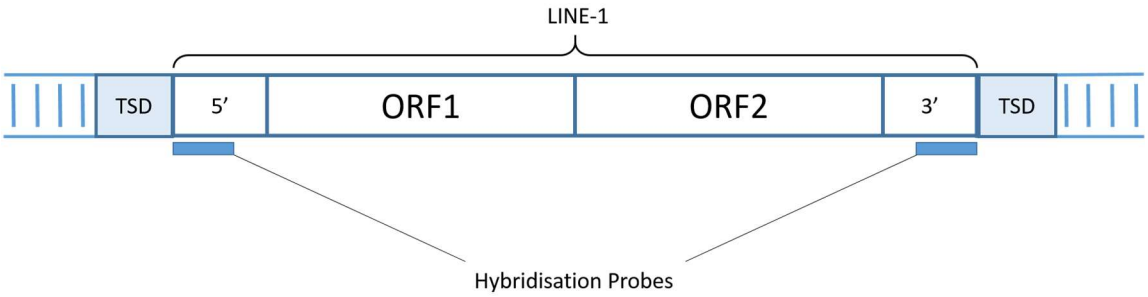


Figure 5.2 – Schematic representation of an *in-situ* LINE-1 element within the human genome. The hybridisation probes indicated bind to the extreme 5' and 3' termini of the L1 which are used to capture the corresponding 5' and 3' ends for sequencing in the RC-Seq protocol allowing for identification of both full length and truncated LINE-1 elements. ORF1 and ORF2 (open reading frame) encode the nucleic acid chaperone and endonuclease/reverse transcriptase L1 machinery, respectively.

RC-Seq libraries were sequenced using the Illumina NextSeq 500 platform in accordance with the Sanchez-Luque *et al*. protocol. The sequencing generated 8 FastQ files per individual (4 forward (R1) and four reverse (R2)) which were analysed using the TEBreak pipeline designed by Adam Ewing and available through GitHub (https://github.com/adamewing/TEBreak) for the detection of transposable elements using next generation sequencing data.

The final filtering parameters used for RC-Seq analysis (post-TEBreak) were used as follows:

- Minimum split read – 8 and 4 for polymorphic and somatic insertion detection respectively
- Minimum Discordant read pair (Disc read) – 4
- Minimum consensus sequence length (Conslen) – 150bp
- Minimum element match (Eltmatch) – 0.90 (90% match)
- Minimum reference genome match (Refmatch) – 0.95 (95% match)
- Maximum number of variants (Maxvar) – 2 SNPs

Bracketed descriptions denote the parameter as viewed within the filtering scripts. The descriptions for each parameter used are outlined in the chapter 5 introductory text (**page 222**).

5.1.3.2 Validation of LINE-1 insertion polymorphisms by PCR

Four truncated non-reference LINE-1 insertions were selected for PCR validation which include two previously reported insertions (chr4:47007784 and chr3:656434) and two novel insertions which were previously unknown (chr2:140179489 and chr8:109058606) (further details in **section 5.1.4.1**). An empty/filled site (ES/FS) PCR approach was used alongside multiplex PCR (multiple primer sets used simultaneously to generate multiple amplicons) for the validation of the four putative polymorphic insertions across 6 patients (5 PD case and 1 control). Validation of one somatic insertion was also attempted within one case sample (PD109) which proved unsuccessful.

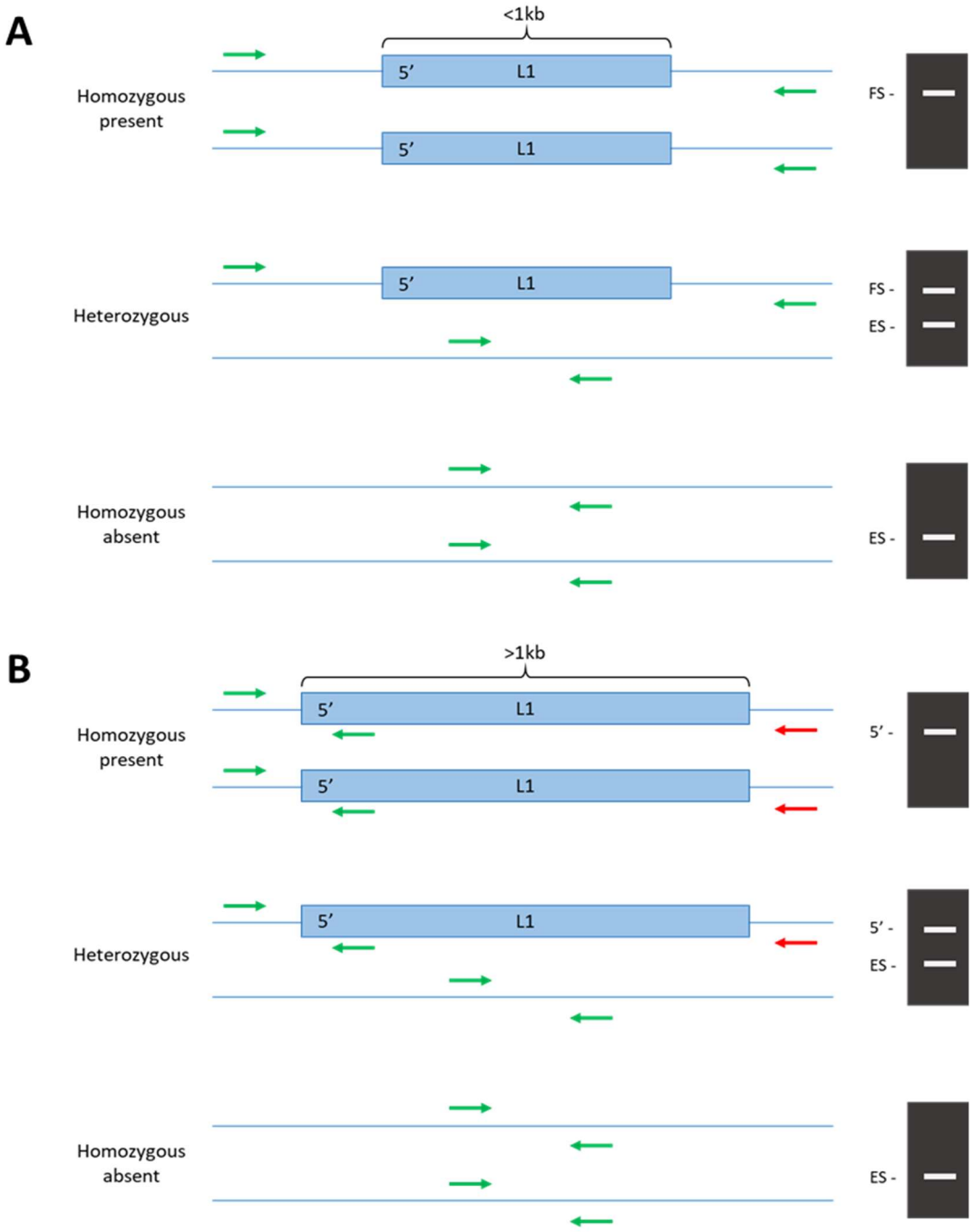


Figure 5.3 – (A) Schematic representation of an empty site/filled site (ES/FS) PCR used for validation of LINE-1 retrotransposon insertion polymorphisms (RIPs) in both the RC-Seq and whole genome sequencing (WGS) protocols. This method was employed when the target L1 insertion was small (<1kb) to allow amplification of both the filled and empty sites using one primer pair. **(B)** Schematic representation of a multiplex PCR in which three primers are used to amplify either the 5′ or 3′ end of an L1 RIP (5′ amplification shown) as well as the empty site, if present. The multiplex PCR method was chosen to validate L1 RIPs that were longer than 1kb in length (up to ~6kb for a full length non truncated LINE-1) as reliable amplification of both long and short (empty site) amplicons was difficult.

Table 5.2 - Primer sets and PCR reaction mixes that were used for the validation of the four non-reference LINE-1 insertions previously described. For multiplex reactions utilised three primers, the internal primer specified was either a 5' or 3' primer which was dependent on the orientation of the target element.

Target	PCR	Primers	PCR reaction mix
Chr4:47007784 (polymorphic)	ES/FS	F – GCCTCCTGAAGATCCAAGGA R – CACAGTTTTGCTAAGCCCCA	- 12.3µl nuclease free water - 4µl green GoTaq buffer (5X) - 1.6µl $MgCl_2$ (25mM) - 0.4 dNTPs (10mM) - 0.3µl F primer (20µM) - 0.3µl R primer (20µM) - 0.1µl GoTaq Hot start polymerase (1U/µl) - 1µl DNA template (5ng/µl)
Chr3:656434 (polymorphic)	ES/FS	F – TCCAGGCATCTTCACACCTT R - CTGATCGCACTGGTCAAACA	
Chr3:656434 (polymorphic)	3'	F – TCCAGGCATCTTCACACCTT R – CACCAGCATGGCACATGTAT	
Chr2:140179489 (polymorphic)	Multiplex	F – GGGGTTTGATTGCCTTGTGAAATGTCC R – AACCACGGGTGTGCCTGTGTAG 5' – AACTCCCTGACCCCTTGC	- 12µl nuclease free water - 4µl green GoTaq buffer (5X) - 1.6µl $MgCl_2$ (25mM) - 0.3µl F primer (20µM) - 0.3µl R primer (20µM) - 0.3µl internal primer (20µM) - 0.1µl GoTaq Hot start polymerase (1U/µl) - 1µl DNA template (5ng/µl)
Chr8:109058606 (polymorphic)	Multiplex	F – GGGGTTTGATTGCCTTGTGAAATGTCC R – AACCACGGGTGTGCCTGTGTAG 5' - AACTCCCTGACCCCTTGC	
Chr7:13890459 (somatic)	Multiplex	F – CACCAGCATGGCACATGTAT R – GAATGGCATGGATGCTACCTCTTCTT 3' – CACCAGCATGGCACATGTAT	

The PCR cycling conditions used for the amplification of the amplification of all polymorphic insertions was as follows: 95°C – 2 mins; 95°C for 30 secs, 60°C for 30 secs, 72°C for 30 secs for 30 cycles; with a final extension at 72°C for 2 minutes. For the validation of somatic insertions, the identical PCR cycling approach was employed with 40 cycles instead of 30 cycles to increase the sensitivity of somatic detection.

5.1.3.3 Haploblock analysis

PD related haploblocks were generated by using the bedtools suite of manipulation tools for bioinformatic analysis of large data sets via the Galaxy hub (https://usegalaxy.org/). The defined list of 90 identified PD GWAS risk SNPs from Nalls *et al.* 2019 was used in conjunction with the human haploblock list defined by Berisa *et al.* 2016 to generate a list of 77 haploblocks that contained PD GWAS risk SNPs [26, 194]. Co-ordinates for all the polymorphic and putative somatic insertions for the PD and healthy aged (HA) groups were independently combined into single bed files and sorted by ascending genomic location using the 'bedtool SortBED' function. The output sorted file was then merged using the 'bedtools MergeBED' function, which combined overlapping genomic intervals into single entries to avoid duplication of results in downstream analysis. The sorted merged bed file was then intersected with the previously generated PD haploblock list using the 'bedtools Intersect Intervals' function, which identified any genomic locations which overlap between the two files and returns single entries for each intersection identified.

5.1.3.4 Pathway analysis of gene sets containing non-reference LINE-1 RIPs using DAVID

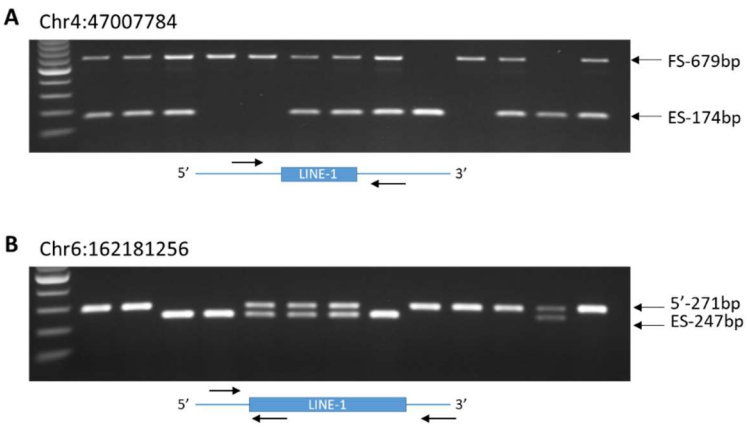
All pathway analysis presented was performed using the database for annotation, visualisation and integrated discovery (DAVID) pipeline (https://david.ncifcrf.gov) [195, 196]. Gene lists were generated consisting of those genes which contained putative LINE-1 non-reference insertions (polymorphic and somatic) identified by TEBreak. Selecting multiple resources within the DAVID analysis to analyse the TEBreak datasets allowed the assessment of multiple facets including:

- KEGG pathway – Provides pathway analysis of genes involved in a variety of processes including metabolism, cellular processes, genetic and environmental information processing, human diseases and drug development (https://www.genome.jp/kegg/pathway.html#genetic).
- Geno ontology (GO) molecular function – molecular level activities performed by gene products such as catalysis or transport.
- Gene ontology (GO) cellular component – gives the cellular location of a gene product i.e. mitochondria, nucleus etc.
- Gene ontology (GO) biological processes – provides information of a genes involvement in large processes involving multiple molecular processes e.g. DNA repair or signal transduction.
- UP tissue – provides information regarding which tissues the inputted genes are expressed in.

5.1.4 Results

5.1.4.1 PCR validation of RC-Seq libraries

Prior to analysis, validation of the RC-Seq libraries is required to ratify that the bioinformatic algorithms employed i.e. TEBreak, are accurate. However, due to limited sample sizes in the PD cohort (n=5), PCR validation of 13 non-reference polymorphic insertions was performed by a colleague, Abigail Savage, on a total of 24 individuals from an ALS cohort containing case and control samples that were prepared in parallel following the same pipeline as used for the PD and healthy aged cohorts. This allowed for a meaningful estimate of the accuracy, sensitivity and specificity of RC-Seq for the detection of true LINE-1 insertion polymorphisms and was to be used as a guideline only. **Figure 5.4** shows the validation of two non-reference polymorphic LINE-1 elements, chr4:47007784 and chr3:162181256, using ES/FS and multiplex PCR respectively which confirmed the genotype of each insertion across 13 samples. Using these approaches a total of 13 insertions were validated across 24 individuals of which the false positive rate (FPR) was determined at 0.03 with a false negative rate (FNR) at 0.14. The FPR of 0.03 gives a 0.97 specificity value that the bioinformatic calls that were made were true. The FNR of 0.14 indicates that there are some insertions being missed during the analysis which could be due to a slightly high stringency and could be reviewed in future work however a high stringency also improves the FPR. These metrics provided confidence that the bioinformatic calling of true insertions was high and the data could be trusted, with only low numbers of true insertions being undetected (indicated by the FNR of 0.14).



C

	L1
False positive rate (FPR)	0.03
False negative rate (FNR)	0.14
Accuracy	0.93
Sensitivity	0.86
Specificity	0.97

Figure 5.4 – PCR validation and summary statistics of the ALS RC-Seq libraries processed concurrently with the PD libraries with the identical processing pipeline being employed. The ALS summary statistics were presented due to the significantly larger sample sizes that were processed compared to the PD group (ALS n=24, PD n=5). Accurate estimations of the PD group RC-Seq library accuracy, sensitivity and specificity were not possible with limited library sample sizes. (**A**) An empty/filled site (ES/FS) PCR of a 5' truncated LINE-1 insertion at chr4:47007784. (**B**) Multiplex PCR validation of a LINE-1 insertion at chr6:162181256 which utilises three primers to amplify the 5' and empty sites but did not produce a filled site amplicon as the product size was too large. (**C**) Summary statistics of 13 polymorphic LINE-1 insertions (10 previously reported and 3 novel previously unknown insertions) which were calculated using the following calculations:

True positive (TP) = present in both TEBreak and PCR

True negative (TN) = absent in both TEBreak and PCR

False positive (FP) = present in TEBreak but absent in PCR

False negative (FN) = absent in TEBreak but present in PCR

False positive rate (FPR) = FP/(FP+TN)

False negative rate (FNR) = FN/(FN+TP)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

The formula's for generating accuracy, sensitivity and specificity were based on those specified by Baratloo *et al.* 2015 [197].

Having demonstrated the TEBreak pipeline and parameters used as being suitable for further analysis, the validation of multiple insertions in the PD datasets was undertaken. Using a PCR approach, validation of four non-reference polymorphic LINE-1 insertions was performed. Two of the insertions validated (chr4:47007784 and chr3:656434) had been previously reported in literature and included one of the previously validated insertions in the TEBreak validation (**figure 5.4**), insertion chr4:47007784. The insertion at chr4:47007784 had been reported in 9 independent studies (Stewart, C. *et al.* 2011, Lee, E. *et al.* 2012, Kuhn, A. *et al.* 2014, Iskow, R.C. *et al.* 2010, Sudmant, P.H *et al.* 2015, Helman, E. *et al.* 2014, Ewing, A.D. *et al.* 2010, Shukla, R. *et al.* 2013 and Tubio, J.M.C. *et al.* 2014) as well as dbRIP (database of retrotransposon insertion polymorphisms in humans - http://dbrip.brocku.ca/) [198-207]. The insertion at chr3:656434 had been reported in 6 independent studies (Lee, E. *et al.* 2012, Kuhn, A. *et al.* 2014, Sudmant, P.H *et al.* 2015, Ewing, A.D. *et al.* 2010, Shukla, R. *et al.* 2013 and Tubio, J.M.C. *et al.* 2014) which gave both of these targets a high chance of validation given they had been reported in a variety of studies and were chosen as the first targets of interest [200, 201, 203, 205-207]. **Figure 5.5** demonstrates the PCR validation of the two polymorphic non-reference LINE-1 insertions that had previously been described in the literature, with the correct validation of all samples with no discordance between TEBreak calling and PCR data. Both ES/FS and 3' amplification was performed on the insertion at chr3:656434 due to unclear amplification of the FS product in the ES/FS PCR. This confirmed the hemizygous presence of the L1 insertion, as predicted by TEBreak.

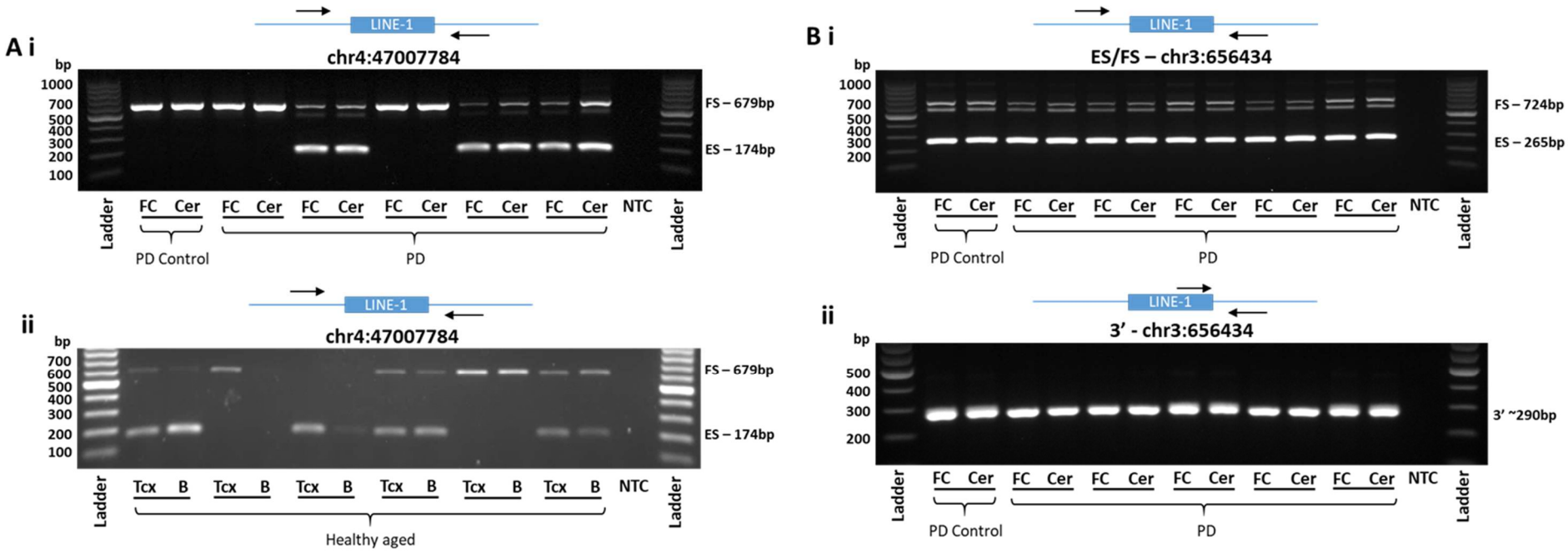


Figure 5.5 – Empty/filled site (ES/FS) PCR validation of two non-reference LINE-1 insertions (**A**-chr4:47007784 and **B**-chr3:656434) in PD and healthy aged samples. Both insertions had been previously reported in multiple studies as non-reference retrotransposon insertion polymorphisms. Due to the non-specific amplification of the filled site amplicon for the chr3:656434 PCR, a 3' only amplification was also performed to validate the presence of a LINE-1 sequence in all samples. No discordances between TEBreak bioinformatic calling and PCR validations were reported (true positive rate = 1.00) for either target. Abbreviations: FC – frontal cortex, CER – cerebellum, Tcx – temporal cortex, B – blood, ES – empty site, FS – filled site and NTC – no template control.

Two further non-reference polymorphic LINE-1 insertions were also validated which had not been previously identified. Insertions at chr2:140179489 and chr8:109058606 were identified in samples PD109 and PD139 respectively. As the insertions were larger than 1kb, multiplex PCR was utilised to validate both insertions which would amplify both the empty site and 5' end of each insertion if present.

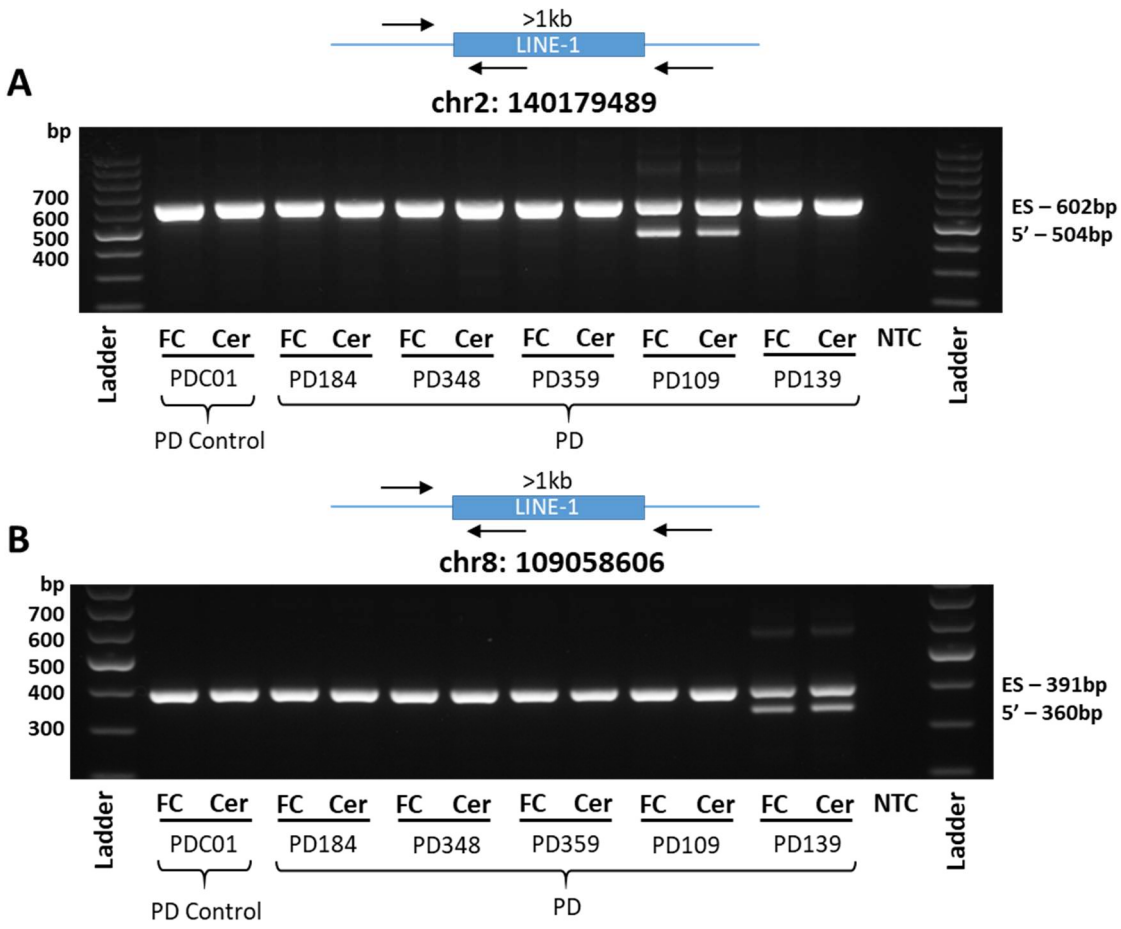


Figure 5.6 – Multiplex PCR validation of two non-reference polymorphic LINE-1 insertions that had not been previously reported. Insertions at chr2:140179489 was reported in sample PD109 and chr8:109058606 was detected in sample PD139 and not reported within the other samples. PCR validation confirmed the presence of these insertions in the predicted samples as indicated by the presence of 5' amplifications in a hemizygous genotype within both samples. Abbreviations: FC – frontal cortex, CER – cerebellum, NTC – no template control and ES – empty site.

Figure 5.6 indicates the validation of both the novel LINE-1 insertions at chr2:140179489 and chr8:109058606, providing evidence of LINE-1 mediated mutations in the brain tissue of Parkinson's disease samples and further increasing the confidence of the datasets produced by the RC-Seq/TEBreak pipeline. The L1 insertion at chr2:140179489 is located within an intergenic region with the nearest gene (NXPH2) being approximately 644kb away whilst the L1 insertion at chr8:109058606 is located within intron 2 of the RSPO2 gene. Interestingly, RSPO2 has a role in the promotion of midbrain dopaminergic neurogenesis which could be important within Parkinson's disease [208].

The validation of one somatic insertion was attempted as a proof of principle to ratify RC-Seq as a valid approach for the detection of somatic insertions. TEBreak identified a putative somatic insertion in the frontal cortex of sample PD109 at chr7:13890459 which could not be validated by multiplex PCR (**figure 5.7**). The putative somatic insertion at chr7:13890459 was located ~40kb downstream of the ETV1 gene which was not strongly associated with any neurological disorder using literature searches and was selected purely as a random insertion for validation. Due to the lack of successful validation of somatic insertions within this study, all somatic insertions detected by TEBreak were labelled as putative until further validation is proven.

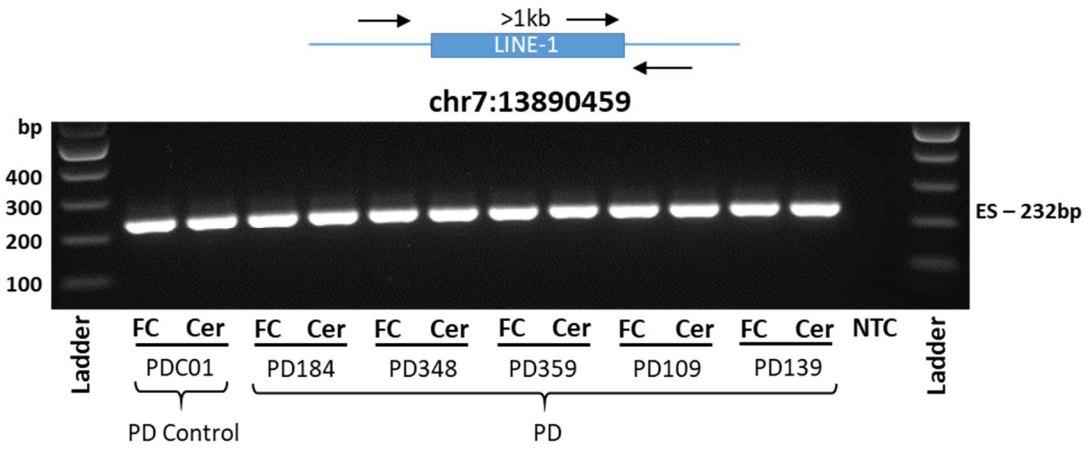


Figure 5.7 – Multiplex PCR of a putative somatic insertion which was identified by TEBreak in the frontal cortex of sample PD109 (FC) at chr7:13890459. The PCR design utilises primer sets to amplify both the empty site and potential 3' end of the putative insertion. The PCR indicated amplification of the empty site only across all samples tested with no somatic insertion detected in the PD109 frontal cortex. FC – frontal cortex, CER – cerebellum, NTC – no template control.

To further improve the credibility the RC-Seq libraries for detecting true L1 retrotransposon insertion polymorphisms, the percentage of fixed (present in the reference genome that are not mobile) full length L1s that were detected within the PD libraries was calculated (n=10, using the two tissue libraries from 5 PD individuals). To do this, genomic co-ordinates of full length human specific L1s (L1HS) were downloaded from UCSC genome browser and curated to remove those insertions that had evidence for being RIPs using data from Stewart *et al* 2011 which produced a list of 252 L1HS elements [199]. From this list two bed files were generated containing 100bp co-ordinates of the 5' and 3' ends of each of the identified L1HS elements. Using the bedtools multicov tool, the numbers of uniquely mapped reads over the generated bed files could be calculated within each bam file for each library. **Figure 5.8** shows the percentages of the L1HS elements which

contained a minimum of 8 uniquely reads over the 5', 3' and 5'+3' ends of the elements. The data indicates that using a combination of 5' and 3' captures to accurately map L1 elements gives approximately 96% coverage of the fixed L1HS elements. This further validates the pipeline used is appropriate for detecting L1 elements and that the data presented within this chapter has a high degree of confidence.

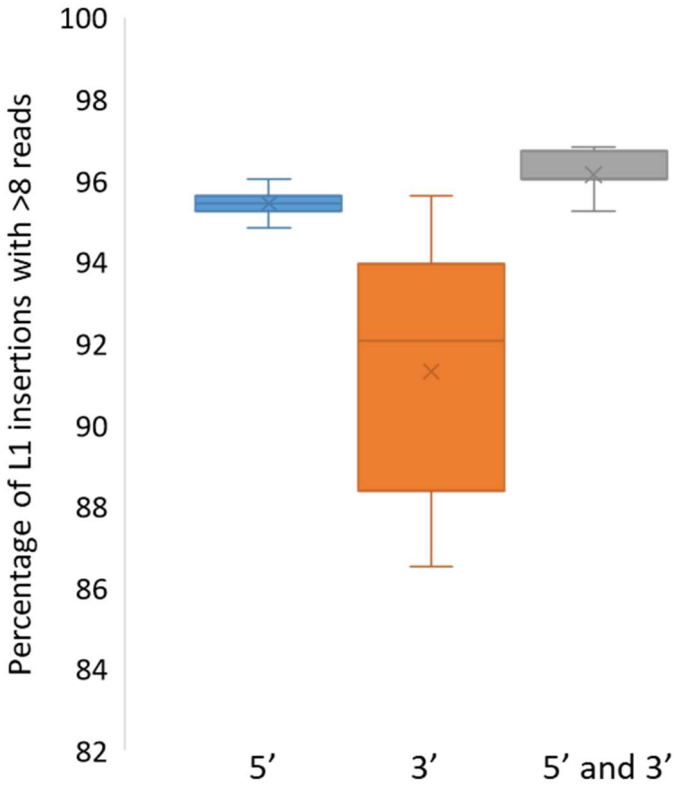


Figure 5.8 – The percentage of full-length fixed human specific L1s (L1HS) that were detected within the PD RC-Seq libraries that had a minimum of 8 uniquely mapped reads (n=10 (frontal cortex and cerebellum libraries from 5 PD individuals)). Average coverage of L1HS elements: 5'=95.4%, 3'=91.3% and 5' + 3'=96.2%.

5.1.4.2 Analysis of the TEBreak identified LINE-1 polymorphic and putative somatic insertions within PD

Primary analysis of the TEBreak output focused on comparing the overall numbers of insertions between the PD samples against the healthy aged control group. The previously described PD control (PDC01) was excluded from analysis due to only having one sample available (n=1) and all further comparisons were made between the PD group (n=5) and the healthy aged control group (n=11). The polymorphic insertions identified were filtered for a minimum cut-off of 8 split reads (higher stringency) in order to reduce the false positive rate. Polymorphic insertions were differentiated from putative somatic ones by using multiple manual filtering parameters which involved selecting those insertions which appeared in both sampled tissues (frontal cortex and cerebellum for PD and temporal cortex and blood for the healthy aged samples). Elements which did not satisfy at least one of the appropriate alignment co-ordinates “TE align start (5’)” and “TE align end (3’)” with values of the 0-54bp and 5291-6038bp ranges were excluded, as these values specify the binding sites of the LNA-probes within the LINE-1 pull-down. Insertions with alignments outside of these ranges would be indicative of noise or false positive elements and were therefore excluded. Elements which were identified as the L1PA2 sub-class of LINE-1 were also removed as these elements are not classified as transposition competent.

The manually curated results indicated no statistical difference in the absolute numbers of polymorphic non-reference LINE-1 insertions between individuals with Parkinson’s disease and healthy aged individuals with no neurological pathologies

(**figure 5.9**). However, there appears to be a trend towards more polymorphic insertions within the PD group compared to the healthy aged group with an 11.4% increase in the average L1 insertions within the PD group. This indicated that there was an increase in retrotransposition events within the PD samples but was not statistically significant due to low N numbers.

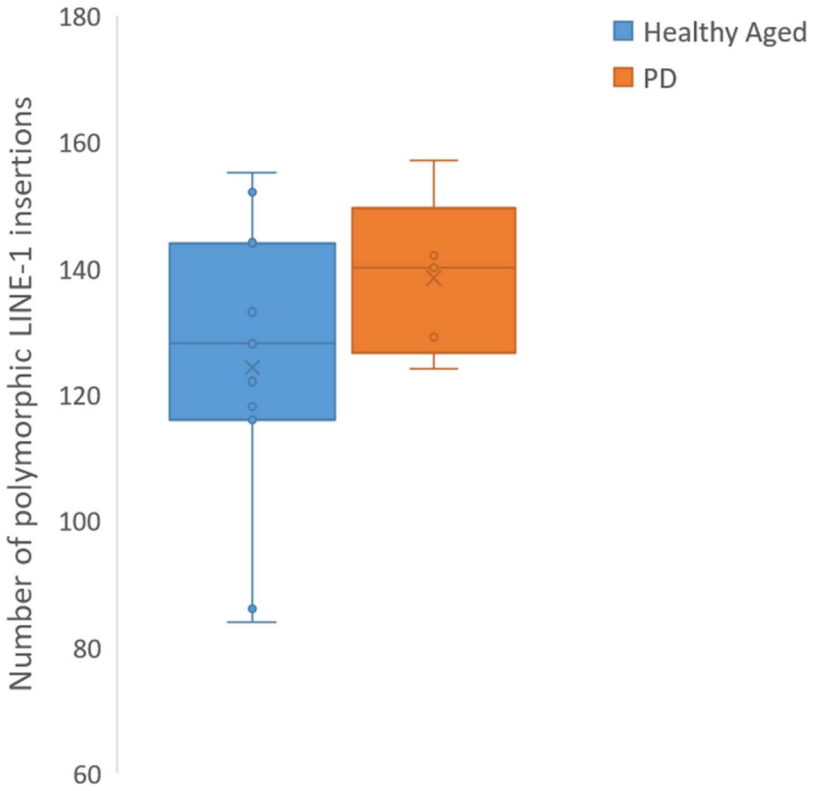


Figure 5.9 – LINE-1 retrotransposon insertion polymorphisms identified within healthy aged (HA) and PD case samples using the RC-Seq/TEBreak pipeline. There was no significant difference in the number of polymorphic LINE-1 insertions between healthy aged (average - 124.3 insertions) and PD (average – 138.4 insertions) samples (p-value=0.142, calculated using a two-sample t-test assuming unequal variance (HA n=11, PD n=5). There was a trend towards more polymorphic L1 insertions within the PD group with an average of 11.4% more L1 insertions present compared to the healthy aged group. Tissues included: HA=temporal cortex and blood, PD = frontal cortex and cerebellum.

In order to assess the numbers of putative somatic insertions present, the split read filtering parameter was altered from 8 split read minimum (used for polymorphic insertions) to a 4-split read minimum, which allowed for a less stringent cut off and a higher true positive rate at the sacrifice of a higher false positive rate also. The hypothesis being that true somatic insertions would be present within the tissue in a subsets of cell populations (potentially one cell for an adult somatic insertion or tissue specific for embryonic somatic insertions) which would be very difficult to detect using high stringency filtering parameters. The same exclusion criteria were used as for the polymorphic LINE-1 elements including filtering out the elements that only appeared in one of the tested tissues, TE alignments for the 5' and 3' which did not include the LNA probe binding sites and any L1PA2 elements. **Figure 5.10-A** indicates no statistical difference (p=0.108) in the numbers of putative somatic insertions between the healthy aged control and the PD group. Comparisons of the means shows an approximate three times increase in the number of putative somatic insertions in the PD group (average – 23.2) compared to the healthy aged control group (average – 7.8). With a higher average number of putative somatic insertions within the case group, the total numbers were broken down to show tissue specificity. **Figure 5.10-B** shows the breakdown of the four tested tissues across the healthy aged and PD groups (HA - temporal cortex and blood, PD – frontal cortex and cerebellum). There was no statistical difference in number of insertions between any of the groups when tested using a two-sample t-test assuming unequal variance. The healthy aged group had a higher variance in both the sampled tissues with several samples being considered outliers (indicated by circles outside of the box plots).

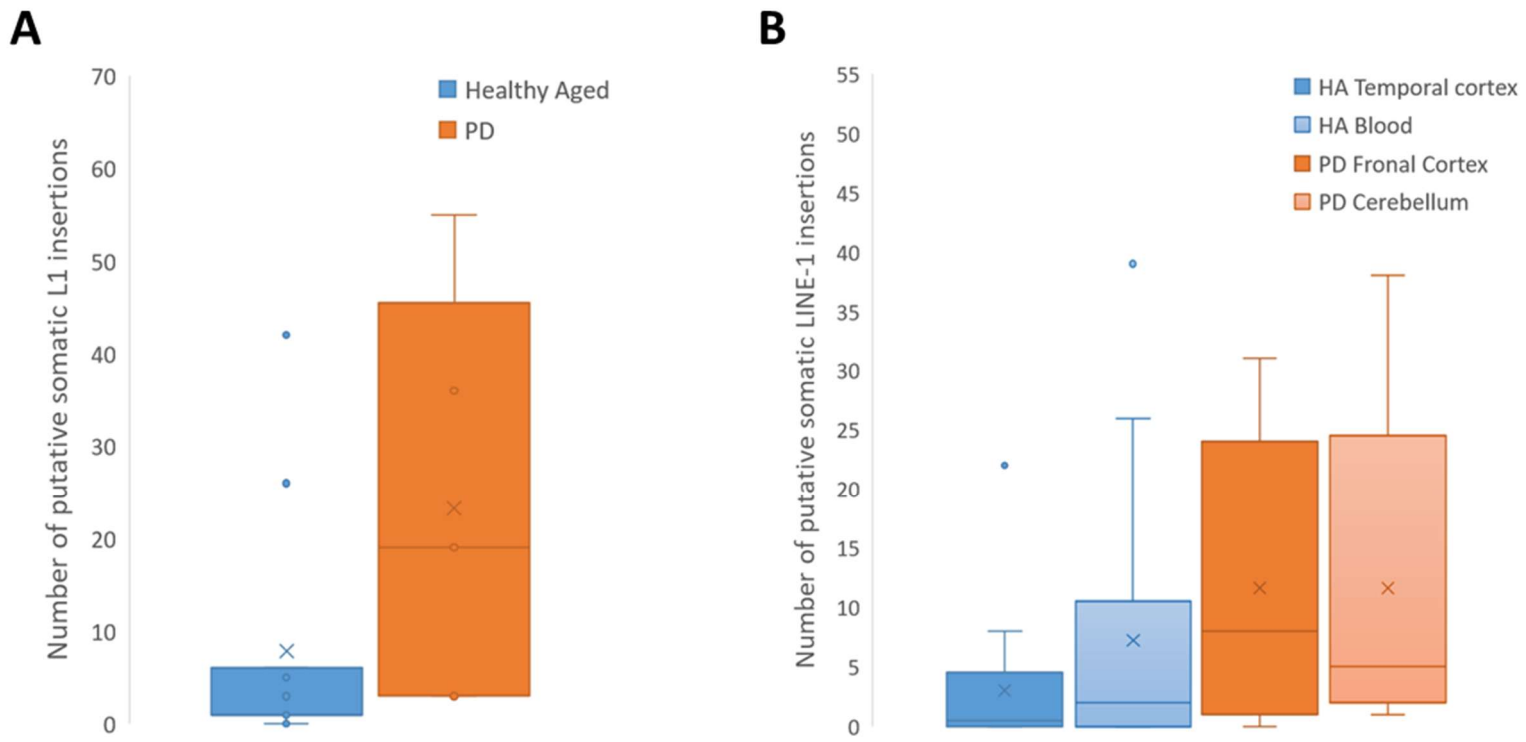


Figure 5.10 – The distribution of putative somatic insertions identified using the RC-Seq/TEBreak pipeline. **(A)** The number of insertions between healthy aged (average – 7.8) and PD (average – 23.2) individuals was not statistically different (p-value = 0.108) using a two-sample t-test assuming unequal variance. **(B)** The breakdown of tissue types indicates no difference between the healthy aged and PD or between the tissue types (temporal cortex (TC), blood, frontal cortex (FC) and cerebellum (CER)). Statistical analysis performed using a two-sample t-test assuming unequal variance (TC/blood p=0.188, TC/FC p=0.121, TC/CER p=0.177, blood/FC p=0.556, blood/CER p=0.607, FC/CER p=1.000). (HA n=11, PD n=5).

5.1.4.3 Functional inferences of non-reference LINE-1 insertions detected by RC-Seq

The analysis of both the number of polymorphic and putative somatic insertions yielded no statistically significant differences between the PD and control groups but did produce trends towards higher numbers of both polymorphic and somatic insertions within the PD group compared to healthy controls. In response to this result, it was decided to further examine potential differences in genomic locations of the insertions within both groups. It was possible that the genomic loci into which the LINE-1 insertions within both tested groups could be different.

The initial analysis measured the percentages of insertions that were found within intragenic (within annotated genes) compared to those found outside of genic areas. The assumption that LINE-1 element insertions within genic areas would have more scope for influencing gene function than those located further away from genes was taken. Interestingly, there were significantly less ($p=0.009$) polymorphic LINE-1 insertions within the PD case samples compared to the healthy aged group suggesting that LINE-1 insertions tended to insert into non-gene rich regions compared to healthy aged individuals (**figure 5.11**). By comparing the average percentages of insertions present within intragenic regions between the two groups, there is a 15.7% decrease of intragenic insertions in the PD group compared to the controls.

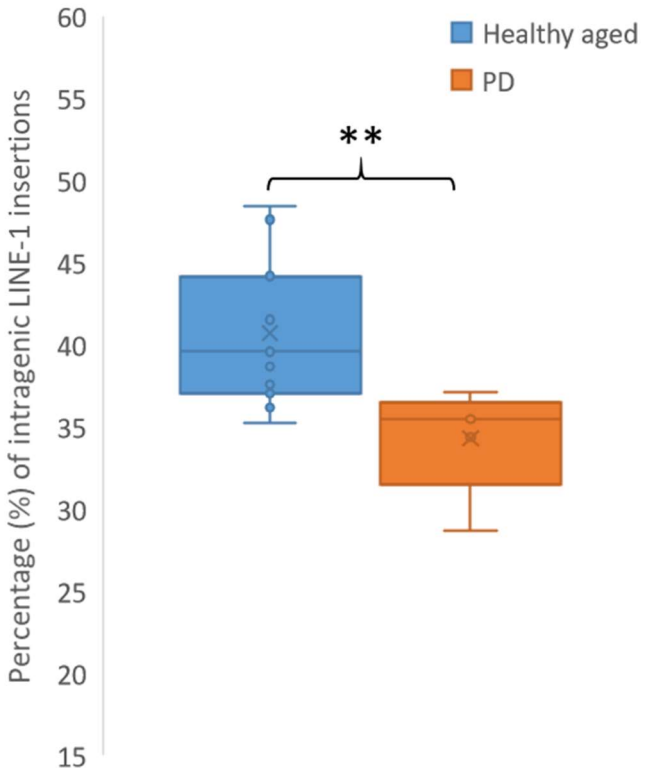


Figure 5.11 – The percentage of non-reference polymorphic LINE-1 insertions that were intragenic in the healthy aged and PD groups. There were significantly more insertions within the healthy aged individuals than the PD group (p=0.009) measured using a two-sample t-test assuming unequal variance. (HA n=11, PD n=5) ** p<0.01.

Having observed a difference in the percentage of LINE-1 insertions within intragenic regions, it was evident there was a difference in genomic location of the polymorphic LINE-1 elements. To study this further, haploblock analysis was undertaken whereby a PD related haploblock list was generated and used to assess if the LINE-1 insertions were located within PD related haploblocks which may give further information on potential function. To create the haploblock list, the defined genomic haploblocks from Berisa *et al.* 2016 were used in conjunction with the defined list of the top 90 PD nominated PD risk GWAS SNPs from Nalls *et al.* 2019 (further details outlined in **section 5.1.3.3**) [26, 194]. By utilising the intersect intervals function in the bedtools suite via the publicly available web tool, Galaxy (https://usegalaxy.org/), a PD related haploblock list containing 77 haploblocks was generated. Less than 90 haploblocks where returned due to some GWAS SNPs being present within the same haploblock (**table 5.3**).

Table 5.3 – The list of 77 PD related Haploblocks generated to be used for analysis of polymorphic and somatic L1 insertions. The list was generated using the bedtools intersect interval algorithm on Galaxy (https://usegalaxy.org/) using the list of the 90 nominated PD risk GWAS SNPs, from Nalls *et al.* 2019, with the human haploblock list defined by Berisa et al. 2016 [26, 194].

77 PD Haploblocks		Associated GWAS SNP	Nominated gene/closest (within 200Kb)
chr1	154770403 156336133	rs114138760, rs35749011, rs76763715	PMVK, KRTCAP2, GBAP1
chr1	159913048 162346721	rs6658353	FCGR2A
chr1	170557776 173097907	rs11578699	VAMP4
chr1	204681068 206073265	rs823118, rs11557080	NUCKS1, RAB29
chr1	226810860 229156248	rs4653767	ITPKB
chr1	232090252 233429284	rs10797576	SIPA1L2
chr2	16329735 18647423	rs76116224	KCNS3
chr2	167355970 169968236	rs2042477	KCNIP3
chr2	95326452 98995201	rs11683001	MAP4K4
chr2	135158578 137042794	rs57891859	TMEM163
chr2	101822329 102688765	rs1474055	STK39
chr3	150252004 151348730	rs73038319	SATB1
chr3	17891118 19125144	rs6808178	LINC00693
chr3	27840910 29142260	rs12497850	IP6K2
chr3	47727212 49316972	rs55961674	KPNA1
chr3	121974097 123517768	rs11707416	MED12L
chr3	159477890 161524504	rs1450522	SPTSSB
chr3	181511166 183769683	rs10513789	MCCC1
chr4	74592390 77130707	rs873786	GAK
chr4	694715 1478711	rs34311866, rs4698412	TMEM175, BST1
chr4	15147446 15927009	rs34025766	LCORL
chr4	17383322 18841874	rs6825004	SCARB2
chr4	77130707 79093979	rs4101061, rs6854006	FAM47E, FAM47E-STBD1
chr4	90231564 91560677	rs356182, rs5019538	SNCA
chr4	113870102 115666246	rs13117519	CAMK2D
chr4	169676825 170776510	rs62333164	CLCN3
chr5	132139649 134777401	rs1867598	ELOVL7
chr5	58524622 60935907	rs26431	PAM
chr5	101578769 103320005	rs11950533	C5orf24
chr6	132765669 134244243	rs4140646	LOC100131289
chr6	26791233 28017819	rs9261484	TRIM40
chr6	29737971 30798168	rs112485576	HLA-DRB5
chr6	31571218 32682664	rs12528068	RIMS1
chr6	71609510 73450097	rs997368	FYN
chr6	110304247 112345014	rs75859381	RPS12
chr7	65689809 68234074	rs199351	GPNMB
chr7	22507629 23471442	rs76949143	GS1-124K5.11
chr8	15991660 17387876	rs1293298	CTSB
chr8	11278998 13491775	rs620513	FGF20
chr8	21661737 22897057	rs2280104	BIN3
chr8	130381139 131639625	rs2086641	FAM49B
chr9	16659655 18661051	rs13294100, rs10756907	SH3GL2
chr9	33578334 34642243	rs6476434	UBAP2
chr10	15026068 16551767	rs896435	ITGA8
chr10	102949239 104380410	rs10748818	GBF1

chr10	120591353	122407323	rs72840788, rs117896735	BAG3, INPP5F
chr11	81266712	84381272	rs7938782	RNF141
chr11	9087317	10952027	rs12283611	DLG2
chr11	133000046	134205993	rs3802920	IGSF9B
chr12	39227169	40816185	rs76904798, rs34637584	LRRK2
chr12	46024229	47714793	rs7134559	SCAF11
chr12	122007651	124977980	rs10847864	HIP1R
chr12	132807034	133841511	rs11610045	FBRSL1
chr13	49383962	51591091	rs9568188	CAB39L
chr13	97519210	98938919	rs4771268	MBNL2
chr14	72889615	76444767	rs12147950	MIPOL1
chr14	35859593	38667725	rs11158026	GCH1
chr14	55233681	56216880	rs3742785	RPS6KL1
chr14	87635341	89497643	rs979812	GALC
chr15	61265836	63215222	rs2251086	VPS13C
chr16	49007926	52035823	rs6497339	SYT17
chr16	18643607	20150571	rs2904880	CD19
chr16	27445755	29036613	rs11150601	SETD1A
chr16	29036613	31382943	rs6500328	NOD2
chr16	52035823	53382572	rs3104783, rs10221156	CASC16, CHD9
chr17	41772087	43056905	rs12600861, rs12951632	CHRNB1, RETREG3
chr17	7317398	8306425	rs2269906	UBTF
chr17	39899810	41772087	rs850738	FAM171A2
chr17	43056905	45876022	rs62053943, rs117615688, rs11658976	CRHR1, WNT3
chr17	59312755	61545589	rs61169879	BRIP1
chr17	76263413	77298636	rs666463	DNAH17
chr18	39892648	42922106	rs1941685	ASXL3
chr18	47730584	51062185	rs12456492	RIT2
chr18	30264066	31780067	rs8087969	MEX3C
chr19	2098396	3019660	rs55818311	SPPL2B
chr20	5477850	7084073	rs77351827	CRLS1
chr21	38711704	40482902	rs2248244	DYRK1A

Using the insect interval feature in the bedtools suite through Galaxy, the total polymorphic and putative somatic LINE-1 insertions were assessed for overlap with the 77 generated PD haploblocks (**table 5.3**) in both the PD and healthy aged groups independently (**figure 5.12**). Fourteen of the seventy-seven PD haploblocks contained polymorphic insertions in the PD group (18.2%) with six haploblocks containing at least one putative somatic insertion. Within the healthy aged group, twelve of the seventy-seven PD haploblocks contained polymorphic L1 insertions (15.6%), similar to that in the PD group (18.2%). Half as many putative somatic insertions were identified within three PD haploblocks within the HA samples (3.9%) compared to the PD group (7.8%). This would suggest a tendency for somatic L1 RIPs

to insert into PD related haploblocks within the PD samples tested compared to the healthy aged controls, however due to the lack of validation for the somatic insertions identified, this result is subject to corroboration and requires further investigation.

No. of PD Haploblocks	Group	Polymorphic L1 insertions			Putative somatic L1 insertions		
		No. of haploblocks	No. of insertions	%	No. of haploblocks	No. of insertions	%
77	PD	14	14	18.18	6	6	7.79
	HA	12	12	15.58	3	3	3.90

Figure 5.12 – Haploblock analysis comparing the numbers of polymorphic and somatic insertions which were located within the PD related haploblock generated. A total of 77 PD related haploblocks were intersected with the total identified polymorphic and putative somatic insertions from both the PD and healthy aged groups. Within the PD group, 14 identified haploblocks contained a total of 14 polymorphic L1 insertions and 6 haploblocks contained a total of 6 putative somatic insertions. Within the healthy aged (HA) group, 12 PD haploblocks contained a total of 12 polymorphic L1 insertions and 3 haploblocks contained 3 putative somatic L1 insertions. (PD n=5, HA n=11).

Following the haploblock analysis which provided some insight into potential differences regarding the genomic locations of polymorphic and putative somatic LINE-1 insertions, pathway analysis using the database for annotation, visualisation and integrated discovery (DAVID) bioinformatic suite was undertaken (https://david.ncifcrf.gov/) (methods for DAVID analysis in **section 5.1.3.4**) [195]. As part of the TEBreak output, genes, pseudogenes and non-coding RNA sequences that contain a LINE-1 insertion are reported. Lists of genes that contained identified polymorphic LINE-1 insertions were generated and submitted to the DAVID tool for

analysis (**table 5.4**). Gene lists containing putative somatic L1 insertions only were also analysed using DAVID, however due to insufficient gene numbers within each group (PD and HA), no enriched pathways were reported due to insufficient statistical power. Gene names that could not be mapped to known entries within the DAVID databases were not analysed (red highlighted names - **table 5.4**). The unmapped gene names contain primarily novel uncharacterised transcripts (i.e. AC and AP transcripts) and long non-coding RNA transcripts (i.e. CTD- and RP11- transcripts) which could be useful for future analyses but were not utilised within this study. This resulted in two gene lists containing 109 and 134 genes for the PD and healthy aged groups respectively that could be processed for pathway analysis. Pathway analysis featured multiple databases of interest which are utilised to interrogate various features of a gene list including: UP tissue, gene ontology (GO) biological process, GO cellular component, GO molecular function and KEGG databases (descriptions of each database provided in **section 5.1.3.4**). **Figures 5.13a** and **5.13b** show the results of the DAVID pathway analysis for PD and healthy aged respectively.

Table 5.4 – Gene lists used for DAVID pathway analysis for both the PD and healthy aged (HA) identified polymorphic L1 insertions. Green highlights represent the genes which appear in both gene lists with red highlights indicating that the genes that could not be mapped by DAVID. The PD gene list contained 153 genes which were submitted to DAVID with 109 genes being correctly mapped. Of the 153 total genes 38 were unique to the PD list. The HA gene list contained 182 genes which were submitted to DAVID with 134 genes being mapped correctly with 67 of the total 182 genes being unique within the HA list.

PD gene list of 153 genes (38 unique genes)					
ADAMTS12	DCC	LINC00954	PRKCA	TYW1	RP11-328J2.1
ADHFE1	DDX58	LRP1B	PRKD1	UGGT2	RP11-328J6.1
ANO3	DSC1	LRRC4C	PTPRD	ZDHHC14	RP11-353N4.5
AOX2P	EFCAB5	LRRIQ1	PTPRR	ZFPM2	RP11-375D13.2
APOLD1	EPHA5	LRRTM3	RCAN2	ZNF33A	RP11-389J22.1
ARHGAP15	ERC2	MAP3K9	ROBO1	AC016723.4	RP11-396J6.1
ARHGAP24	EXOC6	MECOM	RORA	AC018717.1	RP11-403I13.9
ATXN1	F5	MED12L	RSPO2	AC090044.1	RP11-408H20.2
AUH	FBXO33	MYLK	RYR3	AC092657.2	RP11-417J8.3
C3orf67	FDCSP	NAALADL2	SCFD1	AC114765.1	RP11-417J8.6
C8orf44-SGK3	FHIT	NBEA	SCN5A	AP000705.7	RP11-430L16.1
C8orf46	GABRB1	NCOA1	SEMA6D	C9orf156	RP11-436D23.1
CADM2	GMDS	NEDD4	SGCD	CTD-2251F13.1	RP11-499E18.1
CDH12	GPC5	NLGN1	SGK3	CTD-2336H13.2	RP11-551L14.1
CDH23	GPR158	NOS1	SLC10A7	CTD-2378E21.1	RP11-554D15.1
CDH7	GRIK2	NPAS3	SLC39A8	CTD-2544M6.1	RP11-561I11.4
CDRT1	HIVEP3	NRCAM	SLC7A13	CTD-3006G17.2	RP11-594C13.1
CHODL	IGSF10	OCA2	SLC9A9	RP11-1102P16.1	RP11-625L16.1
CLNK	IMMP2L	OPA1	STK38L	RP11-113I24.1	RP11-642D21.2
CLTCL1	KDELC2	OXR1	SYT1	RP11-134F2.2	RP11-649G15.2
COLGALT2	KIAA1549L	PARK2	TBC1D1	RP11-141O11.1	RP11-805F19.2
CPA6	KIF13B	PARN	TDRD5	RP11-167N24.3	RP4-601K24.1
CRYZ	KIF16B	PDE11A	TOX2	RP11-168G16.2	TM4SF2
CSMD1	KLF12	PHACTR1	TRDN	RP1-116K23.1	
CTNNA3	KLHL3	PLD5	TRPC5	RP11-1E3.1	
CTNND2	LINC00355	PNPLA3	TTC37	RP11-307P5.1	

HA gene list of 182 genes (67 unique genes)						
ADAMTS12	CTNNA2	HIVEP3	NBEA	RYR3	ZFPM2	RP11-238I10.1
ADHFE1	CTNNA3	IGSF10	NEDD4	SCFD1	ZMAT4	RP11-260O18.1
APBB2	CTNND2	IMMP2L	NLGN1	SCN5A	ZNF622	RP11-29G8.3
APOLD1	CXADR	KAT2B	NRCAM	SEMA6D	ZRANB2	RP11-303E16.8
ARHGAP15	DACH2	KCNH5	OCA2	SGCD	AC018717.1	RP11-307N16.6
ATRNL1	DCC	KIAA1549L	OPA1	SGK3	AC068490.2	RP11-307P5.1
ATXN1	DDX58	KIF16B	OXR1	SIK3	AC090044.1	RP11-353N4.5
AUH	DLG2	KLF12	PARK2	SLC10A7	AC092657.2	RP11-375D13.2
AUTS2	DMD	KLHL3	PCAT4	SLC39A8	AC114765.1	RP11-396J6.1
BCO2	DNAH14	LINC00355	PCDH15	SLC4A10	AC133680.1	RP11-403I13.9
C3orf67	DNER	LINC00504	PDE11A	SLC7A13	AP000705.7	RP11-420N3.2
C4orf19	EFCAB5	LINC00534	PHACTR1	SLC9A9	C9orf156	RP11-430L16.1
C8orf44-SGK3	EPHA5	LINC00927	PLD5	SPATA13	CTD-2201G3.1	RP11-499E18.1
C8orf46	ERC2	LINC00954	PNPLA3	SPIRE1	CTD-2251F13.1	RP11-541P9.3
CA10	EVC2	LMO7	PRKCA	TAS2R14	CTD-2336H13.2	RP11-551L14.1
CADM2	F5	LRP1B	PRKD1	TDRD5	CTD-2378E21.1	RP11-554D15.1
CADPS2	FBXO33	LRRIQ1	PRR4	TOX2	CTD-2544M6.1	RP11-561I11.4
CDH12	FDCSP	LRRTM3	PTPRB	TRDN	CTD-3006G17.2	RP11-594C13.1
CDH4	FHIT	LSAMP	RAG1	TRIM37	RP11-1102P16.1	RP11-625L16.1
CHODL	GABRB1	MAP3K9	RALYL	TRPC5	RP11-1267H10.4	RP11-642D21.2
CLNK	GCSH	MARCH1	RBFOX1	TSHR	RP11-134F2.2	RP11-649G15.2
CLTCL1	GLIS3	MCC	RCAN2	TTC37	RP11-141O11.1	RP11-720L8.1
COMMD10	GPC5	MECOM	RIMS2	TYW1	RP11-167N24.3	RP11-805F19.2
CPA6	GPHN	MED12L	RNF180	UPK3B	RP11-168G16.2	RP11-90C4.2
CRYZ	GPR158	MGMT	ROBO1	ZDHHC14	RP11-192P3.4	RP4-601K24.1
CSMD1	GRIK2	MYLK	ROR2	ZEB1-AS1	RP11-1E3.1	RP6-114E22.1

Category	Term	Count	%	Fold Enrichment	Bonferroni
Tissue	Brain	67	61.5	1.47	0.0021
Tissue	Skeletal muscle	9	8.3	2.79	0.7559
Tissue	Trachea	6	5.5	2.84	0.9970

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Ion transmembrane transport	7	6.4	5.96	0.5517
Biological process	Homophilic cell adhesion via plasma membrane adhesion molecules	6	5.5	6.78	0.7405
Biological process	Positive regulation of sodium ion transport	3	2.8	28.21	0.9717

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Cell junction	9	8.3	3.54	0.4707
Cellular component	Integral component of membrane	41	37.6	1.43	0.7322
Cellular component	Plasma membrane	34	31.2	1.49	0.8466

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Sodium channel regulator activity	4	3.7	21.75	0.1845
Molecular function	Beta-catenin binding	5	4.9	10.61	0.2727
Molecular function	RNA polymerase II transcription coactivator activity	3	2.8	14.11	0.9927

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Axon guidance	5	4.6	6.30	0.5424
KEGG	Thyroid hormone signalling pathway	4	3.7	5.56	0.9721
KEGG	Calcium signalling pathway	4	3.7	3.57	0.9999

Figure 5.13a – DAVID pathway analysis for the PD gene list containing 109 mapped genes utilising five defined analyses including UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. Green highlighted results indicate statistically significant results with a Bonferroni corrected score of $p<0.05$.

Category	Term	Count	%	Fold Enrichment	Bonferroni
Tissue	Brain	77	57.5	1.41	0.0058
Tissue	Trachea	10	7.6	3.94	0.0973
Tissue	Fetal brain	13	9.7	2.49	0.4569

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Cell junction	12	9.0	3.97	0.038
Cellular component	Adherens junction	5	3.7	15.18	0.057
Cellular component	Postsynaptic membrane	8	6.0	5.76	0.084
Cellular component	Synapse	7	5.2	5.87	0.202

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Nervous system development	8	6.0	4.07	0.9385
Biological process	Startle response	3	2.2	29.20	0.9746
Biological process	Heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	4	3.0	11.68	0.9785

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Beta-catenin binding	5	3.7	9.11	0.4322
Molecular function	Zinc ion binding	16	11.9	2.04	0.9338
Molecular function	Actin binding	7	5.2	3.76	0.9383

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	4	3.0	8.38	0.7063
KEGG	Adherens junction	4	3.0	7.91	0.7619
KEGG	Thyroid hormone signalling pathway	4	3.0	4.88	0.9939

Figure 5.13b – DAVID pathway analysis for the healthy aged (HA) gene list containing 134 mapped genes utilising five defined analyses including UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. Green highlight identifies statistically significant results with a Bonferroni corrected score or <0.05. Green highlighted results indicate statistically significant enrichments (Bonferroni $p<0.05$).

Within both the PD and healthy aged gene lists, the genes harbouring polymorphic L1 insertions were significantly enriched for brain expression pathways (Bonferroni adjusted p value <0.01). This suggests that the polymorphic L1 insertions identified in both PD and healthy aged individuals preferentially insert into genes which are expressed within the brain. This data also suggests that within the healthy aged individuals, polymorphic L1 insertions were enriched within genes involved in cell junction contacts (Bonferroni correct p value=0.038).

Further to the analysis of the total gene lists, the unique genes within both the PD and HA individuals were also analysed. A total of 38 unique genes within the PD list (non-green highlighted genes within **table 5.4 left panel**) and 67 unique genes within the HA list (non-green highlighted genes within **table 5.4 right panel**) were independently inputted to DAVID with no significant results being returned for either list for the UP tissue, GO cellular component, GO biological processes, GO molecular function or KEGG pathway analyses.

5.1.5 Discussion

The aims at the beginning of this study were to provide a meaningful way of analysing mutational insertions driven by LINE-1 retrotransposition in the context of PD and provide a preliminary experimental flow through that would provide a proof of principle for the validity of studying retrotransposition events in neurodegenerative disorders. The hypothesis at the start of this study was that complex diseases such as Parkinson's disease could be caused, in part, to the action of a multitude of retrotransposition events that drive mutation within multiple pathways and that these could cause a cumulative pathological effect as opposed to specific mutations within single genes. In response to this hypothesis, the overall numbers of L1 polymorphic and somatic insertions were to be measured and compared to current literature which implicates increased retrotransposition activity to neurological disorders such as Rett syndrome, autism and schizophrenia amongst others [209-211]. Using RC-Seq and TEBreak bioinformatic analysis, it has been demonstrated in this section that LINE-1 elements are active in the human genome and that insertional polymorphisms exist within the central nervous system within both case and control individuals (**figures 5.5 and 5.6**). The two previously unreported insertions validated (insertions at chr2:140179489 and chr8:109058606) **figure 5.6**) represent novel validated LINE-1 mediated mutations that exist between individuals and provide further evidence for the existence of unique insertions that have the potential to be pathogenic and should be considered more carefully in further genome wide studies for PD. Of the two novel polymorphic RIPs identified within the PD group, the insertion at chr8:109058606 was located within intron 2 of the RSPO2 gene which has been reported to be involved in midbrain dopaminergic neurogenesis

and differentiation [208]. Although there is no direct pathological link between RSPO2 and Parkinson's disease, it is worth noting that polymorphic L1 insertions such as the one reported in RSPO2 were found in individuals with PD and these genes are critically important in the processes that are well characterised within PD such as dopaminergic neuron differentiation and should be explored further.

The validation of somatic insertions proved to be very difficult potentially due to the possibility that somatic LINE-1 insertions may only be present in a very small number of cells within a tissue (potentially a single post-mitotic cell such as a neuron which cannot propagate the mutation via cell division). This would allow a sensitive technique such as RC-Seq to detect such an insertion but would potentially consume the DNA containing the somatic insertion, meaning there would be no DNA left containing the mutation that could be validated by PCR. There are examples of successful somatic insertion validation within the literature using RC-Seq which provide good evidence for the use of RC-Seq for the detection of true somatic insertions. However, the validation methods within these studies rely heavily on the approach taken, for example, Upton *et al.* 2015 attempted the validation of 34 somatic insertions with 13 of these being successfully validated from single cell RC-Seq analysis [212]. Using single cell analysis drastically increases the chances of detection of true somatic insertions compared to bulk tissue analysis due to the whole genome amplification process utilised prior to sequencing which produces a clonal sample with multiple copies of the somatic insertions. For future works to be undertaken for the identification of somatic retroelement insertions within a PD context, it would be desirable to use a single cell approach using different primary neuronal cell types e.g. dopaminergic neurons, glial cells or astrocytes from PD

patients compared to non-neuronal cells e.g. hepatocytes from the same individual. However, whole genome amplification has inherent disadvantages including bias, error rates, yields and robustness depending on the amplification process selected and careful consideration would be necessary prior to analysis using RC-Seq [213]. Utilisation of PD patient derived induced pluripotent stem cells (iPSCs) could also be an appropriate approach to avoid the problems associated with the whole genome amplification used within single cell analysis. This would allow for large quantities of genomic DNA to be utilised for RC-Seq from a single cell source to study somatic retrotransposition in detail. However, it is worth noting that the reprogramming events associated with the production of iPSCs have been reported to induce endogenous L1 retrotransposition which would make distinguishing true insertions from the original tissue difficult from those produced as a result of reprogramming [167]. Due to the lack of validation for the somatic insertions within this chapter, all findings regarding these insertions were reported as putative with more extensive validation attempts required.

A key finding of this study was the lack of evidence that was found to support the hypothesis that a large increase in retrotransposition activity has been reported within other neurodegenerative disorders. Several studies have demonstrated that increases in retrotransposition events and expression of retroelements have been found within neurodegenerative disorders, such as ALS, were both LINE-1 and human endogenous retrovirus-K (HERV-K) elements have been potentially implicated [109, 214-216]. There was no evidence found to support a dramatic increase in L1 copy number within the PD samples compared to neurodegeneration free individuals tested within the scope of this study. However, this does not exclude the potential

importance of L1 somatic driven mutations within these disorders which could provide a large pool of novel mutation within the brain. The estimates of the average numbers of somatic L1 insertions per neuron within brain tissue is highly contested with estimates ranging from ~0.58 to ~13.7 insertions per cell depending on cell type (neuron vs glial cells) which was assessed using single cell analysis [212, 217-220]. It is therefore unsurprising that detection of this level of somatic insertion activity was absent within the experiments performed in this chapter using bulk tissue analysis but cannot be disregarded for future works.

There was a significant decrease in the percentage of polymorphic LINE-1 insertions located within intragenic regions within Parkinson's disease individuals compared to the healthy controls (**figure 5.11**) which provided evidence to suggest there were differences in the novel L1 landscapes that exists between individuals with PD compared to healthy individuals. Using this data in conjunction with the haploblock analysis data which suggested there were more putative somatic insertions within PD haploblocks within the PD group compared to the healthy aged controls, there is evidence to suggest there are novel PD LINE-1 signatures that are driven by unique RIPs as well as somatic insertions.

Haplotype block analysis showed the distribution of insertions between the PD and HA groups to be similar, with a trend towards an increased number of putative somatic insertions being found in PD haploblocks within the PD individuals compared to the healthy controls. There were twice as many putative somatic insertions detected in the PD group compared to the healthy aged group which is interesting when considering the difference in sample sizes (HA n=11, PD n=5). Given that true

somatic insertions should be unique to each individual, it could be an important result that more somatic insertions were reported in the PD group (**figure 5.10**) and that these insertions were more often found within PD related haploblocks (**figure 5.12**). The suggestion that somatic insertions are potentially over-represented within neurological disorders has been suggested within other studies, which suggest that aberrant high expression of LINE-1 elements could be linked to multiple disorders such as frontotemporal lobar degeneration (FTLD) amyotrophic lateral sclerosis (ALS) and schizophrenia via multiple L1 mediated mechanisms [217].

A major flaw within the experimental design within this study includes a small sample size (case n=5, control n=11) and lack of sample details. Given this, it is important to note the findings presented here are intended as a proof of principle preliminary study to validate the use of the RC-Seq/TEBreak pipeline for analysing LINE-1 driven mutations within human DNA extracted from nervous tissue. Further studies of a similar nature should contain more details for the stratification of PD patients to distinguish between multiple factors including pathology severity (severe vs mild cases), age of onset, gender, any known genetic risk factors and age of death/age at sample acquisition. This information was not available for the PD case samples used within this study in conjunction with the lack of larger sample sizes meant it was difficult to provide relevant statistical analysis.

The overall findings suggest that RC-Seq is a powerful tool for the analysis of retrotransposon architecture within a variety of tissue types and can be used to identify both novel retrotransposon insertion polymorphisms and somatic insertions. However, the sensitivity of the method for the detection of somatic insertions is likely

highly dependent on the approach taken, which lacks sensitivity when using bulk tissue rather than a single cell approach.

5.2 Using short read whole genome sequencing as a comparative method to RC-Seq for the identification of polymorphic LINE-1, *Alu* and SVA retrotransposons in Parkinson's disease

5.2.1 Introduction

The data presented in **chapter 5.1** provided evidence for increased retrotransposition events in the PD samples compared to the healthy aged controls with increases in both polymorphic and putative somatic insertions being detected. This was a particularly interesting finding given the links that have already been suggested between LINE-1 activity and neurodegenerative diseases such as multiple sclerosis, ALS and Parkinson's disease [112, 113, 217, 221]. Associations have been reported whereby distressed mitochondria have been shown to activate LINE-1 in the nucleus and also evoke a loss of LINE-1 methylation which led to an increase in activity [113]. Increases in LINE-1 activity have also been reported as being associated with smoking in Parkinson's disease patients whereby the loss of LINE-1 methylation as a hypothesised result from smoking increased retrotransposition activity which may be linked with PD disease progression [112]. The *Alu* family of retrotransposons has also been hypothesised to have importance in neurodegenerative disorders whereby deleterious mutations caused from *Alu* insertions within neuronal genes can lead to disease onset within neuronal cell types [96]. A well characterised example of this is observed in the *TOMM40* gene, where multiple Alu insertion events accumulate within the introns of *TOMM40* with at least one *Alu* variant being linked with late-onset Alzheimer's disease [222]. In addition to the potential deleterious effects of both LINE-1 and *Alu* elements, SVA retrotransposons have been

implicated in multiple disease contexts (**table 1.1**) with one of the most widely accepted mechanisms being the SVA insertion within *TAF1* which leads to X-linked dystonia parkinsonism (XDP) [78]. These examples provide further relevance for studying retrotransposon insertion polymorphisms (RIPs) in neuronal contexts and the need for technologies that can be utilised to characterise multiple elements such as whole genome sequencing (WGS).

De novo retrotransposition is an important source of genetic variation which occurs within the human genome, with approximately one *Alu*, LINE-1 and SVA novel insertions being reported for every 20, 150 and 1000 live births, respectively [223]. Original estimates placed the average numbers of RIPs within the human genome at 1283 *Alu*, 180 LINE-1 and 56 SVA insertions that are classed as presence/absence mutations [224]. One of the aims of this study was to clarify and further characterise these estimates with specific focus on whether the copy number of each RIP sub-class differ between Parkinson's disease individuals and healthy aged controls.

The detection of RIPs by high throughput next generation sequencing has been applied in previous studies and consists of paired short-read (~150bp reads) sequencing for the detection of structural variants [225]. In such, it is extremely difficult to cover entire retrotransposable elements with uniquely mapped read pairs due to their inherent repetitive nature. To overcome this, three mapping tools are used whereby discordant read pairs, clustered split-reads which share common alignment junctions and re-alignment of assembled contigs were used in conjunction to characterise structural variants in short-read paired sequencing data [226]. In this way, identification of large repetitive elements (i.e. ~6kb full length LINE-1 elements)

which are not covered by single overlapping read pairs can be confidently characterised. To further support provenance of potential RIPs identified within the sequencing data, the detection of target-site duplications, which are hallmarks of retroelement insertions and are features resultant of target primed reverse transcription (TPRT) retrotransposition events, are also used [227].

The original reasoning for the use of RC-Seq over conventional WGS methods included the assumption that increased sensitivity should be possible with RC-Seq, allowing for more accurate detection of somatic insertions [140]. However, in the initial findings (**section 5.1**), the detection rate of somatic LINE-1 insertions was low when using bulk tissue, with the validation of putative somatic insertions proving extremely difficult. This led to the exploration of alternative techniques, such as WGS, that may provide a more streamlined pipeline for the detection of polymorphic retrotransposon insertions at similar sensitivities to RC-Seq within bulk tissue. A pipeline was adopted with features similar to that used in RC-Seq where the use of the TEBreak algorithms to bioinformatically analyse RIPs was also employed. However, WGS can be used to detect novel LINE-1, SVA and *Alu* retrotransposon insertion polymorphisms. The WGS workflow is more streamlined than that of the RC-Seq protocol due to significantly less sample preparation required prior to sequencing which allows for a more time and cost-efficient processing of samples. This contrasts to the RC-Seq protocol which was used to only detect LINE-1 elements with high sensitivity with the use of specific LNA probes to pull down LINE-1 sequences (further descriptions of RC-Seq in **section 5.1)**.

5.2.2 Aims and hypothesis

To compare the effectiveness of RC-Seq and short read WGS as methods for the detection of novel retrotransposon insertion polymorphisms (RIPs) within brain tissue of Parkinson's disease and healthy aged individuals.

To characterise the detected LINE-1, SVA and *Alu* RIPs from WGS and assess differences in overall insertion numbers and genomic locations which may infer function in a neurodegenerative context.

Hypothesis: Retrotransposon insertion polymorphisms of the LINE-1, SVA and *Alu* sub-classes have the potential to drive potentially pathogenic mutations via retrotransposition which confer increased risk for Parkinson's disease.

5.2.3 Methods

5.2.3.1 Whole genome sequencing

Short read whole genome sequencing (WGS) of 4 tissue samples from two PD case individuals (frontal cortex and cerebellum) and 18 samples from 9 healthy aged individuals (temporal cortex and blood from each individual) used within the RC-Seq protocols was carried out externally by the Australian Genome Research Facility (AGRF) (sample details provided in **sections 2.1.5, 2.2.14 and table 5.1**). DNA samples were used from the identical stocks used for the RC-Seq analysis to minimise variability. Sequencing was performed at a depth of 40X using 1µg of starting genomic DNA for each tissue on the IlluminaSeq platform. Healthy aged control samples from the Dyne Steele cohort (Manchester brain bank) (**section 2.1.5**) were used as controls to be processed alongside the two PD samples. The WGS libraries were sequenced using the IlluminaSeq platform which provided 2 FastQ files per tissue (1 forward (R1) and 1 reverse (R2)) to be analysed using the same TEBreak pipeline as the RC-Seq workflow. The full methods for the bioinformatic analysis used for WGS is outlined in **section 2.2.15**.

The final filtering parameters for WGS analysis (post-TEBreak) were as follows:

- Minimum split read – 4 reads
- Minimum Discordant read pair (Disc read) – 4 reads
- Minimum consensus sequence length (Conslen) – 150bp
- Minimum element match (Eltmatch) – 0.90 (90% match)
- Minimum reference genome match (Refmatch) – 0.95 (95% match)
- Maximum number of variants (Maxvar) – 2 SNPs

Bracketed descriptions denote the parameter as viewed within the filtering scripts. The descriptions for each parameter used are outlined in the chapter 5 introductory text (pages 203-204).

5.2.3.2 Haploblock analysis

Haploblock analysis was performed following the identical workflow as detailed in **section 5.1.3.3**. Bed files for each of the LINE-1, *Alu* and SVA RIPs were generated from the PD and healthy aged samples and were used to overlap with the 77 PD nominated haploblocks (co-ordinates listed in **table 5.5a**) using the 'intersect intervals' Bedtools function.

5.2.3.3 Pathway analysis of gene lists containing LINE-1, SVA and *Alu* RIPs using DAVID

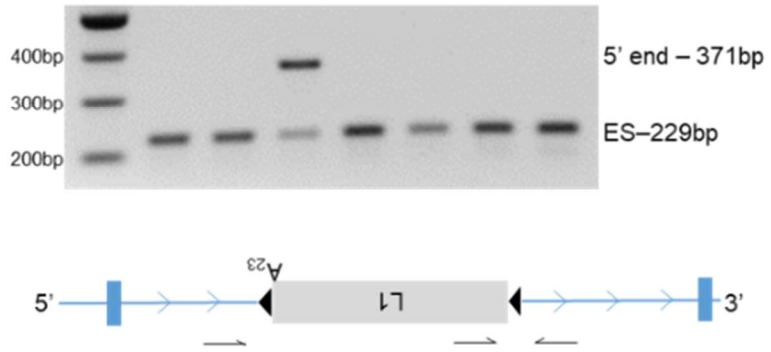
An identical pathway analysis approach used in the RC-Seq analysis (details in **section 5.1.3.4**) was also employed in the WGS analysis. Lists of the genes containing LINE-1, SVA and *Alu* RIPs were generated for both the PD and healthy aged groups and analysed using the database for annotation, visualisation and integrated discovery (DAVID) pipeline (https://david.ncifcrf.gov). The gene lists were analysed using data from the gene ontology (GO) biological processes, GO cellular component, GO molecular function, KEGG and UP Tissue databases (descriptions for each database outline in **section 5.1.3.4**)

5.2.4 Results

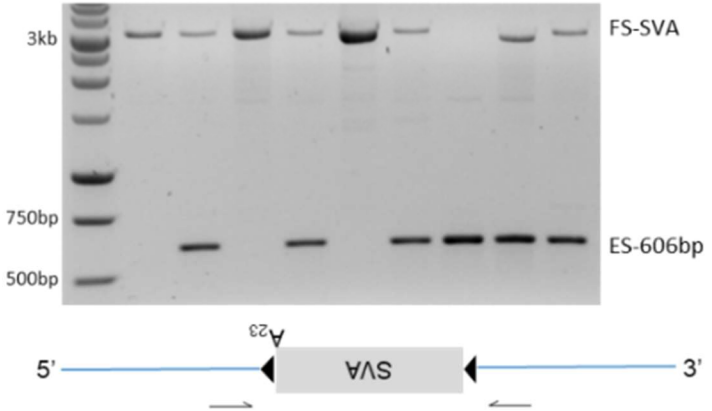
5.2.4.1 PCR validation of WGS libraries with estimates of accuracy, sensitivity and specificity

Due to time constraints upon receiving the WGS raw data for analysis, validation of the WGS libraries for the processed PD or healthy aged samples was not possible. However, validation of samples from an ALS cohort that were processed by a colleague, Abigail Savage, in parallel with the PD and healthy aged samples was performed and provided insight into the levels of accuracy, sensitivity and specificity of retroelement detection within the libraries. **Figure 5.14** shows the PCR validation of 12 *Alu*, 11 LINE-1 and 6 SVA elements within the ALS cohort with a total of 147 Alu, 117 LINE-1 and 71 SVA PCRs being performed. **Figure 5.14-D** indicates the final summary statistics generated for the three elements which showed a perfect false positive rate of 0.00 for each element. Overall, the WGS approach favoured detection of *Alu* elements with the highest rates of accuracy (0.98), sensitivity (0.96) and specificity (1.00) compared to the detection of LINE-1 and SVA RIPs which elicited lower rates of accuracy (LINE-1=0.91, SVA=0.87) and sensitivity (LINE-1=0.79, SVA=0.65).

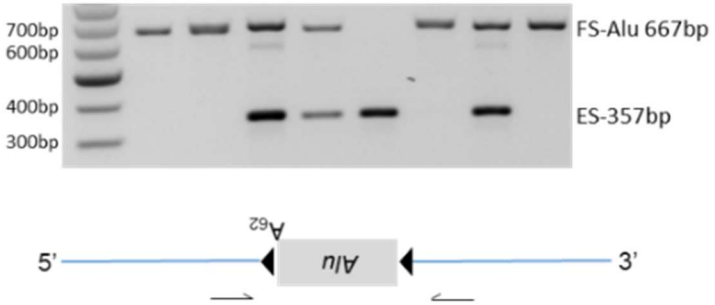
A Chr14:31150808 (SCFD1)



B Chr14:21176120



C Chr13:31150808



D

	Alu	LINE-1	SVA
False positive rate	0.00	0.00	0.00
False negative rate	0.04	0.21	0.35
Accuracy	0.98	0.91	0.87
Sensitivity	0.96	0.79	0.65
Specificity	1.00	1.00	1.00

Figure 5.14 – PCR validation of three retrotransposon insertion polymorphisms (RIPs) with summary statistics within the ALS cohort that were processed for WGS in parallel with the PD and healthy aged samples. **(A)** Multiplex PCR validation of an anti-sense full-length (~6kb) LINE-1 insertion at chr14:31150808 (hg19) within intron 14 of the SCFD1 gene. **(B)** Empty/filled site PCR of an anti-sense full length SVA (~2.5kb) insertion at chr14:21176120 within an intergenic space. An exact predicted amplicon size of the filled site PCR was not possible due to polymorphisms within the primary sequence. **(C)** Empty/filled site PCR of an *Alu* insertion at chr13:31150808 within an intergenic space. **(D)** Summary statistics showing the validation numbers for a total of 147 *Alu*, 117 LINE-1 and 71 SVA RIP PCRs performed for 12 *Alu*, 11 LINE-1 and 6 SVA elements across a minimum of 9 individuals which were calculated using the following calculations:

True positive (TP) = present in both TEBreak and PCR

True negative (TN) = absent in both TEBreak and PCR

False positive (FP) = present in TEBreak but absent in PCR

False negative (FN) = absent in TEBreak but present in PCR

False positive rate (FPR) = FP/(FP+TN)

False negative rate (FNR) = FN/(FN+TP)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

The formulas for generating accuracy, sensitivity and specificity were based on the Baratloo *et al.* 2015 definitions [197].

5.2.4.2 Analysis of the TEBreak identified *Alu*, LINE-1 and SVA retrotransposon insertion polymorphisms

The initial characterisation of retrotransposon insertion polymorphisms (RIPs) in the WGS analysis was performed by comparing the total numbers of RIPs identified across all tested tissues between the PD (frontal cortex and cerebellum) and healthy aged samples (temporal cortex and blood) for each retroelement sub-class (LINE-1, *Alu* and SVA). There was no statistically significant difference between the total number of LINE-1, *Alu* or SVA RIPs identified between the PD and healthy aged individuals when tested using two-sample t-tests assuming unequal variances (**figure 5.15**). However, there were key differences between the average numbers of *Alu* and SVA RIPs between the PD and healthy aged groups, with a 9.5% increase in number of *Alu* RIPs identified within the healthy aged individuals and an 8.3% increase in the number of SVA RIPs within the PD group (**figure 5.15 a and b**). This represents a large pool of genetic diversity between the two groups caused by *Alu* and SVA retrotransposition primarily with on average approximately 115 additional *Alu* retrotransposition events per person within the healthy aged group and 8 additional SVA retrotransposition events within the PD group. The differences between the LINE-1 RIPs identified between the PD and healthy aged group within the WGS libraries was negligible, with only a 2.4% increase in the numbers of LINE-1 RIPs within the PD group.

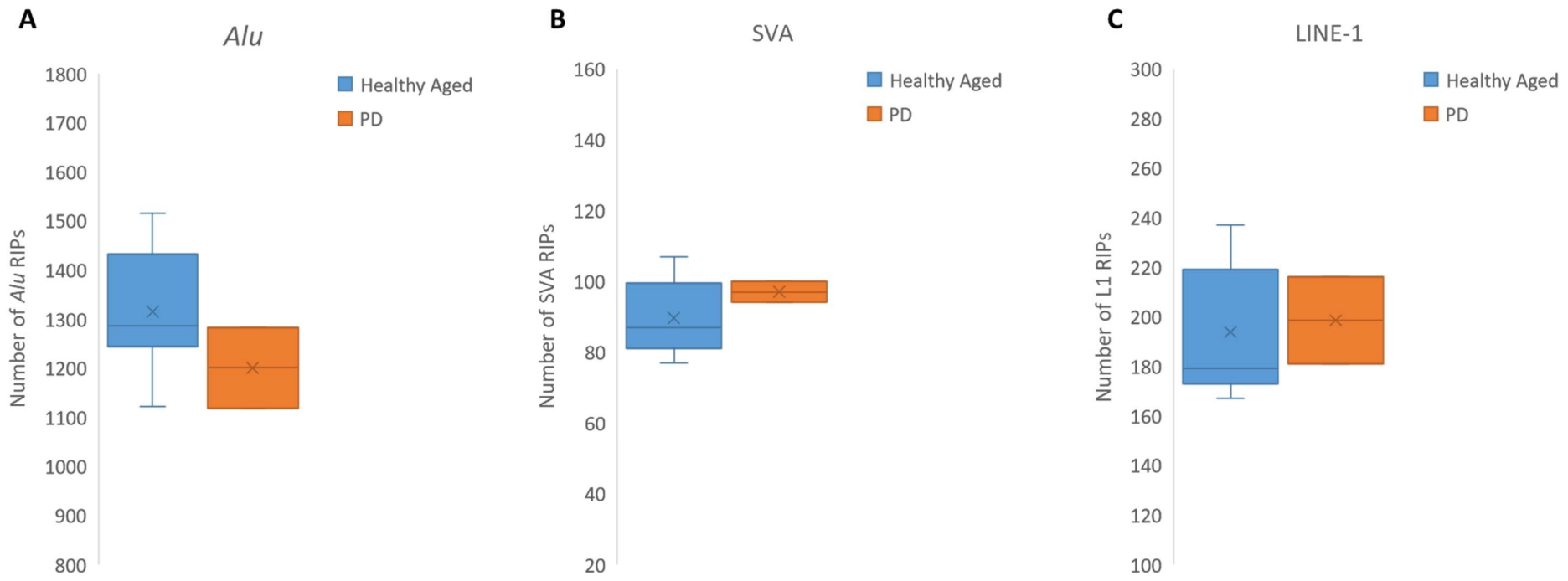


Figure 5.15 – Total numbers of **(A)** *Alu*, **(B)** SVA and **(C)** LINE-1 retrotransposon insertion polymorphisms identified within the healthy aged and Parkinson's disease individuals using with WGS/TEBreak pipeline. There is no significant difference in the number of *Alu* (p-value=0.367), SVA (p-value=0.172) or LINE-1 (p-value=0.837) RIPs between the healthy aged and PD groups using a two-sample t-test assuming unequal variances. (Healthy aged n=9 tissues = temporal cortex and blood, PD n=2 tissues = frontal cortex and cerebellum). X markings indicate the means for each group with horizontal lines indicating median values.

5.2.4.3 Comparison of WGS and RC-Seq for the detection of LINE-1 RIPs

Comparison of the LINE-1 summary statistics presented in **figures 5.4** and **5.14** which assess the quality of the RC-Seq and WGS libraries are summarised in **figure 5.16**. Both techniques presented similar false positive rates, accuracy and specificity of LINE-1 RIP detection with the largest distinctions being between false negative rates (RC-Seq = 0.14, WGS = 0.21) and sensitivity which was lower in WGS than RC-Seq (RC-Seq = 0.86, WGS = 0.79). This highlighted one of the main advantages of RC-Seq over WGS for the sensitive detection of LINE-1 elements with high accuracy and low false negative results.

	Rc-seq	WGS
False positive rate (FPR)	0.03	0.00
False negative rate (FNR)	0.14	0.21
Accuracy	0.93	0.91
Sensitivity	0.86	0.79
Specificity	0.97	1.00

Figure 5.16 – Comparison of the summary statistics regarding false positive and negative rates, accuracy, sensitivity, and specificity of LINE-1 RIP detection in both the RC-Seq and WGS analysis. The two analyses provide very similar false positive rates as well as accuracy and specificity with slightly lower sensitivity for the detection of LINE-1 elements being reported in WGS. Full summary statistics for both RC-Seq and WGS are presented in **figures 5.4 and 5.14,** respectively.

When compared to RC-Seq, WGS detected significantly higher numbers of LINE-1 elements, with 55.4% and 43.4% more insertions being detected within the healthy aged and PD groups, respectively (**figure 5.17**). This suggested that RC-Seq may be missing a significant proportion of LINE-1 RIPs compared to WGS, however, to qualify this finding, PCR validation of the unique RIPs identified in WGS would be required.

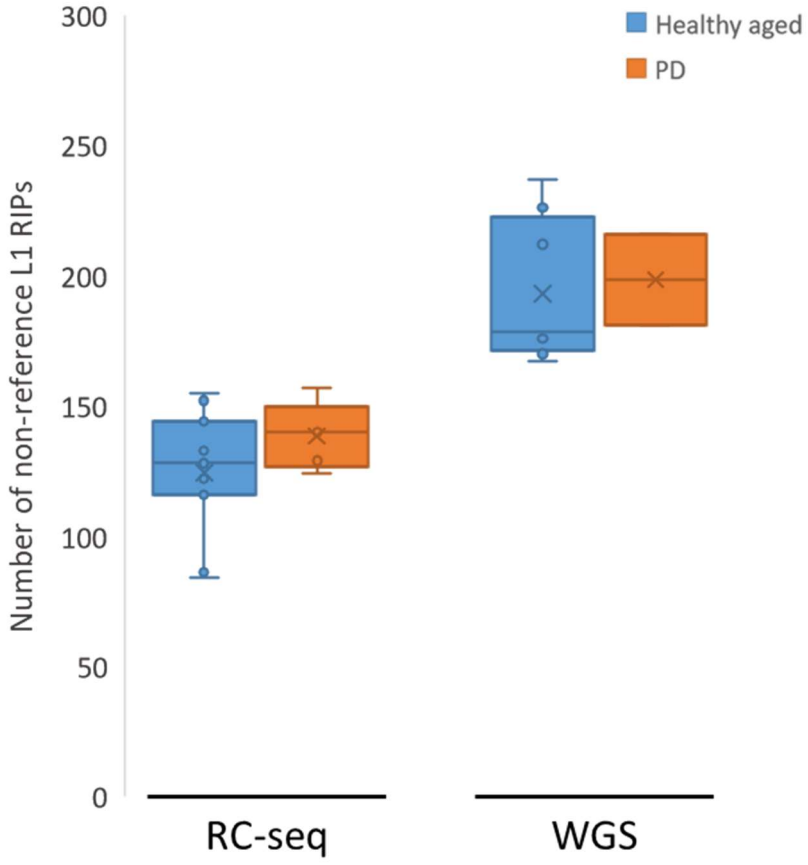
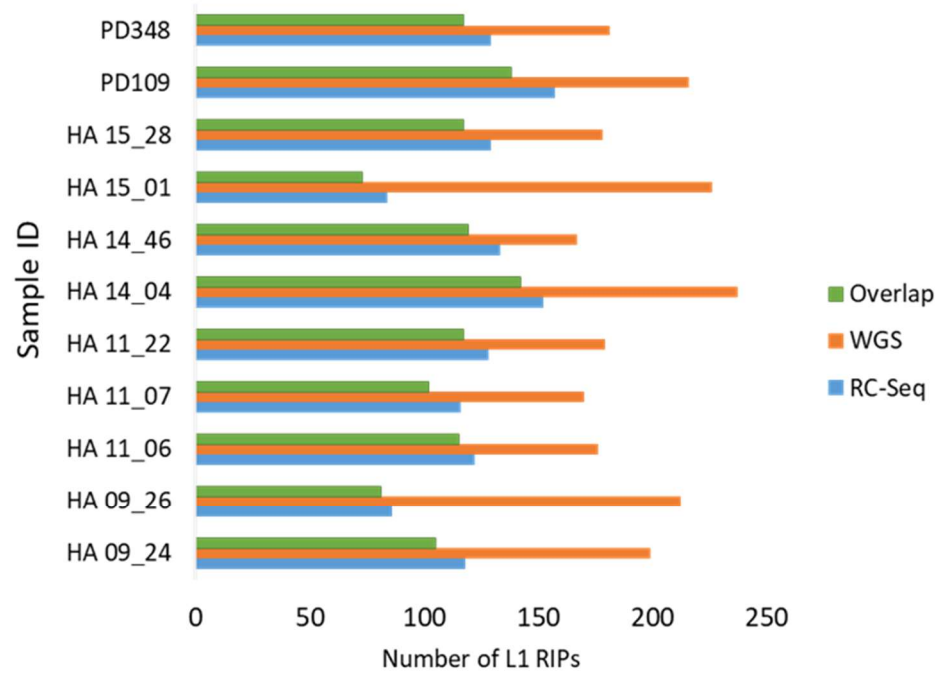


Figure 5.17 – A comparison of the number of non-reference LINE-1 retrotransposon insertion polymorphisms detected by RC-Seq and WGS. The numbers of insertions detected by WGS was 55.4% higher in the healthy aged and 43.4% higher in the PD samples compared to RC-Seq. RC-Seq: healthy aged n=11, PD n=5), WGS: healthy aged n=9, PD n=2). X markings indicate means for each group with horizontal lines representing median values.

To further explore the differences between RC-Seq and WGS as suitable techniques for the identification of LINE-1 RIPs, the specific overlap of discrete insertion polymorphisms was compared rather than only comparing the overall insertion numbers. To do this, bed files for each individual containing the co-ordinates for the identified LINE-1 RIPs within both tissues (PD - frontal cortex and cerebellum and HA – temporal cortex and blood) were generated for both RC-Seq and WGS which were overlapped using the 'interest intervals' function of the bedtools suite through the 'Galaxy' web servers (https://usegalaxy.org/). The results indicated that WGS identified on average 90.5% of the same insertions as RC-Seq, with 72 more insertions being detected per person on average using WGS over RC-Seq (**figure 5.18**). The additional RIPs identified by WGS that were not picked up by RC-Seq would require PCR validation in order to ratify if these insertions are true positives.

A



B

Polymorphic L1 RIPs detected per sample				
Sample	**RC-Seq**	**WGS**	**Overlap**	**%**
PD348	129	181	117	90.7
PD109	157	216	138	87.9
HA 15_28	129	178	117	90.7
HA 15_01	84	226	73	86.9
HA 14_46	133	167	119	89.5
HA 14_04	152	237	142	93.4
HA 11_22	128	179	117	91.4
HA 11_07	116	170	102	87.9
HA 11_06	122	176	115	94.3
HA 09_26	86	212	81	94.2
HA 09_24	118	199	105	89.0
Average	**123.1**	**194.6**	**111.5**	**90.5**

Figure 5.18 – Comparison of the numbers of non-reference L1 retrotransposon insertion polymorphisms (RIPs) between RC-Seq and WGS. (A) represents total non-reference polymorphic L1 RIPs detected by both RC-Seq and WGS with the number of identical insertions detected by both represented as an overlap per individual. (B) Percentage overlap calculations of L1 RIPs detected by RC-Seq and WGS presented as sample breakdown and averages. On average, 90.5% of the RIPs detected by RC-Seq where also detected by WGS.

5.2.4.4 Functional inferences of non-reference LINE-1, SVA and *Alu* insertions detected by WGS

In order to understand the potential functional implications of the LINE-1, *Alu* and SVA RIPs in Parkinson's disease, as identified by WGS, analysis of the genomic locations of each retrotransposon sub-class was performed. The percentage of intragenic RIPs was calculated for each sub-class which showed significantly more SVA RIPs being located within genes within the PD group compared to the healthy aged individuals (p=0.0054) (**figure 5.19**). This analysis also indicated significantly increased intragenic LINE-1 insertions in the healthy aged group compared to the PD cases (p=0.041). There was no difference in the percentage of intragenic *Alu* insertions between the PD and healthy aged groups.

Haploblock analysis was performed following the same pipeline as in the RC-Seq analysis, whereby the nominated list of 77 PD haploblocks was used to identify the proportion of the LINE-1, *Alu* and SVA RIPs which were located within the haploblocks. Two analyses were performed to analyse the distribution of RIPs within the haploblocks (**table 5.5a**) as well as overall summary statistics of the numbers of RIPs per haploblock (**table 5.5b**). The cluster analysis (**table 5.5a**) highlighted a general even distribution of RIPs in both the PD and healthy aged groups, with a few specific blocks which appeared to be hotspots for insertion polymorphisms (highlighted in red in **table 5.5a**).

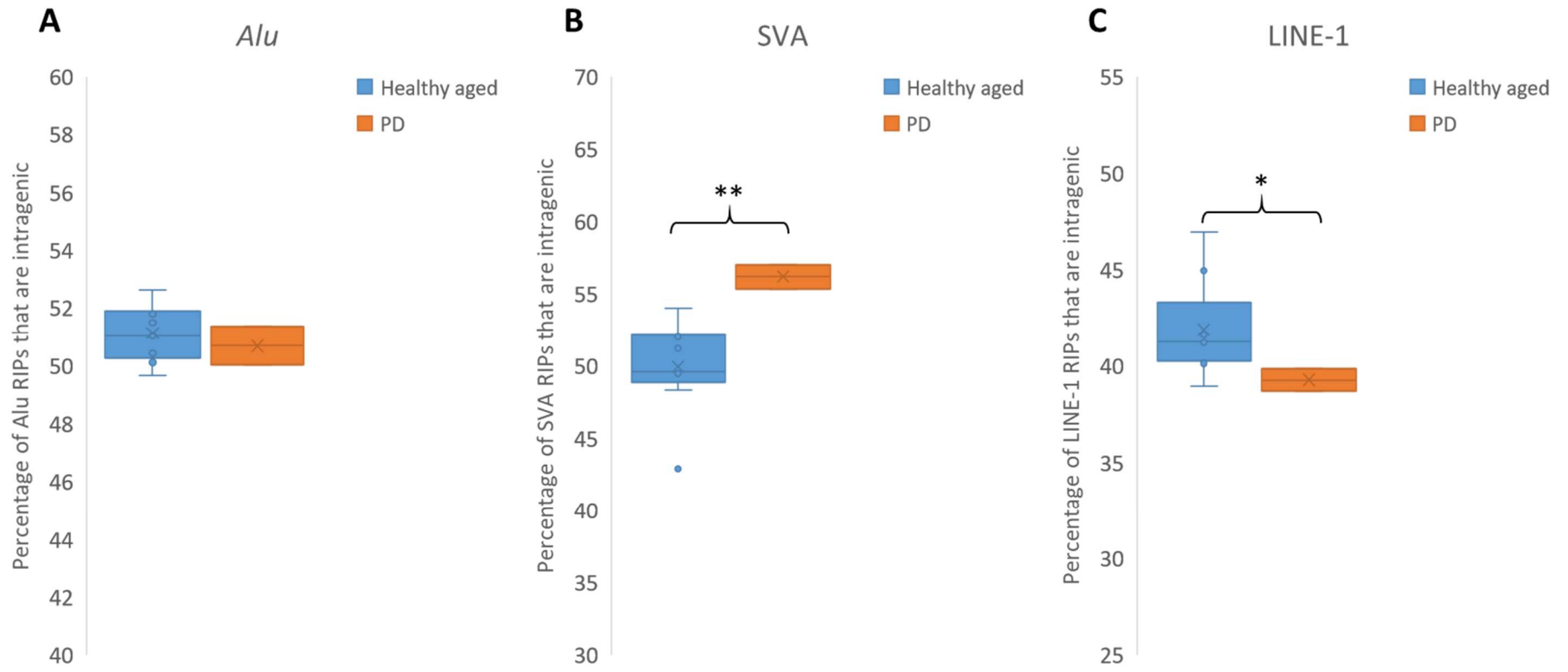


Figure 5.19 – The percentage breakdown of **(A)** *Alu*, **(B)** SVA and **(C)** LINE-1 retrotransposon insertion polymorphisms that are located within intragenic regions (within a defined gene or transcript) as identified using the WGS/TEBreak pipeline. There was significantly more SVA RIPs located within intragenic regions within the PD group compared to the healthy aged group (p-value=0.0054) and more LINE-1 RIPs located within intragenic regions within the PD group compared to the healthy aged controls (p-value=0.041). There was no statistical difference between the percentage of *Alu* RIPs that were intragenic between the two tested groups (p-value=0.624). All statistical analysis performed using two-sample t-test assuming unequal variances (healthy aged n=9, PD n=2).

Table 5.5a – Haploblock cluster analysis indicating the distribution of retrotransposon insertion polymorphisms (*Alu*, LINE-1 and SVA) located in PD haploblocks within the PD and healthy aged groups. Red highlight indicates the haploblock with the largest numbers of RIPs which may be indicative of an insertion polymorphism hotspot. PD n=2, HA n=9.

			LINE-1		SVA		Alu		
	PD Haploblocks		PD	HA	PD	HA	PD	HA	Total RIPs
chr1	154770403	156336133	0	0	0	0	0	2	2
chr1	159913048	162346721	0	0	0	0	1	4	5
chr1	170557776	173097907	0	0	0	0	1	4	5
chr1	204681068	206073265	0	0	0	0	0	0	0
chr1	226810860	229156248	0	0	0	0	2	3	5
chr1	232090252	233429284	0	0	0	0	1	2	3
chr2	16329735	18647423	0	0	0	0	1	2	3
chr2	167355970	169968236	1	0	1	1	1	1	5
chr2	95326452	98995201	0	0	0	0	1	2	3
chr2	135158578	137042794	0	0	0	0	0	0	0
chr2	101822329	102688765	0	0	0	0	0	0	0
chr3	150252004	151348730	1	1	0	0	0	1	3
chr3	17891118	19125144	0	0	0	0	1	1	2
chr3	27840910	29142260	0	0	0	0	3	3	6
chr3	47727212	49316972	0	0	1	1	2	3	7
chr3	121974097	123517768	0	0	0	0	2	3	5
chr3	159477890	161524504	0	0	0	0	0	1	1
chr3	181511166	183769683	0	0	0	0	1	2	3
chr4	74592390	77130707	0	0	0	0	1	4	5
chr4	694715	1478711	0	0	0	0	1	1	2
chr4	15147446	15927009	0	0	0	0	0	0	0
chr4	17383322	18841874	1	1	0	0	0	0	2
chr4	77130707	79093979	0	0	1	1	1	1	4
chr4	90231564	91560677	0	0	0	0	0	1	1
chr4	113870102	115666246	1	1	0	0	1	2	5
chr4	169676825	170776510	1	1	0	0	2	3	7
chr5	132139649	134777401	1	0	0	2	0	4	7
chr5	58524622	60935907	0	0	0	0	1	1	2
chr5	101578769	103320005	0	0	0	0	0	0	0
chr6	132765669	134244243	0	0	1	1	2	2	6
chr6	26791233	28017819	0	0	0	0	0	2	2
chr6	29737971	30798168	0	0	1	1	2	2	6
chr6	31571218	32682664	2	3	0	2	4	14	25
chr6	71609510	73450097	0	0	0	0	3	4	7
chr6	110304247	112345014	0	0	0	2	0	3	5
chr7	65689809	68234074	1	1	0	0	0	1	3
chr7	22507629	23471442	0	0	0	0	1	1	2
chr8	15991660	17387876	0	0	0	0	1	2	3
chr8	11278998	13491775	0	0	0	0	1	3	4
chr8	21661737	22897057	0	0	0	1	1	3	5
chr8	130381139	131639625	0	1	0	0	2	3	6
chr9	16659655	18661051	0	0	0	0	1	3	4
chr9	33578334	34642243	0	0	0	0	0	2	2
chr10	15026068	16551767	0	0	0	0	2	2	4
chr10	102949239	104380410	0	0	0	1	0	1	2
chr10	120591353	122407323	0	0	1	1	1	2	5
chr11	81266712	84381272	0	1	0	0	1	4	6

chr11	9087317	10952027	0	0	0	1	2	4	7
chr11	133000046	134205993	0	0	0	0	0	0	0
chr12	39227169	40816185	0	0	0	0	3	5	8
chr12	46024229	47714793	0	0	0	0	1	1	2
chr12	122007651	124977980	0	0	0	0	4	4	8
chr12	132807034	133841511	0	0	0	0	2	3	5
chr13	49383962	51591091	0	1	0	0	1	4	6
chr13	97519210	98938919	0	0	0	0	3	2	5
chr14	72889615	76444767	0	0	0	0	2	4	6
chr14	35859593	38667725	2	3	0	0	0	2	7
chr14	55233681	56216880	0	0	0	0	1	1	2
chr14	87635341	89497643	0	0	1	0	1	0	2
chr15	61265836	63215222	1	3	0	2	1	3	10
chr16	49007926	52035823	0	0	0	0	1	2	3
chr16	18643607	20150571	0	0	0	0	0	0	0
chr16	27445755	29036613	0	0	0	0	0	1	1
chr16	29036613	31382943	0	0	0	1	2	1	4
chr16	52035823	53382572	0	0	0	0	1	2	3
chr17	41772087	43056905	0	0	0	1	0	1	2
chr17	7317398	8306425	0	0	0	0	0	1	1
chr17	39899810	41772087	0	0	1	2	1	3	7
chr17	43056905	45876022	0	0	3	3	4	3	13
chr17	59312755	61545589	0	1	0	0	0	3	4
chr17	76263413	77298636	0	0	0	0	0	0	0
chr18	39892648	42922106	0	0	1	1	1	2	5
chr18	47730584	51062185	1	2	0	0	0	4	7
chr18	30264066	31780067	0	0	0	0	0	0	0
chr19	2098396	3019660	0	0	0	0	2	3	5
chr20	5477850	7084073	0	0	0	0	1	1	2
chr21	38711704	40482902	0	0	0	0	1	1	2

The cluster analysis indicated one notable difference between the PD and healthy aged groups where the number of *Alu* RIPs present in haploblock chr6: 31571218-32682664 was markedly higher in the healthy aged individuals (HA=14 vs PD=4). Upon further investigation, the chr6: 31571218-32682664 haploblock was found to contain part of the human leukocyte antigen (HLA) super-locus which is known to be highly genetically diverse and contains large numbers of retrotransposon insertion polymorphisms which compliments the findings from this analysis [228, 229]. There were no other observable large differences between the numbers of LINE-1, SVA or *Alu* RIPs between the PD and healthy aged groups across the 77 PD haploblocks tested.

Table 5.5b – Haploblock analysis summary indicating the overall RIPs of each class (LINE-1, SVA and *Alu*) present within the 77 nominated PD haploblocks between the PD and healthy aged (HA) groups. Both the numbers of haploblocks containing at least 1 RIP and the total numbers of RIPs of each class were reported. PD n=2, HA n=9.

No. of PD Haploblocks	Intersect with	L1			SVA			*Alu*		
		No. of haploblocks with >1 L1	No. of insertions	% of haploblocks with >1 L1	No. of haploblocks with >1 SVA	No. of insertions	% of haploblocks with >1 SVA	No. of haploblocks with >1 *Alu*	No. of insertions	% of haploblocks with >1 *Alu*
77	PD insertions	11	13	**14.2**	12	12	**15.6**	51	81	**66.2**
	HA insertions	13	20	**16.9**	18	15	**19.5**	66	166	**85.7**

The overall summary of the haploblock analysis is reported in **table 5.5b**, where the percentage breakdowns of the numbers of RIPs per haploblock is presented. The largest difference between the PD and healthy aged RIP distributions was between the *Alu* breakdowns, where 19.5% more PD haploblocks contained at least 1 *Alu* RIP in the healthy aged individuals compared to the PD group. Furthermore, the difference observed was resultant from a drastic increase in the numbers of *Alu* RIPs within the identified haploblocks, with over twice as many *Alu* RIPs being present (PD=81, HA=166). A similar trend was also observed in the distribution of LINE-1 and SVA RIPs with a less significant difference whereby the healthy aged group contained 2.7% and 3.9% more haploblocks that contained at least one LINE-1 and SVA RIP respectively. However, these results do not reflect the difference in sample sizes used to generate these data and was likely largely influenced by the numbers within each group (PD n=2, HA n=9).

As a final analysis for the inference of potential function of WGS identified RIPs in Parkinson's disease, pathway analysis using the DAVID web tool package was undertaken. Lists of the genes containing LINE-1, SVA and *Alu* RIPs for both the PD and healthy aged groups were generated and analysed using DAVID (method details in **sections 5.1.3.4 and 5.2.3.3**) (**figures 5.20 a – f**).

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Positive regulation of mitochondrial fusion	2	2.33	157.67	0.999
Biological process	Mitochondrion organization	3	3.49	9.21	1
Biological process	Receptor catabolic process	2	2.33	39.42	1

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Cell junction	7	8.14	3.76	0.786
Cellular component	Cytoplasm	31	36.05	1.46	0.899
Cellular component	Plasma membrane	26	30.23	1.55	0.904

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Actin binding	6	6.98	5.20	0.699
Molecular function	RNA polymerase II transcription coactivator activity	3	3.49	19.55	0.885
Molecular function	Calmodulin binding	4	4.65	5.10	1.000

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Calcium signaling pathway	5	5.81	5.49	0.735
KEGG	Thyroid hormone signaling pathway	4	4.65	6.84	0.892
KEGG	Axon guidance	4	4.65	6.19	0.945

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP Tissue	Brain	51	59.30	1.42	0.070
UP Tissue	Trachea	7	8.14	4.20	0.433
UP Tissue	Fetal brain	8	9.30	2.33	0.993

Figure 5.20 a – DAVID pathway analysis of the identified genes containing LINE-1 RIPs within the PD samples (n=2) identified by WGS. A total of 86 mapped genes by DAVID were cross referenced in five defined databases: UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. No results returned significant pathway enrichment with Bonferroni corrected p-values of <0.05.

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Central nervous system development	4	6.06	11.19	0.837
Biological process	Amino-acid betaine catabolic process	2	3.03	335.84	0.871
Biological process	Protein farnesylation	2	3.03	167.92	0.983

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Microtubule associated complex	3	4.55	30.92	0.340
Cellular component	Neuron projection	5	7.58	7.39	0.357
Cellular component	Membrane	14	21.21	2.23	0.479

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Protein farnesyltransferase activity	2.00	3.03	225.08	0.732
Molecular function	Protein binding	32.00	48.48	1.23	1.000
Molecular function	ATPase activity	3.00	4.55	5.53	1.000

Category	Term	Count	%	Fold Enrichment	Bonferroni
-	-	-	-	-	-

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP Tissue	Mammary cancer	3	4.55	10.05	0.919
UP Tissue	Lymph	6	9.09	2.58	0.997
UP Tissue	Teratocarcinoma	5	7.58	3.04	0.997

Figure 5.20 b – DAVID pathway analysis of the identified genes containing SVA RIPs within the PD samples (n=2) identified by WGS. A total of 66 mapped genes by DAVID were cross referenced in five defined databases: UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. No results returned significant pathway enrichment with Bonferroni corrected p-values of <0.05.

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Positive regulation of GTPase activity	41	5.94	2.19	0.012
Biological process	Peptidyl-serine phosphorylation	13	1.88	3.14	0.880
Biological process	Cytoskeleton organization	15	2.17	2.81	0.890

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Postsynaptic membrane	21	3.04	3.04	0.009
Cellular component	Cell junction	32	4.64	2.13	0.051
Cellular component	Sarcolemma	11	1.59	3.95	0.177

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Rac guanyl-nucleotide exchange factor activity	5	0.72	10.79	0.490
Molecular function	GTPase activator activity	21	3.04	2.27	0.530
Molecular function	Ion channel binding	12	1.74	3.21	0.616

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Neuroactive ligand-receptor interaction	21	3.04	2.43	0.079
KEGG	Sphingolipid signaling pathway	10	1.45	2.67	0.939
KEGG	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	7	1.01	3.34	0.982

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP Tissue	Brain	331	47.97	1.22	0.0001
UP Tissue	Lens epithelium	6	0.87	10.18	0.0524
UP Tissue	Retina	26	3.77	2.23	0.0727

Figure 5.20 c – DAVID pathway analysis of the identified genes containing *Alu* RIPs within the PD samples (n=2) identified by WGS. A total of 690 mapped genes by DAVID were cross referenced in five defined databases: UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. Green highlighted results indicate statistically significant results with a Bonferroni corrected score of $p<0.05$.

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Startle response	3	1.88	26.03	0.994
Biological process	Nervous system development	8	5.00	3.63	0.997
Biological process	Angiogenesis	7	4.38	4.09	0.999

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Postsynaptic membrane	9	5.63	5.72	0.034
Cellular component	Cell junction	12	7.50	3.50	0.114
Cellular component	Plasma membrane	48	30.00	1.56	0.145

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Beta-catenin binding	5	3.13	8.23	0.598
Molecular function	Actin binding	8	5.00	3.89	0.729
Molecular function	Ubiquitin conjugating enzyme binding	3	1.88	13.07	0.998

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5	3.13	10.07	0.157
KEGG	Calcium signaling pathway	6	3.75	4.52	0.691
KEGG	Thyroid hormone signaling pathway	4	2.50	4.69	0.999

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP Tissue	Trachea	12	7.50	4.19	0.015
UP Tissue	Brain	79	49.38	1.28	0.256
UP Tissue	Fetal brain	14	8.75	2.37	0.485

Figure 5.20 d – DAVID pathway analysis of the identified genes containing LINE-1 RIPs within the healthy aged samples (n=9) identified by WGS. A total of 160 mapped genes by DAVID were cross referenced in five defined databases: UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. Green highlighted results indicate statistically significant results with a Bonferroni corrected score of $p<0.05$.

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Phospholipid translocation	3	2.42	24.36	0.983
Biological process	Amino-acid betaine catabolic process	2	1.61	178.64	0.999
Biological process	Protein farnesylation	2	1.61	89.32	1.000

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Golgi apparatus	13	10.48	2.83	0.251
Cellular component	Golgi cisterna membrane	4	3.23	9.89	0.684
Cellular component	Microtubule associated complex	3	2.42	16.58	0.879

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	GTPase activator activity	7	5.65	4.37	0.700
Molecular function	Protein farnesyltransferase activity	2	1.61	116.02	0.981
Molecular function	Phosphatidylcholine-translocating ATPase activity	2	1.61	87.02	0.995

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Other types of O-glycan biosynthesis	3	2.42	22.88	0.529
KEGG	ABC transporters	3	2.42	11.44	0.943
KEGG	Legionellosis	3	2.42	9.32	0.985

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP Tissue	Brain	56	45.16	1.21	0.988
UP Tissue	Embryo	6	4.84	3.07	0.989

Figure 5.20 e – DAVID pathway analysis of the identified genes containing SVA RIPs within the healthy aged samples (n=9) identified by WGS. A total of 124 mapped genes by DAVID were cross referenced in five defined databases: UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. No results returned significant pathway enrichment with Bonferroni corrected p-values of <0.05.

Category	Term	Count	%	Fold Enrichment	Bonferroni
Biological process	Positive regulation of GTPase activity	72	5.93	2.17	0.000003
Biological process	Homophilic cell adhesion via plasma membrane adhesion molecules	29	2.39	3.12	0.000458
Biological process	Intracellular signal transduction	52	4.28	2.20	0.000598
Biological process	Regulation of Rho protein signal transduction	18	1.48	3.78	0.012184
Biological process	cAMP-mediated signaling	11	0.91	4.92	0.148562

Category	Term	Count	%	Fold Enrichment	Bonferroni
Cellular component	Postsynaptic membrane	37	3.05	3.08	0.000002
Cellular component	Postsynaptic density	33	2.72	3.15	0.000008
Cellular component	Plasma membrane	301	24.79	1.28	0.000385
Cellular component	Cell junction	54	4.45	2.07	0.000439
Cellular component	Cytoplasm	356	29.32	1.20	0.018882
Cellular component	Receptor complex	20	1.65	2.76	0.057731

Category	Term	Count	%	Fold Enrichment	Bonferroni
Molecular function	Calcium ion binding	75	6.18	1.81	0.0008
Molecular function	Actin binding	38	3.13	2.37	0.0019
Molecular function	Rho guanyl-nucleotide exchange factor activity	17	1.40	3.82	0.0070
Molecular function	GTPase activator activity	36	2.97	2.23	0.0134
Molecular function	Guanyl-nucleotide exchange factor activity	20	1.65	2.93	0.0459
Molecular function	Rac guanyl-nucleotide exchange factor activity	7	0.58	8.66	0.0754

Category	Term	Count	%	Fold Enrichment	Bonferroni
KEGG	Morphine addiction	17	1.40	3.34	0.0091
KEGG	Glutamatergic synapse	18	1.48	2.82	0.0438
KEGG	Dilated cardiomyopathy	15	1.24	3.19	0.0495
KEGG	Insulin secretion	15	1.24	3.15	0.0561

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP Tissue	Brain	580	47.78	1.24	6.828E-10
UP Tissue	Hippocampus	48	3.95	1.81	0.0292
UP Tissue	Amygdala	58	4.78	1.66	0.0524

Figure 5.20 f – DAVID pathway analysis of the identified genes containing *Alu* RIPs within the healthy aged samples (n=9) identified by WGS. A total of 1214 mapped genes by DAVID were cross referenced in five defined databases: UP tissue, gene ontology (GO) biological processes, GO cellular component, GO molecular function and KEGG pathway analysis. Green highlighted results indicate statistically significant results with a Bonferroni corrected score of $p<0.05$.

The results of the pathway analysis indicated significant enrichment of multiple pathways with the most significant results being found within the gene lists containing *Alu* RIPs in the healthy aged group (**figure 5.20 f**). Multiple pathways were indicated which support the notion for *Alu* RIPs to preferentially insert into genes which have key importance in brain and central nervous system function including GTPase activity, Rho protein signal transduction, postsynaptic membrane, postsynaptic density, morphine addiction, glutamatergic synapses and expression in total brain tissue and the hippocampus (**figure 5.20 f**). Similar pathway enrichments were observed for the *Alu* RIPs within genes in the PD group also, although to a lesser extent, where the most significant hits were seen for GTPase activity, postsynaptic membrane localisation and genes that were expressed in the brain (**figure 5.20 c**). The gene list containing LINE-1 RIPs within the healthy aged group was significantly enriched for genes involved in postsynaptic membrane localisation and expression in the trachea. No other significant enrichments were found in the gene lists containing LINE-1 or SVA RIPs within the PD group, or the genes containing SVA RIPs in the healthy aged groups. The difference in the levels of significance likely directly reflects the total numbers of RIPs and therefore genes that were identified, i.e. the pathway analyses for the larger gene lists (HA – *Alu* n=1214 genes, PD – *Alu* n=690 genes) returned the most significant results.

5.2.5 Discussion

Given that there exist multiple examples of disorders caused by single mutational events driven by retroelement insertions, the observed differences between the numbers of *Alu* and SVA RIPs between the PD and healthy aged groups could be a key finding. On average, there were 9.5% less *Alu* insertions and 8.3% more SVA insertions within the PD group compared to the healthy aged group. This difference in the average number of RIPs represents a large source of retrotransposition driven mutation between the two groups that should be highlighted and could be important for understanding disease aetiology. Although these finding were not statistically significant, significance may not be expected given the small sample sizes being tested (PD n=2, healthy aged n=9) and so the trends observed could be equally important for this study.

WGS analysis was able to identify 90.5% of identical LINE-1 insertions compared to RC-Seq with a considerably more streamlined library preparation protocol (**figure 5.18**). However, the validation of both the RC-Seq and WGS libraries (**figures 5.4 and 5.14** respectively) indicated that the WGS/TEBreak pipeline had a lower accuracy (0.93 vs 0.91) and sensitivity (0.86 vs 0.79) than RC-Seq. The sensitivity of SVA RIPs detection within the WGS analysis was poor at only 0.65 (**figure 5.14**) and may suggest that the TEBreak discovery algorithms are not suitable for the characterisation of SVA RIPs. There exist multiple retroelement discovery algorithms, as alternatives to TEBreak, such as mobile element locator tool (MELT), RetroSeq, Mobster, Tangram and TEMP [230-234] all of which have various pros and cons. The MELT algorithms have been suggested to outperform these alternate methods

(RetroSeq, Mobster, Tangram and TEMP) in scalability, sensitivity and specificity and could be a good alternative to TEBreak which could be considered for future studies [234]. It is important to consider the application carefully when selecting suitable techniques and algorithms for the identification of RIPs, with the RC-Seq/TEBreak pipeline being suitable for the detection of novel somatic LINE-1 insertions with high sensitivity and the WGS/TEBreak being suitable for characterisation of primarily LINE-1 and *Alu* polymorphic retrotransposons between individuals.

Haploblock cluster analysis identified the chr6:31571218-32682664 PD related haploblock as having the largest number of RIPs which was largely comprised of *Alu* RIPs present within the healthy aged individuals and to a much lesser extent within the PD samples (**table 5.5a**). This was likely due to difference in sample size with only 2 individuals in the PD group compared to 9 in the healthy aged group. Further analysis of this region indicated that this haploblock contained part of the human leukocyte antigen (HLA) super-locus, which incorporates multiple members of the HLA gene family and has been extensively reported as being highly genetically diverse with many mutations driven by retrotransposon and transposon activity [228]. These regions benefit from the actions of retrotransposon activity to generate diversity within the HLA gene family which is crucial for the formation of the major histocompatibility complex (MHC) in humans. Across the MHC class I genomic region, SVA retrotransposon insertion polymorphisms have been previously reported within different populations including Japanese, Caucasian and African American [229, 235]. These insertions highlight the MHC/HLA locus as a highly targetable region for retrotransposition and reflects the same findings observed within the WGS haploblock analysis in **table 5.5a**.

GWAS variants within the HLA locus have been associated with the risk of Parkinson's disease from two independent studies including the variants rs3129882, rs9268515 and rs2395163 [236, 237]. It would therefore be of interest to explore the potential relationship, if any, of the lower numbers of RIPs, in particular the *Alu* sub-class, observed in the PD groups within the haploblock cluster analysis with the GWAS variants previously described. The final results of the haploblock analysis appeared to suggest that *Alu* RIPs were under-represented within PD haploblocks in the PD individuals compared to the healthy aged group (**table 5.5b**). However, these results were likely directly proportional to the sample size being tested with only two samples in the PD group and nine in the healthy aged group. With the large discrepancy in sample size, the numbers of unique RIPs identified within the PD group is severely reduced and larger sample sizes would be needed for future studies.

The pathway analysis performed highlighted that the genes containing RIPs of the *Alu* and LINE-1 classes were most significantly enriched for multiple pathways related to brain function, including expression within the brain, GTPase activity and postsynaptic membrane localisation (**figures 5.20 c, d and f**). Within the gene lists containing *Alu* RIPs, significant enrichment was found for regulation of Rho protein signal transduction and Rho guanyl-nucleotide exchange factor activity (**figure 5.20 f**). Interestingly, Rho GTPase family members have been extensively reported as key players in neuronal development and neurodegeneration which supports the hypothesis that retrotransposon insertion polymorphisms are located within genes involved with neurodegenerative diseases which could have functional consequences [238]. Several members of the Rho GTPase family are expressed in multiple neuron types and are located at the postsynaptic membrane/plasma

membrane where they transduce signals, which was also detected within the pathway analysis (**figure 5.20 f**) [239]. GTPase activity is strongly associated with Parkinson's disease where the actions of both Rab and Rho GTPases have been implicated via various mechanisms, including the regulation of α-synuclein and VMAT2, neuronal toxicity, dysregulation membrane trafficking and mutations in the GTPase LRRK2 [240-243].

It is crucial to note the power of the pathway analysis using DAVID was dependent on the size of the gene lists inputted, meaning that gene lists containing <100 genes would be unlikely to reach statistical significance unless every gene within the given list was closely related to specific pathways. Given this, the lack of enrichment noted within the gene lists containing SVA RIPs within both the PD and healthy aged groups is likely directly due to those lists only containing 66 and 124 genes respectively. This is both due to the significantly lower copy numbers of SVA RIPs identified compared to LINE-1 and *Alu* elements (**figure 5.15**) and also the small sample sizes used within this study (PD n=2, HA n=9). For further analysis, it may be prudent to use other analysis methods for the identified SVA RIPs such as manual curation of the identified genes containing RIPs, or more direct protein interaction analysis such as STRING (https://string-db.org/) which cross-references known protein interactions with literature searches to estimate functional enrichments.

The purposes of this chapter were to explore the validity of using different next generation sequencing techniques (RC-Seq and WGS) for the detection and characterisation of retrotransposon insertion polymorphisms within human brain tissue and to relate the findings in the context of Parkinson's disease. The ability of

both RC-Seq and WGS to detect global RIPs provides a powerful genomic analysis that provides a wealth of information regarding the genetic diversity between individuals and disorders. From the results presented here, the two tools tested (RC-Seq and WGS) both provide different information and should be used in conjunction to provide more detailed analysis. RC-Seq presents a higher sensitivity technique for the detection of LINE-1 insertion polymorphisms with specific sensitivity for the detection of low abundance somatic insertions. In contrast, short read WGS provides a wider blanket coverage of the genome with the ability to detect multiple element types (only LINE-1, SVA and *Alu* were targeted in this study) with a reduced sensitivity for low abundance targets. Given this, approximately 90.5% of the insertions detected by RC-Seq were also detected by WGS which highlights that the vast bulk of insertions are detected by both techniques despite the lower sensitivity. In conclusion, both techniques provide pros and cons for the studying of novel retroelement insertions and should be selected dependent on the specific biological question.

Chapter 6 - General discussion

6.1 Thesis summary

The aim of this thesis was to further characterise and understand the impacts of retrotransposons within the aetiology of a complex neurodegenerative disease in a fundamental way. To do this, various approaches were taken to study the effects of retrotransposons in both a local and global context with the use of specific case studies (*LRRK2* and the *INPP5F*/*BAG3*/*TIAL1* locus) as well as the use of next generation sequencing to address the modification of those effects by retrotransposon insertion polymorphisms. With the ever-advancing technologies that allow for a deeper understanding of genetic variation within complex disorders comes the need to understand the functional significance of transposable elements. In **chapters 3 and 4**, the characterisation of three SVAs across two loci (*LRRK2* and *INPP5F*) was performed to assess the effects of SVA elements to modify gene expression within their respective loci. Utilising a wide range of techniques including PCR genotyping, luciferase-based reporter gene assays and CRISPR mediated knockouts within HEK293 cells, potential SVA function on gene expression, splicing and novel risk polymorphisms could be assessed. In **chapter 5**, the impact of retrotransposable elements was explored using next generation sequencing methods to view a wider range of retrotransposons including the *LINE-1*, Alu and SVA sub-classes on a genome wide scale within the context of Parkinson's disease.

The *LRRK2* locus has been strongly implicated in the aetiology of PD, from both genetic and proteomic viewpoints, with strong GWAS signals also observed within this locus. The GWAS SNP rs34637584 within *LRRK2* presents the second strongest

global GWAS signal observed within the meta-analysis performed by Nalls *et al.* 2019 for PD with a p-value of $3.61x10^{-148}$, second only to the signal observed within *SNCA* (rs356182: p-value $3.89x10^{-154}$) [26]. This association suggests that the *LRRK2* locus has a strong genetic composition in PD, however the known coding variants can only be attributed to approximately 5-13% of familial PD and 1-5% of sporadic PD cases [244]. One of the aims within this thesis was to address the possibility that a portion of the missing heritability observed within the *LRRK2* locus could be due to non-coding variation; a subject that has attracted much attention in recent years [245]. GWAS has not been able to explain the bulk of observed heritability within Parkinson's disease due to the majority of GWAS signals being reported in non-coding regions [245]. One potential source of non-coding variation are the retrotransposable elements which are often over-looked in GWAS analysis due to their repetitive nature [246]. SVA retrotransposons are located within many PD related genes (**table 1.2**) from which two loci were chosen to study in detail; the *LRRK2* and *INPP5F*/*BAG3*/*TIAL1* loci. The SVAs within these loci were explored for potential functional effects on differential gene expression given previous studies have shown that these elements can act as transcriptional regulators and can alter the expressional characteristics of nearby genes [73, 76].

To characterise variation of the SVA retrotransposons present within the *LRRK2* and *INPP5F* loci, PCR genotyping was utilised which revealed the SVA-C within *LRRK2* and both the SVA-F and SVA-D elements within the *INPP5F* locus all contained primary sequence length polymorphisms within the tested cohorts. Both the VNTR and poly-A tail domains of the *LRRK2* SVA-C were reported to be polymorphic with two alleles identified for each domain. Both the *INPP5F* SVA-F and SVA-D elements were found

to be polymorphic within their VNTR domains with three and five alleles identified for SVA-F and SVA-D, respectively. Genotyping of a PD cohort provided no association ($p>0.05$) for any of the identified alleles of the three SVAs with PD when comparing the allele and genotype frequencies between case and control samples (**figures 3.4, 4.18 and 4.21**). However, these findings only reflect a small case sample size ($N<200$) from a single population of Estonian ethnicity and does not exclude the potential for additional alleles to be present amongst other populations. To efficiently increase the genotyping sample size, bioinformatic approaches can be utilised. A total of seventy-seven tagging SNPs were generated for the *LRRK2* SVA-C poly-A alleles using the PLINK algorithms and PCR genotyping of samples from the North American Brain Expression Consortium (NABEC) cohort (**table 3.2**). The NABEC datasets (dbGaP study accession: phs001300.v1.p1) contain publicly available brain expression and functional data from neurologically healthy individuals collected using techniques such as mRNA and exome sequencing. The tagging SNPs generated for the *LRRK2* SVA poly-A genotypes could be used to computationally correlate with the data available within the NABEC repository to infer potential functional differences between the SVA alleles. This approach could be applied to the other SVA polymorphisms that were identified in the *INPP5F* SVA-F and SVA-D elements in the same manner.

Several previous studies have demonstrated the functional aspects of SVA elements, providing evidence that they possess the capacity to act as regulators of gene function using both *in vitro* and *in vivo* models [73, 76, 80, 247]. To address the functional implications of the *LRRK2* and *INPP5F* associated SVAs, luciferase-based reporter gene constructs were successfully generated. The reporter gene assays conducted for the pGL3 constructs containing the cloned *LRRK2* SVA-C and the

INPP5F SVA-F and SVA-D elements showed that all three SVAs elicited significant repressive properties on the expression of luciferase when transfected within HEK293, SH-SY5Y and SK-N-AS established cell lines (**figures 3.6, 4.4 and 4.5**). It was concluded that the SVAs tested are likely to possess similar mechanisms of action that repress transcriptional activity. One possible mechanism could be due to the repetitive sequences observed in the CT and VNTR domains which may have the potential to form alternative DNA structures which impede effective transcription. An example of this is observed in the pathogenic *TAF1* SVA insertion, whereby the CT element has been hypothesised to form G-quadruplexes (G4) which could stall RNA polymerase II and reduce expression of *TAF1* leading to the onset of XDP [80]. SVAs have also been suggested to be enriched in functional binding sites for the TFs, CTCF and SP1, in the CT and VNTR domains respectively, which can contribute to transcriptional repression [52, 176].

The models that utilised the pSHM06 SVA constructs, to address their role on RNA splicing, containing the cloned *LRRK2* SVA-C (**figure 3.8**) and *INPP5F* SVA-D (**figure 4.5**) elements provided insight into the potential differential effects of intronic and intergenic SVAs. Significant decreases in luciferase expression were observed in both models and may be indicative of intron retention when the pre-mRNA transcripts are being spliced. This would result in decreased expression via the termination of translation due to frameshifts [248]. This mechanism has been proposed by Aneichyk *et al*. as the pathogenic mechanism in XDP whereby the intronic SVA insertion within the *TAF1* gene leads to intron retention and subsequent degradation of *TAF1* mRNA leading to disease onset [78].

The use of iPSC and neuronal type derivatives throughout this thesis have been employed to provide more appropriate models to that of established cell lines which often provide fundamental insights into genetic function. Dopaminergic neuronal lineage differentiated iPSC models would have been a more appropriate model system for testing regulatory elements in the context of Parkinson's disease in contrast to the forebrain cortical neuron differentiation protocols which were employed (protocol details in **section 2.2.9.2.2**). However, the protocols for iPSC dopaminergic neuron differentiation available at the time of processing were very time consuming (~50-70 days), expensive and yielded low efficiency of neuron generation [249]. Hence, forebrain cortical neuron differentiation protocols that had been optimised for a 24-day differentiation were employed in both the *INPP5F* and *LRRK2* SVA experiments. More recently, optimised protocols for midbrain dopaminergic neuron (mDAN) differentiation from iPSCs have been published which can be completed within 16 days with continued culture beyond 50 days improving mDAN generation and could provide better models for studying the potential effects of retrotransposons in Parkinson's disease in future studies [250].

Forebrain cortical neuron differentiation protocols were employed in the reporter gene assays in **chapters 3 and 4** to explore the functions of SVA elements within a neuronal context. Unfortunately, due to time constraints, the efficiency of forebrain cortical neuron generation was not assessed. Given time, this would have been done using immunocytochemistry and confocal imaging to probe for neuronal markers of interest. Suitable neuron specific markers to estimate efficiency of differentiation include NeuN, MAP2 and TUBB3. NeuN is a marker of neuronal cells that is distinguishable from glial cells and is suitable as it should detect most neurons with

some rare exceptions. The exceptions include Cajal-Retzius cells of the neocortex, Purkinje cells, γ-motor neurons in the spinal cord and ganglion cells of the sympathetic chain [251]. Both MAP2 and TUBB3 are neuronal cell specific markers routinely used after differentiation protocols [252, 253]. Using multiple neuronal cell markers allows for increased confidence of neuron generation post-iPSC differentiation. By estimating the percentage of neuronal marker positive cells present in the culture, the percentage of cortical neuron generation can be inferred.

The reporter gene assays performed for the *LRRK2* and *INPP5F* SVAs provided only insight into the role of SVAs as *cis* regulatory elements as they removed the SVAs from their respective genomic contexts. To study the effects of SVA *in situ*, CRISPR technologies were employed to knockout the SVA elements of *LRRK2* and *INPP5F* and compare the effects on gene expression of the differential isoforms of both genes. The CRISPR based functional assays described by our group represent the first successful recorded use of CRISPR for the generation of reference genome SVA knockout cell lines to study their *cis* functional implications on gene expression. Optimisations of the CRISPR protocols outlined by Ran *et al.* 2013 were performed which allowed for the generation of three putative mono and three bi-allelic deletion cell lines for each of the *LRRK2* SVA-C, *INPP5F* SVA-F and SVA-D elements that were taken forward for expression analysis [138]. A combination of RT-PCR and qPCR was performed for the analysis of differential isoform expression for the *INPP5F* SVA-F and SVA-D KO cell lines which showed no significant change in expression of either isoforms 1, 2 or 3 or total *INPP5F* expression in response to SVA knockout (**figure 4.12**). RT-PCR analysis was performed for the *LRRK2* SVA-C KO cell lines which showed no consistent observational changes in gene expression of isoform 1, 2 or 3

or total *LRRK2* expression in response to SVA knockout (**figure 3.12**). To further explore potential functional implications of the *LRRK2* SVA knockout, a serum starvation challenge was carried out. This was selected as an appropriate challenge due to presence of Fos and Jun TF binding sites in the *LRRK2* promoter (**figure 3.13**). Serum starvation and re-introduction after a 24-hour starvation period should induce a multitude of cellular signalling pathways including regulation by the serum response factor (SRF) that could alter *LRRK2* expression in addition to binding of Fos and Jun within the promoter region. SRF is involved in NF-kB, integrin, E-cadherin, Wnt and TGFβ signalling amongst others which allows for a broad challenge to be applied to the KO cell models [254, 255]. The assumption was taken that *LRRK2* expression may be differential between WT and KO cell lines in response to a serum starve challenge if the SVA sequence was involved in *LRRK2* regulation. Observational analysis of the RT-PCRs performed indicated that *LRRK2* isoform 3 expression increased more within the putative mono and bi-allelic KO cell lines than the unmodified WT lines under 24-hour serum starvation conditions with a greater effect observed in the bi-allelic SVA KO lines suggesting the SVA could have a role in regulation of *LRRK2* (**figure 3.15**). Sporadic changes in expression were also noted within isoform 1 and isoform 2 but did not appear to correlate with the presence or absence of the SVA suggesting other factors, not considered within the scope of this study, could have been influencing *LRRK2* transcriptional profiles. Analysis using qPCR would be essential for relative comparisons of isoform expression in response to SVA knockout given that observational analysis using RT-PCR was convoluted, however, was not performed due to time restrictions.

To study retrotransposable element function on a genome wide scale, next generation sequencing techniques were employed. The protocols detailed in this thesis represent the first recorded use (at time of writing) of both WGS and RC-Seq techniques for the identification of RIPs within brain tissue extracted DNA from patients with PD. Multiple studies have provided evidence that correlate an increase in retrotransposon activity with neurological disorders, such as ALS, multiple sclerosis and autism spectrum disorder, resulting in an increase in retrotransposition events of the LINE-1 and *Alu* family members or pathogenic changes in gene expression due to ERV elements [96, 215, 256, 257]. The RC-Seq and WGS analysis reported in **chapter 5** using PD brain extracted DNA provided no statistically significant supporting evidence for increased total numbers of retrotransposition events in PD compared to healthy controls (**figures 5.9, 5.15**). However, the RC-Seq analysis suggested a trend for increased LINE-1 RIPs detected in PD with approximately 10% more insertions detected within the PD samples (**figure 5.9**). The WGS supported this, albeit to a lesser degree, identifying an increase of 2.4% LINE-1 RIPs in the PD group compared to controls. It is possible that with larger sample sizes, statistical significance would be achieved to support the finding of previous studies which suggested increased retrotransposition occurs in neurodegenerative diseases [217]. In addition to the identification of LINE-1 RIPS, the WGS analysis reported in **chapter 5.2** also characterised SVA and *Alu* RIPs to further explore the hypothesis that increased retrotransposition events are correlated with neurodegenerative disease. The results of this analysis did not find any statistically significant increase in the total number of RIPs in PD cases compared to healthy controls, in support of the RC-Seq data, however only a small number of genomes were compared (RC-Seq: PD n=5, HA

n=11, WGS: PD n=2, HA n=9, sample details in **table 5.1**). In fact, the trends observed showed a 9.5% increase in the total numbers of *Alu* RIPs in the healthy individuals compared to PD and an 8.3% increase in the number of SVA RIPs present in the PD group compared to the healthy aged group (**figure 5.15**). The percentage differences between the total retrotransposition events between PD and control group represent a source of genetic diversity between the two groups, with on average 115 more *Alu* insertions per person within the healthy aged group and 8 more SVA insertions per person in the PD group. The implications of the larger numbers of *Alu* elements detected in the healthy aged group by WGS was not clear but could be due to differences in ages between the two groups tested. Using multiple model systems including budding yeast, *C. elegans*, *D. melanogaster*, mice and human cell models, the accumulation of novel retrotransposition events has been suggested as a mechanism for aging [258-263]. This led to the working hypothesis that elevated levels of retrotransposition can cause increased mutations, DNA damage and cell instability that may contribute to the aging process [264, 265]. Unfortunately, the ages of the PD samples were not known, so it was not possible to perform correlative analysis between numbers of RIPs with individual age in the contexts of both healthy aging and PD.

Analysis of the distribution of RIPs identified by RC-Seq and WGS indicated that LINE-1 RIPs were more commonly found within intergenic regions which supports the notion that LINE-1 elements preferentially insert into gene-poor regions (**figures 5.11 and 5.19**) [266]. Interestingly, pathway analysis indicated that in both the PD and healthy aged groups, the genes in which LINE-1 RIPs were located were those expressed within the brain (**figures 5.13a-b**). The WGS pathway analysis indicated

that intragenic *Alu* RIPs were also found in genes expressed within the brain in both the PD and healthy aged groups (**figures 5.20 c and f**). This finding may suggest that RIPs preferentially insert into regions of open chromatin which would reflect the genes that are expressed within the tested sample tissue and hence RIPs present within brain tissue would be more likely to insert into genes expressed within the brain. Both LINE-1 and *Alu* elements have been shown to be critical contributors in the evolution of the human brain and have been implicated in neurological disorders which may explain the findings here which show enrichment of LINE-1 and *Alu* RIPs in brain related genes [267, 268]. SVA RIPs were not found to be enriched within any specific tissue or processes from the pathway analysis, however this is likely due to the low number of genes identified which contained SVA RIPs, as a consequence of the low abundance of SVA elements within the human genome, with only 124 and 66 genes containing SVA RIPs in the healthy aged and PD groups respectively being identified (**figures 5.20 b and e**). Vasieva *et al.* has previously described SVA reference elements as being associated with genes involved in brain function, evolution and cognitive ability using a list of the known reference SVAs, at the time of publication (N=2676) [269]. It is possible that with a larger sample size, these findings would have been reflected within the WGS pathway analysis for SVA insertion polymorphisms presented here.

To conclude, retrotransposable elements of the LINE-1, SVA and *Alu* subclasses provide sources of genetic diversity that can drive genomic evolution in a positive manner. However multiple examples of retrotransposon insertion polymorphisms have been implicated in a wide range of disease phenotypes including neurodegeneration [103]. Within this thesis, SVA retrotransposable elements have

been shown as local *cis* regulatory elements in **chapters 3 and 4** within the *LRRK2* and *INPP5F* loci, respectively. Advances in next generation sequencing, as reported in **chapter 5**, have allowed the detection of novel retrotransposition events in far greater detail than ever before. This has provided the foundation for addressing LINE-1, SVA and *Alu* insertion polymorphisms within brain extracted DNA for studying novel variation within PD for the first time, an area that has attracted more attention in recent years and should be continued to be explored.

6.2 Future work

6.2.1 Utilising the tagging SNPs generated for LRRK2 SVA-C polymorphisms

The generation of tagging SNPs for the identified poly-A polymorphisms within the *LRRK2* SVA-C (**table 3.2**) was produced using genotyping data from the North American Brain Expression Consortium (NABEC). The NABEC datasets contain data from exome sequencing, cap analysis gene expression (CAGE) sequencing, mRNA sequencing and NeuroChip genotyping of known neuropathological variants (dbGaP study accession: phs001300.v1.p1). This array of datasets could be used alongside the identified *LRRK2* SVA-C tagging SNPs to correlate with the expression of genes of interest, other pathological variants via LD analysis and correlation with PD related expression quantitative trait loci (eQTLs).

6.2.2 Quantitative analysis of the *LRRK2* isoform expression profiles in response to CRISPR mediated *LRRK2* SVA-C knockout

Time constraints prevented the quantitative analysis of the differential *LRRK2* isoforms expression in response to SVA knockout under both basal and serum starved conditions using qPCR. This analysis is crucial for the fair comparison of expression pattern differences observed within the RT-PCR data presented in **figures 3.12 and 3.15**, as observational analysis is convoluted and imprecise within complicated data.

6.2.3 Exploring other retrotransposon insertion polymorphism discovery algorithms

Re-analysis of the whole genome sequencing raw data produced for the PD and healthy aged groups could be useful using alternate RIP discovery algorithms such as MELT, in particular for the identification of SVA RIPs. Within the WGS data as analysed by TEBreak, there was a significant lower sensitivity observed for the detection of SVA RIPs compared to LINE-1 and Alu elements (**figure 5.14 D**: SVA – 0.65, LINE-1 – 0.79, *Alu* – 0.96). MELT has been demonstrated by Gardner *et al.* 2017, to have up to 0.96 SVA detection sensitivity at a WGS sequencing depth of 30X which would suggest that MELT is more appropriate for the discovery of SVA RIPs than TEBreak [234].

6.2.4 Mitochondrial sequencing

Mitochondrial sequencing has already been performed on 16 samples (4 PD, 1 control and 11 healthy aged samples) with the bioinformatic analysis currently being processed (sample summary in **table 5.1**). Mitochondrial function has been implicated in the aging process, including age related disorders, and has been the focus of much research into neurodegenerative diseases. Mitochondrial dysfunction is considered as a hallmark of aging and has been associated with the development of neurodegenerative diseases such as Alzheimer's and Parkinson's disease [270, 271]. Given this, it is important to understand if, and how, retrotransposon mediated mutations, via transposition, could give rise to novel regulation of mitochondrial related genes that may influence the normal function of the mitochondria.

References

1. Paisán-Ruíz, C., et al., *Cloning of the Gene Containing Mutations that Cause PARK8-Linked Parkinson's Disease.* Neuron. **44**(4): p. 595-600.
2. Armstrong, M.J. and M.S. Okun, *Diagnosis and Treatment of Parkinson Disease: A Review.* JAMA, 2020. **323**(6): p. 548-560.
3. Mahul-Mellier, A.L., et al., *The process of Lewy body formation, rather than simply alpha-synuclein fibrillization, is one of the major drivers of neurodegeneration.* Proc Natl Acad Sci U S A, 2020. **117**(9): p. 4971-4982.
4. Synofzik, M., *Parkinsonism in neurodegenerative diseases predominantly presenting with ataxia.* Int Rev Neurobiol, 2019. **149**: p. 277-298.
5. Di Maio, R., et al., *LRRK2 activation in idiopathic Parkinson's disease.* Sci Transl Med, 2018. **10**(451).
6. Zimprich, A., et al., *Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology.* Neuron, 2004. **44**(4): p. 601-7.
7. Klein, C. and A. Westenberger, *Genetics of Parkinson's disease.* Cold Spring Harb Perspect Med, 2012. **2**(1): p. a008888.
8. Polymeropoulos, M.H., et al., *Mapping of a gene for Parkinson's disease to chromosome 4q21-q23.* Science, 1996. **274**(5290): p. 1197-9.
9. Shults, C.W., *Lewy bodies.* Proc Natl Acad Sci U S A, 2006. **103**(6): p. 1661-8.
10. Burbulla, L.F., et al., *Dopamine oxidation mediates mitochondrial and lysosomal dysfunction in Parkinson's disease.* Science, 2017.
11. Guo, J.L. and V.M. Lee, *Cell-to-cell transmission of pathogenic proteins in neurodegenerative diseases.* Nat Med, 2014. **20**(2): p. 130-8.
12. Osterberg, V.R., et al., *Progressive aggregation of alpha-synuclein and selective degeneration of lewy inclusion-bearing neurons in a mouse model of parkinsonism.* Cell Rep, 2015. **10**(8): p. 1252-60.
13. de Lau, L.M. and M.M. Breteler, *Epidemiology of Parkinson's disease.* Lancet Neurol, 2006. **5**(6): p. 525-35.
14. Reekes, T.H., et al., *Sex specific cognitive differences in Parkinson disease.* NPJ Parkinsons Dis, 2020. **6**: p. 7.
15. Pezzoli, G. and E. Cereda, *Exposure to pesticides or solvents and risk of Parkinson disease.* Neurology, 2013. **80**(22): p. 2035-41.
16. Vlaar, T., et al., *Association of Parkinson's disease with industry sectors: a French nationwide incidence study.* Eur J Epidemiol, 2018. **33**(11): p. 1101-1111.
17. Langston, J.W., et al., *Chronic Parkinsonism in Humans Due to a Product of Meperidine-Analog Synthesis.* Science, 1983. **219**(4587): p. 979-980.
18. Kopin, I.J., *Features of the dopaminergic neurotoxin MPTP.* Ann N Y Acad Sci, 1992. **648**: p. 96-104.
19. Porras, G., Q. Li, and E. Bezard, *Modeling Parkinson's disease in primates: The MPTP model.* Cold Spring Harb Perspect Med, 2012. **2**(3): p. a009308.
20. Langston, J.W., et al., *Selective nigral toxicity after systemic administration of 1-methyl-4-phenyl-1,2,5,6-tetrahydropyrine (MPTP) in the squirrel monkey.* Brain Res, 1984. **292**(2): p. 390-4.
21. Ritz, B.R., K.C. Paul, and J.M. Bronstein, *Of Pesticides and Men: a California Story of Genes and Environment in Parkinson's Disease.* Curr Environ Health Rep, 2016. **3**(1): p. 40-52.
22. Troncoso-Escudero, P., et al., *Outside in: Unraveling the Role of Neuroinflammation in the Progression of Parkinson's Disease.* Front Neurol, 2018. **9**: p. 860.

23. Ransohoff, R.M., *How neuroinflammation contributes to neurodegeneration.* Science, 2016. **353**(6301): p. 777-783.
24. Le, W.D., J.J. Wu, and Y. Tang, *Protective Microglia and Their Regulation in Parkinson's Disease.* Frontiers in Molecular Neuroscience, 2016. **9**.
25. von Bernhardi, R., L. Eugenn-von Bernhardi, and J. Eugenin, *Microglial cell dysregulation in brain aging and neurodegeneration.* Frontiers in Aging Neuroscience, 2015. **7**.
26. Nalls, M.A., et al., *Expanding Parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk.* bioRxiv, 2019: p. 388165.
27. Bandres-Ciga, S., et al., *Genetics of Parkinson's disease: An introspection of its journey towards precision medicine.* Neurobiol Dis, 2020: p. 104782.
28. Tolosa, E., et al., *LRRK2 in Parkinson disease: challenges of clinical trials.* Nat Rev Neurol, 2020. **16**(2): p. 97-107.
29. Hsu, C.H., et al., *MKK6 binds and regulates expression of Parkinson's disease-related protein LRRK2.* J Neurochem, 2010. **112**(6): p. 1593-604.
30. Boon, J.Y., et al., *Interaction of LRRK2 with kinase and GTPase signaling cascades.* Frontiers in Molecular Neuroscience, 2014. **7**: p. 64.
31. Rui, Q., et al., *The Role of LRRK2 in Neurodegeneration of Parkinson Disease.* Curr Neuropharmacol, 2018. **16**(9): p. 1348-1357.
32. Blanca Ramirez, M., et al., *LRRK2 and Parkinson's Disease: From Lack of Structure to Gain of Function.* Curr Protein Pept Sci, 2017. **18**(7): p. 677-686.
33. Luzon-Toro, B., et al., *Mechanistic insight into the dominant mode of the Parkinson's disease-associated G2019S LRRK2 mutation.* Human Molecular Genetics, 2007. **16**(17): p. 2031-2039.
34. Cardona, F., M. Tormos-Perez, and J. Perez-Tur, *Structural and functional in silico analysis of LRRK2 missense substitutions.* Mol Biol Rep, 2014. **41**(4): p. 2529-42.
35. Verma, M., et al., *Mitochondrial Calcium Dysregulation Contributes to Dendrite Degeneration Mediated by PD/LBD-Associated LRRK2 Mutants.* J Neurosci, 2017. **37**(46): p. 11151-11165.
36. Moehle, M.S., et al., *LRRK2 inhibition attenuates microglial inflammatory responses.* J Neurosci, 2012. **32**(5): p. 1602-11.
37. Balla, T., *Phosphoinositides: Tiny Lipids with Giant Impact on Cell Regulation.* Physiological Reviews, 2013. **93**(3): p. 1019-1137.
38. Trivedi, C.M., et al., *Hdac2 regulates the cardiac hypertrophic response by modulating Gsk3 beta activity.* Nat Med, 2007. **13**(3): p. 324-31.
39. Altomare, D.A. and J.R. Testa, *Perturbations of the AKT signaling pathway in human cancer.* Oncogene, 2005. **24**(50): p. 7455-7464.
40. Kim, H.S., et al., *Inositol Polyphosphate-5-Phosphatase F (INPP5F) inhibits STAT3 activity and suppresses gliomas tumorigenicity.* Sci Rep, 2014. **4**: p. 7330.
41. Nakatsu, F., et al., *Sac2/INPP5F is an inositol 4-phosphatase that functions in the endocytic pathway.* Journal of Cell Biology, 2015. **209**(1): p. 85-95.
42. Drouet, V. and S. Lesage, *Synaptojanin 1 Mutation in Parkinson's Disease Brings Further Insight into the Neuropathological Mechanisms.* Biomed Research International, 2014.
43. Hsu, F.S., F.H. Hu, and Y.X. Mao, *Spatiotemporal control of phosphatidylinositol 4-phosphate by Sac2 regulates endocytic recycling.* Journal of Cell Biology, 2015. **209**(1): p. 97-110.
44. Perrett, R.M., Z. Alexopoulou, and G.K. Tofaris, *The endosomal pathway in Parkinson's disease.* Molecular and Cellular Neuroscience, 2015. **66**: p. 21-28.

45. 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. 989-993.
46. Sturner, E. and C. Behl, *The Role of the Multifunctional BAG3 Protein in Cellular Protein Quality Control and in Disease.* Front Mol Neurosci, 2017. **10**: p. 177.
47. Cao, Y.L., et al., *A role of BAG3 in regulating SNCA/alpha-synuclein clearance via selective macroautophagy.* Neurobiol Aging, 2017. **60**: p. 104-115.
48. Marchese, D., et al., *Discovering the 3' UTR-mediated regulation of alpha-synuclein.* Nucleic Acids Res, 2017. **45**(22): p. 12888-12903.
49. Hancks, D.C. and H.H. Kazazian, *Active Human Retrotransposons: Variation and Disease.* Current opinion in genetics & development, 2012. **22**(3): p. 191-203.
50. Batzer, M.A. and P.L. Deininger, *A human-specific subfamily of Alu sequences.* Genomics, 1991. **9**(3): p. 481-7.
51. 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.
52. Wang, H., et al., *SVA elements: a hominid-specific retroposon family.* J Mol Biol, 2005. **354**(4): p. 994-1007.
53. 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.
54. Li, W., et al., *Human endogenous retrovirus-K contributes to motor neuron disease.* Sci Transl Med, 2015. **7**(307): p. 307ra153.
55. McClintock, B., *The Origin and Behavior of Mutable Loci in Maize.* Proceedings of the National Academy of Sciences of the United States of America, 1950. **36**(6): p. 344-355.
56. Beck, C.R., et al., *LINE-1 Elements in Structural Variation and Disease.* Annual review of genomics and human genetics, 2011. **12**: p. 187-215.
57. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements.* Nat Rev Genet, 2007. **8**(12): p. 973-982.
58. Callinan, P.A. and M.A. Batzer, *Retrotransposable elements and human disease.* Genome Dyn, 2006. **1**: p. 104-15.
59. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
60. Kazazian, H.H., Jr. and J.V. Moran, *The impact of L1 retrotransposons on the human genome.* Nat Genet, 1998. **19**(1): p. 19-24.
61. Claeys Bouuaert, C. and R. Chalmers, *Transposition of the human Hsmar1 transposon: rate-limiting steps and the importance of the flanking TA dinucleotide in second strand cleavage.* Nucleic Acids Research, 2010. **38**(1): p. 190-202.
62. Ayarpadikannan, S. and H.-S. Kim, *The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases.* Genomics & Informatics, 2014. **12**(3): p. 98-104.
63. Belshaw, R., et al., *Long-term reinfection of the human genome by endogenous retroviruses.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(14): p. 4894-4899.
64. Paces, J., A. Pavlícek, and V. Paces, *HERVd: database of human endogenous retroviruses.* Nucleic Acids Research, 2002. **30**(1): p. 205-206.
65. Vassetzky, N.S. and D.A. Kramerov, *SINEBase: a database and tool for SINE analysis.* Nucleic Acids Research, 2012.
66. Wang, H., et al., *SVA Elements: A Hominid-specific Retroposon Family.* Journal of Molecular Biology, 2005. **354**(4): p. 994-1007.
67. Hancks, D.C., et al., *Retrotransposition of marked SVA elements by human L1s in cultured cells.* Hum Mol Genet, 2011. **20**(17): p. 3386-400.

68. Raiz, J., et al., *The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery.* Nucleic Acids Res, 2012. **40**(4): p. 1666-83.
69. Quinn, J.P. and V.J. Bubb, *SVA retrotransposons as modulators of gene expression.* Mobile Genetic Elements, 2014. **4**: p. e32102.
70. Gianfrancesco, O., V.J. Bubb, and J.P. Quinn, *SVA retrotransposons as potential modulators of neuropeptide gene expression.* Neuropeptides, 2017. **64**: p. 3-7.
71. Hancks, D.C., et al., *Exon-trapping mediated by the human retrotransposon SVA.* Genome Res, 2009. **19**(11): p. 1983-91.
72. Savage, A.L., et al., *Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns.* BMC Evolutionary Biology, 2013. **13**(1): p. 1-12.
73. 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**(6): p. e90833.
74. Upton, Kyle R., et al., *Ubiquitous L1 Mosaicism in Hippocampal Neurons.* Cell, 2015. **161**(2): p. 228-239.
75. Hancks, D.C. and H.H. Kazazian Jr, *SVA retrotransposons: Evolution and genetic instability.* Seminars in Cancer Biology, 2010. **20**(4): p. 234-245.
76. Savage, A.L., et al., *Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns.* BMC Evol Biol, 2013. **13**: p. 101.
77. van der Klift, H.M., et al., *Insertion of an SVA element, a nonautonomous retrotransposon, in PMS2 intron 7 as a novel cause of Lynch syndrome.* Hum Mutat, 2012. **33**(7): p. 1051-5.
78. Aneichyk, T., et al., *Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly.* Cell, 2018. **172**(5): p. 897-909 e21.
79. Rakovic, A., et al., *Genome editing in induced pluripotent stem cells rescues TAF1 levels in X-linked dystonia-parkinsonism.* Mov Disord, 2018. **33**(7): p. 1108-1118.
80. 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.* Proc Natl Acad Sci U S A, 2017. **114**(51): p. E11020-E11028.
81. Strichman-Almashanu, L.Z., et al., *A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes.* Genome Research, 2002. **12**(4): p. 543-554.
82. Kejnovsky, E., V. Tokan, and M. Lexa, *Transposable elements and G-quadruplexes.* Chromosome Res, 2015. **23**(3): p. 615-23.
83. David, A.P., et al., *G-quadruplexes as novel cis-elements controlling transcription during embryonic development.* Nucleic Acids Research, 2016. **44**(9): p. 4163-4173.
84. Gu, H.P., et al., *Up-regulating relaxin expression by G-quadruplex interactive ligand to achieve antifibrotic action.* Endocrinology, 2012. **153**(8): p. 3692-700.
85. Wang, X.D., et al., *Turning off transcription of the bcl-2 gene by stabilizing the bcl-2 promoter quadruplex with quindoline derivatives.* J Med Chem, 2010. **53**(11): p. 4390-8.
86. Wang, H.L., et al., *CtIP Maintains Stability at Common Fragile Sites and Inverted Repeats by End Resection-Independent Endonuclease Activity.* Molecular Cell, 2014. **54**(6): p. 1012-1021.
87. Batzer, M.A. and P.L. Deininger, *Alu repeats and human genomic diversity.* Nat Rev Genet, 2002. **3**(5): p. 370-9.

88. Savage, A.L., et al., *Retrotransposons in the development and progression of amyotrophic lateral sclerosis.* J Neurol Neurosurg Psychiatry, 2019. **90**(3): p. 284-293.
89. Mighell, A.J., A.F. Markham, and P.A. Robinson, *Alu sequences.* FEBS Lett, 1997. **417**(1): p. 1-5.
90. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.
91. Bennett, E.A., et al., *Active Alu retrotransposons in the human genome.* Genome Res, 2008. **18**(12): p. 1875-83.
92. Cordaux, R., et al., *Estimating the retrotransposition rate of human Alu elements.* Gene, 2006. **373**: p. 134-7.
93. Hancks, D.C. and H.H. Kazazian, Jr., *Active human retrotransposons: variation and disease.* Curr Opin Genet Dev, 2012. **22**(3): p. 191-203.
94. Zarnack, K., et al., *Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements.* Cell, 2013. **152**(3): p. 453-466.
95. Sela, N., et al., *Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome.* Genome Biology, 2007. **8**(6).
96. Larsen, P.A., et al., *The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease.* Alzheimers Dement, 2017. **13**(7): p. 828-838.
97. Raiz, J., et al., *The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery.* Nucleic Acids Research, 2012. **40**(4): p. 1666-1683.
98. Kim, S., et al., *Structural Variation of Alu Element and Human Disease.* Genomics Inform, 2016. **14**(3): p. 70-77.
99. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
100. 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.
101. Feng, Q., et al., *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell, 1996. **87**(5): p. 905-16.
102. Wei, W., et al., *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol, 2001. **21**(4): p. 1429-39.
103. Hancks, D.C. and H.H. Kazazian, Jr., *Roles for retrotransposon insertions in human disease.* Mob DNA, 2016. **7**: p. 9.
104. Kazazian, H.H., Jr., 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.
105. Wallace, N.A., V.P. Belancio, and P.L. Deininger, *L1 mobile element expression causes multiple types of toxicity.* Gene, 2008. **419**(1-2): p. 75-81.
106. Belancio, V.P., D.J. Hedges, and P. Deininger, *Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health.* Genome Res, 2008. **18**(3): p. 343-58.
107. Gasior, S.L., et al., *The human LINE-1 retrotransposon creates DNA double-strand breaks.* J Mol Biol, 2006. **357**(5): p. 1383-93.
108. Guo, C., et al., *Tau Activates Transposable Elements in Alzheimer's Disease.* Cell Rep, 2018. **23**(10): p. 2874-2880.
109. Krug, L., et al., *Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS.* PLoS Genet, 2017. **13**(3): p. e1006635.

110. de The, F.X.B., et al., *Engrailed homeoprotein blocks degeneration in adult dopaminergic neurons through LINE-1 repression.* Embo Journal, 2018. **37**(15).
111. Jurka, J., *Repbase update: a database and an electronic journal of repetitive elements.* Trends Genet, 2000. **16**(9): p. 418-20.
112. Searles Nielsen, S., et al., *LINE-1 DNA methylation, smoking and risk of Parkinson's disease.* J Parkinsons Dis, 2012. **2**(4): p. 303-8.
113. Baeken, M.W., B. Moosmann, and P. Hajieva, *Retrotransposon activation by distressed mitochondria in neurons.* Biochem Biophys Res Commun, 2020.
114. Neumann, M., et al., *Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis.* Science, 2006. **314**(5796): p. 130-3.
115. Li, W., et al., *Transposable elements in TDP-43-mediated neurodegenerative disorders.* PLoS One, 2012. **7**(9): p. e44099.
116. Liu, E.Y., et al., *Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons.* Cell Reports, 2019. **27**(5): p. 1409-+.
117. Buratowski, S., et al., *Five intermediate complexes in transcription initiation by RNA polymerase II.* Cell, 1989. **56**(4): p. 549-61.
118. Flores, O., H. Lu, and D. Reinberg, *Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH.* J Biol Chem, 1992. **267**(4): p. 2786-93.
119. Matsui, T., et al., *Multiple factors required for accurate initiation of transcription by purified RNA polymerase II.* J Biol Chem, 1980. **255**(24): p. 11992-6.
120. Romanish, M.T., et al., *A novel protein isoform of the multicopy human NAIP gene derives from intragenic Alu SINE promoters.* PLoS One, 2009. **4**(6): p. e5761.
121. Sundaram, V., et al., *Widespread contribution of transposable elements to the innovation of gene regulatory networks.* Genome Res, 2014. **24**(12): p. 1963-76.
122. Nikitin, D., et al., *Profiling of Human Molecular Pathways Affected by Retrotransposons at the Level of Regulation by Transcription Factor Proteins.* Frontiers in Immunology, 2018. **9**.
123. Massari, M.E. and C. Murre, *Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms.* Mol Cell Biol, 2000. **20**(2): p. 429-40.
124. Adhikary, S. and M. Eilers, *Transcriptional regulation and transformation by Myc proteins.* Nat Rev Mol Cell Biol, 2005. **6**(8): p. 635-45.
125. Grandori, C., et al., *The Myc/Max/Mad network and the transcriptional control of cell behavior.* Annu Rev Cell Dev Biol, 2000. **16**: p. 653-99.
126. Kagawa, T., et al., *Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype.* Hum Mutat, 2015. **36**(3): p. 327-32.
127. Sorek, R., G. Ast, and D. Graur, *Alu-containing exons are alternatively spliced.* Genome Res, 2002. **12**(7): p. 1060-7.
128. Mitchell, G.A., et al., *Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation.* Proc Natl Acad Sci U S A, 1991. **88**(3): p. 815-9.
129. Vervoort, R., et al., *A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene.* Human Genetics, 1998. **103**(6): p. 686-693.
130. Jacobs, F.M.J., et al., *An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons.* Nature, 2014. **516**(7530): p. 242-245.
131. Deniz, O., J.M. Frost, and M.R. Branco, *Regulation of transposable elements by DNA modifications.* Nat Rev Genet, 2019. **20**(7): p. 417-431.
132. Yang, X.J., et al., *Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer.* Cancer Cell, 2014. **26**(4): p. 577-590.

133. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-322.
134. Fasching, L., et al., *TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells.* Cell Rep, 2015. **10**(1): p. 20-8.
135. Xicoy, H., B. Wieringa, and G.J.M. Martens, *The SH-SY5Y cell line in Parkinson's disease research: a systematic review.* Molecular Neurodegeneration, 2017. **12**(1): p. 10.
136. Stepanenko, A.A. and V.V. Dmitrenko, *HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution.* Gene, 2015. **569**(2): p. 182-90.
137. Herbst, F., et al., *Extensive methylation of promoter sequences silences lentiviral transgene expression during stem cell differentiation in vivo.* Mol Ther, 2012. **20**(5): p. 1014-21.
138. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nature Protocols, 2013. **8**(11): p. 2281-2308.
139. Krinner, S., et al., *Interplay of Promoter Usage and Intragenic CpG Content: Impact on GFP Reporter Gene Expression.* Human Gene Therapy, 2015. **26**(12): p. 826-840.
140. Sanchez-Luque, F.J., S.R. Richardson, and G.J. Faulkner, *Retrotransposon Capture Sequencing (RC-Seq): A Targeted, High-Throughput Approach to Resolve Somatic L1 Retrotransposition in Humans.* Methods Mol Biol, 2016. **1400**: p. 47-77.
141. Owczarzy, R., et al., *Stability and mismatch discrimination of locked nucleic acid-DNA duplexes.* Biochemistry, 2011. **50**(43): p. 9352-67.
142. Marti-Masso, J.F., et al., *Neuropathology of Parkinson's disease with the R1441G mutation in LRRK2.* Mov Disord, 2009. **24**(13): p. 1998-2001.
143. Zhang, Z.J., et al., *LRRK2 R1628P Variant Is a Risk Factor of Parkinson's Disease Among Han-Chinese from Mainland China.* Movement Disorders, 2009. **24**(13): p. 1902-1905.
144. Trinh, J., et al., *Comparative study of Parkinson's disease and leucine-rich repeat kinase 2 p.G2019S parkinsonism.* Neurobiol Aging, 2014. **35**(5): p. 1125-31.
145. Khursheed, K., et al., *Characterisation of multiple regulatory domains spanning the major transcriptional start site of the FUS gene, a candidate gene for motor neurone disease.* Brain Res, 2015. **1595**: p. 1-9.
146. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-D1012.
147. Pickrell, J.K., et al., *Detection and interpretation of shared genetic influences on 42 human traits.* Nat Genet, 2016. **48**(7): p. 709-17.
148. Lill, C.M., et al., *Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database.* PLoS Genet, 2012. **8**(3): p. e1002548.
149. Bandres-Ciga, S., et al., *The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight.* Mov Disord, 2019. **34**(12): p. 1851-1863.
150. Witoelar, A., et al., *Genome-wide Pleiotropy Between Parkinson Disease and Autoimmune Diseases.* Jama Neurology, 2017. **74**(7): p. 780-792.
151. Brudek, T., *Inflammatory Bowel Diseases and Parkinson's Disease.* Journal of Parkinsons Disease, 2019. **9**: p. S331-S344.
152. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

153. International Parkinson Disease Genomics, C., *Ten Years of the International Parkinson Disease Genomics Consortium: Progress and Next Steps.* J Parkinsons Dis, 2020. **10**(1): p. 19-30.
154. Slatkin, M., *Linkage disequilibrium [mdash] understanding the evolutionary past and mapping the medical future.* Nat Rev Genet, 2008. **9**(6): p. 477-485.
155. Chen, R. and A.J. Butte, *The reference human genome demonstrates high risk of type 1 diabetes and other disorders.* Pac Symp Biocomput, 2011: p. 231-42.
156. Bertrand, N., D.S. Castro, and F. Guillemot, *Proneural genes and the specification of neural cell types.* Nat Rev Neurosci, 2002. **3**(7): p. 517-30.
157. Atchley, W.R. and W.M. Fitch, *A natural classification of the basic helix-loop-helix class of transcription factors.* Proc Natl Acad Sci U S A, 1997. **94**(10): p. 5172-6.
158. Uittenbogaard, M. and A. Chiaramello, *The basic helix-loop-helix transcription factor Nex-1/Math-2 promotes neuronal survival of PC12 cells by modulating the dynamic expression of anti-apoptotic and cell cycle regulators.* J Neurochem, 2005. **92**(3): p. 585-96.
159. Gabay, M., Y.L. Li, and D.W. Felsher, *MYC Activation Is a Hallmark of Cancer Initiation and Maintenance.* Cold Spring Harbor Perspectives in Medicine, 2014. **4**(6).
160. Soh, J.W., et al., *Novel roles of specific isoforms of protein kinase C in activation of the c-fos serum response element.* Mol Cell Biol, 1999. **19**(2): p. 1313-24.
161. Johansen, F.E. and R. Prywes, *Two pathways for serum regulation of the c-fos serum response element require specific sequence elements and a minimal domain of serum response factor.* Mol Cell Biol, 1994. **14**(9): p. 5920-8.
162. Hardison, R.C., *Conserved noncoding sequences are reliable guides to regulatory elements.* Trends Genet, 2000. **16**(9): p. 369-72.
163. Chang, D., et al., *A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci.* Nat Genet, 2017. **advance online publication**.
164. Pugacheva, E.M., et al., *CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention.* Proc Natl Acad Sci U S A, 2020. **117**(4): p. 2020-2031.
165. Rieder, D., Z. Trajanoski, and J.G. McNally, *Transcription factories.* Front Genet, 2012. **3**: p. 221.
166. Ostertag, E.M., et al., *SVA elements are nonautonomous retrotransposons that cause disease in humans.* American Journal of Human Genetics, 2003. **73**(6): p. 1444-1451.
167. Klawitter, S., et al., *Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells.* Nat Commun, 2016. **7**: p. 10286.
168. Thomas, P. and T.G. Smart, *HEK293 cell line: a vehicle for the expression of recombinant proteins.* J Pharmacol Toxicol Methods, 2005. **51**(3): p. 187-200.
169. Loots, G.G. and I. Ovcharenko, *rVISTA 2.0: evolutionary analysis of transcription factor binding sites.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W217-21.
170. Kouwenhoven, W.M. and H.J. van Heesbeen, *BMP/SMAD Pathway and the Development of Dopamine Substantia Nigra Neurons.* J Neurosci, 2018. **38**(28): p. 6244-6246.
171. Jovanovic, V.M., et al., *BMP/SMAD Pathway Promotes Neurogenesis of Midbrain Dopaminergic Neurons In Vivo and in Human Induced Pluripotent and Neural Stem Cells.* J Neurosci, 2018. **38**(7): p. 1662-1676.
172. Duclot, F. and M. Kabbaj, *The Role of Early Growth Response 1 (EGR1) in Brain Plasticity and Neuropsychiatric Disorders.* Front Behav Neurosci, 2017. **11**: p. 35.
173. Friedli, M., et al., *Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency.* Genome Res, 2014. **24**(8): p. 1251-9.

174. Rowe, H.M., et al., *KAP1 controls endogenous retroviruses in embryonic stem cells.* Nature, 2010. **463**(7278): p. 237-40.
175. Wolf, D. and S.P. Goff, *TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells.* Cell, 2007. **131**(1): p. 46-57.
176. 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**(1): p. 35.
177. Soltanian, S., et al., *Expression analysis of BORIS during pluripotent, differentiated, cancerous, and non-cancerous cell states.* Acta Biochim Biophys Sin (Shanghai), 2014. **46**(8): p. 647-58.
178. Pugacheva, E.M., et al., *The structural complexity of the human BORIS gene in gametogenesis and cancer.* PLoS One, 2010. **5**(11): p. e13872.
179. Monk, M., M. Hitchins, and S. Hawes, *Differential expression of the embryo/cancer gene ECSA(DPPA2), the cancer/testis gene BORIS and the pluripotency structural gene OCT4, in human preimplantation development.* Mol Hum Reprod, 2008. **14**(6): p. 347-55.
180. Carta, A.R., *PPAR-gamma: therapeutic prospects in Parkinson's disease.* Curr Drug Targets, 2013. **14**(7): p. 743-51.
181. Perez-Santiago, J., et al., *A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia.* J Psychiatr Res, 2012. **46**(11): p. 1464-74.
182. Knapska, E. and L. Kaczmarek, *A gene for neuronal plasticity in the mammalian brain: Zif268/Egr-1/NGFI-A/Krox-24/TIS8/ZENK?* Prog Neurobiol, 2004. **74**(4): p. 183-211.
183. Krishnan, V. and E.J. Nestler, *The molecular neurobiology of depression.* Nature, 2008. **455**(7215): p. 894-902.
184. Haddley, K., et al., *Behavioural genetics of the serotonin transporter.* Curr Top Behav Neurosci, 2012. **12**: p. 503-35.
185. Guindalini, C., et al., *A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample.* Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4552-7.
186. Brotons, O., et al., *Modulation of orbitofrontal response to amphetamine by a functional variant of DAT1 and in vitro confirmation.* Mol Psychiatry, 2011. **16**(2): p. 124-6.
187. Beck, C.R., et al., *LINE-1 retrotransposition activity in human genomes.* Cell, 2010. **141**(7): p. 1159-70.
188. Vilarino-Guell, C., et al., *DNAJC13 mutations in Parkinson disease.* Hum Mol Genet, 2014. **23**(7): p. 1794-801.
189. Guichard, E., et al., *Impact of non-LTR retrotransposons in the differentiation and evolution of anatomically modern humans.* Mob DNA, 2018. **9**: p. 28.
190. Muotri, A.R., et al., *Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.* Nature, 2005. **435**(7044): p. 903-10.
191. McConnell, M.J., et al., *Mosaic copy number variation in human neurons.* Science, 2013. **342**(6158): p. 632-7.
192. Deng, H., P. Wang, and J. Jankovic, *The genetics of Parkinson disease.* Ageing Res Rev, 2018. **42**: p. 72-85.
193. Kano, H., et al., *L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism.* Genes & Development, 2009. **23**(11): p. 1303-1312.
194. Berisa, T. and J.K. Pickrell, *Approximately independent linkage disequilibrium blocks in human populations.* Bioinformatics, 2016. **32**(2): p. 283-5.

195. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.
196. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
197. Baratloo, A., et al., *Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.* Emerg (Tehran), 2015. **3**(2): p. 48-9.
198. Wang, J., et al., *dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans.* Hum Mutat, 2006. **27**(4): p. 323-9.
199. Stewart, C., et al., *A comprehensive map of mobile element insertion polymorphisms in humans.* PLoS Genet, 2011. **7**(8): p. e1002236.
200. Lee, E., et al., *Landscape of somatic retrotransposition in human cancers.* Science, 2012. **337**(6097): p. 967-71.
201. Kuhn, A., et al., *Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. 8131-6.
202. Iskow, R.C., et al., *Natural mutagenesis of human genomes by endogenous retrotransposons.* Cell, 2010. **141**(7): p. 1253-61.
203. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.
204. Helman, E., et al., *Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing.* Genome Res, 2014. **24**(7): p. 1053-63.
205. Ewing, A.D. and H.H. Kazazian, Jr., *High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.* Genome Res, 2010. **20**(9): p. 1262-70.
206. Shukla, R., et al., *Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma.* Cell, 2013. **153**(1): p. 101-11.
207. Tubio, J.M.C., et al., *Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.
208. Gyllborg, D., et al., *The Matricellular Protein R-Spondin 2 Promotes Midbrain Dopaminergic Neurogenesis and Differentiation.* Stem Cell Reports, 2018. **11**(3): p. 651-664.
209. Muotri, A.R., et al., *L1 retrotransposition in neurons is modulated by MeCP2.* Nature, 2010. **468**(7322): p. 443-446.
210. Jacob-Hirsch, J., et al., *Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders.* Cell Res, 2018. **28**(2): p. 187-203.
211. Bundo, M., et al., *Increased l1 retrotransposition in the neuronal genome in schizophrenia.* Neuron, 2014. **81**(2): p. 306-13.
212. Upton, K.R., et al., *Ubiquitous L1 mosaicism in hippocampal neurons.* Cell, 2015. **161**(2): p. 228-39.
213. Deleye, L., et al., *Performance of four modern whole genome amplification methods for copy number variant detection in single cells.* Sci Rep, 2017. **7**(1): p. 3422.
214. Tam, O.H., et al., *Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia.* Cell Rep, 2019. **29**(5): p. 1164-1177 e5.
215. Tam, O.H., L.W. Ostrow, and M. Gale Hammell, *Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease.* Mob DNA, 2019. **10**: p. 32.

216. Pereira, G.C., et al., *Properties of LINE-1 proteins and repeat element expression in the context of amyotrophic lateral sclerosis.* Mob DNA, 2018. **9**: p. 35.
217. Terry, D.M. and S.E. Devine, *Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders.* Front Genet, 2019. **10**: p. 1244.
218. Evrony, G.D., et al., *Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain.* Cell, 2012. **151**(3): p. 483-96.
219. 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.
220. Sanchez-Luque, F.J., et al., *LINE-1 Evasion of Epigenetic Repression in Humans.* Mol Cell, 2019. **75**(3): p. 590-604 e12.
221. Dunaeva, M., M. Derksen, and G.J.M. Pruijn, *LINE-1 Hypermethylation in Serum Cell-Free DNA of Relapsing Remitting Multiple Sclerosis Patients.* Mol Neurobiol, 2018. **55**(6): p. 4681-4688.
222. Payton, A., et al., *A TOMM40 poly-T variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging.* Neurobiology of Aging, 2016. **39**.
223. Faulkner, G.J., *Retrotransposons: mobile and mutagenic from conception to death.* FEBS Lett, 2011. **585**(11): p. 1589-94.
224. Bennett, E.A., et al., *Natural genetic variation caused by transposable elements in humans.* Genetics, 2004. **168**(2): p. 933-51.
225. Ewing, A.D. and H.H. Kazazian, Jr., *Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans.* Genome Res, 2011. **21**(6): p. 985-90.
226. Ewing, A.D., *Transposable element detection from whole genome sequence data.* Mob DNA, 2015. **6**: p. 24.
227. Sen, S.K., et al., *Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome.* Nucleic Acids Res, 2007. **35**(11): p. 3741-51.
228. Shiina, T., et al., *The HLA genomic loci map: expression, interaction, diversity and disease.* J Hum Genet, 2009. **54**(1): p. 15-39.
229. Kulski, J.K., A. Shigenari, and H. Inoko, *Polymorphic SVA retrotransposons at four loci and their association with classical HLA class I alleles in Japanese, Caucasians and African Americans.* Immunogenetics, 2010. **62**(4): p. 211-230.
230. Zhuang, J., et al., *TEMP: a computational method for analyzing transposable element polymorphism in populations.* Nucleic Acids Res, 2014. **42**(11): p. 6826-38.
231. Keane, T.M., K. Wong, and D.J. Adams, *RetroSeq: transposable element discovery from next-generation sequencing data.* Bioinformatics, 2013. **29**(3): p. 389-90.
232. Thung, D.T., et al., *Mobster: accurate detection of mobile element insertions in next generation sequencing data.* Genome Biol, 2014. **15**(10): p. 488.
233. Wu, J., et al., *Tangram: a comprehensive toolbox for mobile element insertion detection.* BMC Genomics, 2014. **15**: p. 795.
234. Gardner, E.J., et al., *The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology.* Genome Res, 2017. **27**(11): p. 1916-1929.
235. Takasu, M., et al., *Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon.* Tissue Antigens, 2007. **70**(2): p. 144-50.
236. Ahmed, I., et al., *Association between Parkinson's disease and the HLA-DRB1 locus.* Movement Disorders, 2012. **27**(9): p. 1104-1110.
237. Wissemann, W.T., et al., *Association of Parkinson disease with structural and regulatory variants in the HLA region.* Am J Hum Genet, 2013. **93**(5): p. 984-93.

238. Stankiewicz, T.R. and D.A. Linseman, *Rho family GTPases: key players in neuronal development, neuronal survival, and neurodegeneration.* Front Cell Neurosci, 2014. **8**: p. 314.
239. Kalpachidou, T., et al., *Rho GTPases in the Physiology and Pathophysiology of Peripheral Sensory Neurons.* Cells, 2019. **8**(6).
240. Stafa, K., et al., *GTPase Activity and Neuronal Toxicity of Parkinson's Disease-Associated LRRK2 Is Regulated by ArfGAP1.* Plos Genetics, 2012. **8**(2).
241. Zhou, Z., et al., *Rho GTPase regulation of alpha-synuclein and VMAT2: implications for pathogenesis of Parkinson's disease.* Mol Cell Neurosci, 2011. **48**(1): p. 29-37.
242. Hong, L. and L.A. Sklar, *Targeting GTPases in Parkinson's disease: comparison to the historic path of kinase drug discovery and perspectives.* Front Mol Neurosci, 2014. **7**: p. 52.
243. Bonet-Ponce, L. and M.R. Cookson, *The role of Rab GTPases in the pathobiology of Parkinson' disease.* Curr Opin Cell Biol, 2019. **59**: p. 73-80.
244. Drolet, R.E., J.M. Sanders, and J.T. Kern, *Leucine-rich repeat kinase 2 (LRRK2) cellular biology: a review of recent advances in identifying physiological substrates and cellular functions.* J Neurogenet, 2011. **25**(4): p. 140-51.
245. Ohnmacht, J., et al., *Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation.* J Neural Transm (Vienna), 2020. **127**(5): p. 729-748.
246. Bourque, G., *Transposable elements in gene regulation and in the evolution of vertebrate genomes.* Curr Opin Genet Dev, 2009. **19**(6): p. 607-12.
247. Westenberger, A., et al., *A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism.* Ann Neurol, 2019. **85**(6): p. 812-822.
248. Lee, Y. and D.C. Rio, *Mechanisms and Regulation of Alternative Pre-mRNA Splicing.* Annu Rev Biochem, 2015. **84**: p. 291-323.
249. Ma, L., Y. Liu, and S.C. Zhang, *Directed differentiation of dopamine neurons from human pluripotent stem cells.* Methods Mol Biol, 2011. **767**: p. 411-8.
250. Stathakos, P., et al., *Imaging Autophagy in hiPSC-Derived Midbrain Dopaminergic Neuronal Cultures for Parkinson's Disease Research.* Methods Mol Biol, 2019. **1880**: p. 257-280.
251. Gusel'nikova, V.V. and D.E. Korzhevskiy, *NeuN As a Neuronal Nuclear Antigen and Neuron Differentiation Marker.* Acta Naturae, 2015. **7**(2): p. 42-47.
252. Soltani, M.H., et al., *Microtubule-associated protein 2, a marker of neuronal differentiation, induces mitotic defects, inhibits growth of melanoma cells, and predicts metastatic potential of cutaneous melanoma.* Am J Pathol, 2005. **166**(6): p. 1841-50.
253. Latremoliere, A., et al., *Neuronal-Specific TUBB3 Is Not Required for Normal Neuronal Function but Is Essential for Timely Axon Regeneration.* Cell Reports, 2018. **24**(7): p. 1865-+.
254. Sandbo, N., et al., *Critical role of serum response factor in pulmonary myofibroblast differentiation induced by TGF-beta.* Am J Respir Cell Mol Biol, 2009. **41**(3): p. 332-8.
255. Miano, J.M., *Role of serum response factor in the pathogenesis of disease.* Lab Invest, 2010. **90**(9): p. 1274-84.
256. Ochoa Thomas, E., et al., *Awakening the dark side: retrotransposon activation in neurodegenerative disorders.* Curr Opin Neurobiol, 2020. **61**: p. 65-72.
257. Saleh, A., A. Macia, and A.R. Muotri, *Transposable Elements, Inflammation, and Neurological Disease.* Front Neurol, 2019. **10**: p. 894.
258. De Cecco, M., et al., *Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements.* Aging Cell, 2013. **12**(2): p. 247-56.

259. De Cecco, M., et al., *Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues.* Aging (Albany NY), 2013. **5**(12): p. 867-83.
260. Dennis, S., et al., *C. elegans Germ Cells Show Temperature and Age-Dependent Expression of Cer1, a Gypsy/Ty3-Related Retrotransposon.* Plos Pathogens, 2012. **8**(3).
261. Hu, Z., et al., *Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging.* Genes Dev, 2014. **28**(4): p. 396-408.
262. Van Meter, M., et al., *SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age.* Nature Communications, 2014. **5**.
263. Wang, J., et al., *Inhibition of activated pericentromeric SINE/Alu repeat transcription in senescent human adult stem cells reinstates self-renewal.* Cell Cycle, 2011. **10**(17): p. 3016-30.
264. Huang, C.R., K.H. Burns, and J.D. Boeke, *Active transposition in genomes.* Annu Rev Genet, 2012. **46**: p. 651-75.
265. Lopez-Otin, C., et al., *The hallmarks of aging.* Cell, 2013. **153**(6): p. 1194-217.
266. Graham, T. and S. Boissinot, *The genomic distribution of L1 elements: the role of insertion bias and natural selection.* J Biomed Biotechnol, 2006. **2006**(1): p. 75327.
267. Suarez, N.A., A. Macia, and A.R. Muotri, *LINE-1 retrotransposons in healthy and diseased human brain.* Dev Neurobiol, 2018. **78**(5): p. 434-455.
268. Larsen, P.A., et al., *Warning SINEs: Alu elements, evolution of the human brain, and the spectrum of neurological disease.* Chromosome Res, 2018. **26**(1-2): p. 93-111.
269. Vasieva, O., et al., *Potential impact of primate-specific SVA retrotransposons during the evolution of human cognitive function.* Trends in Evolutionary Biology, 2017. **6**(1).
270. Keogh, M.J. and P.F. Chinnery, *Mitochondrial DNA mutations in neurodegeneration.* Biochim Biophys Acta, 2015. **1847**(11): p. 1401-11.
271. Martin-Jimenez, R., O. Lurette, and E. Hebert-Chatelain, *Damage in Mitochondrial DNA Associated with Parkinson's Disease.* DNA Cell Biol, 2020. **39**(8): p. 1421-1430.

Chapter 7 - Appendices

Appendix 1 - PCR reaction setup, primer sequences and cycling conditions for all PCR based reactions.

Appendix table 1 – PCR reaction setup for each genotyping target of SVA targets and CRISPR modified screening PCRs.

PCR target	Component	Volume/reaction	Final concentration
***INPP5F* SVA F Full length**	Nuclease free water	8.8 µL	-
	5M Betaine	4 µL	1M
	10X KOD Hot Start buffer	2 µL	1X
	25mM $MgS0_4$	1.2 µL	1.5mM
	dNTPs (2mM each)	2 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	KOD hot start polymerase (1U/µL)	0.4 µL	0.02U/µL
	Template gDNA (5ng/µL)	1 µL	0.25ng/µl
***INPP5F* SVA F CT element**	Nuclease free water	12.1 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5U/µL)	0.1 µL	0.025U/µL
	Template gDNA (5ng/µL)	1.2 µL	0.3ng/µl
***INPP5F* SVA F poly-A**	Nuclease free water	10.8 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	3.2 µL	4mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.2 µL	0.2 µM
	Rv (3') primer (20µM)	0.2 µL	0.2 µM
	GoTaq Hot Start polymerase (5U/µL)	0.2 µL	0.05u/ µl
	Template gDNA (5ng/µL)	1 µL	0.25ng/ µl
***INPP5F* SVA D Full length**	Nuclease free water	8.9 µL	-
	5M Betaine	4 µL	1M
	10X KOD Hot Start buffer	2 µL	1X
	25mM $MgS0_4$	1.2 µL	1.5mM
	dNTPs (2mM each)	2 µL	0.2mM
	Fw (5') primer (20µM)	0.25 µL	0.25µM
	Rv (3') primer (20µM)	0.25 µL	0.25µM
	KOD hot start polymerase (1U/µL)	0.4 µL	0.02U/µL
	Template gDNA (5ng/µL)	1 µL	0.25ng/ µl
***INPP5F* SVA D CT element**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM

	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µl
	Template gDNA (5ng/µL)	1 µL	0.25ng/ µl
***INPP5F* SVA D poly-A**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µl
	Template gDNA (5ng/µL)	1 µL	0.25ng/ µl
***INPP5F* SVA D VNTR nest**	Nuclease free water	8.8 µL	-
	5M Betaine	4 µL	1M
	10X KOD Hot Start buffer	2 µL	1X
	25mM MgS04	1.2 µL	1.5mM
	dNTPs (2mM each)	2 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	KOD hot start polymerase (1U/µL)	0.4 µL	0.02U/µL
	Template from full length SVA PCR	1 µL	-
***LRRK2* SVA C Full length**	Nuclease free water	8.8 µL	-
	5M Betaine	4 µL	1M
	10X KOD Hot Start buffer	2 µL	1X
	25mM $MgS0_4$	1.2 µL	1.5mM
	dNTPs (2mM each)	2 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	KOD hot start polymerase (1U/µL)	0.4 µL	0.02U/µL
	Template gDNA (5ng/µL)	1 µL	0.25ng/ul
***LRRK2* SVA C CT element**	Nuclease free water	10.9 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	3.2 µL	4mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.2 µL	0.2µM
	Rv (3') primer (20µM)	0.2 µL	0.2µM
	GoTaq Hot Start polymerase (5U/µL)	0.1 µL	0.025u/ µL
	Template gDNA (5ng/µL)	1 µL	0.25ng/µL
***LRRK2* SVA C poly-A**	Nuclease free water	13.35 µL	-
	5X green GoTaq Flexi buffer	5 µL	1X
	25mM $MgCl_2$	4 µL	4mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.2 µL	0.16mM
	Rv (3') primer (20µM)	0.2 µL	0.16mM
	GoTaq Hot Start polymerase (5U/µL)	0.25 µL	0.05U/µL
	Template gDNA (5ng/µL)	1 µL	0.2ng/µL
***LRRK2* SVA C VNTR nest**	Nuclease free water	8.8 µL	-
	5M Betaine	4 µL	1M
	10X KOD Hot Start buffer	2 µL	1X
	25mM MgS04	1.2 µL	1.5mM
	dNTPs (2mM each)	2 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	KOD hot start polymerase (1U/µL)	0.4 µL	0.02U/µL
	Template from full length SVA PCR	1 µL	-

***INPP5F* SVA-F CRISPR KO PCR**	Nuclease free water	1.5µL	-
	2X xtreme buffer	5µL	1X
	dNTPs (2mM)	2µL	0.4mM
	Fw (5') primer (20µM)	0.15µL	0.3µM
	Rv (3') primer (20µM)	0.15µL	0.3µM
	KOD xtreme polymerase (1U/µL)	0.2µL	0.02U/µL
	Template gDNA (10ng/µL)	1µL	1ng/µL
***INPP5F* SVA-D CRISPR KO PCR**	Nuclease free water	1.5µL	-
	2X xtreme buffer	5µL	1X
	dNTPs (2mM)	2µL	0.4mM
	Fw (5') primer (20µM)	0.15µL	0.3µM
	Rv (3') primer (20µM)	0.15µL	0.3µM
	KOD xtreme polymerase (1U/µL)	0.2µL	0.02U/µL
	Template gDNA (10ng/µL)	1µL	1ng/µL
***LRRK2* SVA-C CRISPR KO PCR**	Nuclease free water	1.5µL	-
	2X xtreme buffer	5µL	1X
	dNTPs (2mM)	2µL	0.4mM
	Fw (5') primer (20µM)	0.15µL	0.3µM
	Rv (3') primer (20µM)	0.15µL	0.3µM
	KOD xtreme polymerase (1U/µL)	0.2µL	0.02U/µL
	Template gDNA (10ng/µL)	1µL	1ng/µL
***INPP5F* Isoform 1 RT-PCR for SH-SY5Y and HEK293 endogenous expression**	Nuclease free water	13.85 µL	-
	5X green GoTaq Flexi buffer	5 µL	1X
	25mM $MgCl_2$	4 µL	4mM
	dNTPs (10mM each)	1 µL	0.4mM
	Fw (5') primer (20µM)	0.2 µL	0.16µM
	Rv (3') primer (20µM)	0.2 µL	0.16µM
	GoTaq Hot Start polymerase (5u/ul)	0.25 µL	0.05u/ µL
	Template cDNA (1:10 dilute)	0.5 µL	-
***INPP5F* Isoform 2 RT-PCR for SH-SY5Y and HEK293 endogenous expression**	Nuclease free water	13.85 µL	-
	5X green GoTaq Flexi buffer	5 µL	1X
	25mM $MgCl_2$	4 µL	4mM
	dNTPs (10mM each)	1 µL	0.4mM
	Fw (5') primer (20µM)	0.2 µL	0.16µM
	Rv (3') primer (20µM)	0.2 µL	0.16µM
	GoTaq Hot Start polymerase (5u/ul)	0.25 µL	0.05u/ µL
	Template cDNA (1:10 dilute)	0.5 µL	-
***INPP5F* Isoform 3 RT-PCR for SH-SY5Y and HEK293 endogenous expression**	Nuclease free water	13.85 µL	-
	5X green GoTaq Flexi buffer	5 µL	1X
	25mM $MgCl_2$	4 µL	4mM
	dNTPs (10mM each)	1 µL	0.4mM
	Fw (5') primer (20µM)	0.2 µL	0.16µM
	Rv (3') primer (20µM)	0.2 µL	0.16µM
	GoTaq Hot Start polymerase (5u/ul)	0.25 µL	0.05u/ µL
	Template cDNA (1:10 dilute)	0.5 µL	-
***INPP5F* Isoform 1 RT-PCR (CRISPR experiments)**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***INPP5F* Isoform 2 RT-PCR**	Nuclease free water	12.3 µL	-

(CRISPR experiments)	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***INPP5F* Isoform 3 RT-PCR (CRISPR experiments)**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM $MgCl_2$	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***LRRK2* All isoforms RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***LRRK2* Isoform 1 RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***LRRK2* Isoform 2 RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***LRRK2* Isoform 3 RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***c-FOS* RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL

	Template cDNA (1:10 dilute)	1 µL	-
***TIAL1* All isoforms RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***BAG3* All isoforms RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***ACTB* RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-
***GAPDH* RT-PCR**	Nuclease free water	12.3 µL	-
	5X green GoTaq Flexi buffer	4 µL	1X
	25mM MgCl2	1.6 µL	2mM
	dNTPs (10mM each)	0.4 µL	0.2mM
	Fw (5') primer (20µM)	0.3 µL	0.3µM
	Rv (3') primer (20µM)	0.3 µL	0.3µM
	GoTaq Hot Start polymerase (5u/ul)	0.1 µL	0.025u/ µL
	Template cDNA (1:10 dilute)	1 µL	-

Appendix table 2 – Primer sequences used for all PCR reactions for SVA genotyping, CRISPR modification screening and RT-PCR/qPCR reactions.

PCR target	Forward and reverse primer sequences 5'>3'	PCR cycling conditions	Product size
***INPP5F* SVA F Full length**	F: CCTGGGCAAGAGAAGGAGAC R: CCTGGAGGCTGCTAATGTTG	95°C – 2m 95°C – 20s 64.9°C – 10s 70°C – 1m 4°C - ∞	1836bp
***INPP5F* SVA F CT element**	F: CCTGGGCAAGAGAAGGAGAC R: ATGGCAGCAGTACAGTCCA	95°C – 2m 95°C – 30s 64°C – 10s 72°C – 15s 72°C – 2m 4°C - ∞	117bp
***INPP5F* SVA F poly-A**	F: TCCCATGACCCTGCCAAATC R: CCTGGAGGCTGCTAATGTTG	95°C – 2m 95°C – 30s 64°C – 10s 72°C – 15s 72°C – 2m 4°C - ∞	135bp
***INPP5F* SVA D Full length**	F: TAAAATGTCATCCAGCTCTCCC R: GCCTCAAGTGAATAGCCC	95°C – 2m 95°C – 20s 63.5°C – 10s 70°C – 50s 4°C - ∞	2254bp
***INPP5F* SVA D CT element**	F: TAAAATGTCATCCAGCTCTCCC R: CAGGAGAATCAGGCAGGG	95°C – 2m 95°C – 30s 63°C – 30s 72°C – 15s 72°C – 2m 4°C - ∞	145bp
***INPP5F* SVA D poly-A**	F: ACTATTGTCCTGTGACCCTG R: GCCTCAAGTGAATAGCCC	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 15s 72°C – 2m 4°C - ∞	224bp
***INPP5F* SVA D VNTR nest**	F: CTACAACCTCCACCTCCCAG R: TCTACACAGACACGGCAACC	95°C – 2m 95°C – 20s 62°C – 10s 70°C – 20s 4°C - ∞	1327bp

***LRRK2* SVA C Full length**	F: GGAGATCTGGACATGGCTCCT R: GTGTCCCAGACACAATCCAGC	95°C – 2m 95°C – 20s 64.9°C – 10s 70°C – 1m 4°C - ∞	35X	2061bp
***LRRK2* SVA C CT element**	F: GAGAATCAGGCAGGGAGGTT R: GTGTCCCAGACACAATCCAGC	95°C – 2m 95°C – 30s 60.3°C – 30s 72°C – 30s 72°C – 2m 4°C - ∞	30X	489bp
***LRRK2* SVA C poly-A**	F: GGAGATCTGGACATGGCTCCT R: CCAGGGACACAAACACTACGG	95°C – 2m 95°C – 30s 59.3°C – 30s 72°C – 30s 72°C – 2m 4°C - ∞	30X	314bp
***LRRK2* SVA C VNTR nest**	F: GCCTGTTCTCAATGAGCTGC R: GAGTCTCGTTCACTCAGTGCT	95°C – 2m 95°C – 20s 66°C – 10s 70°C – 15s 4°C - ∞	20X	768bp
***INPP5F* SVA-F CRISPR KO PCR (Hap1 and SH-SY5Y)**	F: GCATTCTGCCCTCATGTTTC R: GGAAGCTGGAAGAGGGTGAT	94°C – 2m 98°C – 10s 64.5°C – 30s 68°C – 2m 4°C - ∞	30X	3249bp
***INPP5F* SVA-F CRISPR KO PCR (HEK293 only)**	F: CCAACATGACAGTCTCCTCAT R: GGTTCTAGCCAGTTCTGTGTG	94°C – 2m 98°C – 10s 64.5°C – 30s 68°C – 2m 4°C - ∞	30X	3926bp
***INPP5F* SVA-D CRISPR KO PCR**	F: TGGAGTGTTTGGTGATGGAA R: CATTCTCGCCTGGATTTGAT	94°C – 2m 98°C – 10s 59°C – 30s 68°C – 2m 4°C - ∞	30X	3510bp
***LRRK2* SVA-C CRISPR KO PCR**	F: GGACAGCTCTCATTTCTTGACT R: CCAGACACAATCCAGCTTTCC	94°C – 2m 98°C – 10s 60°C – 30s 68°C – 2m 4°C - ∞	30X	2690bp
***INPP5F* Isoform 1 RT-PCR for SH-SY5Y and HEK293 endogenous expression**	F: CGATATAACCCAAGACCGCG R: TGTTCCATGACCACTCTCGC	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 30s 72°C – 5m 4°C - ∞	30X	374bp

***INPP5F* Isoform 2 RT-PCR for SH-SY5Y and HEK293 endogenous expression**	F: ATTTTCCGACTGCCTGTTACG R: CGTGAGGTTTACTGCTCTTCC	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 30s 72°C – 5m 4°C - ∞	30X	595bp
***INPP5F* Isoform 3 RT-PCR for SH-SY5Y and HEK293 endogenous expression**	F: CAGAAAGCATTGGTGGGCAA R: AGGCCATCACTTCTTCCCAA	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 30s 72°C – 5m 4°C - ∞	30X	459bp
***INPP5F* Isoform 1 CRISPR RT-PCR (CRISPR experiments)**	F: GAGATTGGTACTCCAGATGTGG R: ATCTACACAGGTGGACTCC	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s 72°C – 2m 4°C - ∞	30X	150bp
***INPP5F* Isoform 2 CRISPR RT-PCR (CRISPR experiments)**	F: GGGATCATGTTTGGCTGATG R: CGTCTCTGTGAGTAGCATC	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s 72°C – 2m 4°C - ∞	40X	177bp
***INPP5F* Isoform 3 CRISPR RT-PCR (CRISPR experiments)**	F: TAGCTTGACCTATGACCTGACC R: GCCATCACTTCTTCCCAAGT	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s 72°C – 2m 4°C - ∞	40X	131bp
***LRRK2* All isoforms CRISPR RT-PCR**	F: TAATGTGGGGAGGATGTGGC R: ACACTTCCACAACAGGGCTA	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s 72°C – 2m 4°C - ∞	35X	186bp
***LRRK2* Isoform 1 CRISPR RT-PCR**	F: ATGAGTGGCAATGTCAGGTGT R: AATGTAAGCCTATGGAGCAAACA	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s 72°C – 2m 4°C - ∞	35X	110bp
***LRRK2* Isoform 2 CRISPR RT-PCR**	F: GTGGAGAGTTTCAGTGCCAG R: TGCTCCTCTTTCCAGACCCA	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s 72°C – 2m 4°C - ∞	40X	139bp

***LRRK2* Isoform 3 CRISPR RT-PCR**	F: GATGGTCTTGGAGGGTCACA R: TGCTTCTCTTGTGGGACTGA	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s (40X) 72°C – 2m 4°C - ∞	107bp
***c-FOS* RT-PCR**	F: TGTCAACGCGCAGGACTTCT R: GGGCTCCTGTCATGGTCTTC	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s (35X) 72°C – 2m 4°C - ∞	227bp
***TIAL1* All isoforms RT-PCR**	F: TGGTTGGGTGGTCGTCAAATC R: CAGACGCAATTCCTCCACAGT	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s (30X) 72°C – 2m 4°C - ∞	160bp
***BAG3* All isoforms RT-PCR**	F: CGACCAGGCTACATTCCCAT R: TCTGGCTGAGTGGTTTCTGG	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s (30X) 72°C – 2m 4°C - ∞	176bp
***ACTB* RT-PCR**	F: ACAGAGCCTCGCCTTTG R: CCTTGCACATGCCGGAG	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s (30X) 72°C – 2m 4°C - ∞	110bp
***GAPDH* RT-PCR**	F: AGATCCCTCCAAAATCAAGTGG R: GGCAGAGATGATGACCCTTTT	95°C – 2m 95°C – 30s 60°C – 30s 72°C – 20s (30X) 72°C – 2m 4°C - ∞	130bp

Appendix 2 – Sequencing primers for validation of constructs

Appendix table 3 – Primers used for the sequencing of generated constructs for pCR-blunt, pGL3, pSHM06 and EF1α-pSpCas9(BB)-2A-GFP based plasmids.

Vector	Primer name	Primer sequence
pCR-blunt	M13 F M13 R	5´GTAAAACGACGGCCAG 5´CAGGAAACAGCTATGAC
pGL3b and pGL3p	GL2 RV3	5'CTTTATGTTTTTGGCGTCTTCCA 5'CTAGCAAAATAGGCTGTCCC
pSHM06	pSHM06seqFw pSHM06seqRv	5'GAACCCACTGCTTACTGG 5'CACTGCGGACCAGTTATC
EF1α-pSpCas9(BB)-2A-GFP	U6 Fw	5'GAGGGCCTATTTCCCATGATTCC

Appendix 3 - Guide RNA sequences used for CRISPR of SVA elements

Appendix table 4 – All guide RNA (gRNA) sequences used to target *INPP5F* SVA-F and SVA-D and the *LRRK2* SVA-C elements. All targeting guides are represented 5' > 3' with the associated adjacent PAM sequence (NGG) in green at the 3' end of each guide. The PAM sequence was not included in the gRNA sequence that was cloned, and instead, was located within the genome immediately adjacent to the gRNA binding site. The non-targeting guides were used as controls within the CRISPR experiments and do not contain PAM sequences as these sequences were not predicted to bind to a specific genomic locus.

Non-targeting gRNA 1	ACGGAGGCTAAGCGTCGCAA
Non-targeting gRNA 2	TACTAACGCCGCTCCTACAG
***INPP5F* SVA-F gRNA Fw 1**	CTCTCATGTGGAGCCGAAGC **TGG**
***INPP5F* SVA-F gRNA Fw 2**	GACCTGCAATCCCGGCACTT **TGG**
***INPP5F* SVA-F gRNA Fw 3**	ACCTGTCAGACCTGCAATCC **CGG**
***INPP5F* SVA-F gRNA Rv 1**	CTGGAGGCTGCTAATGTTGC **AGG**
***INPP5F* SVA-F gRNA Rv 2**	ATGTTGCAGGTCAGCAGTCC **AGG**
***INPP5F* SVA-F gRNA Rv 3**	GATTAGATTTTGGTATCCTC **TGG**
***INPP5F* SVA-D gRNA Fw 1**	CAGTACTCTCCTAATTATGT **TGG**
***INPP5F* SVA-D gRNA Fw 2**	TGGATGCTGAGATTAAGGCT **GGG**
***INPP5F* SVA-D gRNA Rv 1**	CTCAACTCCCAATTGTATTA **AGG**
***INPP5F* SVA-D gRNA Rv 2**	CATAGGTCGGCTTTATTTTC **TGG**
***INPP5F* SVA-D gRNA Rv 3**	ACCACGTAAGGAACGGACAG **AGG**
***LRRK2* SVA-D gRNA Fw 1**	GTGCCTCAACATTAGTTCTG **AGG**
***LRRK2* SVA-D gRNA Fw 2**	TAACCTCAGAACTAATGTTG **AGG**
***LRRK2* SVA-D gRNA Fw 3**	AGTGGATCAATACATTGTGT **TGG**
***LRRK2* SVA-D gRNA Rv 1**	GTTTCAGATGTCATCTTGAT **AGG**
***LRRK2* SVA-D gRNA Rv 2**	AATCTTCTTATATGGTTTGA **AGG**
***LRRK2* SVA-D gRNA Rv 3**	CTAGAACTTGTACAGAATAA **AGG**

Appendix 4 – Scripts used for RC-Seq and WGS analysis

All scripts used for RC-Seq and WGS analysis available upon request and include the following information:

- Adapter trimming
- Alignment of BAM files to the reference genome (hg19)
- Running TEBreak
- Splitting of PICKLE files (optional)
- Resolve scripts
- Final filtering