Structural variation in Parkinson's disease:

focusing on the role of Transposable elements in disease predisposition

and pathogenesis

Thesis submitted in accordance with the requirements

of the University of Liverpool for the degree of Doctor in Philosophy by

Kimberley Jayne Billingsley

May 2019

Acknowledgements

First, I would like to express my utmost gratitude to my PhD supervisors, Prof. John Quinn, Prof. Sulev Kõks, Dr. Jill Bubb and Dr. Andrew Singleton. I would also like to thank Dr. Raph Gibbs and Dr. Mike Nalls for their constant guidance throughout. I could not have wished for a better group of mentors and I am extremely grateful for their support and encouragement over the last three years.

I am particularly grateful for the opportunity I was given to work on projects at NIH, especially given how difficult it was to get me there. I have been constantly surrounded by extremely hardworking and talented scientists, which has been an incredibly inspirational experience. Getting the chance to work at LNG and be part of such a motivating group has been unbelievable!

Not only have I been given such an amazing opportunity, but I've been gifted incredible friends in the process. Especially two very special ladies that I would not of managed the last three years without. Dre, thank you for everything you did to support me through my time back in Liverpool, I am lucky to have such a kind and caring friend. Sara, you've been the best friend I could have asked for in DC and I'm so grateful for how much you've looked after me and let me play Lampito, being part of the weirdos is the best.

I would also like to thank my Quinn lab family past and present, especially Em, Jack, Oly, Ben and Veri for making the PhD so fun and making me feel so loved. Another big thank you to the mathematical wizard for putting up with my rants and filling a lot of my evenings with Beyoncé, keep slaying B. Finally, a big thanks to Chloe, Laura, John and Katie for everything you've done to not let me become a full-blown crazy scientist. Last but not least to a special little brother… thank you for looking after me always.

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder with a complex aetiology including genetic risk factors, environmental exposure and aging. Recent genome wide association studies have been successful at identifying genetic variation that confers a risk for PD, yet despite this it is predicted that the large majority of the genetic attribution to the disease is still unknown. It is also noted that much of the identified risk loci lie within poorly annotated regions of the genome such as those containing repetitive sequences and transposable elements (TE)s, highlighting the importance of further investigation into such regions. Despite many reports that associate TE insertions with PD no study has comprehensively analysed the role of these elements in the disease.

The work presented in this thesis sought to ask three main questions; first, are TE overrepresented at PD risk loci using a haplotype block based genome-wide analysis, second are non-reference TE associated with risk of PD using a newly developed TE detection tool and PD WGS data; and third, are TE differentially regulated in the blood or skin of individuals with PD. This work leveraged genetic and expression datasets to comprehensively address the role of TE in PD. Along with identifying that specific TE are overrepresented at PD risk loci we also show that in the blood specific repetitive elements are differentially expression in PD. Most significantly we characterized known non-reference TE presence/absence polymorphisms in collaboration with the International Parkinson's Disease Genomic Consortium (IPDGC) in PD whole genome sequencing data (WGS) from the Parkinson's Progression Markers Initiative (PPMI) cohort using the TE detection tool MELT. We identify that TE insertions are a heritable and common form of genetic variation that lie within potentially important functional domains of the genome. Not only do many non-reference TE map to PD risk loci, but from our initial study we have identified that non-reference TE's are in moderate linkage disequilibrium with PD risk variants, and thus a candidate causal variant that warrant further study at these loci. In summary, TE insertions are a major source and often overlooked form of genetic variation in the human genome. Collectively the research presented in this thesis suggests that not only could integrating TE variants be a valuable and critical step forward for furthering our understanding of existing risk PD variants, but it could also be valuable for establishing new risk regions.

Publications

1. Billingsley KJ, Manca M, Gianfrancesco O, Collier DA, Sharp H, Bubb VJ, et al. Regulatory characterisation of the schizophrenia-associated CACNA1C proximal promoter and the potential role for the transcription factor EZH2 in schizophrenia aetiology. Schizophr Res. 2018;199: 168–175. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6179964/

2. Billingsley KJ, Bandres-Ciga S, Saez-Atienzar S, Singleton AB. Genetic risk factors in Parkinson's disease. Cell Tissue Res. 2018;373: 9–20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6201690/

3. Bandres-Ciga S, Saez-Atienzar S, Bonet-Ponce L, Billingsley K, Vitale D, Blauwendraat C, et al. The endocytic membrane trafficking pathway plays a major role in the risk of Parkinson's disease [Internet]. Movement Disorders. 2019. pp. 460–468. doi:10.1002/mds.27614 https://www.ncbi.nlm.nih.gov/pubmed/30675927

4. Billingsley KJ, Barbosa I, Bandrés-Ciga S, Quinn JP, Bubb VJ, Deshpande C, Botia JA, Reynol RH, Zhang D, Simpson MA, Blauwendraat C, Gan-Or Z, Gibbs RJ, Nalls MA, Singleton AB, International Parkinson's Disease Genomics Consortium (IPDGC), Ryten M, Koks S. Mitochondria function associated genes contribute to Parkinson's Disease risk and later age at onset. npj Parkinson's Disease. 2019. (Accepted) https://www.biorxiv.org/content/10.1101/475111v1

5. Billingsley KJ , Vitale D, Savage AL, Middlehurst B, Bubb VJ, Kõks S, Nalls MA , Singleton AB, Quinn J, Genome-wide analysis of reference SINE-VNTR-Alu (SVA) elements identifies an overrepresentation at Parkinson's Disease risk loci and overlap with regulatory sites. Parkinsonism & Related Disorders (In review)

Contents

Chapter 1 General introduction

1.1. Overview:

Extensive genome-wide association and meta-analysis initiatives have identified ninety loci that confer a risk to Parkinson's disease (PD)[1]. Despite this success, taken collectively it is predicted that these loci only explain around one third of the heritable component of PD. Therefore, the majority of the common heritability of PD is unknown. Even for the known risk loci the true causal variant is yet to be established and little is understood regarding the underlying cellular and molecular processes through which they act.

Previous analyses have focused on the contribution of single nucleotide polymorphisms (SNP)s to disease risk, however structural variation (SV) is also a huge source and overlooked form of genetic variation in the genome. Comprehensively analysing SV is extremely challenging as the majority of these variants are repetitive and thus very difficult to uniquely map with short-read sequencing approaches. Nonetheless with recent advances in both sequencing methods and bioinformatic programs it is becoming increasingly apparent that SVs contribute to gene regulation and are implicated in disease. This form of genetic variation varies greatly in size and is broadly split into classes: deletions, translocations, inversions, tandem duplications and transposable elements (TE)s. The latter will be the focus of this thesis.

TE derived sequences constitute the majority of the human genome and despite the fact that many of these insertions are ancient, a subset is still capable of mobilization, including the subfamilies; *Alu*, LINE-1 and SVA. Relevant to PD specifically, TEs have already been associated with rare instances of Mendelian PD which are caused by copy number variants (CNVs)[2]. In addition, an SVA insertion into the *TAF1* gene is causative of X-linked dystonia-parkinsonism (XDP)[3]. Despite this, the role of TE in sporadic PD is yet to be addressed. Therefore, in this thesis we leveraged new TE detection and expression quantification tools and PD whole-genome sequencing (WGS) and gene expression datasets and comprehensively analysed the role of TE in sporadic PD.

First, we focused on an already characterized SVA that is in the reference genome and upstream of the PD associated *PARK7* gene. Previous work had demonstrated that two repetitive domains of the SVA were variable in copy number [4,5], giving rise to four distinct alleles. There was also an alternative transcriptional start site adjacent to the SVA which was active in multiple brain regions. Proxy SNPs were generated for three of the four alleles, so utilizing these we explored whether the SVA could be imputed. As SVA are large GC rich repetitive elements PCR genotyping is cost and time intensive, therefore bioinformatically calling the SVA would be beneficial and give power to subsequent association analyses. SVAs in the reference genome have also been shown to drive tissue and allele specific gene expression *in vitro* and *in vivo*[4]. In light of this we also hypothesized that the SVA may act as an expression quantitative trait loci (eQTL).

We identified that the reference *PARK7*-SVA was imputable and utilizing a large WGS dataset demonstrated that it is a common form of SV that is currently uncaptured at the *PARK7* locus (*see chapter 2).* We also ran association analyses using PD WGS but did not identify any significant association between the SVA and risk of PD. EQTL analysis was also applied to brain-specific expression datasets. Although we confirmed that a longer brain specific PARK7 transcript was expressed from the transcriptional start site adjacent to the SVA, we did not identify that this expression was associated with the genotype of the SVA.

Given that SVAs can influence gene expression and have previously been reported to be enriched at Mendelian PD genes, we next performed a genome-wide haplotype-block based enrichment analysis. We also used ENCODE data to annotate whether reference SVA overlapped regulatory sites in the genome. Our goal was to gain insight into the genome-wide regulatory potential of reference SVA and establish whether they were enriched at regions of the genome that contributed to PD risk. We demonstrated that due to their genic nature (*see chapter 3*) reference SVA are overrepresented at current PD risk loci and commonly overlap regulatory sites. Therefore, reference SVA could be involved in disease mechanism at known risk loci.

LINE1 and *Alu* TEs are also known to alter gene expression, so we addressed whether they too were enriched at PD risk loci. Specifically, *Alu* elements have been reported to be enriched within genes that are associated with mitochondria function, which is a known pathway involved in the pathophysiology of PD[6]. Haplotype block-based analyses did not identify enrichment of LINE1 and *Alu* at PD risk loci. However, in support of the literature, *Alu* density positively correlated with mitochondria function gene density (*see chapter 4A*).

Currently it is not possible to address how *Alu* variation contributes to PD risk through the mitochondria function pathway as this variation is not catalogued. So Instead we utilized PD GWAS datasets to gain insight into the overall contribution of the genetic variation within mitochondria function associated genes to risk of sporadic PD.

First, using polygenic risk score analysis we identified that variation within mitochondria function associated genes was significantly associated with risk of PD and later age at onset. We also identified that a proportion of the missing heritability of the PD can be explained by common variation within genes implicated in mitochondrial function. Finally, to identify possible functional genomic associations we implemented Mendelian randomisation which identified that fourteen mitochondrial function associated genes showed functional consequence associated with PD risk. Further analysis suggested that the 14 identified genes are not only involved in mitophagy but implicate new mitochondrial processes (*see Chapter 4B)*. Our data suggests that therapeutics targeting mitochondrial bioenergetics and proteostasis pathways distinct from mitophagy could be beneficial to treating the early stage of PD.

Beyond simple enrichment analysis we next characterized known TE presence/absence polymorphisms in collaboration with the International Parkinson's Disease Genomic Consortium (IPDGC) in PD WGS from the Parkinson's Progression Markers Initiative (PPMI) cohort using the TE detection tool MELT. We show that non-reference TE insertions are a heritable and common form of genetic variation that lie within potentially important functional domains of the genome. Not only do many lie within PD risk regions, but from our initial study multiple risk variants are tagged by TEs. Most significantly a non-reference SVA is in moderate LD with a known PD risk variant and is most commonly inherited with the risk allele (*see chapter 5*).

Finally utilizing RepEnrich a TE expression quantification tool we explored differential expression of not only TE but all repetitive elements (RE) in the blood and skin of PD patients. Despite reports that associate RE expression with PD no study has comprehensively analysed the role of these elements in the disease. Analysis of RNA-sequencing data of 12 PD patients and 12 healthy controls identified tissue-specific expression differences and more significantly, differential expression of four satellite elements; two simple satellite III (repName = CATTC_n and _GAATG_n) a high-copy satellite II (HSATII) and a centromeric satellite (ALR_Alpha) in the blood of PD patients. In support of the growing body of recent evidence associating REs with neurodegenerative disease, this study highlights the potential importance of characterization of RE expression in such diseases (*see chapter 6*). For a general overview of the thesis see Figure 1.1.

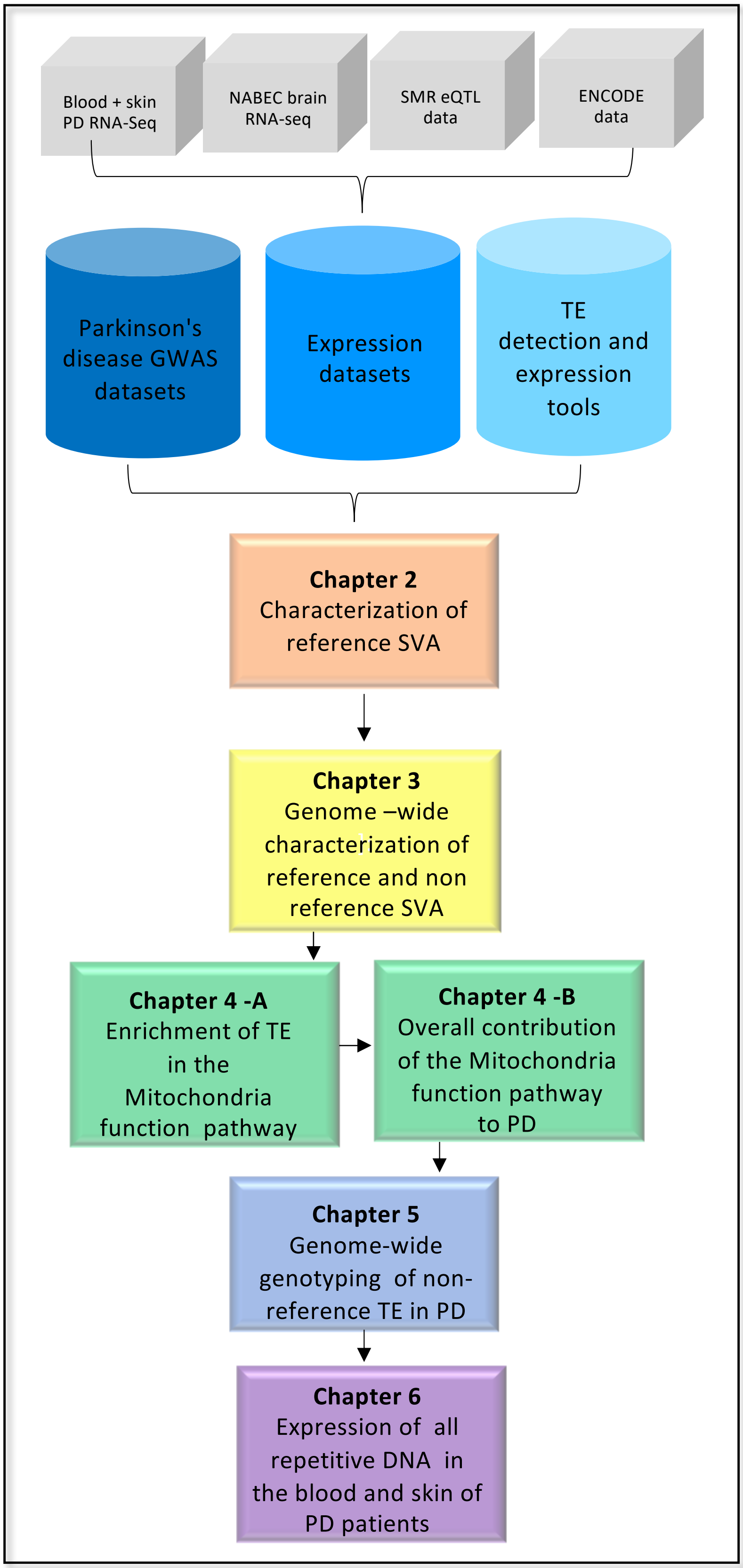

Blood + skin PD RNA-Seq
NABEC brain RNA-seq
SMR eQTL data
ENCODE data
Parkinson's disease GWAS datasets
Expression datasets
TE detection and expression tools
Chapter 2
Characterization of reference SVA
Chapter 3
Genome –wide characterization of reference and non reference SVA
Chapter 4 -A
Enrichment of TE in the Mitochondria function pathway
Chapter 4 -B
Overall contribution of the Mitochondria function pathway to PD
Chapter 5
Genome-wide genotyping of non-reference TE in PD
Chapter 6
Expression of all repetitive DNA in the blood and skin of PD patients

Figure 1.1. General thesis outline

1.2. Parkinson's disease: A brief history

Due to an increasing population of aged individuals the prevalence of neurodegenerative disease is predicted to drastically rise. PD is the second most common neurodegenerative disease after Alzheimer's disease (AD). The prevalence of PD is age-dependent, in which about 1% of the global population at 65 years of age and over, and about 4-5% of individuals at 85 years of age and over affected [7] . PD is also the most frequent movement disorder and the prevalence of the disease in some age groups is likely to double by 2030 [8–10] . This therefore represents a medical and economic challenge for modern society with there being no available treatment that can stop or reverse the neurodegenerative process of the disease.

PD was considered the typical example of a non-genetic disease until only two decades ago. This view was supported by the first published cross-sectional series of twin studies [11,12] and epidemiological studies which linked PD to environmental causes such as viral infection or neurotoxins. This was consistent with, a pandemic influenza virus was strongly associated with post-encephalitic parkinsonism, seen by many as evidence that viral infection may be a major cause of PD. In addition, the observation that drug users exposed to MPTP developed parkinsonian-like features strengthened the notion that PD was an environmental disease [13].

However, evidence supporting a genetic basis of PD aetiology came when molecular genetics were implemented to dissect the underlying genetic cause of several families in which PD was inherited in an autosomal dominant or recessive manner. The first forms of monogenic PD were caused by highly penetrant mutations affecting multiple members per family. Although the insights gained from such mutations were seen as very valuable, these families were extremely rare.

In 1997 the first genetic association with PD was identified with mutations in *SNCA* (encoding α-synuclein)[14]. This first key finding to suggest a heritable component of PD was followed by the identification of additional rare recessive forms associated with early-onset disease. However, all these known monogenic forms combined only explain about 30% of monogenic and 3–5% of genetically complex or “sporadic” cases.

1.3. Parkinson’s disease is a complex genetic disease

PD fits within the wide range of complex polygenic disorders influenced by both genetic and environmental factors. While only a small minority of PD cases are monogenic in nature, sometimes exhibiting variable penetrance, the vast majority of cases are considered to be genetically complex, presenting with multiple clinical presentations. It has been assumed that PD aetiology lies on a continuum, ranging from the monogenic inheritance observed in monogenic disease to complex inheritance associated with an interplay of genetic risk and likely environmental influence [15].

The genetic attribution of PD is often ascribed to two non-mutually exclusive ideas: the common disease common variant (CDCV) hypothesis and the common disease rare variant (CDRV) hypothesis (also known as the multiple rare variant hypothesis). The CDCV hypothesis would accept that the genetic basis of PD is a result of a contributing common variants that each exert relatively small effects on disease risk but that cumulatively confer substantial risk. On the contrary, the CDRV hypothesis considers that a contributing risk component for complex disease will be rare genetic variants of small or moderate/large effect where highly functional, deleterious alleles might exist [16]. This phenomenon may be particularly pronounced in late-onset diseases such as genetically complex PD, where selective pressures are not as profound (described in Figure 1.2).

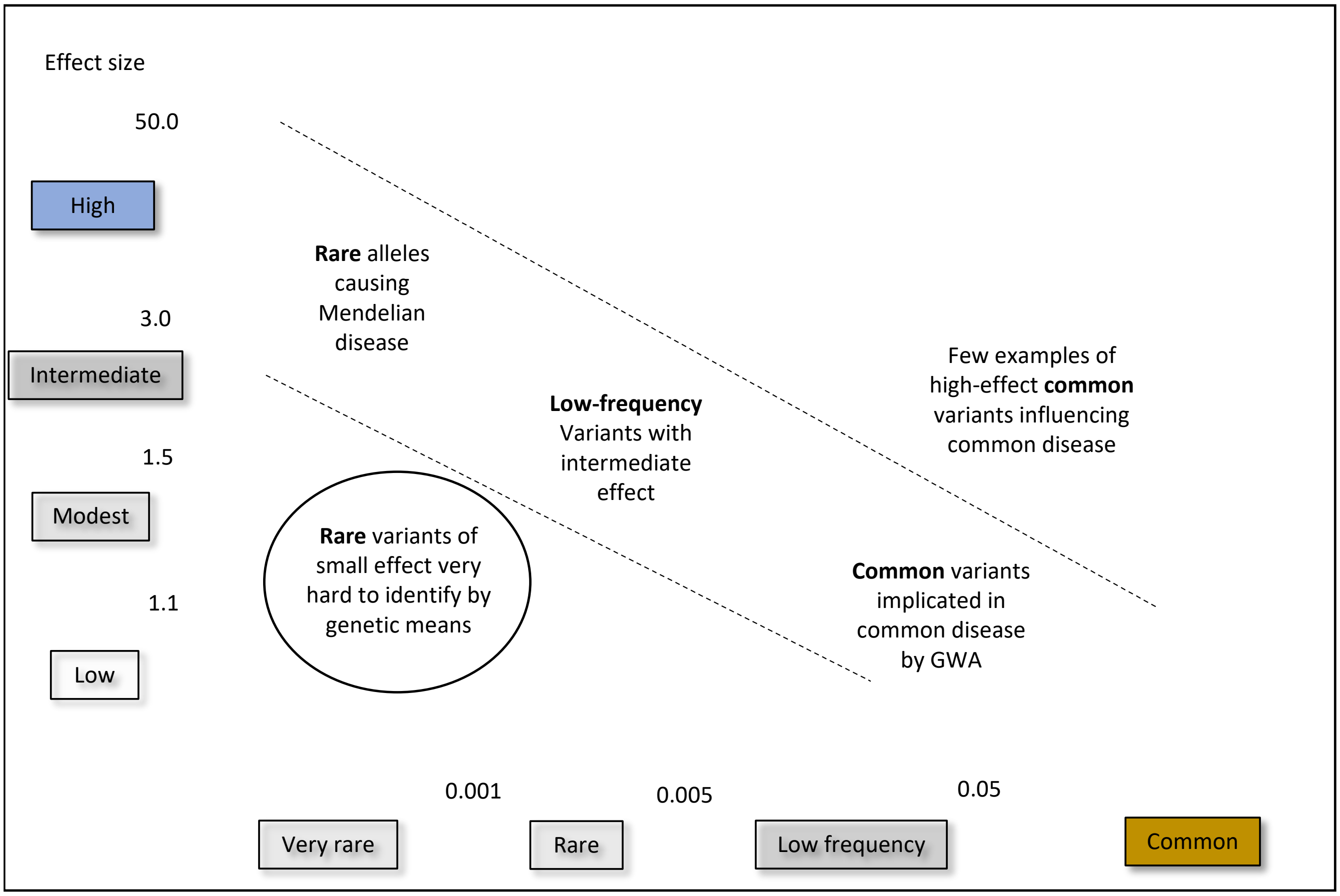

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency Variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common

Figure 1.2 The genetic architecture of complex diseases. (Adapted from *Manolio et al., 2009).*

The development and improvement of technological approaches continues to challenge both paradigms by increasing the identification of very rare causative mutations underlying monogenic forms of disease through whole-genome and whole-exome sequencing (WES) approaches and of common variants with small effects contributing to genetically complex, late-onset disorders through genome-wide association studies (GWAS).

1.4. Currently known common genetic risk factors in PD

Both candidate gene association studies and GWAS continuously validate that the most statistically significant signals associated with PD are common variants located close to *SNCA, LRRK2*, and *MAPT* as well as low-frequency coding variants in *GBA*. These genes are discussed below in more detail.

1.4.1. SNCA

Following the discovery of *SNCA* mutations which were causative of rare monogenic forms of PD, α-synuclein protein aggregates were identified as a major component of Lewy bodies, (a primary pathological hallmark of PD[17]. This finding showed overlap between the pathogenesis of monogenic and genetically complex PD. Interestingly, in the context of risk for PD, *SNCA* is pleomorphic i.e. both rare mutations and common variation at this locus alter risk for disease.

At one end of the risk spectrum deleterious point mutations in and multiplications of this gene cause a severe early-onset form of PD that follows an autosomal dominant pattern of inheritance [18,19]. Copy number variation caused by multiplications of PD associated loci will be detailed in a later section. At the other end of the spectrum, it has been repeatedly reported that non-coding variation within this locus confers risk and predisposes to sporadic PD[1,20]. The first indication that the *SNCA* locus contained risk variants for sporadic PD came from the association between the REP1 (a polymorphic dinucleotide repeat sequence) variant in the promoter region of the gene and PD [21]. Further GWAS signals at *SNCA* showed an association with PD from intron 4 to after the 3′ UTR region [22]. Since then *SNCA* has been overwhelmingly established in GWAS, identifying additional signals and providing further insights about the genetic risk at this locus [20,23–25]. Current research suggests between two to five semi-independent association signals accounting for heritable risk at this locus [1].

1.4.2. LRRK2

Genetic variants in *LRRK2* account for the majority of all known heritable PD. The most common pathogenic variant in *LRRK2*, p.G2019S, is responsible for about 1% of patients with sporadic PD and 4% of patients with a family history of PD. *LRRK2* p.G2019S exhibits incomplete and age-associated differences in penetrance. Collaborative studies have shown that the risk of PD for individuals who inherit the *LRRK2* p.G2019S variant varies from 28% at 59 years to 51% at 60 years reaching up to 75% at 80 years of age.[26] *LRRK2* p.G2019S frequency varies depending on ethnic background, with the highest frequencies among North African Arab and Jewish populations [27,28] . A recent study reported a prevalence of 0.71% among Caucasians, 0.07% among Asians, and 30.2% among individuals of Arab origin with PD[29] . It is thought that *LRRK2* p.G2019S originated from a common founder in the North of Africa and spread globally with the Ashkenazi Diaspora [30].

Similarly, to *SNCA, LRRK2* is also a pleomorphic locus. In addition to several disease-causing mutations characterized by segregation with PD in large families and by functional studies, common variability has been repeatedly reported as a risk factor for PD. There are two lines of evidence supporting the idea that *LRRK2* contains risk-modifying variants. Firstly, it has been widely reported that two polymorphic *LRRK2* variants, p.G2385R and p.R1628P, are associated with a 2-fold risk of PD in Asian populations with a frequency of approximately 6% in cases [31–33]. Secondly, GWAS implicate non-coding variants proximal to *LRRK2* with around a 1.2 fold increase risk for PD. This suggest that this risk may be mediated by an alteration in expression or splicing, although the precise mechanism involved is yet to be established.

1.4.3. GBA

Homozygous mutations observed in *GBA*, are causative of Gaucher's disease, a lysosomal storage disorder with an autosomal recessive pattern of inheritance. A clinical observation in relatives of patients affected with Gaucher's disease identified that first and second degree family members manifested an increased incidence of PD, pointing to *GBA* as a risk factor for PD [34,35] . Further a multicentre study conducted by *Sidransky* et al. showed that heterozygous *GBA* mutations are the largest genetic risk factor for developing PD, enhancing an individual's risk by approximately 5-fold and highlighting the importance of the lysosomal pathway in the pathogenesis of PD[36]. As *GBA* variants can appear with frequencies <5%, it was initially filtered out of GWAS analyses, (highlighting a possible limitation of GWAS). Only following a candidate gene approach was GWAS able to confirm its clear significance as a PD risk factor. *GBA* encodes a lysosomal glucocerebrosidase enzyme responsible for the synthesis of ceramide [37]. A reduced expression level of *GBA* in addition to a significant decrease in the enzyme activity has been observed in PD patients carrying heterozygous mutations in *GBA.* Further, decreased rates in glucocerebrosidase activity have been found in the substantia nigra of PD patients in comparison with other brain regions [37,38].

Multiple reports suggest that up to 10% of PD patients carry a *GBA* mutation and it has been reported that the penetrance and lifetime risk of developing PD for these *GBA* carriers varies in an age-dependent fashion from 20% at 70 years to 30% at 80 years [39] . Therefore, *GBA* mutations are a substantial common risk factor for PD. However, the frequency of *GBA* variants varies according to different ethnicities, being particularly frequent among Ashkenazi Jewish subjects. For example, the most common *GBA* variant (p.N370S) is present among those of European, American, or Middle Eastern origin but is not typically seen in Chinese or Japanese populations [40].

1.4.4. MAPT

Mutations that are dominantly inherited in *MAPT* were first associated with forms of Frontotemporal dementia and Parkinsonism linked to chromosome 17[41]. *MAPT* mutations and tau pathology have been predominantly associated with dementias, therefore making the association observed between PD and the locus harbouring *MAPT* of particular interest. Along with the monogenic forms, several studies have deeply studied *MAPT* for variability that may infer risk for PD. There are two major haplotypes at the *MAPT* locus: the directly oriented haplotype H1 and the H2 haplotype, which has an inverted chromosome sequence [42]. The H2 haplotype is present in approximately 20% of the European population and shows very limited genetic variability contrary to the H1 haplotype [43]. To note, this locus represents the largest area of linkage disequilibrium (LD) known in the human genome, [44] which makes identifying the true causal variation at this locus difficult.

Within the past decade a growing body of evidence has suggested that the *MAPT* H1 and its sub-haplotype H1c are associated with increased risk for PD. It is suggested that haplotype-specific differences in expression and potentially alternative splicing of *MAPT* transcripts affect cellular functions at different levels, which eventually increases susceptibility to PD[45]. Further, *MAPT* is one of the top GWAS hits, although it seems to be limited to Europeans and not Asian populations. However, it is a challenge to determine how common variants at the *MAPT* locus increase the risk for PD. This is due to the locus harbouring many genes and the extended LD means that it is difficult to localize the genetic signal. Thus, while *MAPT* is a strong candidate as the effector gene at this locus, it cannot be confirmed that this is the true biological mediator of risk. Therefore, dissecting this locus and identifying the true causal variant is still an ongoing problem.

1.5. Identifying risk loci from GWAS: Where we are at

The underlying idea of a GWAS is based on the CDCV paradigm with the objective of detecting common variants (MAF > 1%) in ethnically homogeneous populations. While critics suggest a genome-wide fishing expedition, the overwhelming majority of the genetics community would argue that the results gathered from such studies have marked a significant advancement from candidate gene studies and have driven the new era and concept of PD genetics. These advances are based on the premise that risk variants may occur within haplotype blocks shared with common variants through LD. Since common variants can be tagged through genotyping marker arrays, risk variants in linkage disequilibrium should manifest an association by proxy with tagged common variants and ultimately with PD. By increasing sample size and genotype marker frequency, lower-risk variants with a lower population-attributable risk can be detected. As the field has progressed, several GWAS [22,24,46–50] and meta-analyses [1,20,25] have been key at identifying common risk variability associated with sporadic PD. At present following the most recent meta-analysis involving over one million individuals 90 risk variants have been established, across 78 loci [1] . The current known PD risk loci are shown the Manhattan plot in (Figure 1.3).

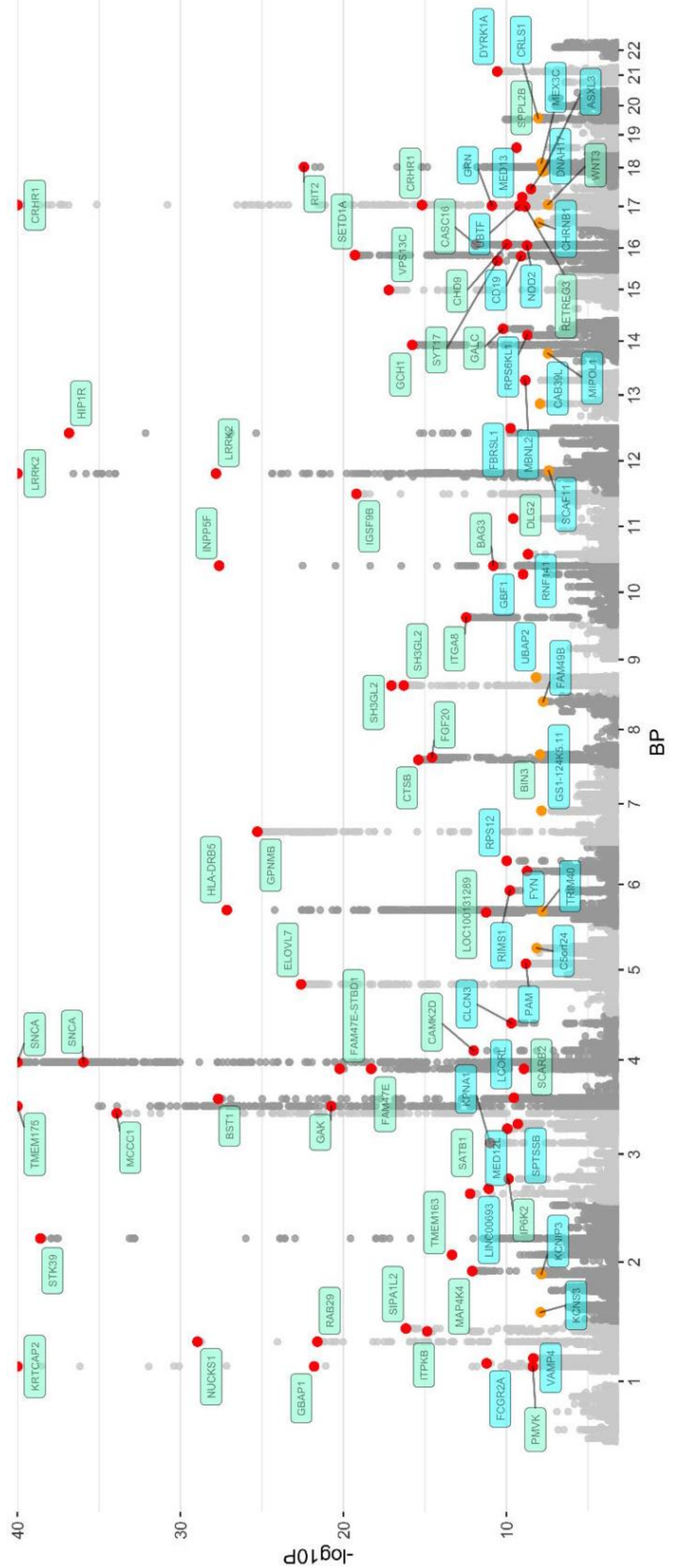


Figure 1.3 Manhattan plot. The nearest gene to each of the 90 significant variants. –log10P values were capped at Variant points are colour coded red and orange, with orange representing significant variants at P 5E-08 and 5E-9 and red representing significant variants at P < 5E-9. The X axis represents the base pair position of variants from smallest to largest per chromosome (1-22) (*Nalls et al 2019*).

Despite the considerable success of PD GWAS studies, only a relatively small proportion of the heritable component of PD can be explained from the ninety identified risk loci1. Genetic heritability is a measure of the extent to which genetics is involved in a given trait or phenotype[51]. Twin studies are a frequently used method for studying the heritability of a given phenotype, taking advantage of the nearly 100% shared genetic data of monozygotic (MZ) twin pairs and nearly 50% shared genetics data for dizygotic (DZ) twin pairs. Both are assumed to share the same environment and assuming that this unique environment equally contributes to the development of the phenotype of interest in MZ and DZ pairs, it is possible to estimate the variance contributed by genetic effects by comparing the phenotypic correlation in MZ and DZ pairs. The resulting estimate is a measure of "broad sense" heritability[52].

Historically, using twin studies to study the heritability of apparently sporadic PD has been problematic, with several early twin studies failing to show differences in concordance rates, as they were generally limited by small sample size and cross-sectional design[**53**]. As previously mentioned, monogenic familial PD is often early onset and it has been suggested that this form of PD has a dissimilar aetiology and greater heritability component than the later onset forms of PD, which appear to be sporadic [54,55]. Given this factor a more recent PD twin study adjusted for age and showed significant rates of concordance for MZ pairs with early age at onset of PD, compared to a near lack of concordance for individuals with later onset PD. This strongly supports the role of genetics in early-onset PD; however again this study was limited by a small sample size[56].

Narrow-sense heritability is another estimate used to understand the heritability of a given phenotype and captures the proportion of genetic variation that is due to additive genetic values. For PD specifically, utilizing existing well powered PD GWAS datasets narrow-heritability estimates have identified that the heritable component of PD due to common genetic variability is estimated to be around 22%[1]. This means that for every individual with PD around 22% of the disease can be explained by shared common genetic variability. However, if this is calculated using only the 90 known risk loci (opposed to the former that was calculated with all the GWAS variants) then only a fraction of this 22% can be explained1. Therefore, there are many loci that contribute to PD risk that are still unknown and this is termed in the literature the “missing heritability” of PD[57]. It is highly likely that an important part of the “missing heritability” exists in rare variants with low or high degrees of risk and structural variation, both of which are difficult to detect using traditional GWAS methods. In addition to not capturing the latter variation, calculating heritability it this way is also limited due to the fact it does not take into consideration dominant or epistatic effects. Despite the limitations of heritability estimates, they illustrate that genetic variation significantly contributes to PD aetiology and most importantly they emphasize that there are still many genetic risk factors that are yet to be identified.

1.6. Structural Variation

The majority of PD risk variants lie within non-coding regions of the genome with no known function and taken collectively they can only account for around 30% of the heritable component of the disease. Therefore, it is predicted that a portion of this "missing heritability" will likely be within regions that are currently not covered by such GWAS, such as structural variants (SV).

The human-genome can differ by a single nucleotide polymorphism (SNP) to large chromosomal events. Following huge advancements in sequencing and bioinformatic techniques it is now apparent that human genomes differ more as a consequence of structural variation than as a result of a point mutation [52–57]. Originally SV's were defined as deletions, insertions and inversions greater than 1 kb [58]. However following the recent advances in SV discovery technologies the detection of smaller events is possible. As a result, the operational range of SVs and copy number variants (CNVs) has widened to include much smaller events (e.g. >50 bp in length). Evidently SVs greatly vary in size and are broadly split into classes of structural variation: deletions, translocations, inversions, TEs, tandem duplications and novel insertions (Figure 1.3).

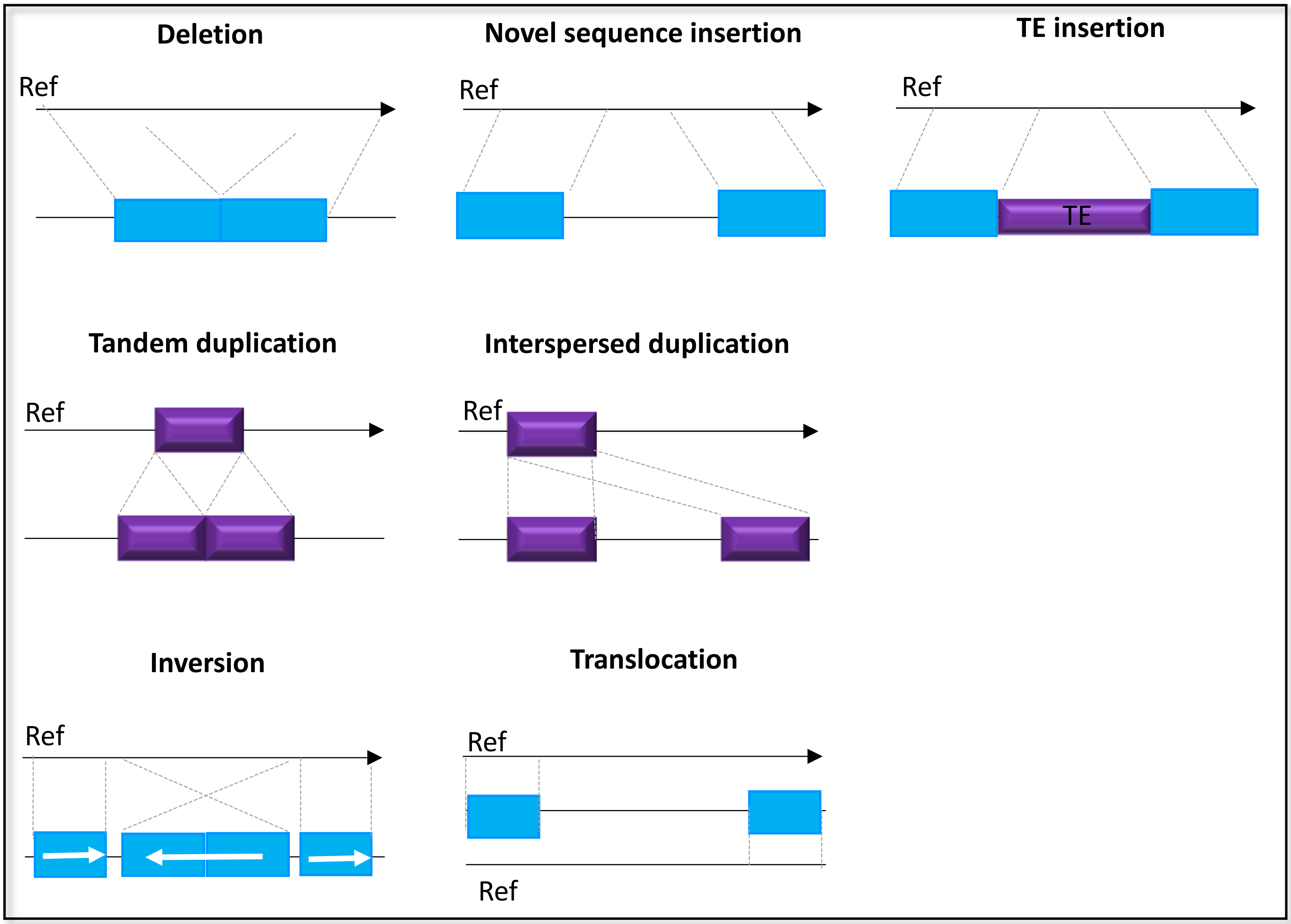


Figure 1.4.The five classes of structural variation in the genome. Structural variation refers to genomic alterations that are larger than 1 kb in length, but advances in discovery techniques have led to the detection of smaller events. Currently, >50 bp is used as a standard cut off between indels and copy number variants (CNVs). The schematic shows deletions, novel sequence insertions, mobile-element insertions, tandem and interspersed segmental duplications, inversions and translocations in a test genome (lower line) when compared with the reference genome (Adapted from Alkanet al).

Although SVs are a huge source of genetic variation in the genome there are many challenges in characterising this form of variation in standard genetic analysis. Therefore, the successes in identifying SVs associated with disease is currently mainly limited to rare Mendelian forms of disease.

There are now two distinct models proposed regarding the way in which SVs are associated with disease:

1) Large variants (typically gains and loss of several hundred kb) that are rare in the population (MAF >1%) collectively account for a significant amount of the heritable component of that specific disease. This form of SV consequence has been observed in developmental diseases (such as autism [59] and intellectual disability [59,60]) and other genetic disorders [61,62].
2) Copy number variants (CNV) of multicopy gene families that contribute to risk of disease. This has been reported in phenotypes associated to immune gene function [63,64]. CNVs can influence phenotype through several mechanisms such as; influence gene expression through simple gene dosage effect, insertions or deletions of regulatory regions and alterations of chromatin architecture [65]. Significantly this is also observed in rare forms of Mendelian PD.

1.7. Structural variation associated with Parkinson’s Disease

There are several genes associated with Mendelian PD including; SNCA, PARK2, PINK1 and PARK7, have also been identified to contain CNVs that are causative of the disease in rare cases. The first report of a PD causing CNV was in 2003 with the significant discovery of a genomic multiplication at the *SNCA* locus [19]. The genomic triplication at this locus causes a rapidly progressive form of autosomal dominant PD[19]. However duplications of *SNCA* resemble sporadic PD phenotypically, with late age at onset and slower disease progression [66]. Recent studies have indicated that there is a reduced penetrance of disease in duplication carriers after observing several asymptomatic duplication carriers over 70 years of age without any signs of PD[66].

In addition, CNVs are also reported to cause PD in a recessive manner, which is observed with multiplications of PARK2, PINK1 and PARK7. PARK2 is one of the largest known genes in the genome (1.4 Mb genomic region) and resides in a region of high deletion frequency [67]. Around one third of all pathogenic PARK2 variants are CNVs occurring between exons 2 and 5, which form a recombination hotspot [68]. Although much rarer PD causing multiplications have also been reported in PINK1, including even deletions of the entire gene[69]. Finally very rare instances of early-onset PD causing CNVs in PARK7 have been reported [61,70].

But what causes the observed CNVs at Mendelian PD loci? One suggested instigator is transposable elements (TEs). *Bose et al* identified that globally TEs (specifically a class of TE called *Alu*) are enriched at CNV breakpoints and suggested that this increases the regions susceptibility to genomic instability [71]. Evidently an enrichment of TE at PD associated CNV breakpoints has been identified in *SNCA, PARK2 , PINK1* and *PARK7*. Ross and colleagues reported that the presence of *Alu* and LINE1 elements at the *SNCA* locus may contribute to the genomic instability at this region which induces the disease causing copy number variation [72]. Further two *AluJo* repeats have been suggested to enclose putative breakpoints that are in part involved in complex rearrangement that causes the PD causing PINK1 CNV. In regards to the PARK7 PD causing CNV *Alu* repeat elements flank the deleted sequence of PARK7 on both sides, suggesting that unequal crossing-over was likely at the origin of this genomic rearrangement [73]. Finally at the PARK2 locus in multiple patients *Alu* elements have mediated non-allelic homologous recombination which is the suggested causative mechanism of these events of CNV [74]. Therefore, it is evident from these events that enrichment of TEs at specific regions at PD loci could induce CNV causative of the disease.

1.8. Transposable elements

In 1948 “jumping genes” or TEs were first discovered by molecular geneticist pioneer Barbara McClintock [75]. Decades later following huge advances in genetic technologies it is now established that TE derived sequence constitutes over half of the human genome [76]. These elements are capable of moving to a new location in the genome and be subdivided based on their method of replication, i.e via an RNA (retrotransposable elements (RTEs)) or DNA (DNA transposons) intermediate. DNA transposons insert into the genome in a “cut-and-paste” manner. On the other hand, RTEs follow a so called “copy-and-paste” mechanism whereby the RTE is first transcribed into RNA and then inserted back into a different location of the genome. For a broad description of the subclasses of TE elements see (Figure 1.5).

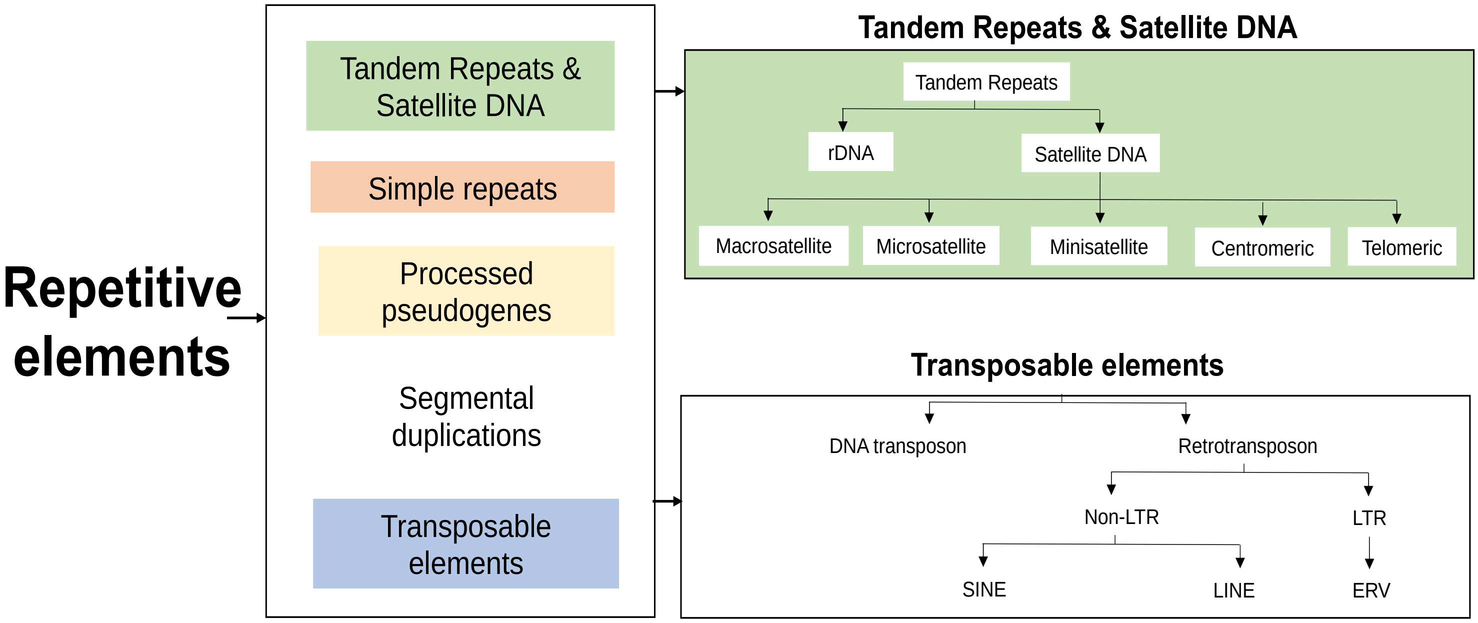


Figure 1.5 Repetitive DNA classes in the human genome. The major class of repetitive DNA are TEs, which can be further divided into DNA transposons or RTE according to their mechanism of transposition, i.e through an RNA or DNA intermediate. Retrotransposons are the most abundant class in the human genome and can be further divided into long terminal repeats (LTR) and non –LTR retrotransposons. Non-LTR elements have the ability to mobilise and can be further subdivided into SINE (e.g. *Alu* elements) and LINE (e.g. LINE-1 elements). The LTR class of RTE contains endogenous retroviruses (ERVs) such as HERV-K.

1.9. Retrotransposons

RTE are the main focus of this thesis and are further classified into long terminal repeats (LTR) or non-(LTR) elements. The later resemble integrated mRNAs and have a distinct mechanism of transposition. Non-LTR RTE can be further classified as either long interspersed nuclear elements (LINEs) or short interspersed nuclear elements (SINEs). Collectively LINE and SINE elements comprise over 34% of the human genome.

The majority of TE in the genome (~99%) have accumulated truncation events or mutations rendering them incapable of transposition. Despite this several subclasses of TEs are still actively transposing, including; LINE-1 [77,78], *Alu* [79,80] and SVA (SINE-VNTR-*Alu*)[81,82] . SVA and *Alu* elements are non-autonomous TEs, which are transposed in *trans* by the LINE1 machinery [83,84]. Not only can LINE 1 mobilize adjacent non-LTR TEs but the LINE-1 replication machinery can also facilitate the duplication of non-TE transcripts, typically protein coding genes. Further through the mechanism of retroduplication LINE-1 can generate processed pseudogenes (PPG)s[85] in the genome. At present approximately 130 pathogenic TE variants caused by these transposition events have been associated with rare Mendelian forms of disease [86]. Consequently, as a result of this constant transposition, non-LTR are a huge source of uncaptured structural variation in the human genome.

1.9.1. Non-LTRs retrotransposons

1.9.1.1. Long interspersed nuclear elements (LINE1)

LINE-1 are the most abundant class of non-LTR and constitute nearly one fifth of the human genome. There are around half a million copies of LINE-1 in the genome and full length they are around 6kb. The structure of a full-length LINE-1 is shown in Figure 1.6.

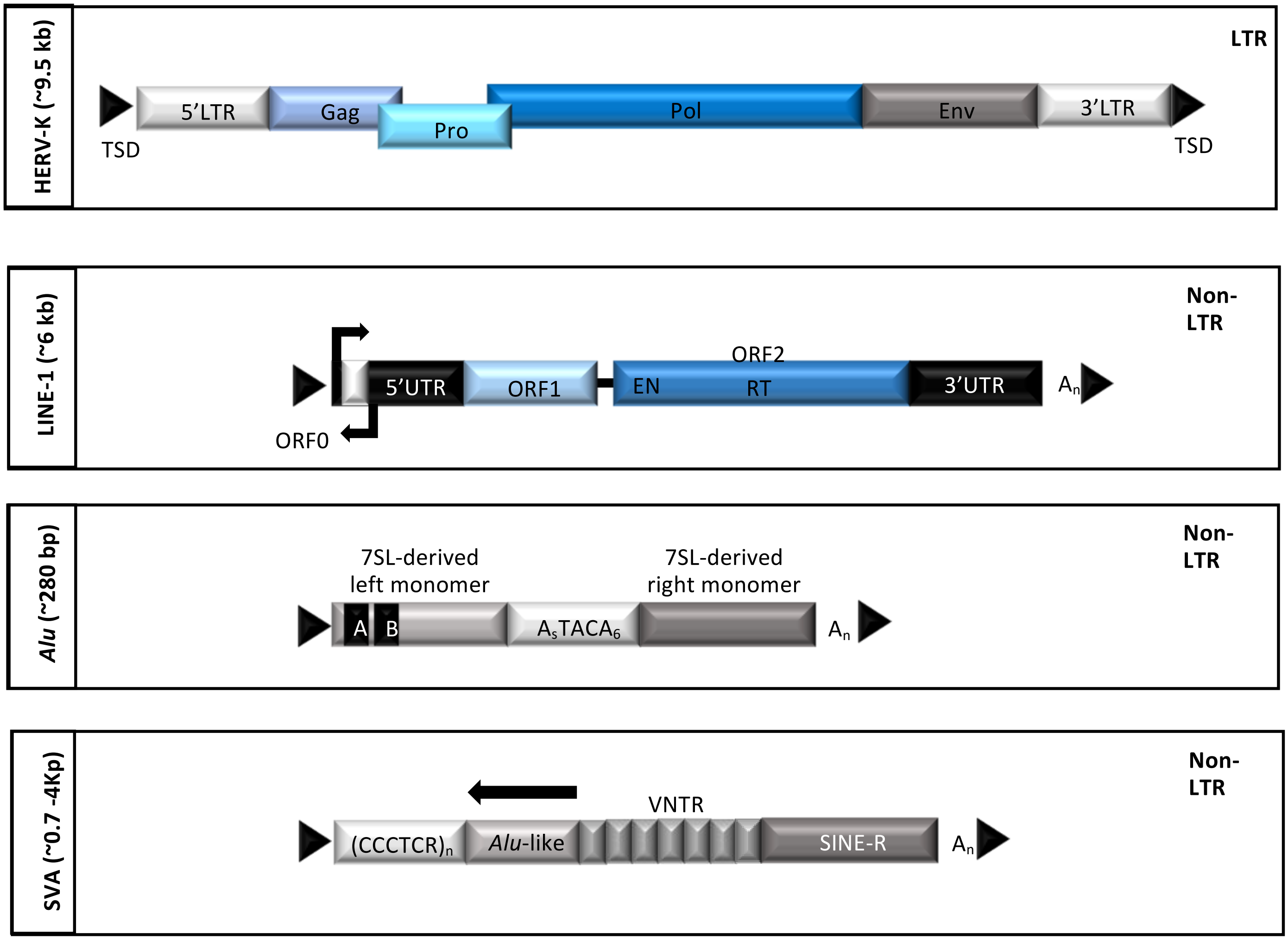


Figure 1.6 . Structure of transposable elements in the human genome (adapted from Savage et al 2019)

Typically LINE-1 consist of two open reading frames (ORF)[87] and both ORF encoded proteins are required for transposition [88]. The mechanism by which LINE-1 transposition occurs is a process called target primed reverse transcription (TPRT)[89]. In brief the LINE-1 RNA is transcribed by RNA polymerase II which is regulated by a promoter within the 5'UTR of the LINE-1[90]. This is then exported into the cytoplasm of the cell where the ORF1 and ORF2 proteins are translated. Next the ORF1p, ORF2p and LINE-1 RNA then form a ribonucleoprotein complex (LINE-1 RNP) which is transported back in to the nucleus. With its endonuclease activity the ORF2p nicks the bottom strand of the DNA at the consensus sequence 5'TTTTAA 3', exposing a 3' hydroxyl group. The ORF2p then uses as a primer to reverse transcribe the LINE-1 RNA, nicks the top strand of the DNA and then the newly reverse transcribed cDNA of the LINE-1 is integrated into the genome. Finally, the complementary strand of DNA is synthesised. TPRT can result in target site duplications (TSDs), 5' truncations, 3 'transductions and internal rearrangement and inversions.

Over 99.9% of LINE-1s in the human genome are in-active due to mutation and translocation events. Despite this a recent study has reported that there are many intact LINE1 that still have the ability to facilitate their own transposition. Originally *Brouha et al* identified 90 L1 elements with intact ORFs from the 2001 working draft of the haploid human genome sequence. 82 of these elements were assayed for their transposition capabilities which identified that around half of were active in a cell culture transposition assay. This led to the initial prediction that there were 80-100 LINE-1 elements that are transposition competent in a given human genome, often termed "hot" LINE-1 [77].

Expanding on this analysis *Yang et al* undertook a systematic characterization of LINE-1 in the genome and identified that over two hundred LINE-1 were full-length, which has more than doubled the number of LINE-1 identified as still capable of transposition. These "hot" LINE-1 not only expand the human genome through their own replication but can also mobilise non-autonomous retrotransposons including *Alus*, SVA and PPGs.

1.9.1.2. *Alu*

There are over 1.1 million *Alu* elements in the human genome [76]. Termed *Alu* due the presence of the *AluI* restriction enzyme site in their sequence, *Alus* also contain an internal RNA polymerase III promoter to regulate their transcription [91]. A conical *Alu* is typically around 300bp (approximately 280bp with a poly A tail) and are derived from 7SL RNA (Figure 1.6). Due to their small size *Alus* are usually well-tolerated when inserting into the genome and this also means they are easier to detect bioinformatically[92].

*Alu*s can be broadly divided into five subfamilies based on evolutionary age. This has led to the suggestion that *Alu* subfamilies have originated through successive waves of fixation from active *Alu* sequences. The oldest *Alu* elements are the monomeric *FAM*, *FRAM* and *FLAM* sequences. The oldest *Alu* dimeric subfamilies are *Alu* -Jo and *Alu* -Jb which are estimated to be around eighty million years old. The intermediately aged *Alu* subfamilies belong to the *Alu* -S class, which can be further subdivided into families Sx, Sp, Sq, Sg and Sc, which are estimated to be 30–50 million years old. Finally the youngest subfamilies belong to the *Alu* -Y class, which are less than 15 million years old [93]. Recent studies have shown that *Alu* insertion variants can disrupt gene expression through a number of mechanisms such as alternative splicing and exonization[94].

1.9.1.3. SINE-VNTR-Alu

SVAs are the most recently evolved family of active non-LTR transposable elements, with approximately 2600-3000 SVA copies in humans [76]. These hominid-specific non-autonomous elements contain consensus sequences for LINE-1 endonuclease recognition, and rely directly on active expression of the LINE-1 machinery, ORF1p and ORF2p, in order to be mobilised in trans [95,96]. The SVA family of transposable elements can be further divided into second subfamilies, named A – F1 in order of evolutionary age. SVA A is the oldest subfamily in evolutionary terms (~13.6 million years) and SVA E, F (~3.5 and ~3.2million years, respectively), and F1 the youngest, whilst subfamilies D and B are the most abundant in the genome, accounting for ~40% and ~15% of the total number of SVA elements. The youngest subfamilies, SVA E, F, and F1, are all human specific [82]. In addition to this, a recent study that characterized SVA-D (~9.6million years) identified that the large majority of SVA-D (78%)are also human-specific[97] .

Structurally, a canonical SVA is comprised of five main components (Figure 1.6), beginning with (1) a simple hexamer repeat of (CCCTCT)n at the 5' end, which may be variable in copy number, followed by (2) an Alu-like region made up of two antisense Alu fragments separated by a region of intervening sequence, (3) one or two variable number tandem repeat (VNTR) regions, typically with a repeating sequence between 35

– 50 bp, (4) a SINE region derived from the 3’ LTR of the retroviral HERV-K10 element, and finally (5) a 3’ poly-A signal[82]. The seventh SVA family, known as F1, lacks the 5’ CCCTCT hexamer repeat, instead containing a 5’ transduction of exon 1 of the MAST2 gene[98].

The insertion of SVA elements into the human genome has influenced our evolution through mechanisms such as insertional mutagenesis, recombination events, exonisation and modulation of gene expression [99,100]. But these insertions can be friend or foe depending on where about in the genome they insert. On the one hand it is hypothesized that the presence of SVAs at neuropeptide gene loci points to a retrotransposon-mediated evolutionary mechanism which may have contributed to the development of human behavioural traits [101]. However, SVA insertions can also be pathogenic, to date there are 8 disease-specific SVAs that have been implicated in conditions including, Fukuyama-type congenital muscular dystrophy, cystic fibrosis, haemophilia and several cancers [102]. These pathogenic SVA insertions are disease causing due to various mechanisms such as exon skipping and decreased mRNA production.

A relevant example of this can be seen with the disease-specific SVA-F insertion in intron 32 of the Transcription initiation factor TFIID subunit 1 (*TAF1*) gene which is causative of X-Linked Dystonia Parkinsonism (XDP). Not only does the SVA cause disease but it is also disease modifying as the size of the SVAs hexanucleotide CT repeat domain inversely correlates with XDP age at onset [103]. The disease-specific SVA-F insertion was previously found to alter sequence within TAF1 introns causing abnormal mRNA expression and significant dysregulation of a neural-specific TAF1 isoform, N-TAF1 in XDP causative relative to control brain tissue [104]. In addition, generated XDP and matched control induced pluripotent stem cell (iPSC) lines confirmed TAF1 transcript dysregulation and also found a significant decrease in expression of TAF1 transcript fragments that span the region of the SVA (intron 32-36). Remarkably in the most recent analysis CRISPR/Cas9 excision of the SVA rescued the aberrant transcriptional signature and normalized expression of TAF1 in patient-derived iPSCs [105].

SVAs preferentially insert into gene dense and high GC regions of the genome [4,82], which consequently means they have a greater potential to affect gene regulation. Our group first supported this with analysis focussed on human specific reference SVAs i.e. ones that are "fixed" in the genome, identifying that fixed SVA have the ability to differentially affect transcription. First this was demonstrated with the SVA-D 9.9kb upstream of the major transcriptional start site of the fused in sarcoma (*FUS)* gene. Genetic mutations in *FUS* have been associated with ALS and Frontotemporal Lobar Degeneration [5,106–108]. The SVA was shown to act as a classical transcriptional regulatory domain in the context of a reporter gene construct both *in vitro* in the human SK-N-AS neuroblastoma cell line and *in vivo* in a chick embryo model [5]. Further a SVA 8kb upstream of the major transcriptional start site of the PD associated *PARK7* gene was shown to modulate gene expression of a reporter gene [4]. Both of the mentioned studies also identified the reference SVAs that were characterised were variable in sequence length of repeat domains. This variation could contribute to differential expression at these loci through mechanisms as shown in Figure 1.7.

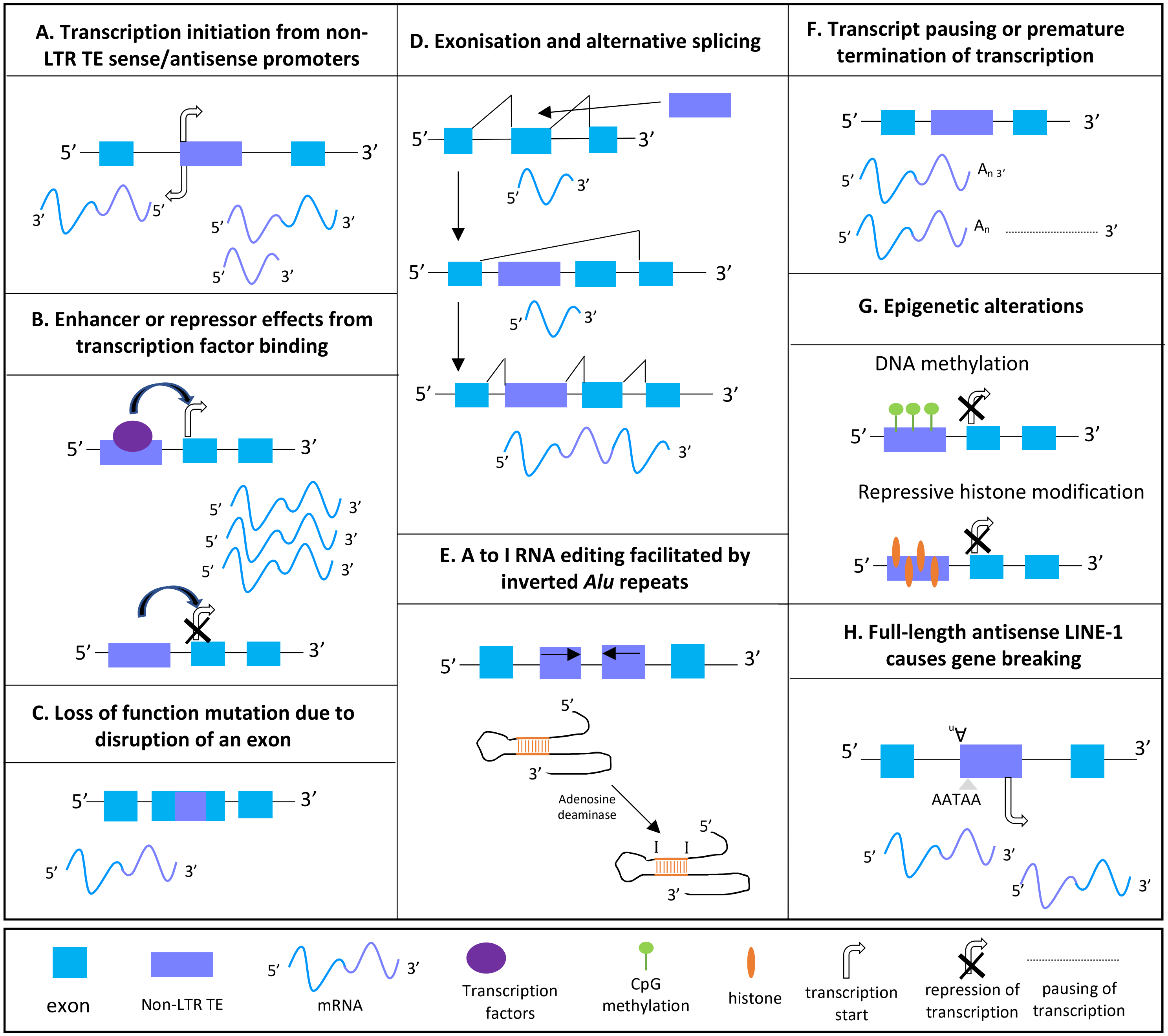

A. Transcription initiation from non-LTR TE sense/antisense promoters
5'
3'
5'
3'
5'
3'
5'
3'
D. Exonisation and alternative splicing
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
F. Transcript pausing or premature termination of transcription
5'
3'
5'
A_n 3'
5'
A_n
3'
B. Enhancer or repressor effects from transcription factor binding
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
G. Epigenetic alterations
DNA methylation
5'
3'
Repressive histone modification
5'
3'
E. A to I RNA editing facilitated by inverted *Alu* repeats
5'
3'
5'
3'
Adenosine deaminase
I I
5'
3'
H. Full-length antisense LINE-1 causes gene breaking
5'
3'
A_n
AATAA
5'
3'
5'
3'
C. Loss of function mutation due to disruption of an exon
5'
3'
5'
3'
exon
Non-LTR TE
5'
3'
mRNA
Transcription factors
CpG methylation
histone
transcription start
repression of transcription
pausing of transcription

Figure 1.7. The effects of non-LTR TE insertions on host gene expression. Non-LTR retrotransposon insertions can affect gene expression through multiple mechanisms and the major mechanisms are outlined. (A) The control of host gene expression by sense and/or antisense promoters of neighbouring non-LTR retrotransposon insertions has been reported for LINE-1 and Alu elements, and the initial 328 bp of a specific subtype of SVAs has been reported to harbour promoter activity. Transcriptional start sites have been identified within LINE-1, Alu and SVA sequences.(B) LINE-1, Alu and SVA sequences contain binding sites for transcription factors that can have either positive or negative regulatory effects on the expression of neighbouring host genes depending on the protein complexes bound.(C) LINE-1, Alu and SVA insertions into exons can cause loss of function mutations.(D) Splice sites within LINE-1, Alu and SVA insertions residing in introns can result in new exons within host genes and alternative splicing (E) When two Alu repeats are inserted in the opposite orientation in close proximity, base pairing between the two repeats can occur in the mRNA forming double-stranded RNA. Adenosine deaminases can bind to the double-stranded RNA and deaminate adenosine to inosine (A to I editing) affecting transcript stability. As inosine is recognised by the translational and splicing machinery as guanosine, this RNA editing could lead to an amino acid substitution (if it occurs in the coding sequence), alternative splicing or modification of microRNA binding.(F) The adenosine-rich nature of LINE-1, Alu and SVA transcripts can introduce premature polyadenylation and/or RNA polymerase II transcriptional pause sites into genes, thereby resulting in termination of transcription within the retrotransposons' sequence or reducing their expression.(G) Epigenetic alterations at the integration site of a new retrotransposon insertion can restrict retrotransposon expression and include DNA methylation (LINE-1, Alu and SVAs contain multiple CpG sites) and heterochromatin formation which can also lead to the repression of neighbouring genes.(H) Full-length LINE-1 insertions that insert in the antisense orientation into an intron of a cellular gene can split the gene’s transcript into two smaller transcripts through a mechanism known as gene breaking. LINE-1, long interspersed nuclear element-1; LTR, long terminal repeat; SINE, short interspersed nuclear elements; SVA, SINE-VNTR-Alu; VNTR, variable number tandem repeat. (adapted from Savage et al 2019)

1.10. Regulation of transposable elements in the genome

As previously mentioned in isolated cases TE insertion events have been identified as causative of rare Mendelian forms of disease. Thus, with active TE this constant transposition can pose a continuous pathogenic threat to the human genome. Despite this threat many TEs have been “domesticated” throughout evolution bringing beneficial traits to the host. In regards to the human genome specifically It is now evident that TE have been widely recruited and integrated and provide functional species-specific regulatory networks [109,110]. Thus, TEs provide an abundant source of *cis* regulatory sequences in the human genome, including promoters [111–113] and enhancers [114–117]. In fact, around one quarter of all experimentally validated human promoters contain TE derived sequence [112] and over fifty thousand ERV sequences were found to initiate transcription [111]. TEs also contribute to trans-regulatory elements such as; small RNAs [118–120], transcription terminators [121] and most significantly, with the recent advancement of chromatin sequencing, it is clear that TE play a significant role in establishing chromatin architecture [61,122,123].

As TEs can be highly disruptive to transcription, the human genome has developed multiple mechanisms to tightly suppress TEs pre and post transcription. TE Transcription is mainly suppressed as a result of epigenetic silencing through different chromatin modulations such as; histone modification, DNA methylation and chromatin remodelling [124].

TEs can be highly disruptive to transcription, therefore the human genome has developed multiple mechanisms to tightly suppress TEs pre and post transcription. TE transcription is mainly suppressed as a result of epigenetic silencing through different chromatin modulations such as; histone modification, DNA methylation and chromatin remodelling [124]. TEs also contribute to trans-regulatory elements such as; small RNAs[125–127], transcription terminators[128] and most significantly, with the recent advancement of chromatin sequencing, it is clear that TE play a significant role in establishing chromatin architecture [67,129,130].

Recent studies identified that TEs are associated with the three-dimensional organization of chromatin in the nucleus, such as the intrachromosomal colocalization of similar repetitive elements [131] , or the occurrence of TEs in domains or at domain boundaries [132–135]. Chromosomal organization within the nucleus is strongly associated with cell-specific transcriptional activity. Globally, transcriptional repression or activation is accompanied by nuclear relocation of chromatin in a cell-type-specific manner. This process forms chromatin compartments of coordinated gene silencing or expression. Locally, chromatin is organized into sub-mega base pair domains of self-contained chromatin proximity which are termed as topologically associated domains (TADs). TADS encompass interactions between regulatory elements such as enhancers and promoters, as well as between coregulated genes, which reflects cell-type-restricted programs[136]. These regions can be further divided into smaller, nested sub-TADs. Going beyond algorithmic or data quality differences, there are several lines of evidence that suggest that TADs are functionally distinct from sub-TADs. Perhaps the best current insights into these issues relate to how each of these features are conserved between different cell types. On one hand there is little evidence that TADs vary between cell types, suggesting that they are largely invariant feature of genome organization [132,137,138]. However, on the contrary sub-TADs appear to differ, at least partially, between different cell lineages and the cell-type-specific organization of sub-TADs appear to be related to cell type specific regulatory events. In this regard, it is believed that TADs represent a larger, more invariant feature of genome organization, within which cell type specific structures can form to influence lineage specific genome regulation[139].

Chapter 2 Characterisation of the variation and regulatory properties of the reference PARK7-SVA

2.1 Introduction

SVAs are one of the most polymorphic sources of recent structural variation in the human genome [101,125]. Yet this form of genetic variation is currently completely uncaptured in any reference panel. These elements preferentially insert into gene dense and active regions of the genome; hence they have a greater potential to modulate gene expression and can do so in an allele-like manner. Due to their high GC sequences (60-70%) SVAs can be seen simplicity as large CpG islands, meaning that they can act as novel promoters or cryptic splice sites and have been shown to effect transcription of their neighbouring genes through these mechanisms [82].

Consistent with this our group previously demonstrated that a polymorphic SVA-D upstream of *PARK7*, also termed *DJ-1* (a gene associated with early-onset monogenic PD[73]) can direct gene expression within a reporter gene construct in two cell-lines (SK-N-AS, a human neuroblastoma cell line and MCF-7 a breast cancer cell line)[126,127]. In addition, different domains of the SVA, such as the VNTR and SINE regions have different regulatory properties. Further characterization of the variation of the *PARK7* SVA identified that there were four alleles, which were polymorphic in the VNTR (variable number tandem repeat) and hexamer repeat domains.

In relation to VNTR variation, it has previously been shown by our group that VNTRs can be both differential regulators and biomarkers of disease based on genotype of the repeat. This variation can direct differential response to an environmental stimulus through affecting epigenetic parameters which we have shown with the serotonin transporter (*SLC6A4*) VNTR. Not only is the *SLC6A4* VNTR a genetic risk factor for a number of neurological conditions, but it has also been shown that the repeat copy number mediates differential transcription factor binding, which causes a differential response to cocaine [128]. Therefore, one could imagine that variation within the sequence of existing reference SVAs (which is variation not currently captured in any reference panel) could influence gene expression (shown in Figure 2.1) and be associated with disease in a similar manner [126,127].

In the following chapter we focus on the *PARK7* SVA-D cited above, which is located ~700bp upstream of the *PARK7* minor transcriptional start site and from here will be termed the "*PARK7*-SVA".

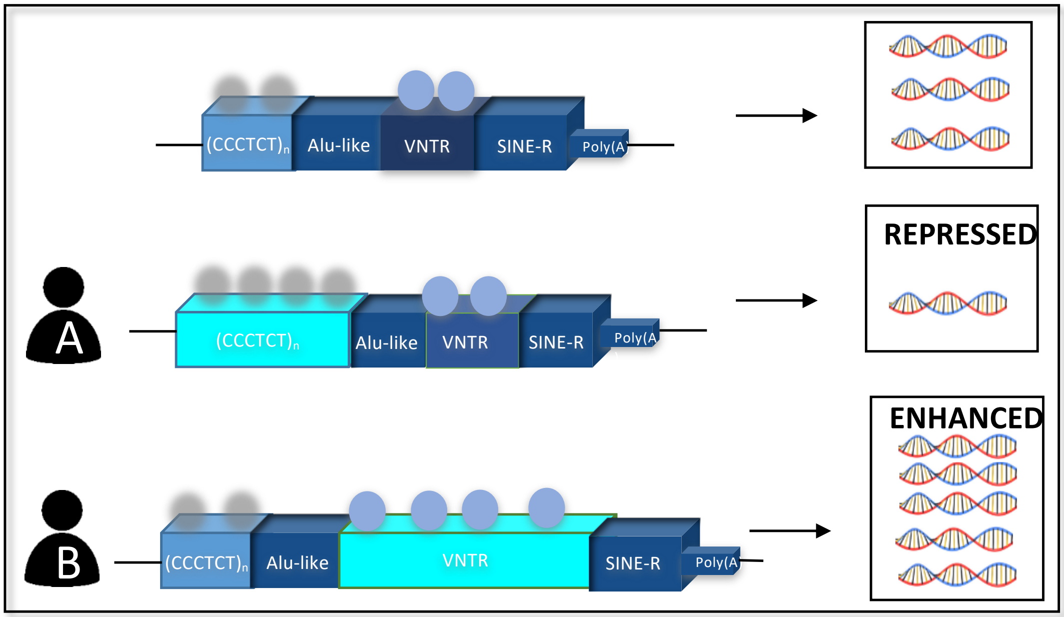
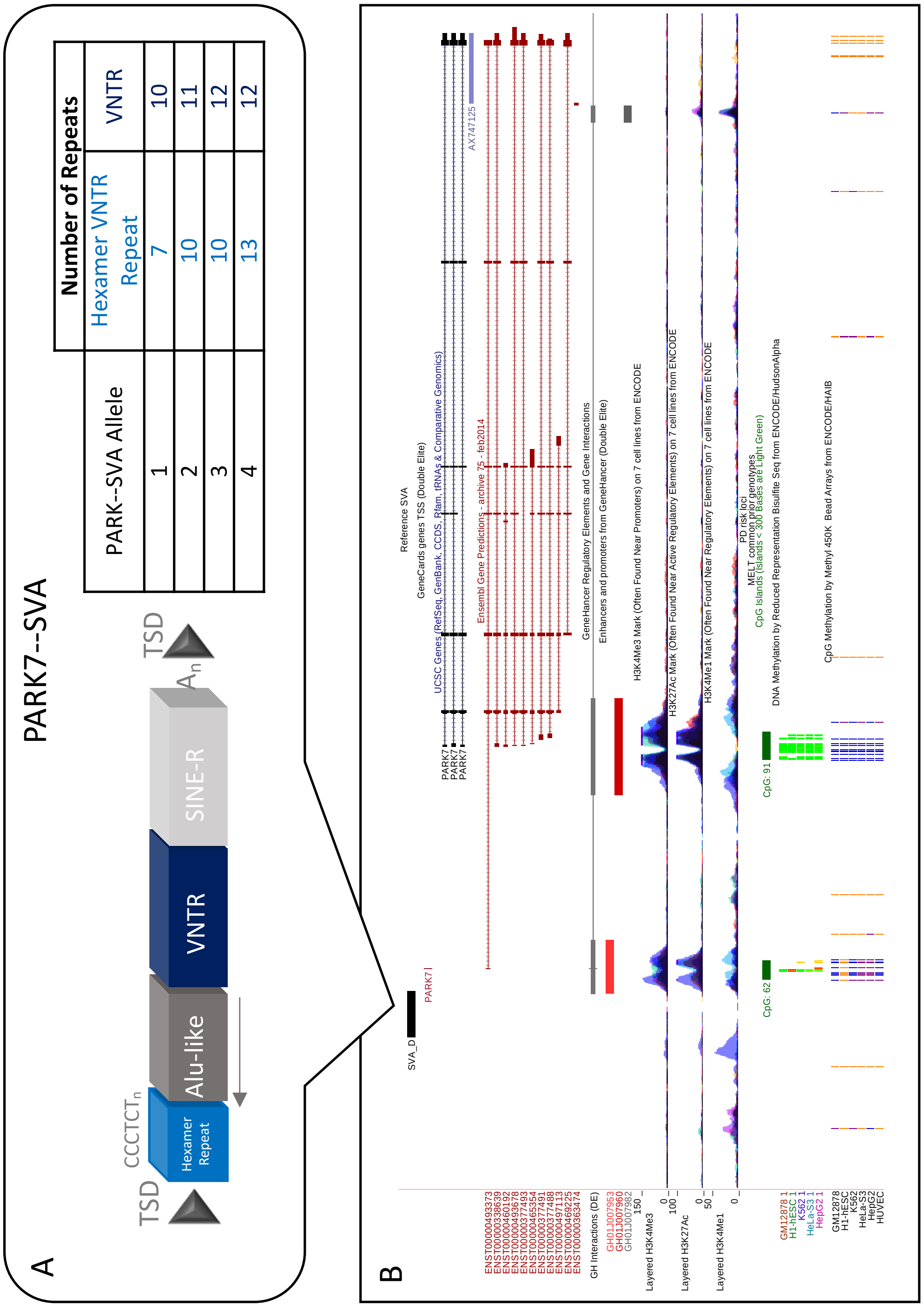


Figure 2.1.SVAs can act as regulatory elements and affect expression of nearby genes in an allele-like manner. An example of this mechanism is shown above. The top SVA is the SVA which has the same sequence to as the reference genome SVA. Individual A has a variant of the SVA that has a longer hexamer repeat domain. This encodes binding of different transcription factors to composite sequences which repress gene expression. Individual B has a variant of the SVA that has a longer VNTR domain, which acts as in enhancer. Overall this source of variation results in varied gene expression between individuals, yet this is isn't catalogued in the genome.

Along with identifying that the *PARK7*-SVA directed gene expression in an allele-like and tissue-specific manner, *Savage et al* also validated that a novel transcript was expressed from the distal promoter. Further, utilizing affymetrix human exon array data from an extended probe set they confirmed that there was significant expression detected at the probes located over the *PARK7*-SVA in all brain regions analysed (cerebellum, frontal cortex, hippocampus, medulla, occipital cortex, putamen, substantia nigra, temporal cortex, thalamus and white matter). Consequently proxy SNPs were developed so that in future studies genotyping and expression quantitative loci (eQTL) analysis could be performed bioinformatically [127]. The PARK7 locus is shown in detail below in Figure 2.2.


A
PARK7--SVA
TSD
CCCTCT$_n$
Hexamer Repeat
Alu-like
VNTR
SINE-R
A_n
TSD
B
SVA_D
PARK7
Reference SVA
GeneCards genes TSS (Double Elite)
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
AX747125
Ensembl Gene Predictions - archive 75 - feb2014
GeneHancer Regulatory Elements and Gene Interactions
Enhancers and promoters from GeneHancer (Double Elite)
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
PD risk loci
MELT common prior genotypes
CpG Islands (Islands < 300 Bases are Light Green)
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB
CpG: 62
CpG: 91
ENST00000493373
ENST00000338639
ENST00000460192
ENST00000493678
ENST00000377493
ENST00000465354
ENST00000377491
ENST00000377488
ENST00000497113
ENST00000469225
ENST00000363474
GH Interactions (DE)
GH01J007953
GH01J007960
GH01J007982
150 _
Layered H3K4Me3
0 _
100 _
Layered H3K27Ac
0 _
50 _
Layered H3K4Me1
0 _
GM12878 1
H1-hESC 1
K562 1
HeLa-S3 1
HepG2 1
GM12878
H1-hESC
K562
HeLa-S3
HepG2
HUVEC

PARK--SVA Allele	Number of Repeats	
	Hexamer VNTR Repeat	VNTR
1	7	10
2	10	11
3	10	12
4	13	12

Figure 2.2 The PARK7 locus: A) Schematic of the reference PARK7-SVA. The SVA-D is variable in the hexamer repeat and VNTR domains (blue). The repeat variation for each allele is described in the adjacent table. B) The PARK7-SVA region is highlighted in red. A novel transcript arises from the region which is shown by the Ensembl Gene Predictions track. The presence of a distal promoter at this region is supported by the GeneHancer track (which detects an additional promoter (red)) and the ENCODE functional annotation (showing histone peaks over the region). Encode 450k methylation data shows that the CpG probe adjacent to the promoter is differentially methylated.

As bioinformatic and sequencing technologies advance it is becoming increasingly apparent that overlooked structural variants (such as SVAs) are in fact involved in disease and can act as major regulators of gene expression in the human genome. With that in mind in this chapter we utilized the previously generated *PARK7*-SVA proxy SNPs to call the SVA genotype and then comprehensively analysed possible association between the variation in the SVA sequence and PD. Further we implemented eQTL analysis and assessed whether it modulated gene expression in an allele-like manner.

2.2 Aims

- Identify if a reference SVA is imputable using previously generated proxy SNPs for the *PARK7*-SVA
- Run association analysis to address whether the SVA is associated with risk of PD
- Run QTL analysis to identify if the SVA directs gene expression in the normal brain in an allele-like manner.

2.3. Methods

2.3.1. Subjects

2.3.1.1. Parkinson's Progression Markers Initiative (PPMI) cohort

PARK7 encodes DJ-1 which is a protein that protects cells from oxidative stress (a major pathogenesis of PD) and mutations within *PARK7* cause an extremely rare form of early onset PD. Therefore, we wanted to address whether the *PARK7*-SVA was associated with risk of PD. For this we used the publicly available and very characterized PPMI cohort:

https://www.ppmi-info.org/access-data-specimens/download-data/

2.3.1.2. North American Brain Expression Consortium (NABEC) cohort:

We utilized the publicly available NABEC cohort to characterize the variation of the *PARK7*-SVA and address whether the SVA is an eQTL in the brain. The NABEC cohort data is an extensive resource generated from neurotypical individuals that includes genotyping, RNA-sequencing, CAGE-sequencing, and CpG DNA methylation data. Available here:

https://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs001300.v1.p1

As the later do not contain the exact same individuals a detailed description of the datasets used will be given in subsequent sections. To note, the NABEC WGS that was utilized in this study is not currently publicly available but will be deposited on dbGaP when data release is possible

2.3.2. Samples and quality control

WGS was generated in the following manner. The DNA was received at NIH Laboratory of Neurogenetics and prepared using picogreen quantification, to a total of 2ug of DNA, concentrated at 30ng/ul, with a minimum volume of 40ul and sent for sequencing to Macrogen. At Macrogen one microgram of each DNA sample was fragmented by Covaris System and further prepared according to the Illumina TruSeq DNA Sample preparation guide to obtain a final library of 300-400 bp average insert size. Libraries were multiplexed and sequenced on Illumina HiSeq X platform.

Paired-end read sequences were processed in accordance with the pipeline standard developed by the Centers for Common Disease Genomics[129] . This standard allows for whole-genome sequence (WGS) processed by different groups or centers to generate "functionally equivalent" (FE) results[130] . The GRCh38DH reference genome was used for alignment as specified in the FE standardized pipeline[131] . The Broad Institute's implementation of this FE standardized pipeline, which incorporates the GATK (2016) Best Practices[132], is publicly available and is used for WGS processing. Single-nucleotide polymorphisms (SNP) and InDel variants were called from the processed WGS data following the GATK (2016) Best Practices using the Broad Institute's workflow for joint discovery and Variant Quality Score Recalibration (VQSR)[133] .

For quality control each sample was checked using common methods for genotypes and sequence related metrics. Plink v1.9 [134] was used to check each sample's genotype missingness rate (< 95%), heterozygosity rate (exceeding +/- 0.15 F-stat), and gender. The King v2.1.3 kinship tool [134,135] was used to check for the presence of duplicate samples and to check concordance with existing genotypes for the PPMI subjects for whom data are available. Sequence and alignment related metrics generated by the Broad's implementation of the FE standardized pipeline were inspected for potential quality problems. This included the sample's mean sequence depth (< 30X) and contamination rate (> 2%), as reported by VerifyBamID, and SNP count as reported by Picard's CollectVariantCallingMetrics (< 3 StDev)[136] based on the sample's genomic VCF (gVCF).

2.3.3. Genotyping of the *PARK7*-SVA

Savage et al previously developed proxy SNPs for the *PARK7*-SVA by genotyping the SVA in CEU HapMap DNA which had corresponding SNP data. The analysis of the individuals in the CEU HapMap cohort with alleles 1 and/or 3 (the most common of the alleles) identified that at the SNP rs2493215 a genotype of G was a proxy for the *PARK7*-SVA allele 1 and a genotype of A was a proxy for *PARK7*-SVA allele 3 with a (r2 =0.909). When the analysis was expanded to include individuals with allele 2 and 4 it was shown that the A genotype also corresponded to these alleles. The genotype data was analysed further in Haploview software to include all four alleles of the *PARK7*-SVA, which identified a SNP (rs226476) that would tag allele 2 with a r2 = 0.903 a genotype of T corresponds to allele 2 and a genotype of G corresponds to alleles 1, 3 and 4. Using these two SNPs in combination alleles 1, 2 and 3 can be tagged by their specific genotypes. We extracted rs226476 and rs2493215 from both the PPMI and NABEC individuals WGS and from these genotypes typed the genotype of the *PARK7*-SVA.

2.3.4. Association analysis with the PARK7-SVA

All statistical analyses were performed in the R Statistical environments (R version 3.4.1 (2017-06-30). Principal components (PCs) were created from the directly assayed genotypes using Plink v1.9 10 . For the PC calculation, variants were filtered for MAF (>0.01), genotype missingness (<0.05) and Hardy–Weinberg equilibrium (P =>1E-6). To assess whether the genotype of the *PARK7*-SVA was associated with PD risk, a linear regression model using the following formula was used (adjusting for usual covariates; age sex and PC1-PC5):

PD ~ PARK7-SVA_genotype + age + sex + PC1-PC5

2.3.5. *PARK7* expression data from the frontal cortex of the NABEC cohort

2.3.5.1. RNA-seq data

A full description of the NABEC frontal cortex RNA-Seq data generation is given by *Gibbs et al*[137]. Reads were processed using two different pipelines to assess whether the *PARK7*-SVA was an eQTL for the reference *PARK7* transcripts and also the novel brain specific transcript (described by *Savage et al.).*[126,127]

2.3.5.1.1. Refseq annotation quantification

Profiling of 22,184 mRNA transcripts was performed using HumanRef-8 Expression BeadChips (Illumina) as previously described [138]. Raw intensity values for each probe were transformed using the rank invariant normalization method [139–141] and then log2 transformed for mRNA analysis.

2.3.5.1.2. Ensembl annotation quantification

The standard Illumina pipeline was used to generate fastq files. Ensembl GRCh37 annotated transcript abundance was quantified using Salmon[142] in a non-alignment-based mode, and quantile normalized TPM (Transcript per million) values were used for covariates adjustment by PEER tool.[143,144]

2.3.5.2. Methylation data

To determine if the *PARK7*-SVA was an mQTL we accessed the NABEC methylation data. For this CpG methylation status was determined using HumanMethylation27 BeadChips (Illumina), which measured methylation at 27,578 CpG dinucleotides at 14,495 genes[138].

2.3.5.3. CAGE-Seq

CAGE is a gene expression technique used to produce a snapshot of the 5′ end of the messenger RNA population in a sample. A full description of the generation of the NABEC CAGE-Seq used for our eqtl analysis has been previously described by *Blauwendraat et al*[145]. In brief frozen human frontal lobe material was collected for the 106 NABEC individuals. Total RNA was extracted from the frontal lobe of each individual using Life Technologies TRIzol. Libraries were constructed using a published CAGEseq protocol adapted for next generation sequencing [146] CAGEseq data were processed using a previously described analysis pipeline [147]. Finally, mapped CAGEseq reads were grouped into CAGE-clusters using a series of Python scripts designed at the RIKEN Omics Science Center[148] and single base pair promoters within 20 bp of each other were merged into one CAGE-cluster. Raw counts were normalized dividing the number of CAGEseq reads observed at a given CAGE-cluster by the total number of mapped tags in the library and multiplied by 1 million (tags per million, tpm).

2.4. Results

2.4.1. The PARK7-SVA represents a common form of uncaptured structural variation at the PARK7 locus

Using the proxy SNPs developed by *Savage et al* the *PARK7*-SVA was genotyped in the PPMI and NABEC cohorts. The PPMI cohort was utilized to be able to identify association with risk of PD and the NABEC was used to allow downstream etql analysis. Savage *et al* previously genotyped the *PARK7*-SVA with PCR in the CEU HAPMAP individuals. Although we were unable to call the rarest allele (allele 4) in our analysis, when comparing allele frequencies between this present study and Savage et al, our data was consistent with that reported in the CEU HAPMAP population (Table 3.1 & Table 3.2).

Table 2.1. Genotype Frequencies of the PARK7-SVA variants in the PPMI NABEC and CEU HAPMAP cohorts. Table displaying the frequency of each genotype within the 87 individuals genotyped from the CEU Hapmap cohort (Savage et al) which coincides with the genotyping of the 386 individuals in the NABEC cohort. It was not possible to generate a tagging SNP for allele 4 therefore it was not included in the bioinformatic genotyping of the NABEC and PPMI cohorts.

PARK7-d-p-SVA genotype	**Genotype frequency (%)**		
	Previously reported CEU HAPMAP(n=87) (*Savage et al*)	**NABEC (n=386)**	**PPMI (n=562)**
1/1	21.8	24.1	126
1/2	4.6	2.6	26
1/3	40.2	38.9	241
1/4	4.6	-	-
2/2	4.6	0.0	2
2/3	3.4	4.4	17
2/4	1.1	-	-
3/3	18.4	30.1	150
3/4	1.1	-	-
4/4	0.0	-	-

Table 2.2. Allele frequencies the PARK7-SVA variants observed in the previously reported CEU Hapmap cohort and the NABEC PPMI cohorts

Allele	**Allele frequency (%)**		
	Previously reported CEU HAPMAP(n=87) (*Savage et al*)	**NABEC (n=386)**	**PPMI (n=562)**
1	46.6	44.8	46.2
2	9.2	3.5	4.2
3	40.8	51.7	49.6
4	3.4	-	-

2.4.2. The variants of the *PARK7*-SVA are not associated with risk of Parkinson disease

Using the PPMI *PARK7*-SVA genotypes next R was used to identify if the *PARK7*-SVA variants were associated with risk of PD. Using a linear regression model that adjusted for known covariates (sex, age and P1C-PC5) no significant association was found (Table 3.3).

Table 2.3.Initial analysis suggests there is no association between the PARK7-SVA variants and risk of PD in the PPMI cohort. Reporting p value of association analysis and standard error (SE).

PARK7-SVA genotype	P	SE
1/2	1.90E-01	1.28E+00
1/3	9.70E-01	-3.00E-02
2/2	9.90E-01	2.00E-02
2/3	6.60E-01	-4.40E-01
3/3	6.30E-01	4.80E-01

2.4.3. Initial analysis suggests the *PARK7*-SVA is not a significant eQTL for the Refseq *PARK7* transcripts in the frontal cortex of the brain

The NABEC cohort is a publicly available expression data set. The RNA-seq was extracted and was aligned to the three Refseq *PARK7* transcripts shown in Figure 3.3. (Refseq ID: UC001aou (AOU), UC001aox (AOX), UC001aov (AOV)) and quantified for each NABEC individual. Linear regression was used to identify if the *PARK7*-SVA was an eQTL in the normal brain, i.e. did the different alleles of the SVA drive differential expression of the RefSeq transcripts. Although there were observed differences in expression the PARK7-SVA was not a significant eQTL (Figure 2.3) for the RefSeq transcripts.

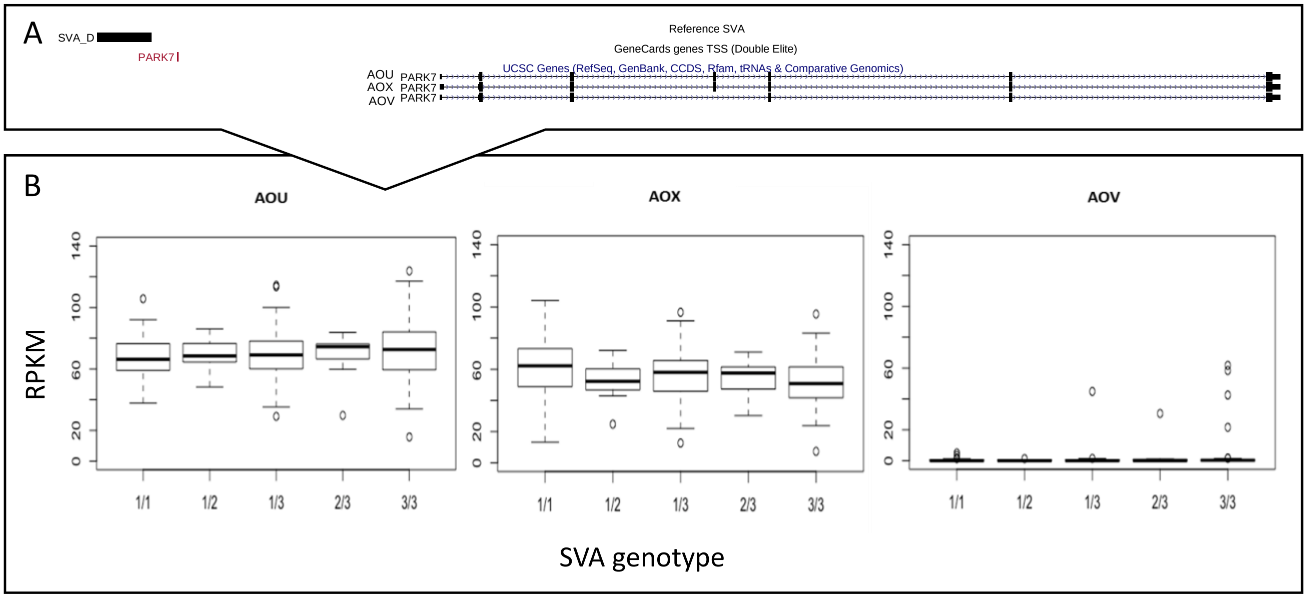


Figure 2.3. Initial analysis suggests the PARK7-SVA is not a significant eQTL in the frontal cortex for the RefSeq transcript A) UCSC image showing the three Refseq transcripts analysed and SVA D. B) Box-plots showing the expression levels of the three reference PARK7 transcripts (AOU, AOX, AOV) for the NABEC cohort. There is no significant association between the PARK7-SVA genotype and expression.

2.4.4. The longer PARK7 transcript is highly expressed in the frontal cortex of the brain but initial analysis suggests the PARK7-SVA is not a significant eQTL

The *PARK7*-SVA is adjacent to the distal promoter/minor transcriptional start site of the *PARK7* longer transcript. Previous studies using Affymetrix human exon arrays report that the longer transcript can be detected in the brain (Ensembl ID: ENST00000493373)[126], however no other literature or data was found to support the expression of this transcript. Therefore, to expand upon this in this present study the GTEX portal was used to further delineate tissue-specific expression of *PARK7*. As shown in Figure 2.4 the longer brain specific transcript is the second most highly expressed PARK7 transcript in the human brain [149].

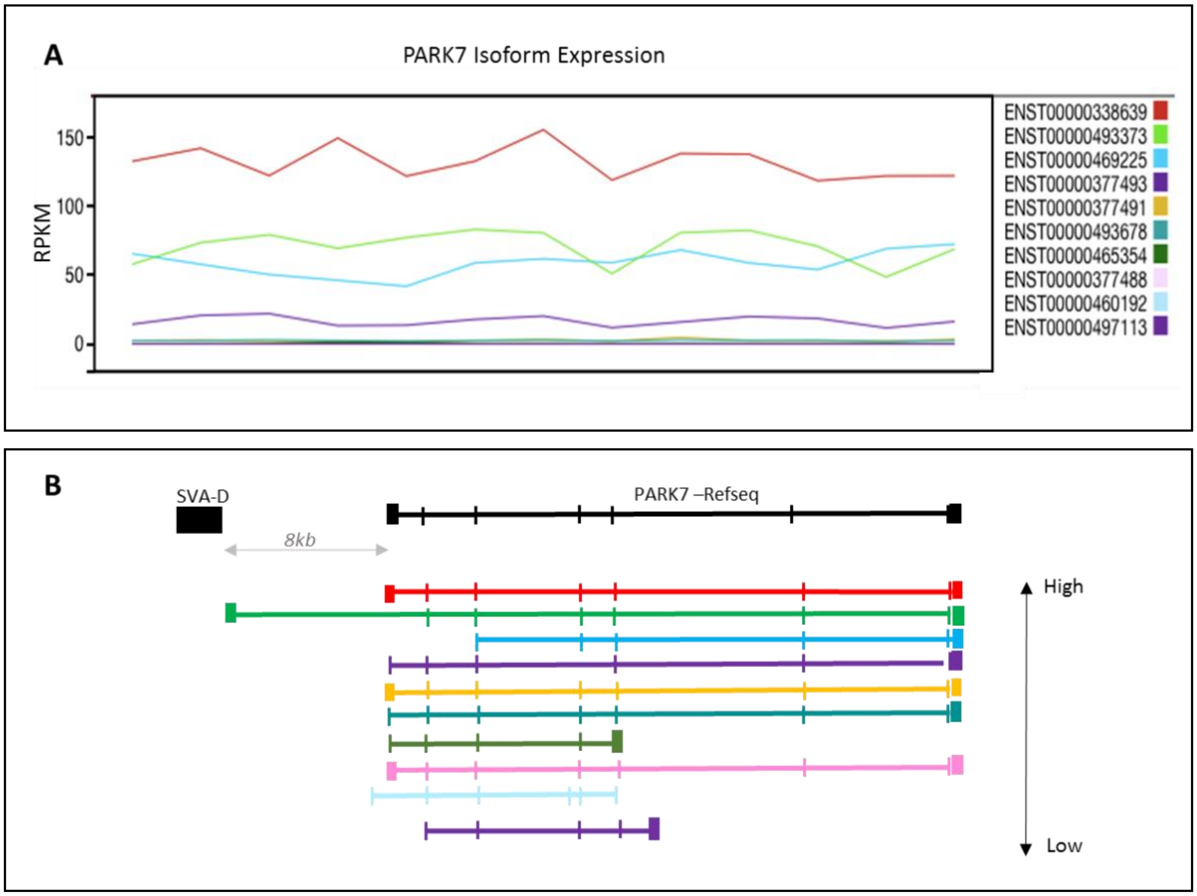


Figure 2.4.The PARK7 longer transcript that originates from the PARK7-SVA region is the second most highly expressed PARK7 transcript in brain tissue according to GTEX A) GTEX portal generated data showing expression levels in brain tissue of the PARK7 transcripts B)Schematic demonstrating the position of the SVA-D and the PARK7 reference transcript and the structure of the transcripts in order of how highly they are expressed in brain tissue according to GTEX (colour coordinated with A). Vertical lines represent exons, orientation is 5’ -> 3’.

Reference SVAs have been shown to act as eQTLs to neighbouring genes. In light of this eQTL analysis was implemented to identify if the *PARK7*-SVA directed expression of the brain-specific longer transcript in the brain in an allele-like manner. Using Salmon[142] a non-alignment-based mode and the longer transcript was quantified in the NABEC frontal cortex RNA-Seq data (Ensembl ID:ENST00000493373). In the NABEC we can confirm that the longer transcript is highly expressed in the frontal cortex of the brain however the *PARK7*-SVA is not a significant eQTL as shown in Figure 2.5.

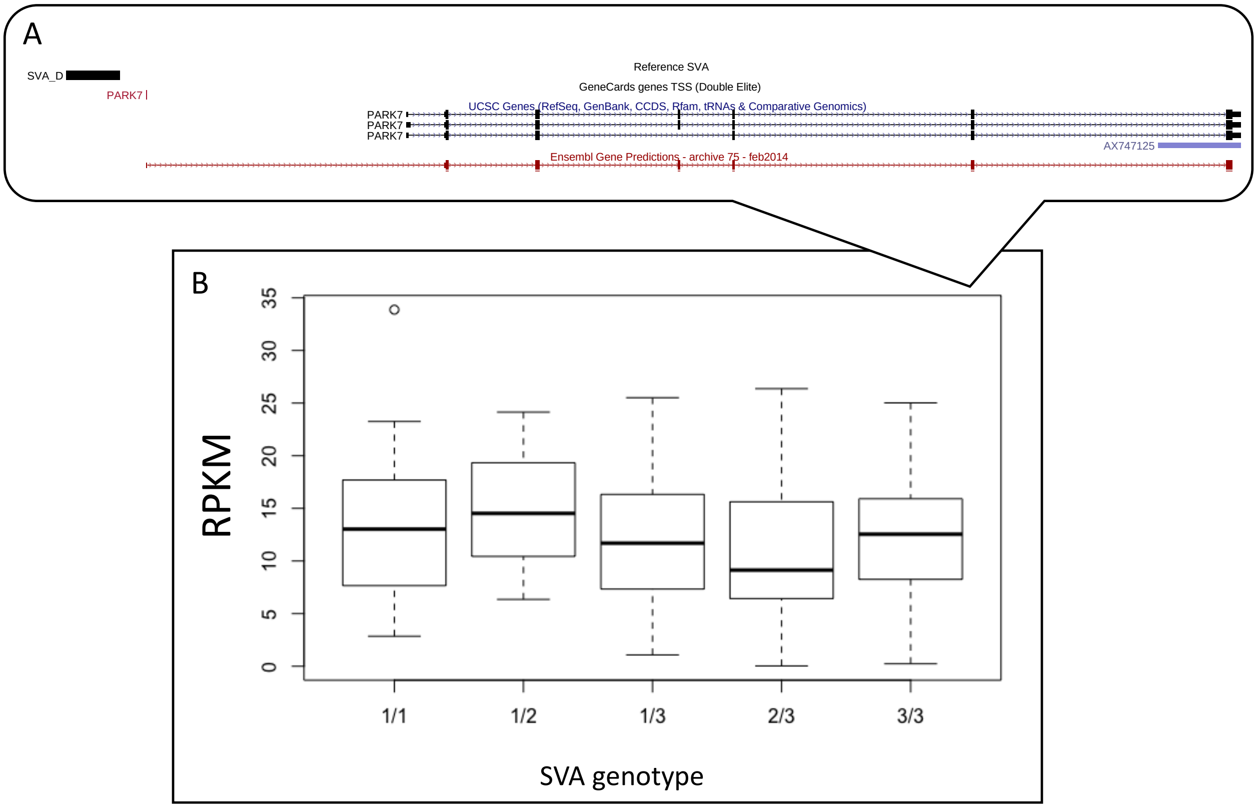


Figure 2.5. Initial analysis suggests the PARK7-SVA is not a significant eQTL for the longer PARK7 transcript that originates from the alternative transcriptional start site A) UCSC image showing the Ensembl annotated longer transcript (highlighted in Red) (Ensembl ID: ENST00000493373) and SVA-D. B) Box-plots showing the expression levels of the longer PARK7 transcript for the NABEC cohort. There is no significant association between the PARK7-SVA genotype and expression.

2.4.5 Initial analysis suggests the *PARK7*-SVA is not a significant mQTL for the *PARK7* promoter in the frontal cortex of the brain

Due to the repetitive nature of SVAs they encode multiple sites for methylation. Thus, if the SVA is variable within the repetitive domains then this could alter methylation in that region. The *PARK7*-SVA is variable within its VNTR therefore we reasoned that this may cause differential methylation at the minor transcriptional start site. 450k methylation data for the NABEC cohort was utilized and the cg probe most adjacent to the SVA was analysed. To note, typically arrays are designed to avoid probes being placed within repetitive regions like an SVA, hence there is no cg probe within the *PARK7*-SVA itself. As shown in Figure 2.6 there was no significant association between the *PARK7*-SVA genotype and methylation status of the adjacent cpg probe. However, the data does show that the region is hypomethylated supporting the UCSC, GTEX and RNA-Seq data that indicate that the region is transcriptionally active in the brain

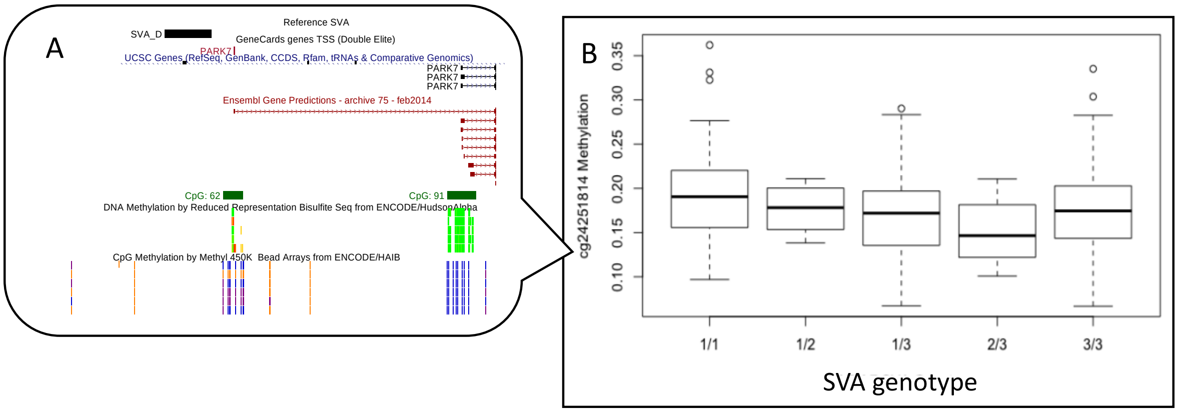


Figure 2.6. The Cg24251814 probe proximal to the PARK7-SVA is hypo methylated but the SVA is not a significant mQTL. The Cg24251814 probe proximal to the PARK7-SVA is hypo methylated but the SVA is not a significant mQT A) UCSC image showing the location of the PARK7-SVA (highlighted red) and the location of the CpG tested (green :CpG:62) which is in the region of the PARK7 minor transcriptional start site. The ENCODE methylation tracks also show that there is differential methylation over the site. B) Boxplot of the methylation over the distal promoter and SVA genotype. The PARK7-SVA is not a significant mQTL at these regions.

2.4.6. CAGE-sequencing identifies that the longer brain-specific *PARK7* transcript is highly expressed in the frontal cortex, but initial analysis suggests the *PARK7*-SVA is not an eQTL for its expression

The aim of this chapter was the explore a novel transcriptional start site and further identify if the PARK7-SVA was an eQTL for a novel brain-specific PARK7 transcript. At the beginning of our analysis the NABEC RNA-Seq was the only expression data-set available. RNA-Seq is not the ideal data to use when studying the effect of TEs on gene expression as you cannot discover novel transcriptional start sites, which are believed to be a major consequence of TE variation. This is because when RNA-Seq is generated the reads are randomly fragmented and are then aligned to a reference. CAGE-seq is a method which offers highly accurate and detailed gene expression analysis by targeting transcription start site (TSS) rather than whole genes. For a subset of the NABEC individuals CAGE-seq data was available (n=106). Therefore, we took advantage of newly available CAGE-seq data as it is particularly useful for our analysis as it captures the 5' regions of mRNA which allows for the identification of novel transcriptional start sites. The CAGE-seq was extracted for the PARK7 region and further analysis identified transcriptional activity at the proximal and distal promoter regions (Figure 2.7). EQTL analysis was implemented to identify if the SVA variation directed gene expression in an allele-like manner. As shown in Figure 2.7 differential expression was observed, however it was not statistically significant.

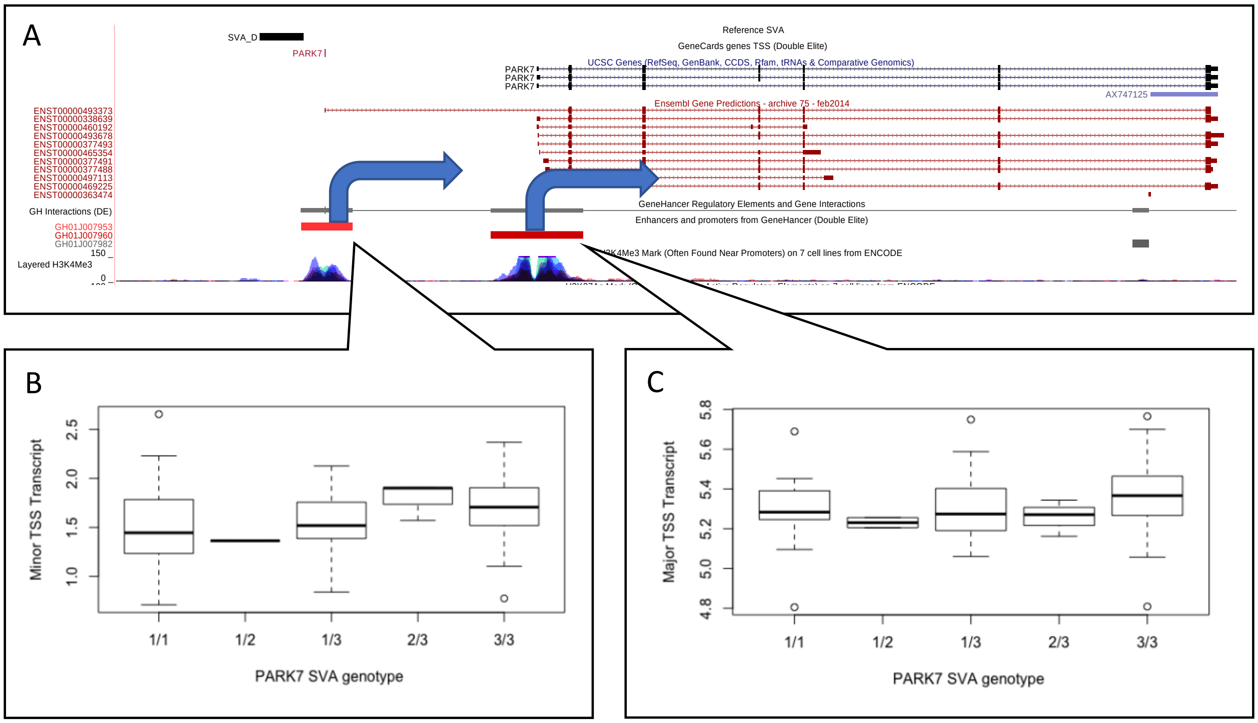


Figure 2.7. CAGE-Seq of the NABEC cohort (n=106) confirmed expression from the minor (distal promoter) and major (proximal promoter) transcriptional start site*. The detected minor TSS (distal promoter) was located 604bp downstream to the* PARK7 *SVA-D (chr1:8,014,244. Hg/19) A) UCSC image of the* PARK7*-SVA/*PARK7 *region. The blue arrows indicate the minor (distal promoter) and major (proximal promoter). B) Boxplot showing the expression over the distal promoter Vs SVA genotype. The* PARK7*-SVA is not a significant eQTL at these regions. C) Boxplot showing the expression over the proximal promoter Vs SVA genotype. The* PARK7*-SVA is not a significant eQTL at these regions.*

2.5. Discussion:

PARK7 is a causative gene for early onset familial PD[150]. In regard to its relationship with sporadic PD, although *PARK7* is not a reported risk locus [151], excess oxidation of *PARK7* (which renders *PARK7* inactive) has been observed in patients with sporadic forms of the disease, suggesting that *PARK7* may also participate in the onset and pathogenesis of the sporadic form of the disease [152]. Therefore, characterizing existing variation at the *PARK7* locus which contributes to differential transcription of *PARK7* could be of importance of further understanding the underlying disease aetiology.

SVAs are known to disrupt transcription through a number of mechanisms such as causing alternative splicing, exon shuffling, formation of secondary structure, recombination events and generation of differentially methylated regions[102]. Because of this they are assumed to be silenced in the genome. At the *PARK7* locus a reference SVA-D lies 8kb upstream of the major transcriptional start site of *PARK7.* Previous work has extensively characterized the uncatalogued variation with the SVA, identifying four alleles of the SVA which are variable in its VNTR and hexamer repeat domains. *In vivo* this variation directed expression of a reporter gene in an allele-like manner. Further bioinformatic analysis of the SVA region showed characteristics such as a CpG island, active histone marks and DNase 1 hypersensitivity clusters indicating a distal promoter in this region adjacent to the *PARK7*-SVA. The presence of a distal promoter was supported by Affymetrix data which detected expression in multiple brain tissues and Ensembl predictions indicated a novel transcript originating from this site[126]. In this chapter we set out to expand on the previous analysis leveraging newly available brain and PD specific datasets to identify if this variation caused differential expression in the human brain and whether this variation was associated with risk of PD.

First, using GTEX, RNA-Seq, CAGE-Seq and methylation data we confirmed that a brain-specific *PARK7* transcript, which originates from the minor transcriptional start site, was highly expressed in the frontal cortex. Following extensive eQTL analysis we did not find a significant association between *PARK7* expression and SVA genotype for the Refseq or novel transcripts. Though it should be noted that this present study could not test the status of the rarest variant (allele 4) as our study bioinformatically genotyped the SVA using the proxy SNPs previously generated by *Savage et al. Savage et al* utilized the HAPMAP CEU GWAS data which did not detect rarer variants (previously as defined MAF>5%) therefore it was not possible to establish a proxy for allele 4 which had a MAF of 3.4%. Evidently the 4th allele is masked within our obtained genotypes and therefore this could bias the current data[126]. Taking this into account although our current analysis suggests that the SVA alleles 1-3 are not associated with differential expression of *PARK7* in the brain or associated with PD, further characterization including the rare variant (allele 4) is needed. Large WGS datasets are now available which have more power to accurately call rare variants compared to existing GWAS data. Therefore, if it is possible to generate proxy SNPs for the fourth allele, then the analysis undertaken in this chapter should be repeated.

In addition, we identified that the PARK7-SVA was not significantly associated with risk of PD. Although this is not surprising given the fact that this gene was originally associated with risk of PD through a candidate gene approach and has not been identified as a risk locus following recent GWAS's. We focused on this locus to bypass months/years of functional studies, taking advantage of the fact that the SVA variation and function had already been deeply characterized by Savage et al. The main advantage being that proxy SNPs had been generated which allowed for bioinformatic analysis in well-powered datasets, meaning we could explore the effects of reference SVA variation on gene expression. However, the work carried out by Savage et al was conducted before the first PD GWAS, when the only "PD associated" genes were genes that had been associated with monogenic forms of disease. Hence PARK7 was only nominated as a gene associated with apparently sporadic PD risk following a candidate gene approach. A candidate gene approach is in essence, the antithesis of genome wide unbiased approaches. As illustrated by numerous failed candidate gene based PD studies the likelihood of nominating the correct gene to be tested and testing the right variants within it is next to none, particularly as through GWAS it has been identified that the typical risk effect sizes associated with variants are too small to be seen by the majority of studies.

It should be emphasized that the work in this chapter illustrates that reference common SVA variation is imputable from current WGS data. This is evident as our observed allele frequencies were in line with previous studies of the same population, supporting the fact that bioinformatically we were able to accurately call the SVA. Initial characterization of the polymorphism within specific reference SVAs is a challenge due to their high GC content and repetitive nature of these elements. Reference SVA commonly lie within genic regions of the genome and can modulating gene expression

[126,153]. Through this SVA could be involved in complex disease but this variation is not currently in any reference. Fortunately, here we were able to take advantage of the already extensively studied *PARK7* SVA but to comprehensively address the impact of reference SVA variation in the genome it would not be as straight forward. Each individual harbour a minimum of 2600 reference SVAs in their genome and the variation within these domains is completely uncatalogued. Thus, to assess the impact of reference SVA variation genome-wide this analysis would require extensive resources. The work in this chapter illustrates that this can be minimized if the genome wide SVA analysis is performed in DNA that has existing WGS. For example, one could characterize a subset of SVA of interest in a small sample size and if proxy SNPs can be generated then this would allow for imputation and hence allow accurate calling of the SVA in large-scale WGS dataset. This would mean the analysis was scalable, able to include rarer variants and also that the analysis had the power to address association with disease.

In summary in this chapter we expanded on the existing characterization of a variable reference SVA-D at the PD associated *PARK7* locus. We confirmed that a longer, brain-specific transcript was highly expressed from the distal *PARK7* promoter which is adjacent to the SVA-D. In addition, we addressed whether the SVA could direct differential PARK7 expression in the brain in an allele-like manner. Although our initial analysis did not find the SVA to be a significant eQTL or associated with PD risk we could not test the status of the rarest allele and therefore future studies using WGS are needed. Most importantly this chapter highlights that reference SVA variation can be imputed which will allow for large-scale analysis in future studies.

Chapter 3

A haplotype-block based genome-wide analysis of SVAs in the reference and non-reference genome

3.1 Introduction

Previous studies have established that SVAs preferentially insert into genic regions of the genome, with reports that ~ 60% reside in a gene or +- 10kb[154]. In addition, SVAs are reported to be enriched in regions of high GC density, which in all, may suggest that they preferentially insert into transcriptionally "active" regions. SVAs are the most recently evolved non-LTR TEs and therefore they are a huge source of human-specific variation in the human genome [97,154]. Human-specific TE insertions are a considered to be on the two key driving forces in evolution of human-specific regulatory networks[155]. This evolution of more complex gene regulatory mechanisms in humans is likely to contribute to species and tissue specific gene regulation that allowed diversity with regard to epigenetic modulation and response to environmental changes [156]. Currently there are at least 2600 known SVA in the reference that are "fixed" in every individual's genome and as we have described in the later chapter, reference SVA harbour common variation that is yet to be catalogued.

In addition to the SVA in our reference genome (that can be variable in repeat size), with the advances of bioinformatic and sequencing techniques, it is now evident and well documented that many more "non-reference" SVAs exist[157]. These non-reference SVAs are usual rare in the human population and the observed variation is predominantly present/absent in an individual rather than overall length of SVA. Non-reference TE insertions have been implicated in over one hundred different forms of Mendelian disease [157,158]. Further non-reference TE have been shown to act as eQTL[159] and can contribute to complex genetic disease risk [160].

In regard to non-reference SVAs specifically, a relevant example of how a non-reference SVA insertion can contribute to disease is the *de novo* SVA insertion in intron 32 of the TAF1 gene which is causative of XDP. XDP is a form of Parkinsonism that is endemic to the Philippines. The SVA insertion causes aberrant TAF1 transcription through alternative splicing and intron retention which leads to the disease [161]. In addition, not only is the SVA shown to be disease causing but the variation within the SVA itself is shown to be disease modifying. *Bragg et al* directly linked sequence variation within the XDP-specific SVA sequence to phenotypic variability in clinical disease manifestation. Therefore, non-reference TEs can contribute to both Mendelian and complex forms of genetic disease and can be both disease causing and modifying [103].

Sporadic PD is a complex genetic disease for which the underlying genetic mechanism is still unknown. Pre GWAS, only the genes that harboured mutations that caused Mendelian forms of PD were reported as being "PD associated". Noting that a reference SVA was located upstream of *PARK7* (a PD associated gene) *Savage et al* performed enrichment analysis and showed that reference SVAs were over-represented at the five recognized Mendelian PD genes[162].

Now, following the most recent PD GWAS meta-analysis, which consisted of over one million individuals, 90 independent risk loci have been identified[1]. Although this international-based effort more than doubled the number of PD variants identified, it is still not understood how these variants contribute to disease risk. In fact, the majority of the hits lie within non-coding regions of unknown function. Therefore, further characterization of the risk loci is crucial to identifying how these regions are involved in PD.

In light of the growing body of evidence that identifies that TEs are associated with differential gene expression and Mendelian and complex genetic disease, in this chapter we set out to 1) deeply characterize reference and non-reference SVAs in the human genome to identify for the first time if the latter follow the same distribution pattern and 2) using the later information identify if SVA insertions are enriched at PD risk loci.

3.2 Aims

- Assess the genome-wide distribution of reference SVA using a haplotype block analysis to address insertion preferences
- Replicate using non-reference SVA to identify if this human specific genetic variation shares the same insertion preference as reference SVA
- Annotate reference SVA to gain insight on the global regulatory potential
- Identify distribution of reference and non-reference SVA at PD risk loci

3.3 Methods

3.3.1. Haplotype block analysis

To characterise SVA distribution across the genome and to assess whether insertions were enriched in specific regions (such as PD risk loci) we split the genome into haplotype blocks. Our approach differed from other studies that rather divide the genome into uniform 1MB regions. Haplotype blocks for the human genome, GRch37/hg19, were previously defined by *Berisa et al*, which were extracted and used for the analysis[163], for the European population, as this is the origin of PD risk meta-analysis that identified the 90 risk locus. This resulted in 1703 blocks in total with an average block size of 1.6 Mb (SD = 1.2Mb).

3.3.1.1 Calculating reference SVA and non-reference RIP SVA density

For the reference SVAs the positions of all repetitive elements in the genome were generated using the RepeatMasker GRch37/hg19 Library downloaded from the UCSC genome browser:

http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/chromOut.tar.gz

Here the GRch37/hg19RepeatMasker track was used to annotate SVA position rather than the most recent annotation (GRC38, hg38) as the most recent annotation contains many SVA “fragments” which are < 1kb in length. This likely means that these are not conical SVAs and do not contain the functionally important repetitive domains such as the VNTR region. Therefore, for the basis of this initial analysis which is focused on the potential regulatory potential of these elements the fragments should be excluded. The coordinates for SVAs were extracted from the RepeatMasker track, which gave a total of 2676 elements. The SVA positions were next overlaid with the defined haplotype blocks and the number of SVAs per haplotype block calculated using the 'countOverlap' function in the R package 'GenomicRegions'. As the haplotype blocks differed in size, the number of SVAs per block was scaled, whereby the number of SVAs was divided by the encompassing block size (bp). This was repeated for the non-reference TE elements which were instead extracted from the Ewing non-reference TE resource, with a total of 640 non-reference SVA[157]:

https://figshare.com/articles/Additional_file_2_Table_S1_of_Transposable_element_detection_from_whole_genome_sequence_data/4418360/1

3.3.1.2 Calculating GC content

GC content data for every haplotype block in the genome was downloaded from the UCSC genome browser GRch37/hg19 in which GC content was calculated per every 5bp of the genome.

http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.gc5Base.wib

Thus, to determine the GC content for each haplotype block all the encompassing 5bp regions within a block were combined and the mean of each block was calculated with R.

3.3.1.3 Calculating open chromatin region density

To assess SVA distribution in regions of open chromatin we utilised the ENCODE open chromatin synthesis dataset generated by; Duke University's Institute for Genome Science and Policy, University of North Carolina at Chapel Hill, University of Texas at Austin, European Bioinformatics Institute and University of Cambridge, Department of Oncology and CR-UK Cambridge Research Institute.

This data consisted of DNase 1 sites across 15 cell lines using DNase-chip. All regions with $p < 0.05$ were extracted and classed as significant DNase1 hypersensitive sites. All hypersensitive sites for all samples were pooled together and merged used bedtools[164]. The positions were next overlaid with the defined haplotype blocks and the number of open chromatin regions per haplotype block was calculated using the 'countOverlap' function in the R package 'GenomicRegions'. Again, as the haplotype blocks differed in size, the number of open chromatin regions per block was scaled by dividing total number of open chromatin regions with block size.

3.3.1.4 Calculating gene density

To calculate gene density per haplotype block RefSeq gene annotations were downloaded from the UCSC genome browser (GRch37/hg19) and intersected with the haplotype block regions using the 'countOverlap' function in the R package GenomicRegions and was scaled by dividing the number of genes in each block by the block size (bp).

3.3.1.4 Defining PD risk regions

The positions of the 90 independent PD risk variants were extracted and overlaid with haplotype blocks so that blocks could be classified as PD and non-PD associated risk regions. In some instances, multiple variants lay in the same haplotype block. Consequently 85 haplotype blocks of the 1703 genome-wide were classified as 'PD associated risk regions.' Descriptive statistics for the PD and non-PD blocks are given in Table 3.1

Table 3.1. Descriptive statistics for PD and non-PD haplotype blocks

		Non-PD	PD
	N	1618	85
Average	Block size (Mb)	1.6 (1.18)	1.8 (0.75)
	SVA (per Mb)	0.80 (1.13)	1.30 (1.30)
	Gene density (per Mb)	23.30 (2.27)	37.80 (35.60)
	Open chromatin regions	566.01(202.54)	625.13(168.50)
	GC %	42.02 (3.56)	41.48 (2.81)

All measures in mean (standard deviation)

3.3.1.5 Statistical analysis

All statistical analyses were performed in the R Statistical environments (R version 3.4.1 (2017-06-30). Once all data was incorporated into the linear regression model (scaled SVA, scaled open chromatin density, scaled gene density and mean GC content) the association between these factors and SVA density was assessed with the following formula:

scaled_reference_sva ~ scaled_open_chromatin_density + scaled_gene_density + mean_GC_content

In addition to identify if SVAs were enriched in the defined PD risk associated regions the following formula was used:

scaled_reference_sva ~ PD_risk_region + scaled_gene_density + mean_GC_content +
scaled_open_chromatin_density

3.3.2. ANNOVAR annotation

SVAs were annotated with ANNOVAR software (v2.1.1) or gene and region-based annotation. The gene-based annotation was used to infer whether the variant was exonic , intronic, splicing, 3'- untranslated region (UTR), 5'-UTR, or intergenic[165]. Region-based annotation was used to infer whether the SVA overlapped with variants already associated with a phenotype through existing GWAS studies using the GWAS catalogue and also whether the SVA overlapped with ENCODE-annotated regions. The ENCODE -annotation is a powerful resource which we used to identify if the SVA were located in regulatory regions such as, enhancers, repressors, promoters, or insulators etc. The chromHMM predictions were downloaded for the GM12878 B-lymphocyte cell line which was generated by the International HapMap Project (accessible through GEO Series accession number (GEO: GSE53628). The chromHMM data annotates 15 possible states which are described in (Table 3.2)

Table 3. 2 Descriptive of HMM chromatin state

HMM Chromatin state
State 1 - Active Promoter
State 2 - Weak Promoter
State 3 - Inactive/poised Promoter
State 4 - Strong enhancer
State 5 - Strong enhancer
State 6 - Weak/poised enhancer
State 7 - Weak/poised enhancer
State 8 – Insulator
State 9 - Transcriptional transition
State 10 - Transcriptional elongation
State 11 - Weak transcribed
State 12 - Polycomb-repressed
State 13 - Heterochromatin; low signal
State 14 - Repetitive/Copy Number Variation
State 15 - Repetitive/Copy Number Variation

3.4 Results

3.4. Genome-wide analysis of the distribution of reference and non-reference SVA elements

It has been repeatedly reported that SVA elements insert into regions of the genome that have high GC content, are gene dense and contain regions of open chromatin. To date, all previous studies into insertion preference have focused on the reference SVA elements that are "fixed" in the genome and so are defined by the GRch37/hg19 reference annotation. However, following the improvement of TE detection tools and an increased recognition of the importance of TEs, databases and resources have now been curated that inform of the presence of newly identified TEs that have been detected through thousands of independent NGS initiatives. Non-reference TE are a major source of human-specific variation in the genome and have recently been associated with already identified disease loci. Ewing *et al* have provided one of the most extensive and descriptive non-reference TE resources. Utilizing this resource and using a haplotype block-based genome-wide analysis we extracted the SVA non-reference TEs to determine if this human-specific form of TE variation followed the same distribution bias as the known reference SVA elements.

To asses TE insertion preference previous studies have traditionally divided the genome into 1MB regions for investigation. However, in our analysis we focused on splitting the genome into haplotype blocks (defined by *Berisa et al)*, with the rationale that defining regions based on LD-aware cut offs, rather than the uniform "1MB" would be more suitable approach for investigating genomic factors. Next SVA coordinates were extracted from UCSC to note the “reference” SVA locations and the Ewing non-reference TE resource was used to extract the “non-reference” SVA locations. For each haplotype block; mean GC content, gene density, number of open chromatin regions, reference SVA content and non-reference SVA content was calculated. Finally, a linear regression model was constructed in order to investigate possible association between the stated genomic factors and SVA content.

Following our haplotype block analysis, we report that our data is in line with what have been previously reported for insertion distribution for reference SVA elements, i.e. SVA elements are positively correlated with gene density and open chromatin regions. However, we did not detect a significant association between GC content and SVA density in our model (p= 0.125, β = - 0.035). The most influential genomic factor for determining SVA content was gene density (p =2.34E-50, β = 0.036), followed by number of regions of open chromatin (p =6.64E-04, β = 0.079) (Table 3.3 & Figure 3.1).

Table 3.3. Linear regression model showing reference and non-reference SVAs are most enriched in gene dense regions of the genome

	Reference SVA		Non-reference SVA	
	p	β	p	β
Gene density	2.34E-50*	0.364	4.72E-20*	0.230
Regions of open chromatin	6.64E-04*	0.079	5.06E-02*	0.001
GC	1.26E-01	-0.035	9.58E-01	0.047

*** Indicates significance (p-value < 0.05)**

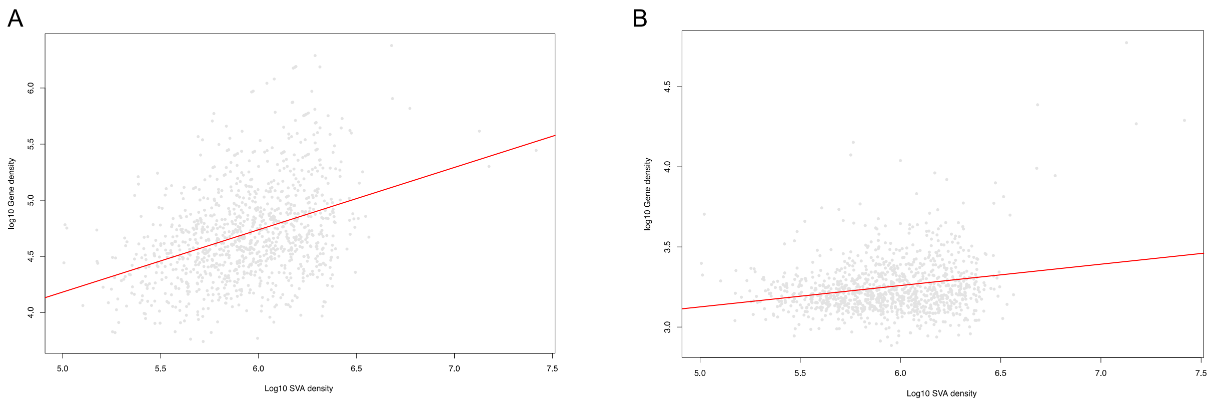


Figure 3.1. Gene (A/y-axis) and open chromatin regions density (B/y-axis) are positively correlated with reference SVA density (x-axis). Grey dots represent the log10 of SVA density for each haplotype block in the genome and the red line represent the line of regression.

Using the same approach, we constructed a model with non-reference SVA to assess if these elements mirrored this distribution bias. Although less significant, we observed the same insertion preference as the most influential genomic factor for determining non-reference SVA content was gene density (p =4.72E-20, β = 0.023), followed by number of regions of open chromatin (p =5.06E-02, β = 0.001). A list of the ten most SVA and non-reference SVA dense haplotype blocks, i.e the blocks in which SVAs most cluster can be found in (Table 3.4 & Table 3.5).

Table 3.4. The top ten most reference SVA dense haplotype blocks in the genome (hg/19).

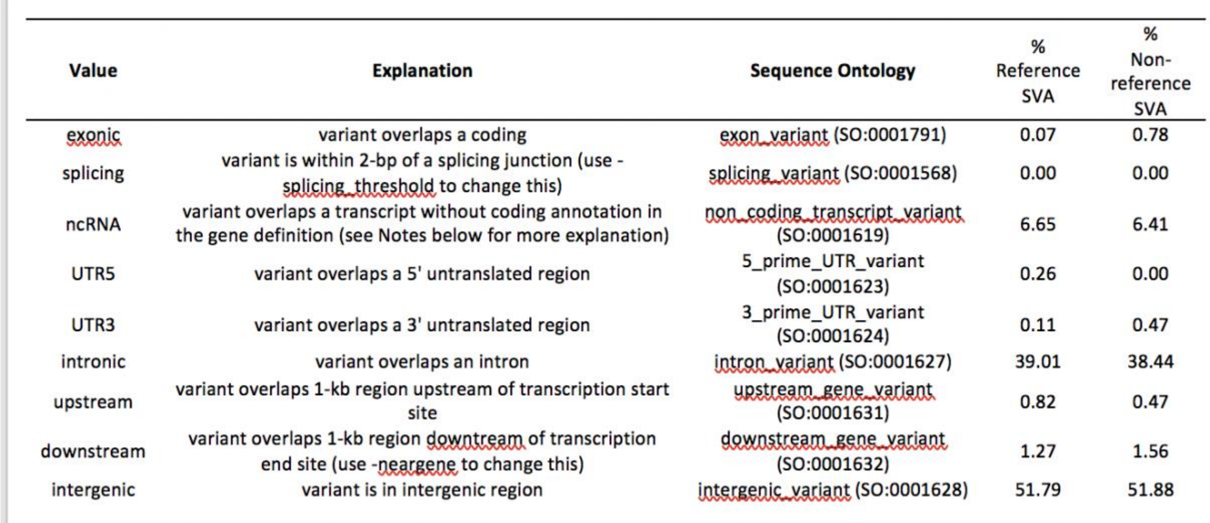
Chr	Start	Stop	Blocksize	SVA count	Gene count	Gc Mean	Scaled SVA	Scaled gene density	Scaled open Chromatin density
chr19	20905757	22732896	1827139	18	66	47	9.85E-06	3.61E-05	4.00E-04
chr19	19877471	20905757	1028286	10	17	47	9.72E-06	1.65E-05	4.74E-04
chr19	22732896	23467746	734850	7	13	47	9.53E-06	1.77E-05	1.97E-04
chr7	972752	1353067	380315	3	33	41	7.89E-06	8.68E-05	5.15E-04
chr9	6557589	7154923	597334	4	11	41	6.70E-06	1.84E-05	9.11E-04
chr20	31614823	32813441	1198618	8	42	46	6.67E-06	3.50E-05	6.15E-04
chr7	139933177	140235210	302033	2	11	43	6.62E-06	3.64E-05	4.44E-04
chr1	44969183	46899501	1930318	12	115	40	6.22E-06	5.96E-05	6.51E-04
chr4	10240	694715	684475	4	56	38	5.84E-06	8.18E-05	5.13E-04
chr7	2062398	2772227	709829	4	44	40	5.64E-06	6.20E-05	5.95E-04

Table 3.5. The top ten most non-reference SVA dense haplotype blocks in the genome (hg/19).

Chr	Start	Stop	Blocksize	Scaled non-ref SVA	Gene count	Gc Mean	Scaled open Chromatin density
chr7	5416232	5854526	438294	4.56E-06	11	40	6.62E-04
chr19	36469295	37527033	1057738	3.78E-06	109	48	4.59E-04
chr11	70926292	72286017	1359725	3.68E-06	77	41	5.24E-04
chr4	9326479	10699152	1372673	3.64E-06	76	38	4.20E-04
chr6	32682664	33236497	553833	3.61E-06	51	39	7.53E-04
chr6	31571218	32682664	1111446	3.60E-06	168	39	6.87E-04
chr10	69900148	70195991	295843	3.38E-06	37	43	6.66E-04
chr19	19877471	20905757	1028286	2.92E-06	17	47	4.74E-04
chr5	79393144	80481471	1088327	2.76E-06	27	40	7.04E-04
chr11	8333274	9087317	754043	2.65E-06	41	41	7.36E-04

To further understand on a gene-level if non-reference TE distribution followed the same pattern as reference SVAs we used ANNOVAR to annotate the positions of both the reference and non-reference TE SVAs. ANNOVAR annotates whether the SVAs lie within exons, intergenic regions, introns, or non-coding RNA genes. In support of previous studies our data identified that reference SVAs predominantly reside within intergenic regions of the genome (n = 1386, 51.79%), followed by introns (n = 1044, 39.01%) and finally exons (n = 2, 0.07%) (Table 3.6).

Table 3.6. ANNOVAR gene-based annotation for reference and non-reference SVAs.



Value	Explanation	Sequence Ontology	% Reference SVA	% Non-reference SVA
exonic	variant overlaps a coding	exon_variant (SO:0001791)	0.07	0.78
splicing	variant is within 2-bp of a splicing junction (use -splicing_threshold to change this)	splicing_variant (SO:0001568)	0.00	0.00
ncRNA	variant overlaps a transcript without coding annotation in the gene definition (see Notes below for more explanation)	non_coding_transcript_variant (SO:0001619)	6.65	6.41
UTR5	variant overlaps a 5' untranslated region	5_prime_UTR_variant (SO:0001623)	0.26	0.00
UTR3	variant overlaps a 3' untranslated region	3_prime_UTR_variant (SO:0001624)	0.11	0.47
intronic	variant overlaps an intron	intron_variant (SO:0001627)	39.01	38.44
upstream	variant overlaps 1-kb region upstream of transcription start site	upstream_gene_variant (SO:0001631)	0.82	0.47
downstream	variant overlaps 1-kb region downtream of transcription end site (use -neargene to change this)	downstream_gene_variant (SO:0001632)	1.27	1.56
intergenic	variant is in intergenic region	intergenic_variant (SO:0001628)	51.79	51.88

We also report that 34 reference SVAs (1.27%) reside upstream of a gene (+ 1kb) and 22 downstream (0.82%) (-1kb). In addition, we identify that many reference SVAs are located within non-coding RNA genes (n = 178, 6.65%) and 3’ and 5’ UTRs (n = 7, 0.26% and n =3, 0.11% respectively). This distribution pattern was mirrored by non-reference SVAs, although to note, more SVAs were found within exons (n= 5, 0.78%). This observed increase in insertion into exons is most likely due to the fact that the non-reference resource has been curated from thousands of NGS studies, which traditionally have a bias towards the study of exomes. Therefore, these regions have higher sequencing depth and coverage and so are more likely to detect TE variants.

3.4. SVA elements lie within regulatory regions of the genome

SVAs have been shown to act as independent regulatory domains that drive differential gene expression in an allele and tissue-specific manner. Although this has mainly been demonstrated with studies focused on characterizing the regulatory potential of specific reference SVAs on an individual basis. Therefore, to gain further insight into the potential influence of all reference SVAs we conducted a genome-wide analysis utilising chromatin state ENCODE data to begin to understand the global regulatory potential of these elements.

The regions of the SVAs in the reference genome (NCBI build 36.1/hg18) were annotated using ANNOVAR against the ENCODE ChromHMM predictions for the GM12878 cell line. Although the GM1878B-lymphocyte cell line is not the most relevant for addressing transcriptional activity in the brain, it was used in this analysis it is a well-curated resource that would allow general transcriptional activity genome-wide. This epigenetic resource is used to identify non-coding variation that disrupts enhancers, repressors and promoters by classifying regions into the 15 core chromatin states described in (Table 3.7). As expected, our data identifies that the majority of reference SVA elements (64.74%) lie within regions of heterochromatin in this particular cell line and so in theory should be heavily silenced in the genome. Interestingly nearly a quarter of all SVAs are weakly transcribed (24.97%) and 3.57% lie within active regulatory regions such as promoters, enhancers and repressors. Finally, the remaining 2.52% were classed as “repetitive/copy number variation” regions, which by ChromHMM classification means that there is an abundance of nearly all marks and the regions typically fall within repetitive sequences.

Table 3.7. ANNOVAR functional annotation for reference SVAs.

HMM Chromatin state	% Reference SVA
State 1 - Active Promoter	0.08
State 2 - Weak Promoter	0.26
State 3 - Inactive/poised Promoter	0.08
State 4 - Strong enhancer	0.04
State 5 - Strong enhancer	0.08
State 6 - Weak/poised enhancer	0.90
State 7 - Weak/poised enhancer	1.31
State 8 – Insulator	0.15
State 9 - Transcriptional transition	0.19
State 10 - Transcriptional elongation	4.21
State 11 - Weak transcribed	24.97
State 12 - Polycomb-repressed	0.49
State 13 - Heterochromatin; low signal	64.74
State 14 - Repetitive/Copy Number Variation	2.14
State 15 - Repetitive/Copy Number Variation	0.38

3.4. Known GWAS variants are located within reference SVA elements

We annotated reference SVAs using the GWAS catalogue to identify if they mapped to regions that’s have already been associated with a given phenotype by previous GWASs. It should be mentioned here that although we have highlighted that the allele length of SVAs has been shown to be both diseases associated and modifying, this form of variation is currently not captured in any existing reference panel. Therefore, the variation that is reported in the GWAS catalogue is most likely a single nucleotide polymorphism rather than an association due to sequence length of the SVA. Nonetheless assessing overlap with existing GWAS hits is relevant as 1) the majority of GWAS hits for complex diseases represent loci rather than specific base pair changes 2) allele-size of the SVA could still be tagged by the SNP if that is the true signal and 3) the individual regulatory domains of an SVA such as the VNTR or CT element could be modified if there is just a single base pair change and this could affect function.

According to the ANNOVAR GWAS catalogue annotation we report that reference SVAs harbour 77 variants previously reported by GWASs as associated with a phenotype. A full list of the overlapping GWAS associations can be found in Table 3.8. The list includes multiple associations with risk of neurological (e.g. Schizophrenia, Post-traumatic stress disorder and Alzheimer's (AD)), autoimmune and inflammatory disease. The list also includes associations that have been reported as disease modifying, such as age of onset and accelerated cognitive decline in AD.

.

Table 3.8. ANNOVAR GWAS catalogue annotation. Reference SVAs harbour 77 variants previously reported by GWASs as associated with a phenotype

		hg/18 ref SVA coordinates	
GWAS catolog reference	**CHR**	**start**	**stop**
Inflammatory skin disease	chr1	12067823	12069304
Hair shape,Male-pattern baldness	chr1	152116612	152118218
Inflammatory skin disease	chr1	152167947	152169745
Daytime sleep phenotypes	chr1	172540262	172541856
Pediatric autoimmune diseases	chr1	197374169	197375722
Mean corpuscular volume	chr1	198552755	198553613
Post-traumatic stress disorder	chr1	202159959	202161610
Serum alkaline phosphatase levels	chr1	21349715	21351405
Cognitive empathy	chr1	247194848	247196269
Heel bone mineral density	chr1	27143733	27145832
	chr10	35264856	35266229
Mean corpuscular hemoglobin,Red blood cell count,Cholesterol, total	chr10	46006824	46008669
DNA methylation variation (age effect)	chr10	97568477	97570296
Inflammatory skin disease	chr10	99911664	99913385
Heel bone mineral density	chr11	47118555	47119876
Depressed affect	chr11	47913823	47915299
Schizophrenia,Autism spectrum disorder or schizophrenia	chr11	57483053	57485227
Sum eosinophil basophil counts,Allergic rhinitis,Food allergy,Eosinophil counts,Asthma,Allergic sensitization,Allergic disease (asthma, hay fever or eczema),Eosinophil percentage of white cells,Neutrophil percentage of granulocytes,Peanut allergy,Eosinophil percentage of granulocytes,Allergy	chr11	76292966	76294455
Lung function (FEV1)	chr11	86442568	86444195
Magnesium levels	chr12	48944995	48946287
Anorexia nervosa	chr12	56468875	56470558
Monocyte count,Granulocyte percentage of myeloid white cells,NT-proBNP levels in acute coronary syndrome	chr12	89896855	89898398
Accelerated cognitive decline after conversion of mild cognitive impairment to Alzheimer's disease (Alzhiemer's diagnosis trajectory interaction)	chr13	33289761	33291843
White blood cell count (basophil)	chr14	23576903	23577709
Breast cancer	chr14	32596377	32597593
Tonsillectomy,Intraocular pressure	chr14	38023856	38025752
Lung adenocarcinoma	chr15	49841523	49843482
Post bronchodilator FEV1/FVC ratio in COPD,Post bronchodilator FEV1,Nicotine dependence,Post bronchodilator FEV1/FVC ratio	chr15	78812597	78813341
Schizophrenia	chr16	89872756	89873422
Highest math class taken	chr17	45497278	45499010
Schizophrenia	chr17	78510656	78511882
3-hydroxypropylmercapturic acid levels in smokers	chr18	14009955	14011906
Granulocyte percentage of myeloid white cells,Monocyte count,Monocyte percentage of white cells	chr19	18115825	18117465
Late-onset Alzheimer's disease	chr19	20174330	20176102
Self-reported math ability	chr19	38613904	38615679
Anti-saccade response	chr19	39345610	39347376
Blood protein levels	chr19	45424339	45425280

Blood protein levels	chr19	55419344	55420010
Sporadic neuroblastoma	chr2	215699019	215701152
Mean corpuscular hemoglobin concentration	chr2	25945507	25947023
Heel bone mineral density	chr2	26132611	26134535
Aspartate aminotransferase levels	chr2	27779578	27781830
Heel bone mineral density,Total body bone mineral density	chr2	42240869	42242785
Glaucoma (primary open-angle)	chr2	55932823	55933562
Dysmenorrheic pain	chr2	85714115	85716186
Glomerular filtration rate in chronic kidney disease	chr2	86775310	86777587
Heel bone mineral density	chr20	32286283	32288520
Cancer,Cancer (pleiotropy)	chr20	32719596	32721485
Blood protein levels	chr20	32816398	32819224
Blood protein levels	chr20	37006728	37008588
Blood protein levels	chr22	36686483	36687742
Interleukin-1-receptor antagonist levels	chr3	129057288	129059709
Depressive symptoms (MTAG),Depressive symptoms	chr3	174805020	174805714
White blood cell count	chr3	196488023	196489567
Ankle injury	chr3	20588777	20590323
Coronary artery calcified atherosclerotic plaque (90 or 130 HU threshold) in type 2 diabetes	chr3	39696718	39698146
DNA methylation variation (age effect)	chr3	50348025	50350444
Blood protein levels	chr3	9981015	9982467
Triptolide cytotoxicity	chr4	64902315	64903851
Hepatitis A	chr5	157463380	157465107
Post bronchodilator FEV1/FVC ratio in COPD	chr5	43745274	43747246
Pulse pressure	chr5	43823986	43825510
High light scatter reticulocyte percentage of red cells,High light scatter reticulocyte count	chr5	53676571	53678077
Interleukin-18 levels	chr5	68516327	68517028
Lung function (FVC),Lung function (FEV1)	chr5	77391828	77394120
Depressive symptoms (MTAG)	chr6	105438437	105439902
Heel bone mineral density	chr6	152219535	152221254
Photic sneeze reflex	chr6	165156588	165158141
Autism spectrum disorder or schizophrenia	chr6	27620654	27622181
Itch intensity from mosquito bite	chr6	29899781	29901531
General cognitive ability	chr6	30189273	30190496
Blood protein levels	chr6	31943432	31944815
Plateletcrit	chr7	129302607	129304028
Eotaxin levels	chr7	75581285	75582657
Facial morphology (factor 5, width of mouth relative to central midface)	chr9	136979898	136984303
Alzheimer disease and age of onset	chr9	79675073	79676561
Eosinophil percentage of white cells	chr9	86418137	86420398
Facial morphology (factor 14, intercanthal width),Eye morphology	chrX	72287130	72290389

3.4. PD risk loci are enriched with SVA elements due to their genic nature

It has previously been shown that SVA elements are over-represented at genes that can cause monogenic forms of PD[4]. However, if all known monogenic forms are combined this only explains around 30% of Mendelian and 3–5% of genetically complex PD cases. Fortunately for the remaining (genetically unexplained) cases substantial progress in understanding the genetic basis of the disease has been made in recent years. As a result, 90 PD associated independent risk loci have now been identified. Despite this significant step forward in our understanding of these genetically complex cases, at this stage little is known about how these risk loci contribute to disease mechanism. We have just shown that these elements have regulatory potential genome-wide therefore we characterized the SVA content within PD risk loci. We hypothesised that it could be possible that SVAs could be playing a role, even a concerted role, in regulation of these identified risk loci and thus the genetic mechanism of PD.

Using the haplotype block analysis, we determined that overall PD loci (defined by the encompassing haplotype block of the 90 risk variants) contained more reference SVAs than non-PD blocks (Figure 3.2), which was not observed with non-reference SVAs, although this list could be incomplete. However, as previously described the most important factor in determining SVA content is gene density. Therefore, we reasoned that this over-representation could simply be due to the nature of previous PD GWAS studies, which have been bias towards covering genic regions. In light of this we adjusted our model for the known insertion preferences which included; gene density, GC content and regions of open chromatin. Consequently, when we adjusted for these covariates, we found no significant association between reference SVA content and PD risk loci.

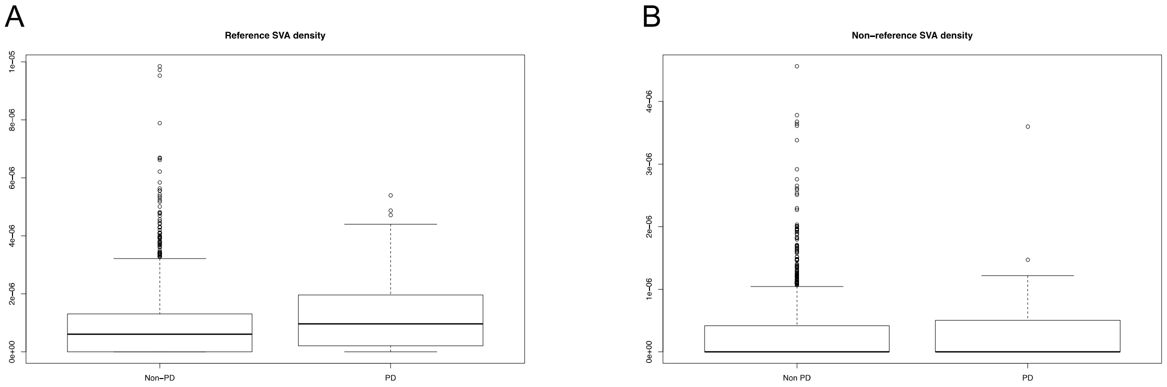


Figure 3. 2. Reference SVA are over-represented at PD risk loci A) Plot of the SVA density within PD risk loci Vs non-PD associated haplotype blocks. y=scaled SVA count (SVA count per block/ size of block). B) in comparison non-reference SVAs are not over-represented (p=4.58E-01).

Table 3.9. Linear regression model showing reference SVAs are most enriched in high gene and chromatin site density

	Ref SVA		Non-ref SVA	
	P	β	P	β
Gene density	2.14E-50*	3.61E-01	4.40E-20*	2.48E-01
Regions of open chromatin	6.40E-04*	7.92E-02	4.66E-02*	7.92E-02
PD meta 5 risk loci	2.73E-01	2.48E-02	4.79.E-01	-1.68E-02
GC	1.47E-01	-3.35E-02	5.05E-01	1.68E-02

*** Indicates significance (p-value < 0.05)**

So, in sum, PD blocks contain more reference SVAs than non-PD blocks but do not contain significantly more than other regions of similar gene density and open chromatin (Table 3.9) (p=2.73E-01). Regardless, this does not suggest that SVAs aren’t biologically relevant at these loci. On the contrary it is apparent that SVAs within PD risk loci are in regulatory regions and therefore could be playing a regulatory role as demonstrated in (Figure 3.3-3.6). But this does suggest that all currently known GWAS loci will be “enriched” for these elements, not necessarily due to a disease-specific phenomenon, but more likely due to current GWASs being bias towards covering genic regions.

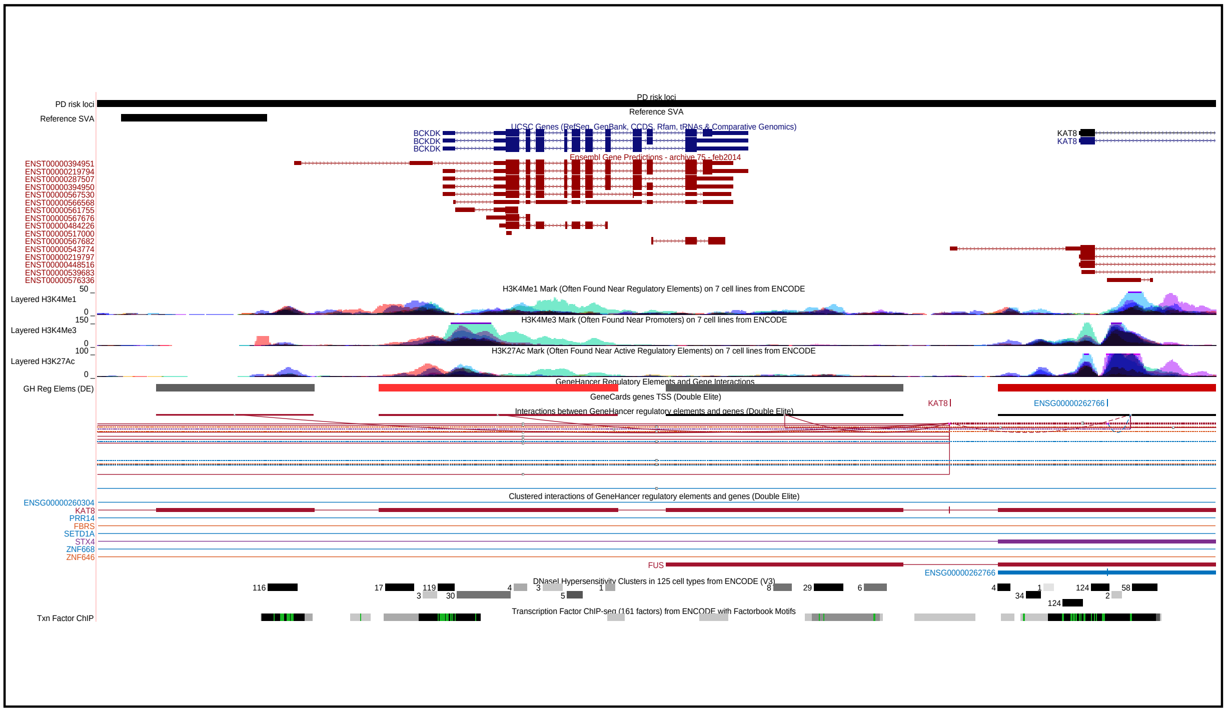

PD risk loci
PD risk loci
Reference SVA
Reference SVA
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
BCKDK
BCKDK
BCKDK
KAT8
KAT8
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000394951
ENST00000219794
ENST00000287507
ENST00000394950
ENST00000567530
ENST00000566568
ENST00000561755
ENST00000567676
ENST00000484226
ENST00000517000
ENST00000567682
ENST00000543774
ENST00000219797
ENST00000448516
ENST00000539683
ENST00000576336
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
50
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
150
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
100
0
GeneHancer Regulatory Elements and Gene Interactions
GH Reg Elems (DE)
GeneCards genes TSS (Double Elite)
KAT8
ENSG00000262766
Interactions between GeneHancer regulatory elements and genes (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
ENSG00000260304
KAT8
PRR14
FBRS
SETD1A
STX4
ZNF668
ZNF646
FUS
ENSG00000262766
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
116
17
119
3
30
4
3
5
1
8
29
6
4
34
1
124
124
58
2
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP

Figure 3.3. UCSC image showing the reference SVA D at n the BCKDK PD risk locus. The UCSC genome browser image highlights the region of the SVA (red). Each track is named. Showing; PD risk region, SVA region, Refseq and Ensembl gene annotation. ENCODE functional annotations of histone marks and Dnase Hypersensitivity also shown. The GeneHancertrack shows interaction between the SVA and the BCKDK gene and adjacent KAT8 gene.

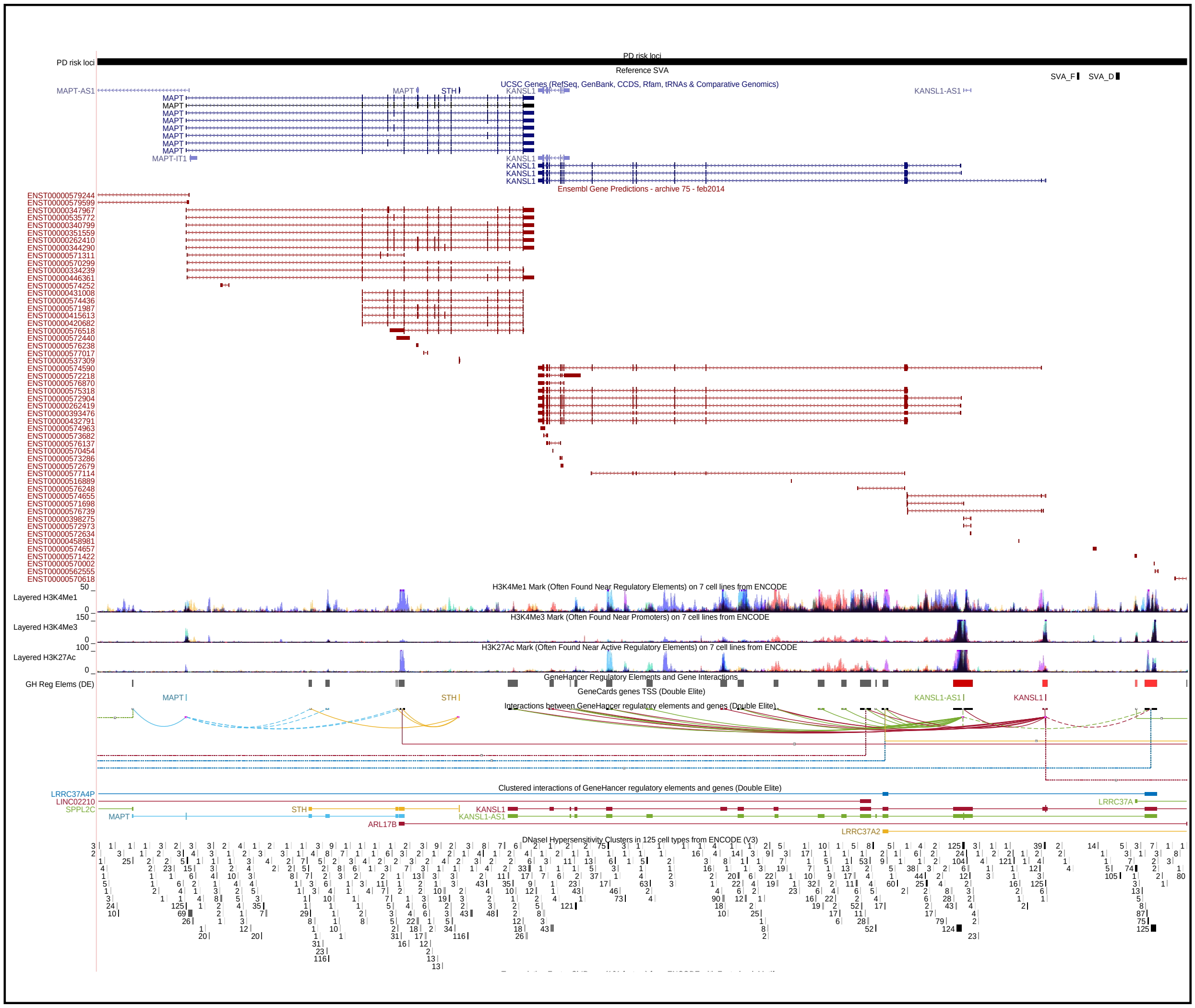


Figure 3. 4. UCSC image showing the reference SVA F and SVA D at the MAPT/KANSL1 PD risk locus: The UCSC genome browser image highlights the region of the SVAs (red). Each track is named. Showing; PD risk region, SVA region, Refseq and Ensembl gene annotation. ENCODE functional annotations of histone marks and Dnase hypersensitivity also shown. The GeneHancertrack shows interaction between the SVA D and the PD associated gene LRRC37A.

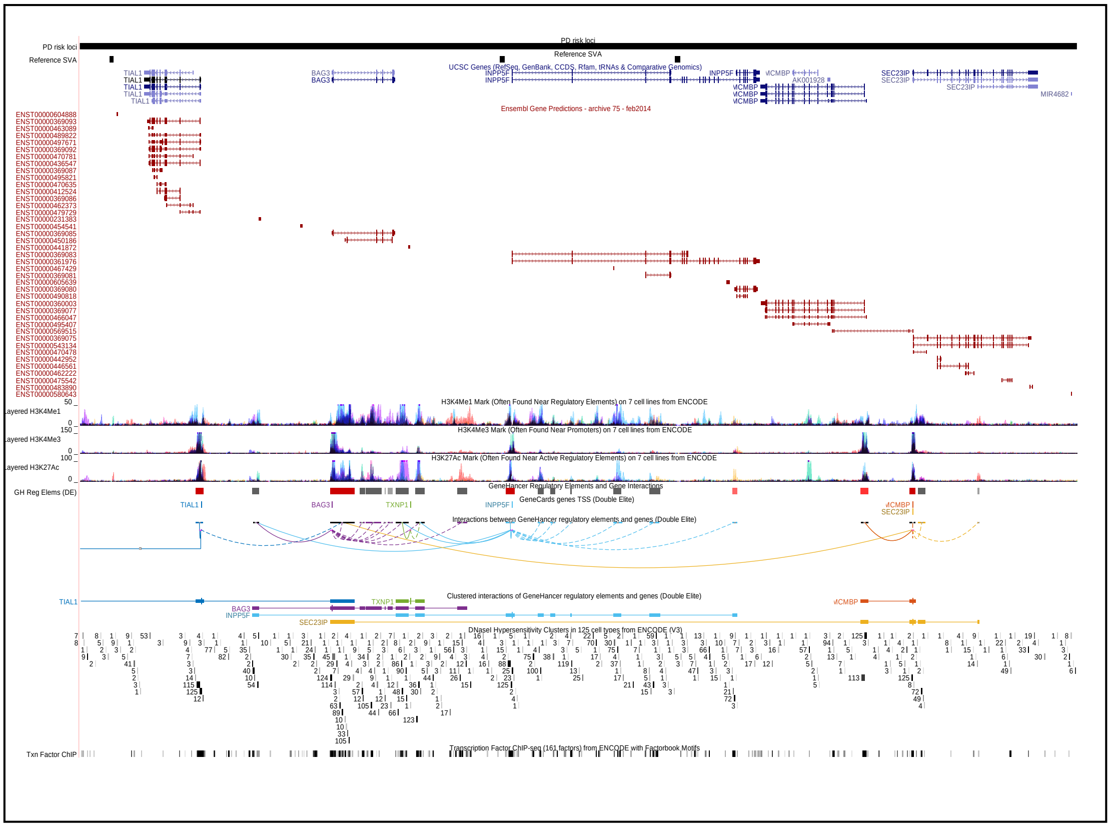

PD risk loci
Reference SVA
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
TIAL1
BAG3
INPP5F
MCMBP
AK001928
SEC23IP
MIR4682
Ensembl Gene Predictions - archive 75 - feb2014
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
GeneHancer Regulatory Elements and Gene Interactions
GH Reg Elems (DE)
GeneCards genes TSS (Double Elite)
TXNP1
Interactions between GeneHancer regulatory elements and genes (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP

Figure 3. 5. UCSC image showing the reference (from left to right) SVA C, SVA F and SVA D at the BAG3/INPP5F risk locus: The UCSC genome browser image highlights the region of the SVAs (red). Each track is named. Showing; PD risk region, SVA region, Refseq and Ensembl gene annotation. ENCODE functional annotations of histone marks and Dnase Hypersensitivity also shown. The SVA C lies upstream of the RGS10 gene promoter a PD associated gene. The SVA F is within the promoter region of the INPP5F gene and the SVA D lies within intron 4.

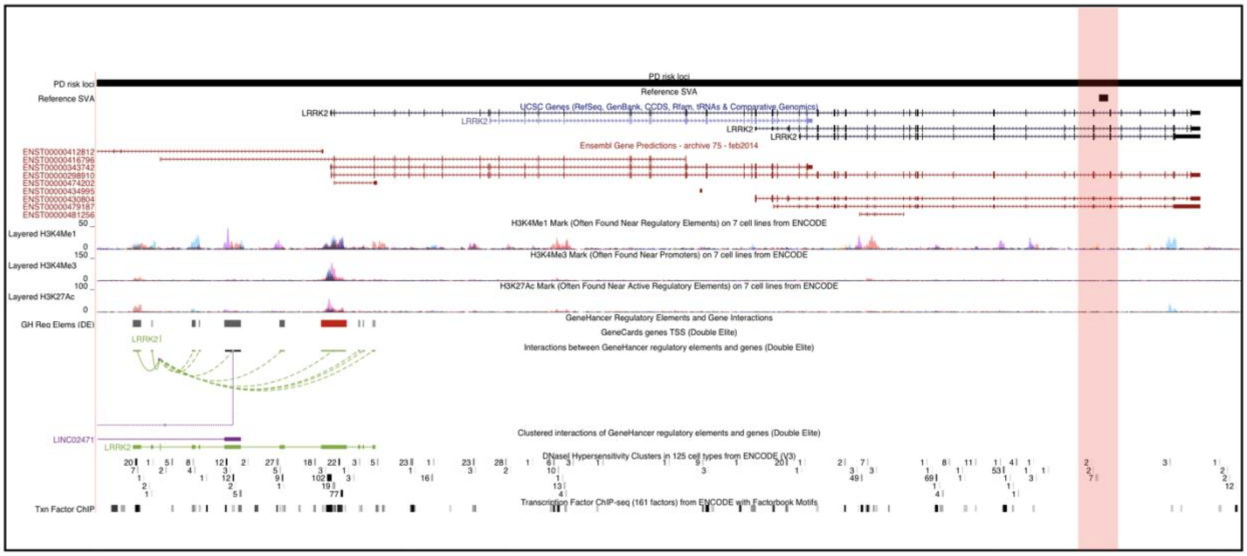

PD risk loci
PD risk loci
Reference SVA
Reference SVA
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
LRRK2
LRRK2
LRRK2
LRRK2
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000412812
ENST00000416796
ENST00000343742
ENST00000298910
ENST00000474202
ENST00000434995
ENST00000430804
ENST00000479187
ENST00000481256
50
Layered H3K4Me1
0
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
150
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
GH Reg Elems (DE)
GeneHancer Regulatory Elements and Gene Interactions
GeneCards genes TSS (Double Elite)
LRRK2
Interactions between GeneHancer regulatory elements and genes (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
LINC02471
LRRK2
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP

Figure 3.6. UCSC image showing the intronic reference SVA C at the LRRK2 PD risk locus: The UCSC genome browser image highlights the region of the SVA (red). Each track is named. Showing; PD risk region, SVA region, Refseq and Ensembl gene annotation. ENCODE functional annotations of histone marks and Dnase Hypersensitivity also shown.

3.5 Discussion

SVAs are not randomly distributed in the human genome. Previous studies have reported that they preferentially insert into regions of high gene density and GC content [82,126]. However, the majority of studies that report this preference have split the genome into uniform one mega base regions. In this present study we adopted a different approach and divided the genome into haplotype blocks with the rationale that defining regions based on LD-aware cut offs, rather than the uniform "1MB" would be more suitable approach for investigating genomic factors. With our haplotype block genome-wide enrichment analysis we confirmed that gene density was the most influential factor for determining SVA density. In addition, it has been suggested that SVA's preferentially insert into regions of active chromatin [166]. Although collectively TEs are reported to preferentially insert into regions of open chromatin this had not yet been systematically addressed for reference SVAs. Thus, in this chapter we built a comprehensive genome-wide map of open chromatin sites and report that SVA density positively correlates with open chromatin site density, consistent with the idea that SVA inserting into active regions of the genome.

As described in the previous chapter for the individual reference SVAs that have already been characterized, SVA are polymorphic regulatory domains that can direct differential gene expression in a tissue and allele specific manner, as exemplified by the SVA-D upstream of PARK7[5,126]. In light of this, in this present analysis all reference SVA were annotated for possible functional consequence, to gain an understanding of the global regulatory potential of these elements. ANNOVAR annotation identified that the majority of reference SVA lie in regions of heterochromatin (64%) implying they are silenced in the genome. However, the analysis also identified that many of the SVA lie within transcriptionally active regions such as sites of strong enhancers. Significantly we also report that one quarter of reference SVA reside in regions that are weakly transcribed, from the GM12878 cell-line ENCODE annotation, which suggests that globally many SVA could be involved in gene regulation of their neighbouring gene. As a limitation to this analysis it should be noted that the current annotation is based on data from only one cell line (GM1878 B-lymphocyte). However, this ENCODE generated cell-line dataset was chosen as it is well-curated and sufficient to gain general insight into the global regulatory potential of reference SVAs. Further studies utilizing and comparing regulatory overlap in different cell-lines would aid in the understanding of the tissue-specific nature of SVA regulatory properties. Reference SVAs were also annotated for GWAS hits to identify if they had already been associated with a phenotype in previous studies. As shown in Table 3.8 reference SVA harbour over 70 existing GWAS variants which have been associated with risk of multiple neurological conditions (such as Schizophrenia, Post-traumatic stress disorder and AD), autoimmune and inflammatory diseases and therefore could be involved in disease mechanism and further analysis into this variation is needed.

For the newly described non-reference SVA's, we identified that the newer SVAs followed the same distribution pattern as the reference SVA. The haplotype block-based genome-wide analysis confirmed that the new non-reference SVA follow the same distribution pattern as reference SVA i.e. into active regions that are gene dense. The polymorphism for non-reference SVA is presence/absence polymorphism rather than difference in repeat size. Therefore, one could imagine that between individuals the insertion of a 3kb regulatory domain has more potential to have an impact on gene expression at that region.

In this chapter we have demonstrated that globally SVA correlate with transcriptionally active regions of the genome. As of yet the variation within the reference SVAs is not incorporated into any reference panel, neither is the huge source of human-specific genetic variation contributed by the non-reference SVA. Expanding on the analysis performed by *Savage et al*, which identified that reference SVAs were over-represented at monogenic PD genes [167], we hypothesized that SVAs more generally could be involved in the aetiology of sporadic PD and set out to systematically identify any potential correlation between SVAs and the currently known PD risk loci. Using regression modelling for all the haplotype blocks in the genome we show that reference SVA are over-represented at PD loci due to their genic nature.

Further in-depth analysis of the SVA containing PD risk loci identified that reference SVA lie within potentially important regulatory domains of PD associated candidate genes, such as BCKDK (branched chain ketoacid dehydrogenase kinase). As shown in Figure 3.3 at the BCKDK locus an SVA-E is located 400bp upstream of the major transcriptional start site. Also as shown in Figure 3.5 three reference SVA are located at the INPP5F/BAG3 (inositol polyphosphate-5-phosphatase F and BCL2 Associated Athanogene 3 respectively) LOCUS; First an SVA C lies upstream of the RGS10 gene promoter a PD associated gene. In addition, an SVA F is within the promoter region of the INPP5F gene and the SVA D lies within intron 4. LRRK2 is a monogenic and sporadic associated PD gene and as shown in Figure 3.6 contains an intronic SVA-C./ Savage et al have already identified that the SVA is polymorphic and therefore this could be contributing to the alternative splicing transcripts at this locus. Finally, the MAPT/KANSL1 locus harbours two reference SVAs upstream of the promoter of KANSL1; an SVA-F and SVA-D (Figure 3.4). SVA are both over-represented and lying in regions that could have an impact on regulatory networks, to further understand how these loci could be contributing to disease risk a large-scale high-throughput analysis of the variation at these loci is required.

In this chapter we have comprehensively assessed the distribution of reference SVAs and confirmed that they can reside in active regions of the genome and therefore have global regulatory potential. We have also shown that many of the reference SVAs have already been associated with risk of disease as they already harbour GWAS variants. Using regression modelling we also show that the newer non-reference SVA follow the same distribution bias as reference SVA. This is significant as these elements are presence/absence in the genome and therefore have a greater potential to impact on gene expression in a distinct manner. Finally, we explored the relationship between SVA and known PD risk loci, expanding on a previous analysis that identified enrichment of reference SVA at monogenic PD genes. We report that due to their genic nature PD loci have a higher number of SVA than non-PD and lie within functional regions that could be involved with disease. Reference and non-reference SVA variation has not been catalogued, so their association with disease is yet to be established. This chapter highlights that to fully understand how PD loci are contributing to disease it is crucial to begin to study and incorporate this source of genetic variation into genetic analysis.

Chapter 4

Mitochondria function associated genes are enriched for young *Alu* elements and these genes contributes to Parkinson disease risk and later age of onset

4.1. Introduction

In the previous chapter, reference and non-reference SVA were characterised at PD risk loci demonstrating that reference SVA were over-represented at-risk loci and potentially involved in regulatory mechanisms at these regions. This analysis was an expansion of previous work from our group that identified that SVA were enriched at monogenic PD genes and capable of directing gene expression in an allele-specific manner. Together with SVA, *Alu* and LINE1 elements also make up the non-LTR TEs, which are capable of modulating transcription and epigenetic parameters. Yet these elements too are completely uncatalogued in the genome.

In this chapter we focus on *Alu* elements specifically, in genes associated with the mitochondria function pathway. The deleterious activity of *Alu* elements has already been implicated as a contributing factor for the manifestation of disease and for many of these disorders this activity is operating on genes that are essential for proper mitochondrial function, which is major pathogenesis of PD.

It has recently been demonstrated that genes that have been associated with mitochondria function are enriched for primate specific *Alu* elements. In light of this *Parson et al* proposed the "*Alu*-neurodegeneration" hypothesis, which suggests that this *Alu* enrichment could in part explain human-specific sporadic neurodegenerative disorders, such as AD and PD[6,168]. The hypothesis being that A*lu* elements can influence efficient and accurate transcription and/or post-transcription of genes through several transposon-induced mechanisms, therefore this enrichment would make this pathway more vulnerable to this *Alu* specific effect (Figure 4.1). An example of this activity is observed in the *TOMM40* gene, encoding a b-barrel protein critical for mitochondrial preprotein transport which plays an essential role in mitochondrial stability. Variants within *TOMM40* have been implicated in a number of neurologic disorders, ranging from mild cognitive impairment to major neurodegenerative diseases including LOAD and PD[169,170]. This region is hypothesized to be vulnerable to Alu-related mechanisms that contribute to genomic instability including alternative splicing and modification of pre-mRNA transcripts which has been associated with disease[171].

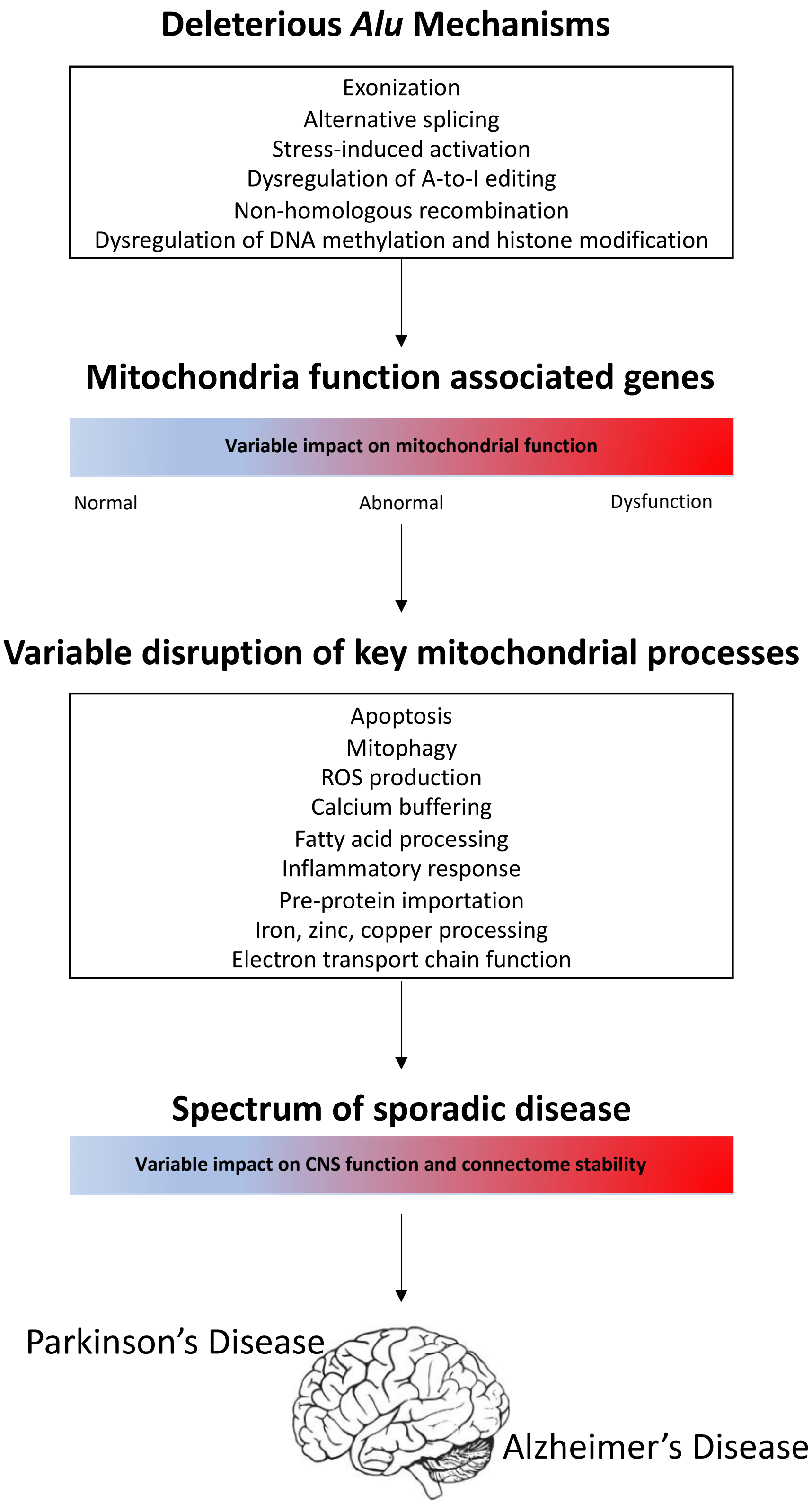

Deleterious *Alu* Mechanisms
Exonization
Alternative splicing
Stress-induced activation
Dysregulation of A-to-I editing
Non-homologous recombination
Dysregulation of DNA methylation and histone modification
Mitochondria function associated genes
Variable impact on mitochondrial function
Normal
Abnormal
Dysfunction
Variable disruption of key mitochondrial processes
Apoptosis
Mitophagy
ROS production
Calcium buffering
Fatty acid processing
Inflammatory response
Pre-protein importation
Iron, zinc, copper processing
Electron transport chain function
Spectrum of sporadic disease
Variable impact on CNS function and connectome stability
Parkinson's Disease
Alzheimer's Disease

Figure 4.1 Deleterious Alu activity impacting on genes associated with mitochondria function can disrupt mitochondrial function in the CNS and contribute to a number of diseased phenotypes. Figure adapted from (*Larsen et al* 2018).

In relation to the mitochondria pathway and PD specifically, although there have been great advances in understanding both the genetic architecture and cellular processes involved in PD, the exact molecular mechanisms that underlie PD remain unknown [172]. However, it is suggested that PD has a complex aetiology, involving several molecular pathways, and understanding these specific pathways will be key to establishing mechanistic targets for therapeutic intervention. While several key pathways are currently being investigated, including autophagy, endocytosis, immune response and lysosomal function, [20,25,173,174]mitochondrial function was the first biological process to be associated with PD [175,176].

An interest in mitochondrial function and PD began with the observation that exposure to the drug 1-methyl-4-phenyl-1,2,3,4-tetrahydropyridine (MPTP) can cause rapid parkinsonism and neuronal loss in the substantia nigra (SN) of humans, and that this is mediated through inhibition of complex I of the mitochondrial electron transport chain [174,177,13]. Subsequent work suggested that individuals with sporadic PD have reduced complex I activity not only in the SN, but in other brain regions and peripheral tissues [178]. Genetic studies focusing on monogenic forms of PD provided further support for the involvement of mitochondrial dysfunction in the disease. Pathogenic mutations that lead to autosomal recessive forms of PD have been reported in *PINK1*, *PARK2, PARK7*, *CHCHD2*and *VPS13C* and the proteins they encode are all now known to be involved in the mitochondrial quality control system and in particular mitophagy [179–182].

In this chapter we have two aims a) identify if *Alu* elements were enriched at mitochondria associated genes using newly developed datasets and b) comprehensively address the role of mitochondrial function in sporadic PD by leveraging improvements in the scale and analysis of PD genome wide association study (GWAS) data with recent advances in our understanding of the genetics of mitochondrial disease. The availability of large scale genome wide association data in PD cases and the rapid identification of genetic lesions that underlie mitochondrial disease provide an opportunity to systematically assess the role of genetic variability in mitochondrial linked genes in the context of risk for PD[183].In this chapter we combine these new resources with current statistical tools, such as polygenic risk scoring and Mendelian randomization, to explore the role of mitochondrial function in both PD risk and age at onset of disease to obtain novel insights.

4.2. Aims

Section A

- Generate a comprehensive resource of young *Alu* elements
- Using the young *Alu* resource validate whether *Alu* are enriched in regions of the genome that are mitochondria function associated gene dense

Section B

- Run polygenic risk score analysis to identify whether collectively mitochondria function associated genes are associated with sporadic PD risk.
- Run polygenic risk score analysis Identify whether collectively mitochondria function associated genes are associated with sporadic PD age at onset.
- Calculate heritability estimates to identify if novel PD heritability lies with mitochondria function associated genes.
- Implement Mendelian Randomisation to establish if expression of any of the mitochondria function associated genes is associated with PD risk.

4.3. Methods

4.3.1. Generation of mitochondria associated gene lists

Gene lists were built to encompass different levels of evidence for involvement of the respective protein products in disease phenotypes that relate to mitochondrial function. The list of genes implicated in genetic mitochondrial disorders ("primary" gene list, n=196) (see *Appendix* 1) has the most stringent criteria of evidence that the respective genes is related to mitochondrial dysfunction. It consists of 102 nuclear genes listed in MITOMAP (downloaded 2015) and 94 sourced from literature review as containing mutations that cause with mitochondrial disease.

The list of genes implicated in mitochondrial function ("secondary" gene list, n = 1487 (see *Appendix* 2) was constructed using the OMIM API to identify all genes for which the word "mitochondria" (or derivatives) appeared in the free-text description, and by combining this information with MitoCarta v2.0 genes with no OMIM phenotype. This therefore gathered a list of plausible biological candidate genes, i.e. genes that are functionally implicated in mitochondrial function and morphology for which we may lack genetic evidence for mitochondrial disease association.

4.3.2. Generation of the young *Alu* resource

As PD is a human-specific disease[184], we reasoned that the youngest and most polymorphic *Alu* insertions in the human genome were more likely be involved in disease aetiology. To systematically assess this, we curated a list that contained all of the currently known young *Alus* in the reference and non-reference genome.

For the reference genome *Alus*, the positions of all repetitive elements were generated using the RepeatMasker GRch37/hg19 Library downloaded from the UCSC genome browser:

http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/chromOut.tar.gz

Next, the coordinates of all "*AluY*" were extracted from the RepeatMasker track, which gave a total of 97823 elements. For the non-reference genome *Alus*, using the previously described (Chapter 3) Ewing non-reference "RIP" TE resource:

https://figshare.com/articles/Additional_file_2_Table_S1_of_Transposable_element_detection_from_whole_genome_sequence_data/4418360/1

We extracted the positions of all the known non-reference *Alu* insertions (n=24,294) and concatenated these with the extracted reference *AluY* which gave a total of 122,117 *Alus*.

4.3.3. Haplotype-based analysis of *Alu* enrichment at mitochondria genes

In Chapter 3 we used a haplotype-based genome-wide model to assess SVA distribution bias and enrichment of SVA at PD loci. Utilizing the same approach, we used our haplotype-based data to assess whether *Alu* elements were enriched at mitochondria associated genes. Similar to SVAs, *Alus* are known to insert into regions of open chromatin and that are GC rich and gene dense. Therefore, this data was extracted from the previous model. In addition, we integrated *Alu* and mitochondria specific data as described below:

4.3.3.1. Calculating mitochondria function specific gene density

Both mitochondria gene lists (1 and 2 as described above) were combined to give a comprehensive mitochondrion function associated gene list (n =1432 unique genes). Gene regions for the corresponding genes were extracted from the UCSC genome browser (GRch37/hg19) RefSeq gene annotation. To incorporate these into the haplotype model these regions were then intersected with the haplotype block regions as defined by *Berisa et al* and described in Chapter 3 using the 'countOverlap' function in the R package GenomicRegions. The number of mitochondrion function associated genes was then scaled by dividing the number of mitochondria function specific genes in each block by the block size (bp).

4.3.3.2. Calculating young *Alu* content

As detailed above, we generated a list of all known young *Alu* elements in the reference and non-reference genome and then overlaid these coordinates with the defined haplotype blocks. The number of young *Alus* per haplotype block was calculated using the 'countOverlap' function in the R package 'GenomicRegions'. As the haplotype blocks differed in size, the number of young alus per block was scaled, whereby the number of young alus was divided by the encompassing block size (bp).

4.3.3.3. Statistical Analysis

All statistical analyses were performed in the R Statistical environments (R version 3.4.1 (2017-06-30). Once all relevant data was incorporated into the linear regression model (scaled young *Alu* content, scaled mitochondria gene density, scaled open chromatin density, scaled gene density, and mean GC content) the following formula was used to assess if young *Alu* elements were enriched in mitochondria associated gene regions(adjusting for all already known insertion bias' as covariates):

```
scaled_young_alu ~ scaled_mitochondria_associated_genes + scaled_ _genes + scaled_open_chromatin_density + mean_GC_content
```

4.3.4. Understanding the overall role of mitochondria function in sporadic PD with current GWAS data

4.3.4.1. Samples and quality control of IPDGC datasets

All genotyping data was obtained from IPDGC datasets, consisting of 41,321 individuals (18,869 cases and 22,452 controls) of European ancestry. Detailed demographic and clinical characteristics are given in **Table 4.1** and are explained in further detail in along with detailed quality control (QC) methods[173,185]. For sample QC, individuals with low call rate (<95%), discordance between genetic and reported sex, heterozygosity outliers (F statistic cut-off of > -0.15 and < 0.15) and ancestry outliers (+/- 6 standard deviations from means of eigenvectors 1 and 2 of the 1000 Genomes phase 3 CEU and TSI populations from principal components [186]) were excluded. Further, for genotype QC, variants with a missingness rate of > 5%, minor allele frequency < 0.01, exhibiting Hardy-Weinberg Equilibrium (HWE) < 1E-5 and palindromic SNPs were excluded. Remaining samples were imputed using the Haplotype Reference Consortium (HRC) on the University of Michigan imputation server under default settings with Eagle v2.3 phasing based on reference panel HRC r1.1 2016[187,188].

Table 4.1. Demographic and clinical characteristics for all IPDGC genotyping data

Study	Cases (n)	Controls (n)	Total (n)	Case age at onset (mean, SD in years)	Control age at last exam (mean, SD in years)
IPDGC NeuroX (Nalls et al 2015, PMID:25444595)	5533	5853	11386	61.22 (12.64)	64.34 (14.82)
WTCCC PD GWAS (PMID:21044948)	1609	5195	6804	64.07 (12.04)	NaN (NA)
NIA PD GWAS (Simón-Sánchez et al 2009, PMID:19915575)	883	3009	3892	58.21 (12.89)	63.29 (10.04)
Spanish Parkinson's (IPDGC) part1	1920	1164	3084	60.07 (12.70)	69.02 (9.95)
Dutch GWAS (PMID:21248740)	768	1987	2755	54.83 (11.1)	53.53(5.98)
PROBAND	1815	NA	1815	66.24 (9.20)	NaN (NA)
Myers-Faroud (PMID:22451204)	873	850	1723	NaN (NA)	NaN (NA)
German GWAS (PMID:19915575)	741	944	1685	55.76 (11.55)	47.42 (12.38)
McGill Parkinson's	583	906	1489	65.71 (9.78)	55.79 (10.69)
Harvard Biomarker Study (HBS)	541	473	1014	66.25 (9.97)	69.93 (9.04)
Baylor College of Medicine / University of Maryland	789	195	984	64.9 (10.11)	65.45 (8.31)
Oslo Parkinson's Disease Study	476	462	938	65.32 (9.28)	61.85 (11.06)
Vance (dbGap phs000394)	621	303	924	77.44 (8.41)	81.88 (12.73)
Finnish Parkinson's	386	493	879	55.27 (5.64)	92.35 (3.86)
Parkinson's Disease Biomarker's Program (PDBP)	543	284	827	64.59 (9.34)	62.23 (10.70)
Parkinson's Progression Markers Initiative (PPMI)	363	165	528	64.24 (9.65)	63.79 (10.59)
Spanish Parkinson's (IPDGC) part2	200	169	369	67.25 (10.55)	58.09 (13.98)
PROPARK	235	NA	235	55.69 (9.96)	NaN (NA)
TOTAL	18879	22452	41331		

4.3.4.2. Curation of genes implicated in mitochondrial disorders and associated with mitochondrial function

Mitochondria specific gene lists were generated as described above. To report only novel PD associations the 349 genes identified to be in LD with the PD risk variants of interest in the most recent PD meta-analysis were removed from both lists. Following the removal, the PD-associated genes the gene lists were n= 178 and n=1328 respectively**.**

4.3.4.3. Cohort-level heritability estimates and meta-analysis

Genome-wide complex trait analysis (GCTA) was used to calculate heritability estimates for the four largest IPDGC cohorts (UK_GWAS, SPAIN3, NIA, and DUTCH) using non-imputed genotyping data for all variants within both mitochondria gene lists using the same workflow as [51]. GCTA is a statistical method that estimates phenotypic variance of complex traits explained by genome-wide SNPs, including those not associated with the trait in a GWAS. Genetic relationship matrices were calculated for each dataset to identify the genetic relationship between pairs of individuals. Genetic relationship matrices were then input into restricted maximum likelihood analyses to produce estimates of the proportion of phenotypic variance explained by the SNPs within each subset of data. Principal components (PCs) were generated for each dataset using PLINK (version 1.9). In order to adjust for factors accounting to possible population substructure, the top twenty generated eigenvectors from the PC analysis, age, sex and prevalence were used as basic covariates. Disease prevalence standardized for age and gender based on epidemiological reports was specified at 0.002[51,189–192]. Summary statistics from these estimates were produced for all four datasets and were included in the meta-analyses. Random-effects meta-analysis using the residual maximum likelihood method, was performed using R (version 3.5.1) package metafor to calculate p-values and generate forest plots[193].

4.3.4.4. Risk profiles versus disease status and age at onset

Previous risk profiling methods have calculated polygenic risk scores (PRS) using only variants that exhibit genome-wide significant associated with disease risk. However, in the most recent PD meta-analysis, it was shown that using variants at thresholds below genome-wide significance improves genetic predictions of disease risk ([51,173]). Mirroring this workflow, but rather using only variants within gene regions outlined in both the primary and secondary gene lists, the R package PRsice2 was used to carry out PRS profiling in the standard weighted allele dose manner. In addition, PRsice2 performs permutation testing and p-value aware LD pruning to facilitate identifying the best p-value threshold for variant inclusion to construct the PRS. External summary statistics utilized in this phase of analysis included data from leave-one-out meta-analyses (LOOMAs) that exclude the study in which the PRS was being tested, avoiding overfitting/circularity to some degree. LD clumping was implemented under default settings (window size = 250kb, r^2> 0.1) and for each dataset 10,000 permutations of phenotype-swapping were used to generate empirical p-value estimates for each GWAS derived p-value threshold ranging from 5E-08 to 0.5, at a minimum increment of 5E-08.

Each permutation test in each dataset provided a Nagelkerke's pseudo r^2 after adjustment for an estimated prevalence of 0.005 and study-specific PCs 1-5, age and sex as covariates. GWAS derived p-value threshold with the highest pseudo r^2 was selected for further analysis. Summary statistics were meta-analysed using random effects (REML) per study-specific dataset using PRSice2 [194]. For the age at onset risk profiling, the same workflow was followed, however instead, age at onset was used as a continuous variable, as previously reported[185]

4.3.4.5. Mendelian randomization to explore possible causal effect of mitochondria function genes

MR uses genetic variants to identify if an observed association between a risk factor and an outcome is consistent with causal effect [195]. This method has been implemented in several recent genetic studies to identify association between eQTL, to more accurately nominate candidate genes within risk loci. Therefore, for this study, in the aim of identifying whether changes in expression of mitochondria function genes are potentially causally related to PD risk, two-sample MR was implemented. Both mitochondria gene lists were combined, and all genes already associated with PD (i.e. that have been identified to be in LD with PD risk loci in the last meta-analysis) were removed, leaving 1432 unique mitochondria function gene regions. We utilized four large-scale methylation and expression datasets through the summary data-based Mendelian randomization (SMR) (*http://cnsgenomics.com/software/smr*) resource. Summary statistics were compared to PD outcome summary statistics for the mitochondria variants of interest (extracted from [22,24,48,173,196–199]) to identify possible association using the R package TwoSampleMR.

Tissues were selected based on their relevance to the pathobiology of PD, which ultimately consisted of tissues from 10 brain regions, whole blood, skeletal muscle, and nerve. For the methylation QTLs "middle age" data was used, which was the oldest available time point. For each mitochondria function variant of interest (considered here the instrumental variable), wald ratios were generated for each variable that tagged a cis-QTL (probes within each gene and meeting a QTL p-value of at least 5E-08 in the original QTL study) and for a methylation or expression probe with a nearby gene. Using the default *SMR* protocols, linkage pruning, and clumping were implemented. Finally, for each dataset p-values were adjusted by false discovery rate to account for multiple testing.

4.4. Results

4.4.1. Enrichment analysis of young *Alus* in mitochondria associated genes: Young *Alu* elements are significantly enriched at mitochondria function associated genes

We utilized our haplotype block-based enrichment model (already defined in Chapter 3) to assess whether mitochondria associated genes were enriched for young *Alu* elements. In addition, we curated and incorporated both; an extensive mitochondria gene list and a young *Alu* resource that contained all known young *Alu* in the reference and non-reference genome. Using linear regression modelling and adjusting for known insertion preferences (such as scaled mitochondria gene density, scaled open chromatin density, scaled gene density, and mean GC content) we determined that young *Alu* content and mitochondria associated gene density are positively correlated in the human genome (p = 1.17E-06, β =0. 122). To note, with the understanding that all non-LTR could impact on gene expression, reference and non-reference young SVA and LINE1 enrichment was also tested. For reference and non-reference SVA there was no significant enrichment at mitochondria function associated genes (p = 9.56E-02, β =0.049). LINE1 TEs are enriched in supposed gene deserts in the genome [200]. In line with this we found a negative correlation between young L1 and mitochondria function associated genes (p = 7.20E-03, β =-0.071).

4.4.2. Characterization of the overall contribution of genes associated with the mitochondria pathway in sporadic PD using current GWAS datasets

It is hypothesized that *Alu* enrichment at mitochondria function associated genes could contribute to human-specific sporadic neurodegenerative diseases (such as PD) through *Alu*-mediated mechanisms promoting differential mitochondria function associated gene transcription and translation. However, we cannot currently establish the contribution this variation has to sporadic PD predisposition and pathogenesis, as this source of genetic variation is not captured on any array or reference panel. Nonetheless, we can begin to assess the overall role that mitochondria function plays in sporadic PD by utilizing large-scale GWAS datasets.

We comprehensively assessed the role of mitochondrial processes in sporadic PD by leveraging improvements in the scale and analysis of PD genome wide association study (GWAS) data with recent advances in our understanding of the genetics of mitochondrial disease. The availability of large scale genome wide association data in PD cases and the rapid identification of genetic lesions that underlie mitochondrial disease provided an opportunity to systematically assess the contribution of these mitochondrial linked genes to risk for PD[183].

4.4.2.1. A component of the “missing heritability” of PD lies within mitochondria function genes

The general workflow for the genetic analysis used in this chapter is shown in Figure 4.2. First, to study the importance of mitochondrial function in sporadic PD, we investigated the heritability of PD specifically within genomic regions that contained genes annotated as important in mitochondrial function. The construction of this annotation was driven by the principle that genomic regions, which are known to be the sites of mutations in individuals with rare mitochondrial diseases or are candidate regions for such mutations provide the best evidence for involvement in mitochondrial function.

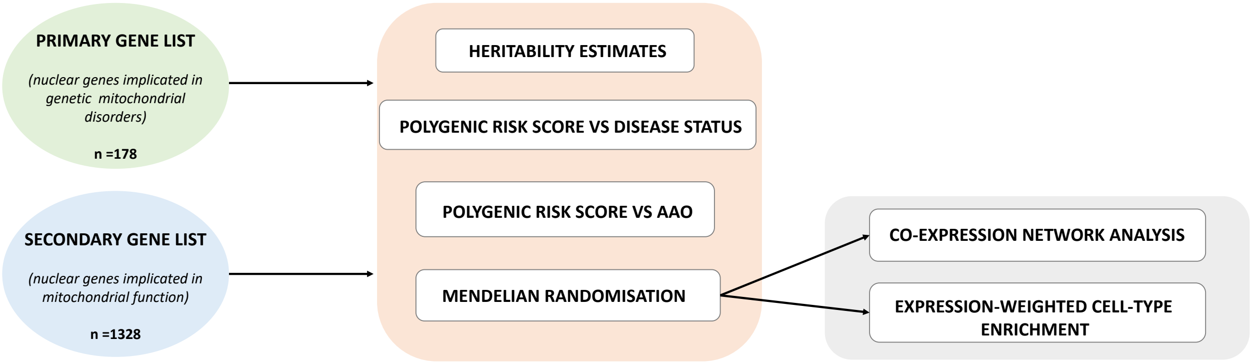

PRIMARY GENE LIST
(nuclear genes implicated in genetic mitochondrial disorders)
n =178
SECONDARY GENE LIST
(nuclear genes implicated in mitochondrial function)
n =1328
HERITABILITY ESTIMATES
POLYGENIC RISK SCORE VS DISEASE STATUS
POLYGENIC RISK SCORE VS AAO
MENDELIAN RANDOMISATION
CO-EXPRESSION NETWORK ANALYSIS
EXPRESSION-WEIGHTED CELL-TYPE ENRICHMENT

Figure 4.2. Workflow of genetic analyses to address the contribution of the mitochondrial-function pathway to PD risk and age at onset

Using GCTA, heritability estimates were first calculated for the four largest IPDGC GWAS datasets and including all variants (UK_GWAS, SPAIN3, NIA, DUTCH). Due to the low number of included cases, the heritability estimates in the other IPDGC datasets were deemed less reliable. Consistent with previous heritability estimates from both Keller and colleagues (2012; 24%) and Chang and colleagues (2017; 21%), our random effects meta-analysis for the four datasets identified 23% (95% CI 12-34, p= 2.72E-05) phenotypic variance associated with all PD samples (Table 4.2 & 4.3).There was a high degree of consistency across the cohorts.

Table 4.2. Cohort level heritability analysis for the primary and secondary mitochondrial gene lists. reporting estimates for the WTCCC PD GWAS (PMID:21044948), Spanish Parkinson's (IPDGC) part2, NIA PD GWAS (PMID:19915575), Dutch GWAS (PMID:21248740) cohorts. Showing heritability estimates generated using GCTA and standard error of the estimates (SE).

	Primary		Secondary	
	Heritability estimate	SE of heritability estimate	Heritability estimate	SE of heritability estimate
WTCCC PD GWAS (PMID:21044948)	0.00321	0.00277	0.00563	0.00688
Spanish Parkinson's (IPDGC) part2	0.00027	0.00314	0.00629	0.00932
NIA PD GWAS (PMID:19915575)	0.00945	0.00540	0.03616	0.01365
Dutch GWAS (PMID:21248740)	0.00000	0.00530	0.03562	0.01681

Table 4.3. Summary of random-effects meta-analysis for the primary and secondary mitochondrial gene lists. Here we show the random-effects meta-analysis of heritability estimates for; all SNPs in the genome (All SNPs), estimate calculated with for the SNPs within the primary mitochondria list genes (Primary) and the SNPs within the secondary mitochondria list genes (Secondary).

	Heritability Estimate from random-effects	Lower 95% confidence interval	Upper 95% confidence interval	*P*-value from random effects	Heterogeneity of variance from random effects (%)	Heterogeneity *P*-value
All SNPs	0.2313	0.1233	0.3393	2.72E-05	0.0100	3.00E-03
Primary	0.0026	-0.0011	0.0062	1.66E-01	0.0000	4.85E-01
Secondary	0.0167	0.0007	0.0328	4.10E-02	0.0001	9.63E-02

After establishing the consistency of our heritability estimates we next calculated heritability using only variants located within genic regions specified as being of primary (n=176) or secondary (n=1463) importance in mitochondrial function. Genes within the primary or secondary lists, which had already been identified in the most recent PD meta-analysis were excluded [173]. Initially, to assess the full contribution of mitochondrial processes we ran the analysis including and excluding known PD risk genes. 166 However, as shown in Supplementary Fig. 1 there was little difference overall in the heritability estimates. Therefore, we chose to catalogue mitochondrial-specific genetic risk outside of known loci and focused on the analysis excluding these genes.

The heritability estimate using a random-effects meta-analysis for the primary gene list was estimated to be a modest 0.26% (95% CI -0.11-0.66, p=0.166). However, the heritability estimate using a random-effects meta-analysis for the secondary list, namely genes implicated in mitochondrial function or morphology as well as disease, was estimated to be 1.67% (95% CI -0.07-0.32, p=0.041).

4.4.2.2. Mitochondria function specific polygenic risk score is significantly associated with disease status

We calculated PRS to capture the addictive effect of all common variants within genes implicated in mitochondria function on PD risk. PRS is a particularly powerful approach in this context because it is able to efficiently incorporate information from all hits including sub-significant hits, which may nonetheless be etiologically relevant. Again, initially we ran the analysis including and excluding the known PD risk genes

(Supplementary Fig. 2), but in order to ensure that we are reporting novel associations we focused on the lists excluding the known PD risk genes.

Using this approach, the primary and secondary mitochondrial genomic annotations were found to be significantly associated with PD disease status. Remarkably, the primary gene list consisting of only 176 genes implicated in Mendelian mitochondrial disorders, was associated with PD with an odds ratio of 1.12 per standard deviation increase in the PRS from the population mean corresponding to an overall AUC of 0.53 (random-effects p-value = 6.00E-04, beta = 0.11, SE = 0.03). The secondary gene list, which also included genes implicated in mitochondria function or morphology, was associated with PD with a higher odds ratio of 1.28 per standard deviation increase in the PRS from the population mean corresponding to an overall AUC of 0.56 (random-effects p-value =1.9E-22, beta = 0.25, SE = 0.03) (Figure 4.3). Together, these analyses not only provide further support for importance of mitochondrial processes in PD, but potentially provide a tool for identifying PD patients most likely to benefit from treatments specifically targeting mitochondrial function.

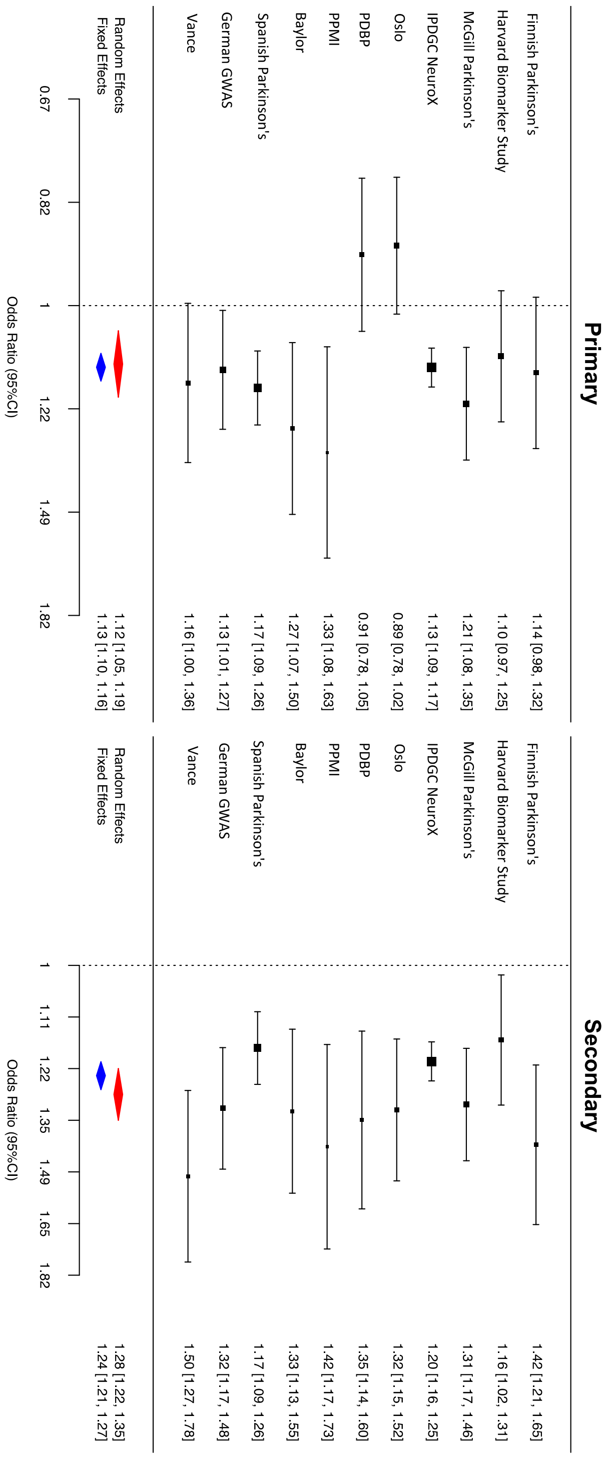


Figure 4.3. Forest plots of PRS for Parkinson's Disease across cohorts. Random effect meta-analysis results are shown as red diamonds and fixed effects are shown as blue, with the centreline of each diamond representing the summary PRS for that dataset.

4.4.2.3. Mitochondria function-specific polygenic risk score is significantly associated with later age at onset

Although multiple lines of evidence point to the importance of mitochondrial dysfunction as a primary cause of PD, impaired mitochondrial dynamics appears to be common to a wide range of neurodegenerative diseases including Huntington's disease[201,202], amyotrophic lateral sclerosis[203,204]and Alzheimer's disease[205–208]. The latter suggests that even when impaired mitochondrial function is not the primary event in disease pathogenesis; it is a common outcome and could contribute to disease progression. We sought to test this hypothesis by investigating the importance of common variation within our mitochondrial gene lists in determining the age at onset of PD (AAO). Given the significant lag period between PD pathophysiology and symptoms, AAO was used as an indirect measure of disease progression. This analysis was performed using PRS since it has been consistently found to be the main genetic predictor of AAO [173,209210,211] with higher genetic risk scores being significantly associated with an overall trend for earlier AAO of disease. While the primary mitochondrial gene list was not significantly associated with AAO of disease, the secondary gene list was correlated with AAO. Contrary to expectation, the cumulative burden of common variants within the 1326 genes comprising the PRS for PD risk, were positively correlated with AAO of PD. After meta-analysing, we found that each 1SD increase in PRS, led to a 0.55 year increase in the AAO of disease corresponding to an overall AUC of 0.51 (summary effect = 0.211, 95%CI (0.141-0.970), $|^2$ =68.49%, p-value=9.00E-03, Figure 4.4). As the forest plots demonstrate, although there was a relatively high heterogeneity across studies, the directionality and magnitude of effect on AAO were in concordance with the meta-analysis with the exception of the Oslo cohort. This could suggest that firstly, disease causation and progression are genetically separable processes in PD and that secondly the role of mitochondrial dysfunction in PD is likely to be highly complex with multiple distinct mitochondrial processes likely to be involved at different disease stages.

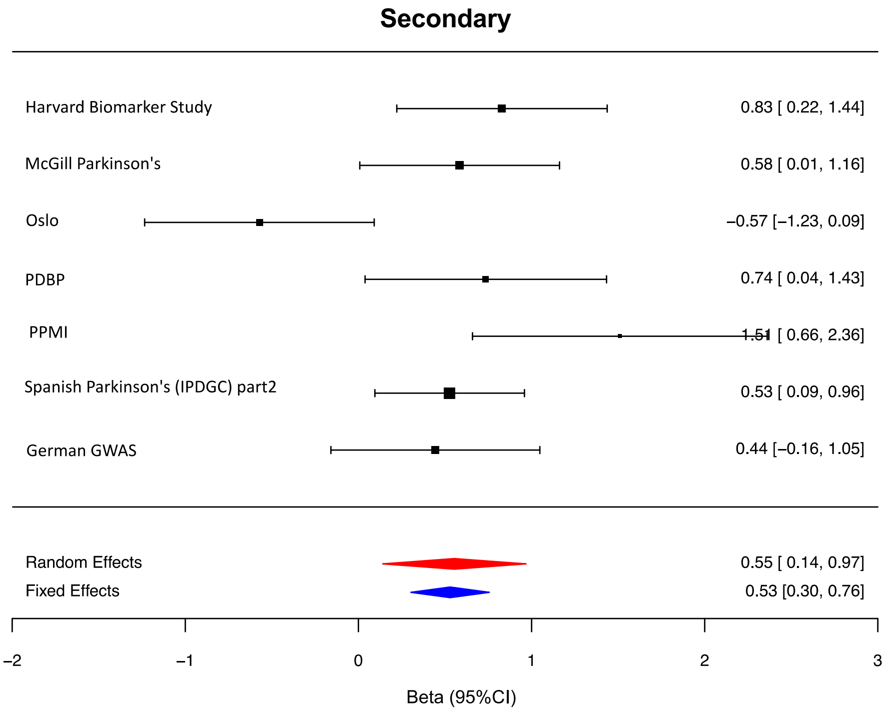


Figure 4.4. Forest plots of PRS for the age at onset of Parkinson's Disease across cohorts. Random effect meta-analysis results are shown as red diamonds and fixed effects are shown as blue, with the centreline of each diamond representing the summary PRS for that dataset.

4.4.2.4. Mendelian Randomisation suggests potential causal association of fourteen novel mitochondria function genes with PD risk

Given the robust evidence for the involvement of mitochondrial function in sporadic PD, next we used two sample MR analysis to identify specific genes likely to be important in PD risk. Since we wanted to identify novel associations, we excluded genes already linked to PD through the most recent GWAS meta-analysis[173]. This resulted in the exclusion of 31 genes linked to mitochondrial function and in linkage disequilibrium with the top PD risk variants. Analysis of the remaining 1432 genes (generated by combining the primary and secondary gene lists) resulted in the identification of fourteen novel genes linked to mitochondrial function and causally associated with PD risk (**Table 4.4**). Of the fourteen genes, the expression of 5 genes (*CLN8, MPI, LGALS3, CAPRIN2* and *MUC1)* was positively associated with PD risk in blood. Similarly, in brain PD risk was associated with increased expression of *ATG14, E2F1*, and *EP300* in brain. However, negative associations in brain and blood expression were observed for *MRPS34*and *PTPN1 and LMBRD1* respectively. Finally, elevated methylation of *FASN* in the brain was found to be positively associated with PD risk but elevated methylation of *CRY2* was found to be inversely correlated.

Six of the fourteen novel PD risk genes we identified (*CLN8*, *EP300*, *LMBRD1*, *MPI*, *MRPS34* and *MUC1*) are already associated with a monogenic disorder. We noted that neurological abnormalities were a feature of the condition in five of the six cases with Combined Oxidative Phosphorylation Deficiency 32 due to biallelic mutations in *MRPS34* being perhaps of particular interest. In common with PD, this condition is associated with abnormalities of movement, including dystonia and choreoathetoid movements. Mutations causing this condition result in decreased levels of MRPS34 protein causing destabilisation of the small mitochondrial ribosome subunit and suggesting the involvement of mitochondrial processes distinct from mitophagy and mitochondrial homeostasis in PD. In this context, it is noteworthy that *MRPL43,* another nuclear gene encoding for a component of the large mitochondrial ribosome subunit is also highlighted by the MR analysis. Thus, this analysis not only enabled us to identify specific genes of interest, but also pointed to the role of multiple mitochondrial processes in PD distinct from mitophagy.

Table 4.4. Significant functional associations of mitochondrial function associated genes via two-sample Mendelian randomization. Multi-SNP eQTL Mendelian randomization results focusing on the mitochondria associated genes (combining the primary and secondary gene lists). Showing the fourteen mitochondria function associated genes that are significantly associated with PD risk after FDR adjustment.

Gene	Beta	SE	P, FDR, adjusted	Probe	Data source	Analyte	Top QTL SNP	CHR, top QTL SNP	BP, top QTL SNP	Associated phenotype in OMIM	Neurological phenotypic features	Treatment response
LMBRD1	-0.173	0.050	4.71E-04	ENSG00000117245	Expression (Qi et al., 2018)	Brain	rs7763213	6	70416294	Methylmalonic aciduria and homocystinuria, CBLF type; MAHCF	Hypotonia, Lethargy, Developmental delay, Impaired coordination	Responsive to vitamin B12 therapy
ATG14	0.113	0.028	4.21E-05	ENSG00000171612	Expression (Qi et al., 2018)	Brain	rs11621265	14	55822095	NA	NA	NA
E2F1	0.121	0.032	1.52E-04	ENSG00000159423	Expression (Qi et al., 2018)	Brain	rs57668191	20	32289763	NA	NA	NA
EP300	0.151	0.042	2.80E-04	ENSG00000090432	Expression (Qi et al., 2018)	Brain	rs139486	22	41627654	Rubinstein-taybi syndrome 2; RTS2	Behavioral difficulties, Mental retardation (mild to moderate), Low-normal intelligence, Autism spectrum disorder (in some patients), Delayed psychomotor development, Delayed gross motor development, Speech delay	NA
MRPS34	-0.353	0.100	4.04E-04	ILMN_2210482	Expression (Westra et al., 2013)	Blood	rs2575369	16	1817431	Combined oxidative phosphorylation deficiency 32; COXPD32	Delayed psychomotor development, Lack of speech, Inability to walk, Spasticity, Hyperreflexia, Dystonia, Choreoathetoid movements , Abnormal T2-weighted signals in the basal ganglia and brainstem	NA
PTPN1	-0.090	0.022	5.27E-05	ILMN_1681591	Expression (Westra et al., 2013)	Blood	rs17788127	20	49166548	NA	NA	NA
MRPL43	0.047	0.014	8.78E-04	ILMN_1652147	Expression (Westra et al., 2013)	Blood	rs2863095	10	102746503	NA	NA	NA
CLN8	0.148	0.047	1.54E-03	ILMN_1701094	Expression (Westra et al., 2013)	Blood	rs3812477	8	1734452	Ceroid lipofuscinosis, neuronal, 8; CLN8	Developmental regression, Seizures, Ataxia, Speech and language difficulties, Myoclonus, EEG abnormalities, Cerebral atrophy, Cerebellar atrophyAutofluorescent lipopigment in neurons	NA
MPI	0.214	0.068	1.52E-03	ILMN_1761262	Expression (Westra et al., 2013)	Blood	rs4886636	15	75196176	Congenital disorder of glycosylation, type Ib; CDG1B	Hypotonia	Responsive to oral mannose therapy
LGALS3	0.307	0.085	3.14E-04	ILMN_1803788	Expression (Westra et al., 2013)	Blood	rs7157678	14	55548739	NA	NA	NA
CAPRIN2	0.337	0.101	8.49E-04	ILMN_2345739	Expression (Westra et al., 2013)	Blood	rs11051061	12	30914668	NA	NA	NA
MUC1	0.487	0.118	3.39E-05	ILMN_1756992	Expression (Westra et al., 2013)	Blood	rs6427184	1	155122783	Medullary cystic kidney disease 1; MCKD11	NA	NA
CRY2	-0.054	0.015	1.82E-04	ch.11.939596F	Methylation (Qi et al., 2018)	Brain	rs72902436	11	45881792	NA	NA	NA
FASN	0.068	0.019	4.47E-04	cg03407524	Methylation (Qi et al., 2018)	Brain	rs9905991	17	80052073	NA	NA	NA

4.5 Discussion:

Reference and non-reference *Alu* variation is yet to be catalogued and incorporated into standard genetic analysis of complex genetic disease. Therefore, the extent to which this variation contributes to disease phenotypes is yet to be determined. It has previously been reported that *Alu* elements are enriched in mitochondria function associated genes [6], which is supported in this chapter by our haplotype block analysis. *Alu*s can have a deleterious impact on neighbouring/encompassing genes by altering splicing and affecting mRNA isoform prevalence. As we could not currently explore the role of *Alu* variation in mitochondrial associated genes (or how this contributes to disease risk), instead in Section B of this chapter existing GWAS datasets were utilized to assess overall the role of this pathway in sporadic PD.

Utilising large-scale GWAS datasets, we first demonstrate that a proportion of the "missing heritability" of sporadic PD can be explained by additive common genetic variation within genes implicated in mitochondrial disease and function, even after exclusion of genes previously linked to PD through linkage disequilibrium with the top risk variants [20,25,49,50,151,212–214]. In fact, using PRS, which efficiently incorporates information from sub-significant hits, we demonstrate that cumulative small effect variants within only 196 genes linked to monogenic mitochondrial disease significantly increased PD risk (with odds ratios of 1.12 per standard deviation increase in PRS from the population mean). These findings are important for two main reasons. Firstly, given that risk profiling performed in the recent PD meta-analysis did not identify a significant association with mitochondrial function [24,48–50,151,198,199,212183]. Secondly, since the primary gene list consisted solely of the 196 genes mutated in monogenic mitochondrial disorders, this analysis highlights the increasingly close relationship between Mendelian and complex disease[1].

In order to maximise the utility of this study, we used MR which identified 14 specific mitochondrial genes of interest with putative functional consequences in PD risk. We found that although a number of the genes we identified are clearly linked to known PD-related pathways, such as lysosomal dysfunction in the case of *CLN8* and *LMBRD1* or autophagy in the case of *ATG14*, others appeared to point towards new processes. In particular, this analysis highlighted the mitochondrial ribosome through the identification of the genes, *MRPL43* and *MRPS34*, encoding components of the large and small mitochondrial ribosome subunits respectively. Interestingly, biallelic mutations in *MRPS34* are known to cause a form of Leigh syndrome, characterised by neurodegeneration in infancy with dystonia and choreoathetoid movements due to basal ganglia dysfunction. Furthermore, we note that a recent study that utilized whole exome sequencing (WES) data from two PD cohorts to investigate rare variation in nuclear genes associated with distinct mitochondrial processes, not only provided support for the involvement of mitochondrial function in sporadic PD, but also identified the gene, *MRPL43,* which encodes a component of the large mitochondrial ribosomal subunit[215]*.* Consequently, these data implicate entirely distinct mitochondrial processes in PD risk.

Finally, and perhaps most remarkably using our mitochondrial gene lists we observe clear differences between disease causation and AAO in PD. Although PRS of the primary mitochondrial gene list was not significantly associated with AAO, the PRS of the secondary mitochondrial gene list was positively correlated (p value =3.6E-05), indicating association with later age at onset. This result is some-what surprising given that previous studies have shown that risk variants in key genes associated with mitochondria function, such as *SNCA, LRRK2* and *PRKN* are associated with earlier age of onset. One possible, albeit speculative explanation for this observation could be that this variation is disease modifying only in conjunction with normal aging when mitochondria dysfunction naturally occurs. This contradiction highlights the need for replication of this correlation in future studies that leverage both; better powered AAO PD datasets and more detailed mitochondria function gene lists. However, given these findings it seems plausible that some mitochondrial processes may contribute to PD risk. Thus, this analysis is consistent with the findings of the most recent and largest AAO PD GWAS, which reported that not all the well-established risk loci are associated with AAO and suggested a different mechanisms for PD causation and AAO [185].

Although in this study we comprehensively analysed the largest PD datasets currently available with very specific and inclusive mitochondrial function gene lists, there are a number of limitations to our analyses. Firstly, there was a relative amount of heterogeneity in age at PD diagnosis within the AAO GWAS studies used. This was due to certain cohorts AAO being self-reported and other cohorts specifically recruiting younger onset cases. Nonetheless, the highly significant p-value we obtain for the association mitochondrial genes and AAO of PD (p-value=3.56E-05) and the recognized importance of mitochondrial function in aging would suggest that this finding is likely to be robust. Furthermore, it is important to recognize that our understanding of mitochondrial biology is far from complete and this is evident by the fact that many individuals with probable genetic forms of mitochondrial disease remain undiagnosed. Also, a major limitation to this analysis is that we have only been able to assess the contribution of genetic variation within nuclear encoded genes, meaning their still remains a huge gap in our knowledge concerning how variation within mitochondria DNA (mtDNA) contributes to PD risk. The importance of further understanding mtDNA variation is supported by recent studies that have shown that there are significantly elevated levels of heteroplasmic mtDNA mutations in dopaminergic neurons in the substantia nigra of PD patients compared to controls at very early pathological stages of PD231Finally, the statistical tools we have used in these analyses are currently limited. For example, MR relies on the availability of sufficient quantities of high quality eQTL data. However in PD research and in particular the IPDGC, there is a future focus to; increase data-set sample size, report and characterize phenotypes such as AAO more accurately and continue to increase the number of identified mitochondrial disease and function genes, we will be able to further explore the role of specific mitochondrial processes in more detail and identify their distinct contribution to disease causation and progression.

In summary, in this chapter we provide robust evidence for the involvement of mitochondrial processes in sporadic PD, as opposed to its defined and well-established role in the monogenic forms of the disease. In relation to the 14 novel mitochondrial function genes that we have identified, our data suggests that it is not only mitochondrial quality control and homeostasis which contributes to PD risk but other key mitochondrial processes, such as the function of mitochondrial ribosomes, mirroring the biological complexity of mitochondrial disorders. Evidently in this analysis we have only been able to address the contribution of SNP variation, however we have now demonstrated that uncatalogued *Alu* variation is also enriched at these loci. This illustrates that it will be important in future studies to incorporate this type of variation into genetic analysis. Not only is it possible that *Alus* could in part explain the association signals (and effect regulatory networks associated with dysfunctional mitochondria function in disease) but in addition they could be independent hits.

Chapter 5

Non-reference transposable elements colocalize at PD risk loci and are in moderate linkage disequilibrium with known PD risk variants

5.1 Introduction

TE derived sequences have been reported to comprise over two thirds of the human genome but despite this they are poorly understood [216]. In fact, TEs are commonly ignored in most genetic analysis. The majority are ancient and “fixed” in place in the genome. However, non-LTR elements (which are comprised of; *Alu*, LINE1 and SVA) and a small number of LTR HERV-K endogenous retroviruses still possess the capability to mobilize [217]. Due to this ongoing mobilization they are the largest source of human-specific variation in the genome. Further, in rare instances non-LTR mobilization events can cause Mendelian forms of disease such as; XDP (SVA), cystic fibrosis (*Alu*) and haemophilia A (LINE1)[86] .

We have previously described how reference non-LTR TEs can be polymorphic in sequence length and noted that this source of genetic variation is currently uncaptured in any reference panel. However, in the following chapter we leverage recently developed TE detection tools and instead focused on another form of uncaptured variation; the presence/absence of non-reference TE. The form of variation that we focus on in this chapter is further explained in Figure 5.1.

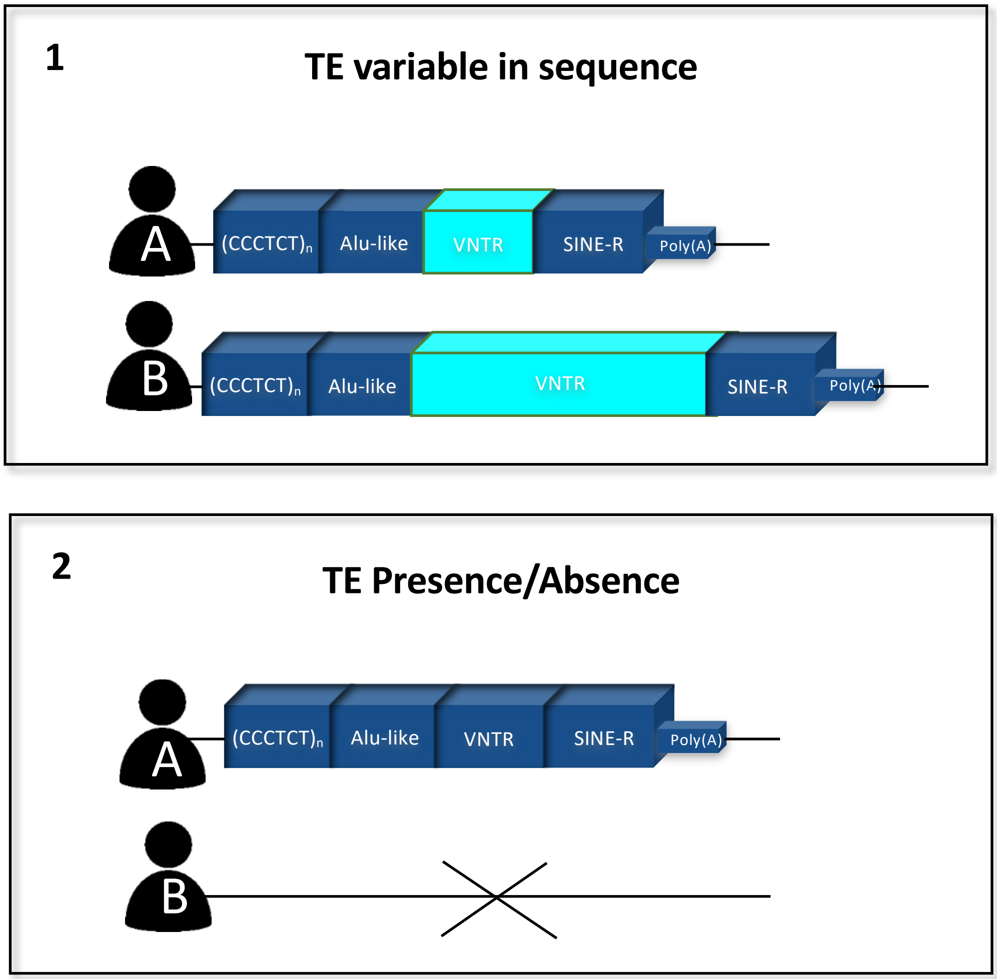


Figure 5.1. Observed TE variation in the human genome 1) TEs can be variable in primary sequence. An example is shown here where individual B has a higher copy number of the tandem repeats of the VNTR region of an SVA (turquoise). This is the most common form of variation for reference TE elements. 2). TEs can also be presence/absence in the genome. This is more commonly observed with newer non-reference TE. However, it is also now reported in certain reference TE. In this example individual A has the SVA insertion and individuals B does not. The MELT analysis here focusses on the later form of variation.

The field has witnessed the recent development of new technologies that can now comprehensively call non-reference TE insertions genome-wide. This improvement in non-reference detection tools, in conjunction with the wealth of WGS data now available, has started to shed light on the sheer scale that these elements contribute to interindividual variation and health and disease. An example of this came after the release of the 1000 genome project, which allowed for the detection of non-reference TE. These analyses identified that the 2504 individuals included harboured > 16 000 non reference TE variants [186,218,219]. Hence, that is potentially more than sixteen thousand uncatalogued variants that exist within the population that have never been included in any GWAS or other large-scale genetic analysis.

In regards to PD specifically, if non reference TE contribute to PD then why have they not been previously associated with risk of disease after extensive genetic study? Well the simple answer in part is that they have. In rare cases of Mendelian forms of PD deleterious *Alu* activity at PD loci such as *PARK7/DJ-1*[73] and *PARK2*[74] has been reported to be causative the disease. Another rare mutation that has been implicated in the autosomal dominant form of the disease is duplications (220 to 394 kb) and a triplication (1.61 to 2.04 Mb) of the *SNCA* gene. Ross and colleagues reported that the presence of *Alu* and LINE1 elements at the *SNCA* locus may contribute to the genomic instability at this region which induces the disease causing copy number variation [72]. Therefore, if non-LTR elements mobilize and cause major disruption within PD associated genes they can cause Mendelian forms of PD.

Concerning sporadic PD, as non-reference TE variation is filtered out of most genetic analysis their role is yet to be established. Currently over 90 risk variants have now been identified after extensive meta-analyses involving over one million individuals. Despite this effort, the majority of the common genetic variation that contributes to the heritability of PD is still unknown (~70-80%)[1]. PD risk variants predominantly reside within non-coding regions of no clear function and are hypothesized to contribute a small effect to PD risk by influencing allele-specific gene expression. Determining the functional and eventual causal mechanisms underlying these relationships has proven difficult. This is in part because it is still unknown whether a nominated hit is the true causal variant and also the target gene has not yet been established. Therefore, one possible source of genetic variation that could be contributing to PD risk is non-reference TEs. In addition to their impact through mobilization events non-reference TEs can also influence gene expression by providing a wide variety of regulatory sequences such as promoters, enhancers, transcription terminators and several classes of small RNAs and recent studies have shown that non-reference TEs often act as cis and trans eqtls[219].

Given the established link between rare non-reference TE insertions and Mendelian forms of PD and their ability to affect gene expression, we reasoned that common non-reference TE could be contributing to sporadic PD. Previously, it has not been possible to include non-reference TE calling in genetic analysis due to the lack of technology to call these variants in multiple genomes. For the current study we utilized newly developed non-reference TE detection tools to integrate non-reference TE and

SNP variants in PD WGS data in the aim of characterizing the role that non-reference TE could be playing in PD.

5.2 Aims

- Run a TE detection tool to call non-reference TE in PD WGS data
- Colocalization analysis to identify if any non-reference TE map to known PD risk loci
- Linkage disequilibrium analysis to address whether only of the non-reference TE are in linkage with known GWAS variants
- Linkage disequilibrium analysis to address whether only of the non-reference TE are in linkage with known PD risk variants
- Initial association analysis to identify if any variants are associated with PD at genome-wide significance

5.3 Methods

5.3.1. Samples and quality control

A total of 790 individuals were included in this study from the PPMI and NABEC cohorts, 382 PD cases and 408 controls. The NABEC individuals were included to facilitate downstream (brain-specific) eQTL analysis. Originally there were 820 individuals included but to identify ancestry outliers PCs 1-10 were calculated in PLINK and using an in-house script in R, samples were then clustered using principal component analysis (PCA) to evaluate European ancestry as compared to the HapMap3 CEU/TSI populations (International HapMap Consortium, 2003). Subsequently any individual of non-European ancestry was removed. Therefore, the final 790 individuals were of European ancestry and did not carry mutations known to cause PD. A description of how the WGS was generated and the quality control pipeline followed is described in (Chapter 2.3).

5.3.2. TE detection and calling

To detect and call non-reference TEs we originally tested the TE detection tool TEbreak. But despite a substantial effort we were unable to adapt the tool to run with our already generated GRCh38/hg38 BAMs and re-generating all the existing BAMs would not have been cost or time appropriate. Instead we used the previously published Mobile Element Locator Tool (MELT) software package:

http://melt.igs.umaryland.edu/manual.php

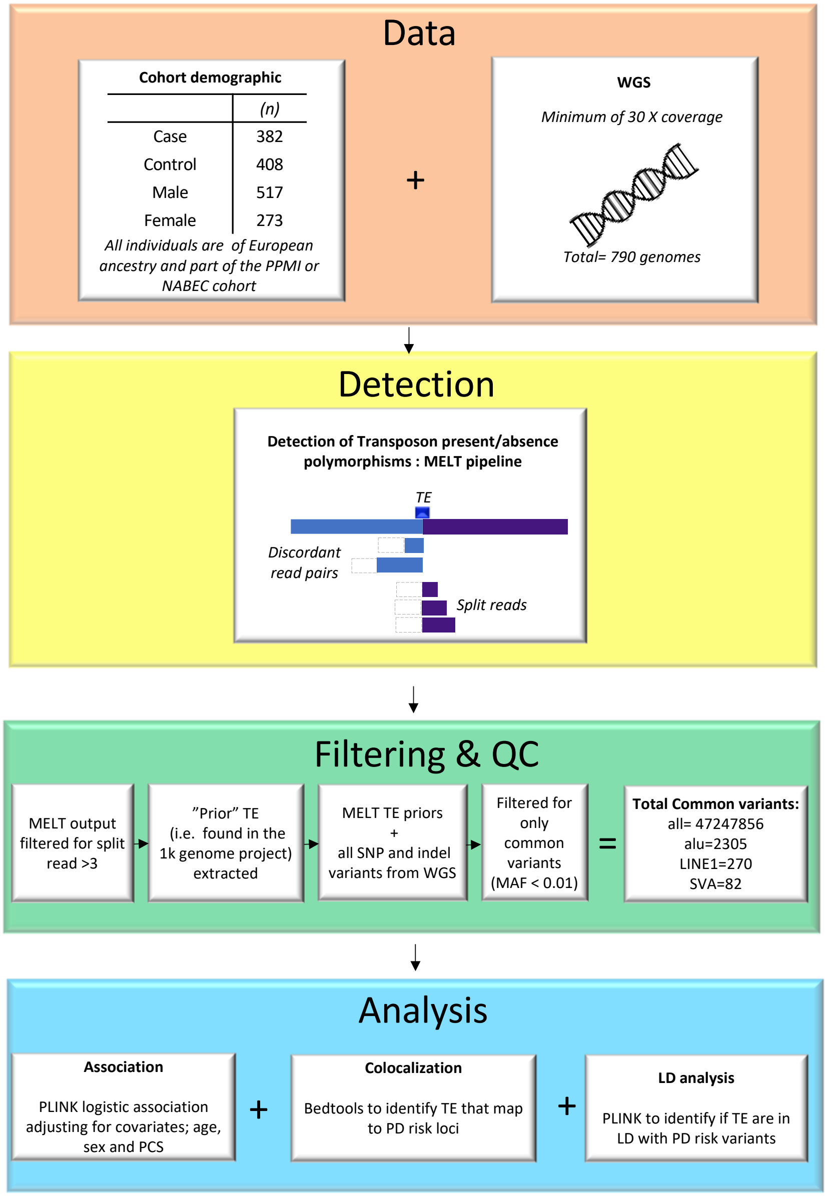
MELT is a widely used tool that was built to identify non-reference present/absent TEs in large genome sequencing projects. Specifically, MELT was developed as part of the 1000 Genomes Project and is currently one of the top TE detection tools that has high scalability. In addition, due to the desired downstream analyses and large sample size, we deemed MELT a more suitable tool. This was because, unlike TEbreak, MELT outputs non-reference TE calls in a VCF format which can then be inputted straight into existing NGS analysis pipelines in a similar fashion as the output of SNP variant callers.

Subsequently, MELT was run with the PPMI and NABEC datasets using default parameters. Due to time limitations (and apparent licensing restraints that stopped MELT being run on google cloud, (the platform where the BAMs were stored)) the MELT-SINGLE method was used for each BAM individually. The -t option was used to increase the total number of genotyped sites and the accuracy of the genotype. This option uses a collection of non-reference TE sites that were previously identified in the 1000 genome project and known as "priors". The priors file was available through the MELT software in GRCh37/hg19 build, therefore for our analysis the coordinates were extracted and lifted over to GRCh38/hg38 using the UCSC genome browser LiftOver tool. Finally, variants that had > 3 split reads were filtered.

Calls were obtained for presence/absence *Alu*, LINE1, SVA and HERV-K. Unlike the non-LTR TEs, HERV-K was not called in the 1000 genome project and therefore they did not have "prior" sites established. As the HERV-K consensus sequence has very high sequence homology the accuracy of the calls is reported to be much lower than the non-

LTR elements. Therefore, we did not include the HERV-K genotypes in the current analysis.

The non-reference TE calls (*Alu*, LINE1 and SVAs) for each individual were merged using bcftools v1.9[220]. This non-reference TE VCF was then merged with the SNP and indel calls to create one VCF which contained all variants for all individuals (n = 10109859 variants) (Alus, LINE1s, SVAs, indel and SNPs). Next, Plink v1.9 [221] was used to extract common variants (MAF >0.01) (n=47247856 variants)to ensure both the confidence of the genotype calls and the reliability of the association analyses. The overall workflow for the non-reference TE analysis is summarized in (Figure 5.2).



	(n)
Case	382
Control	408
Male	517
Female	273

All individuals are of European ancestry and part of the PPMI or NABEC cohort
+
WGS
Minimum of 30 X coverage
Total= 790 genomes
Detection
Detection of Transposon present/absence polymorphisms : MELT pipeline
TE
Discordant read pairs
Split reads
Filtering & QC
MELT output filtered for split read >3
"Prior" TE (i.e. found in the 1k genome project) extracted
MELT TE priors + all SNP and indel variants from WGS
Filtered for only common variants (MAF < 0.01)
=
Total Common variants:
all= 47247856
alu=2305
LINE1=270
SVA=82
Analysis
Association
PLINK logistic association adjusting for covariates; age, sex and PCS
+
Colocalization
Bedtools to identify TE that map to PD risk loci
+
LD analysis
PLINK to identify if TE are in LD with PD risk variants

Figure 5.2. Workflow and summary of non-reference TE PD analysis.

5.3.3. LD analysis and non-reference TE annotation

To identify LD between the common non-reference TEs and known PD risk loci, Plink v1.9 was used with the "--show-tags" option. Strong LD was defined as r2 ≥ 0.8 and D′ ≥ 0.8. The detected common non-reference TE and there tagging SNPs (generated by the --show-tags option with Plink v1.9 [221]) were annotated with ANNOVAR v2018-04-16 software gene-based annotation to infer whether the non-reference TEs variant was exonic, intronic, splicing, 3’- untranslated region (UTR), 5’-UTR, or intergenic, or for the tagging SNP variants a known GWAS variant.

5.3.4 Colocalization with PD risk loci

A variant does not need to be in LD with a disease associated SNP to be biologically important to disease mechanism. Therefore, we characterized the non-reference TE content within already known PD risk loci. PD risk loci were defined as the 85 haplotype blocks which incorporated the 90 PD risk variants. The coordinates of these regions were extracted and bedtools v2.27.1 was used to intersect these positions with the common non-reference TEs.

5.3.5. Preliminary non-reference TE association analysis

We ran an initial association analysis to identify if any of the non-reference TEs were significantly associated with PD risk. For the association analysis, SNPs were included to use as a control so that the association results could be compared with the results from the last PD meta-analysis to confirm directionality. Principal components (PCs) were created from the directly assayed genotypes using Plink v1.9 [221] . For the PC calculation, variants were filtered for MAF (>0.01), genotype missingness (<0.05) and Hardy–Weinberg equilibrium (P =>1E- 6 PCs were re-calculated in using Plink v1.9237 for the 790 individuals of European ancestry and using these a logistic regression model adjusted for sex, age and PCs 1-5 was used to estimate risk associated with the disease for each variant. Where possible, age of onset was defined based on patient report of initial manifestation of parkinsonian motor signs (tremor, bradykinesia, rigidity or gait impairment).

5.4 Results

5.4.1. Non-reference TEs are common structural variants

Overall 7478 non-reference TEs were detected in the NABEC and PPMI individuals that have been previously described in the 1000 genome project. Of these, 2657 variants were common (MAF >0.01). Consistent with previous results, the majority of non-reference TE variants show low allele frequencies, suggesting that the majority of TE insertions can be highly disruptive [219,222](Figure 5.3).

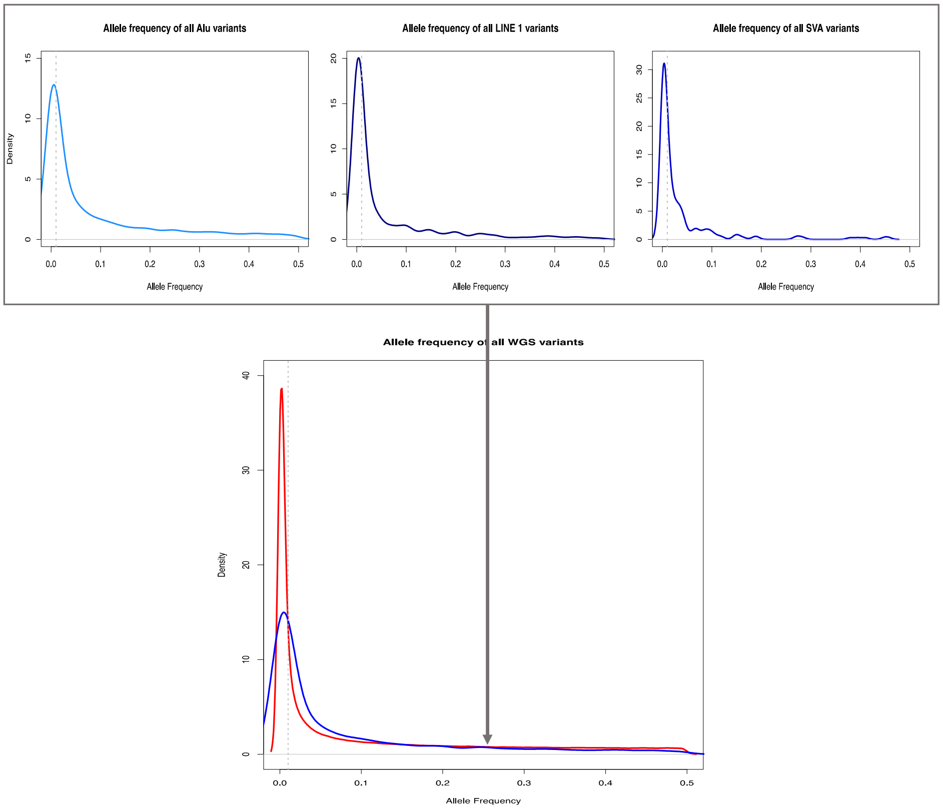


Figure 5.3. The distribution of MELT detected known non-reference TEs in WGS. *(Top) Blue shaded lines represent non-reference TEs; Alu, LINE1, SVA (Bottom) red line represents allele frequencies of SNP variants from corresponding WGS. Grey dotted line represents MAF cut off (0.01)*

In line with what is observed in the reference genome, for the common non-reference TE variants (n =2657) the most abundant class was *Alu* (n =2305/86.75%), followed by LINE1 (n=270/10.16%) and SVA (n=82/3.09%). The majority of non-reference insertions were intergenic (n =1442) and intronic (n =1136), however many were also located within exons (n=32) 5'UTRs (n=6), 3'UTRs and (n=16) upstream (n=15) and downstream (n=10) of genes (+-1kb) (Figure 5.4). But the reported gene annotation could potentially be skewed by the fact that the original VCF merge did not carry over the predicted sequence length. Therefore, the annotation only currently represents the region of the base change in the reference genome rather than the full region spanning the new insertion site.

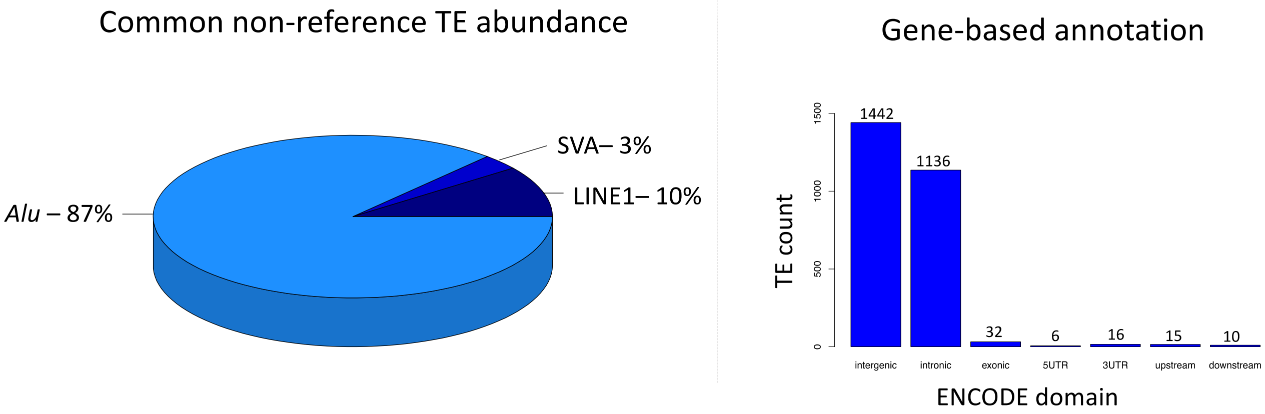


Figure 5.4. Characterization of common non-reference TEs in the genome. a) Overall abundance of common non-reference by TE type b) locations of common non-reference TE annotated with ANNOVAR package for exonic, intronic, intergenic, 5'UTR, 3'UTR, and whether +1kb (Upstream)or -1kb (Downstream) from a gene.

5.4.3. Prior non-reference TEs are not in strong LD with GWAS loci

Currently the majority of GWAS associated variants represent an association signal, not a causal variant. In fact, for many loci very little is understood about how the signal contributes to risk of disease. Identifying the true causal variant is a difficult process, especially as reference panels do not currently include structural variants, such as non-reference TEs, which are a significant and disease relevant source of genetic variation in the human genome.

Recent studies have started to assess the extent to which non-reference TEs tag GWAS associated variants, with many now reporting that several non-reference TEs are in strong LD with GWAS associated loci. Further they show that in specific cases, following functional analysis the non-reference TE has been identified as the likely causal variant. In light of this, using our MELT generated non-reference TE calls for the PPMI and NABEC cohorts, we addressed possible linkage disequilibrium between these variants and known GWAS hits.

Our approach involved an integrative analysis of non-reference and SNP variants. The LD structure of the resulting common TE insertions with adjacent common SNPs was then defined using Plink. From the 2657 common prior non-reference TEs tested we did not find any in strong linkage disequilibrium (defined as r2 ≥ 0.8 and D' ≥ 0.8) with known GWAS variants (according to ANNOVAR annotation) in this initial analysis.

5.4.3. Prior non-reference TEs are in moderate linkage disequilibrium with the known PD risk variants

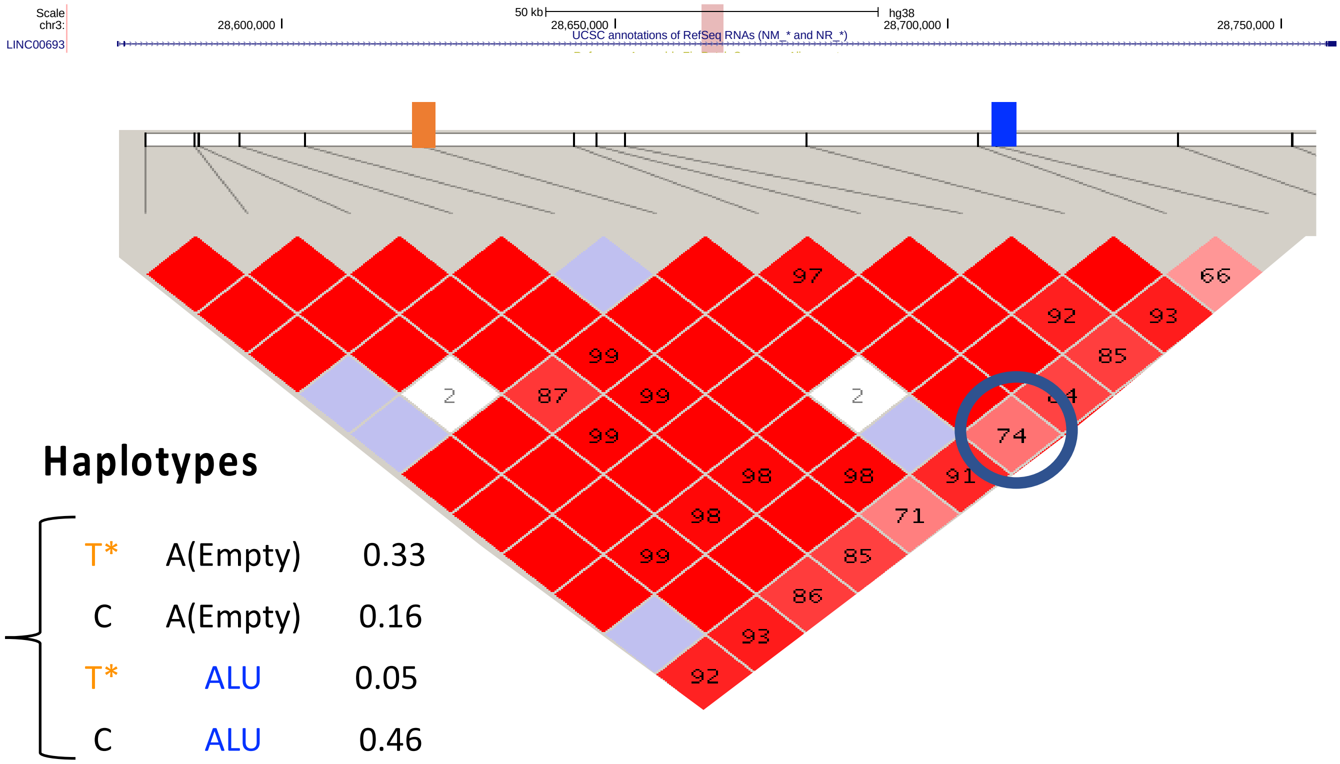
The most recent meta-analysis of PD GWAS data (which was the largest study of PD genetics to date and involved over one million individuals) defined 90 genome-wide significant risk loci. Although this large-scale international-based effort more than doubled the number of known risk loci, the field have still not established how the majority of these hits contribute to disease risk. Many of these variants lie within non-coding regions of no clear function.

Although each of the identified risk loci are currently the strongest candidate for the causal variant at that region, it is possible (and one could argue likely) that this signal could in fact represent a signal from a variant that is not currently detected on an array; such as a structural variant. Therefore, in light of this, we next assessed the genetic relationship between the common prior non-reference TEs and the known PD risk loci. LD was calculated between the 2657 common prior non-reference TE variants and 90 known PD risk loci using Plink. We did not find any strong correlation between the common prior non-reference TEs and PD risk variants (r^2> 0.4). However, three of the non-reference TEs are in moderate LD with loci associated with PD (Table 5.1).

Table 5.1. Descriptive statistics of the three non-reference TEs that are in linkage disequilibrium with Parkinson's disease associated variants

Nearest gene	CHR	BP	PD SNP	Beta	P-value	polyTE	polyTE MAF	R2	D'
RPS12	6	133210361	rs6808178	-0.22	1.04E-10	ALU_2247	0.49	0.34	0.74
CRHR1	17	43744203	rs62053943	-0.27	3.58E-68	SVA_704	0.19	0.38	0.67
LRRK2	12	40713873	rs11176013	-	-	ALU_9158	0.47	0.39	0.72
LRRK2	12	40716510	rs10878371	-	-	ALU_9158	0.47	0.39	0.72

Two of the prior non-reference TEs are in weak LD with two a PD risk loci rs6808178 and rs62053943. First the rs6808178 risk variant (p= 1.04E-10, β =-0.221) is weakly tagged by an *Alu* element (r^2=0.34, D'= 0.74), which is only ~1kb upstream of the hit and intronic of *LINC00693* (the noted nearest gene) (Figure 5.5).

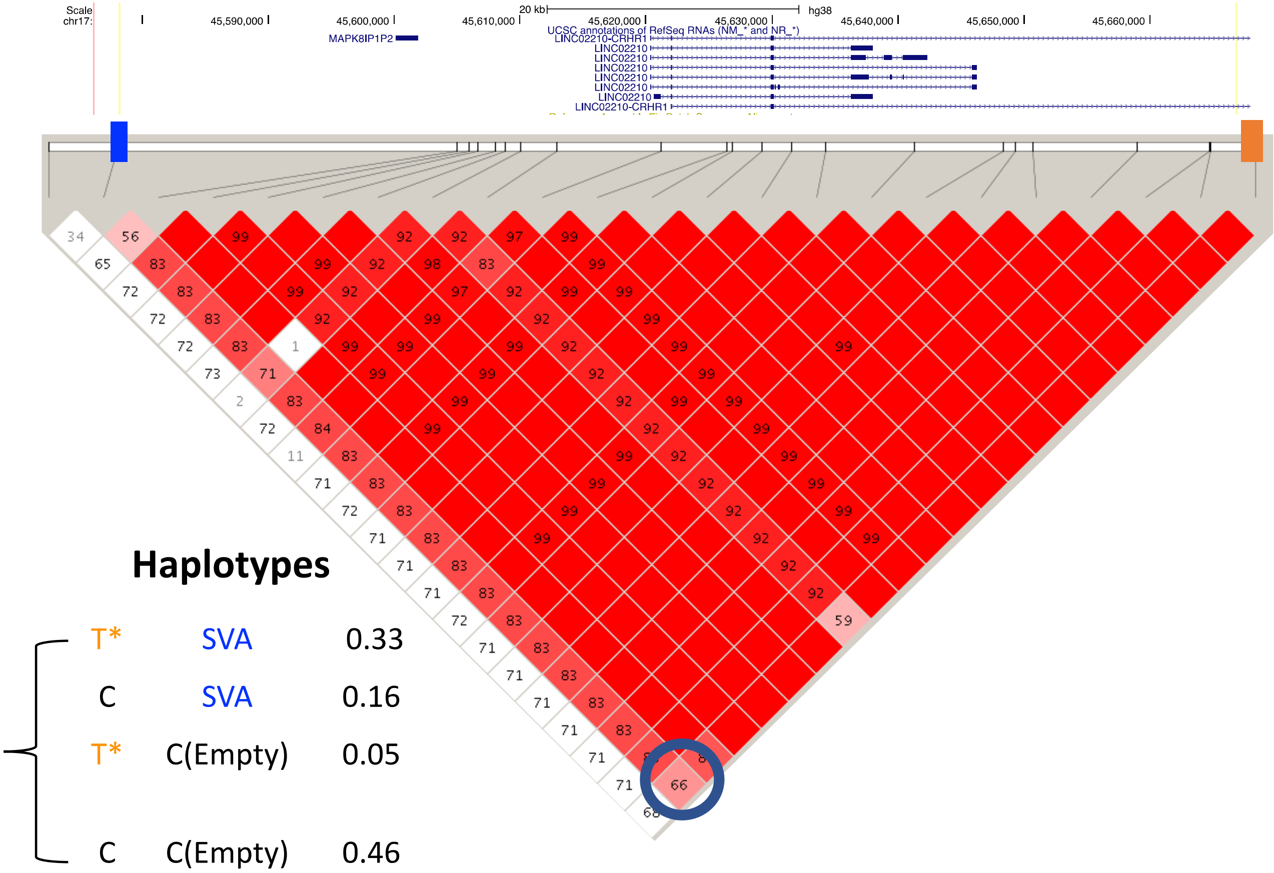


Haplotypes

T*	A(Empty)	0.33
C	A(Empty)	0.16
T*	ALU	0.05
C	ALU	0.46

Figure 5.5. A common Alu non-reference TE is in linkage disequilibrium (LD) with a PD risk SNP (rs6808178, p=1.04E-10, β=-0.22). From top : genomic position of the non-reference TE(blue) and risk variant (orange). Below= the LD block and genetic relationship of the variants of interest. * denotes effect allele.

In addition, the rs62053943 risk variant (p=3.58E-68, β =-0.270) is moderately tagged by an SVA at the MAPT locus (r^2=0.38, D’= 0.66) (Figure.5.6).



Haplotypes

T*	SVA	0.33
C	SVA	0.16
T*	C(Empty)	0.05
C	C(Empty)	0.46

Figure 5.6. A common SVA non reference TE is in moderate linkage disequilibrium with a PD risk SNP (rs62053943 p=3.58E-68, β=-0.27). From (top): genomic position of the non-reference TE (blue) and risk variant (orange). (Below) = the LD block and genetic relationship of the variants of interest. * denotes effect allele.

Finally, at the LRRK2 locus an *Alu* element is in weak LD with two SNPs that are associated with autosomal dominant PD (r^2=0.39, D'=0.72) and these two variants (rs11176013 and rs10878371) are in complete LD (r^2=1, D'=1) (Figure 5.7).

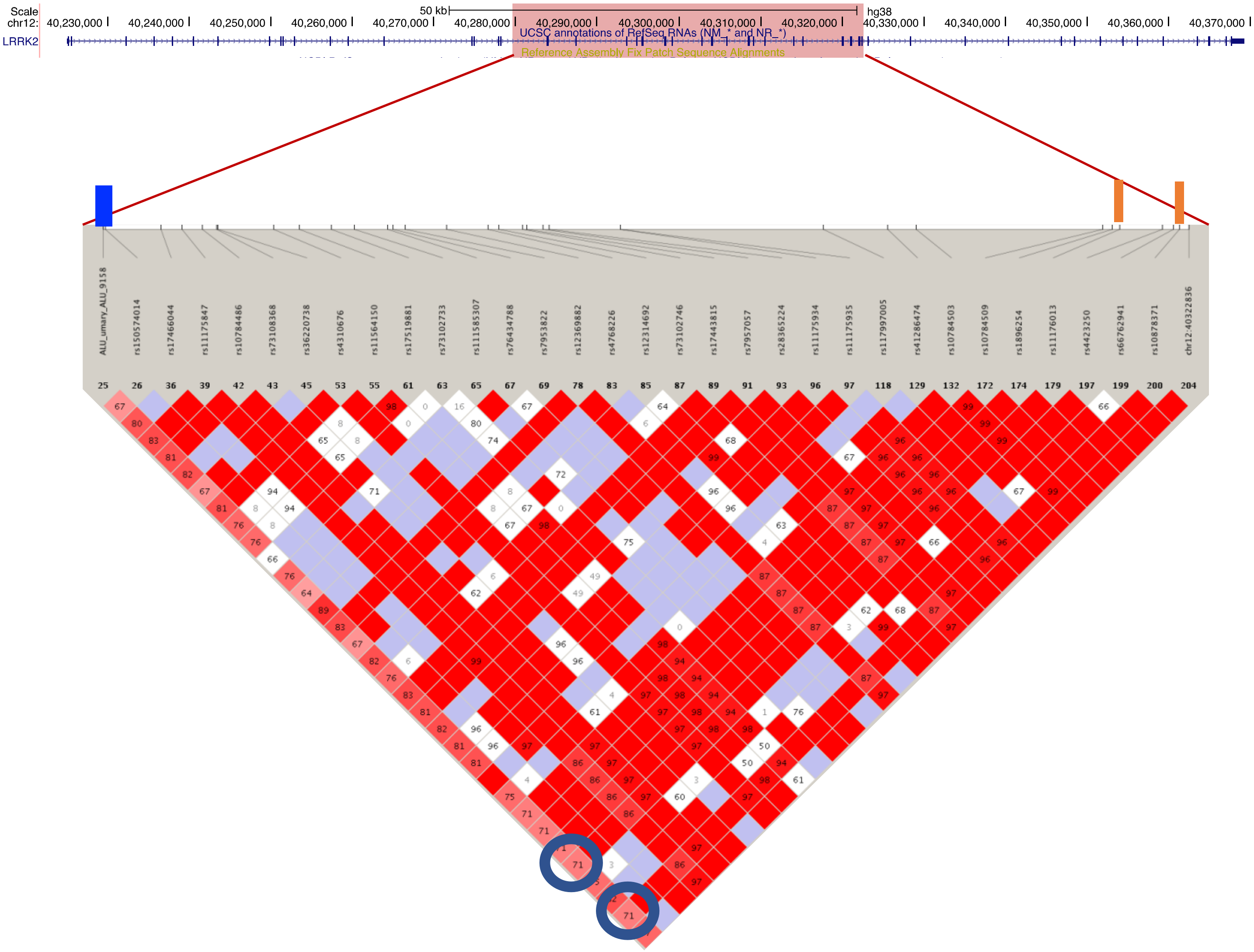


Figure 5.7. A common *Alu* non-reference TE is in linkage disequilibrium with PD associated SNPs (rs11176013 & rs10878371 at the *LRRK2* locus. From top: genomic position of the non-reference TE (blue) and PD associated variants (orange). Below= the LD block and genetic relationship of the variants of interest. * denotes effect allele.

5.4.4. Colocalization of common non-reference TE at PD risk loci

GWAS loci are enriched for TEs due to their genic nature, as reported by previous studies and our own data described in earlier chapters (Chapter 2 & 3). As TEs are yet to be catalogued at these loci, they are therefore an ignored form of genetic variation that could be involved in disease mechanism. Although we have identified that the common prior non-reference TEs were not in strong LD with the known PD loci they may still be involved in disease mechanism. Therefore, we assessed whether the common prior non-reference TEs mapped within the known PD risk loci. We report that overall, of the prior non-reference TEs detected in the NABEC and PPMI individuals, 367 elements mapped to known PD risk loci. Of these, 302 were *Alu*, 44 LINE1 and 21 SVA. When these variants were filtered to focus on common variants only (MAF >0.01), overall 165 mapped to known PD risk loci (146 were *Alu*, 13 LINE1 and 6 SVA) Figure 5.8.A. Out of the 85 known PD risk haplotype blocks, 37 (44%) contained at least one common non-reference TE while the mean of number of common non-reference TE for all PD blocks was 1.6 (0-5). As shown in in Figure 5.8.B we observe an enrichment non-reference TE at PD risk loci, which isn't surprising given the nature of GWAS and the bias they have towards genic regions. This supports what we noted in Chapter X whereby we identified that TE were enriched at genic regions of the genome and followed the same insertions pattern as the reference TEs. To note, we also show that many non-reference TE are located within top PD risk loci at potentially important functional domains. An example of this is a non-reference TE *Alu* located at the INPP5F/BAG3 PD risk locus between the two genes as seen in Figure 5.8.C. Another example is shown in Figure 5.8.D where two non-reference TE *Alu*s lie within the HLA locus.

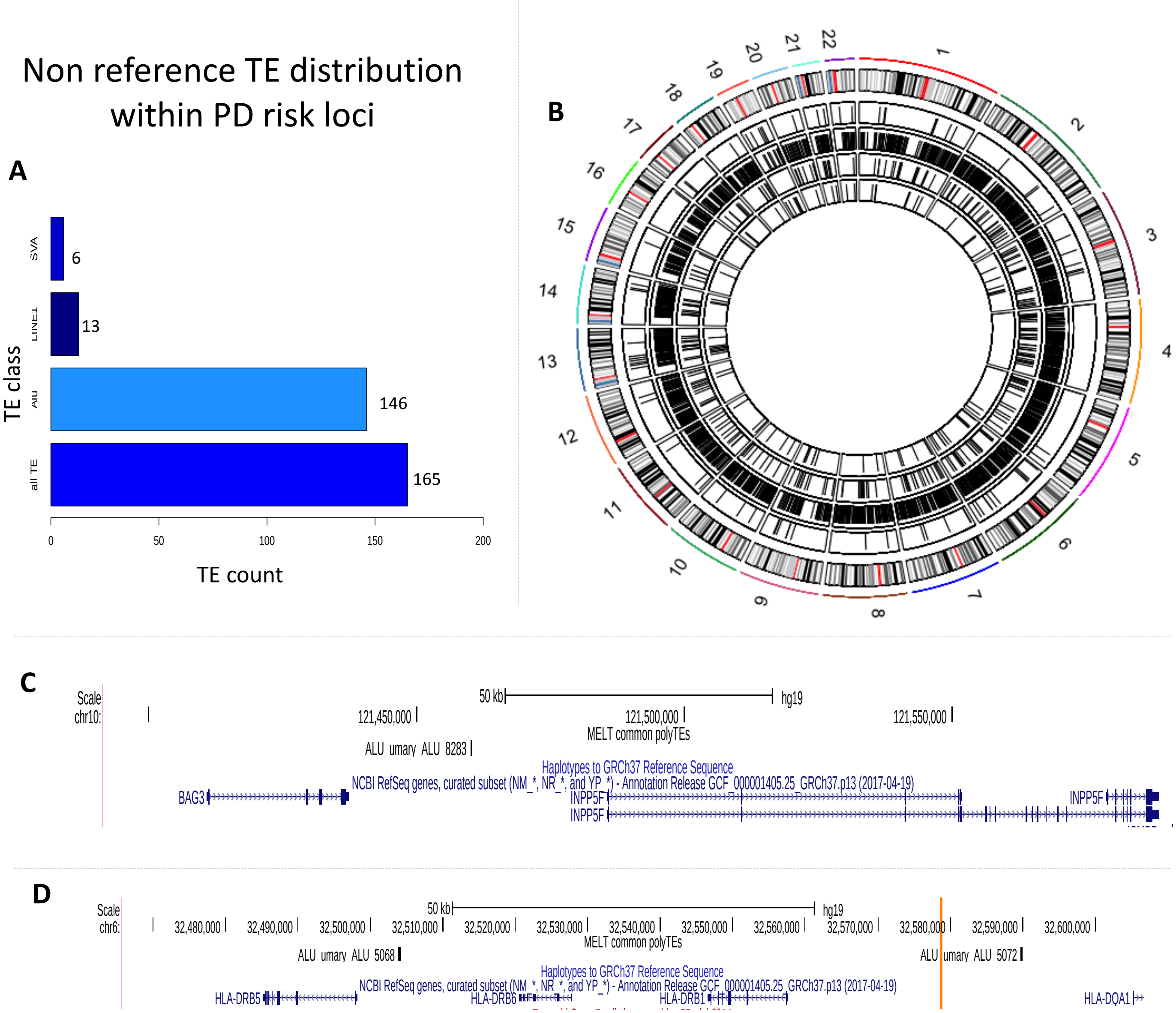


Figure 5.8. Non-reference TEs colocalize with PD risk loci and lie within potentially important regulatory domains A: breakdown non-reference TE distribution for each class at PD risk loci, B: Plot of chromosomal locations of the non-reference TE. From outside-in; PD risk loci, non-reference; Alu, LINE1, SVA, C: Alu non-reference TE at the BAG3/INPP5F locus **D**: Alu non-reference TE at the HLA locus, orange highlight shows the position of the PD risk SNP at the HLA locus.

5.4.5. Genome-wide association analysis

Descriptive statistics of the PPMI/NABEC cohort are summarised in Table 5.2 below.

Table 5.2. Demographic characteristics of the PPMI/NABEC cohort.

Case	Control	Cases, % female	Controls, % female	Case, age at onset in years, mean (SD)	Control, age at ascertainment in years, mean (SD)
382	408	55%	35%	62 (9)	59(23)

To identify if any variant was associated with risk of PD, GWA was performed with all called variants (SNP, indel and non-reference TE variants) adjusting for the appropriate covariates; study-specific PCs 1-5, age and sex. As shown by the Manhattan plots in Figure 5.9, we did not detect any genome-wide significant variants at p >5E-08, although this is not surprising given the low number of individuals analysed. To ensure that our results were in line with previous PD GWAS, the beta-values of the known genome PD risk hits were plotted from the most recent meta-analysis and our current GWAS (PPMI/NABEC). As shown in Figure 5.10 the beta-values were positively correlated (p=8.05E-04), supporting our association analysis.

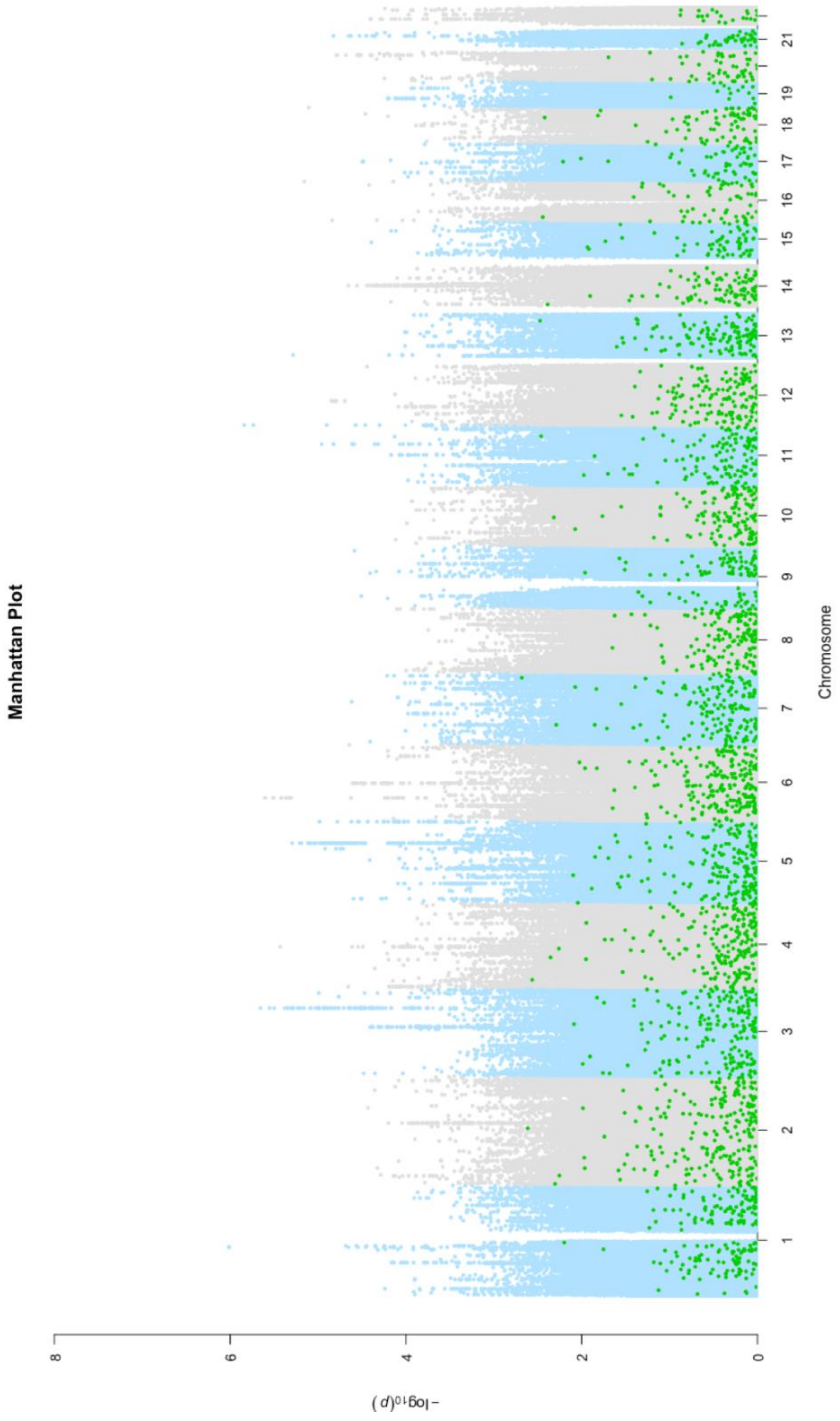


Figure 5.9. Manhattan plot: All WGS SNP variants are shown in blue and grey, no variant of genome wide significance was detected. Non-reference TEs variants are shown in green. The X axis represents the base pair position of variants from smallest to largest per chromosome (1-22).

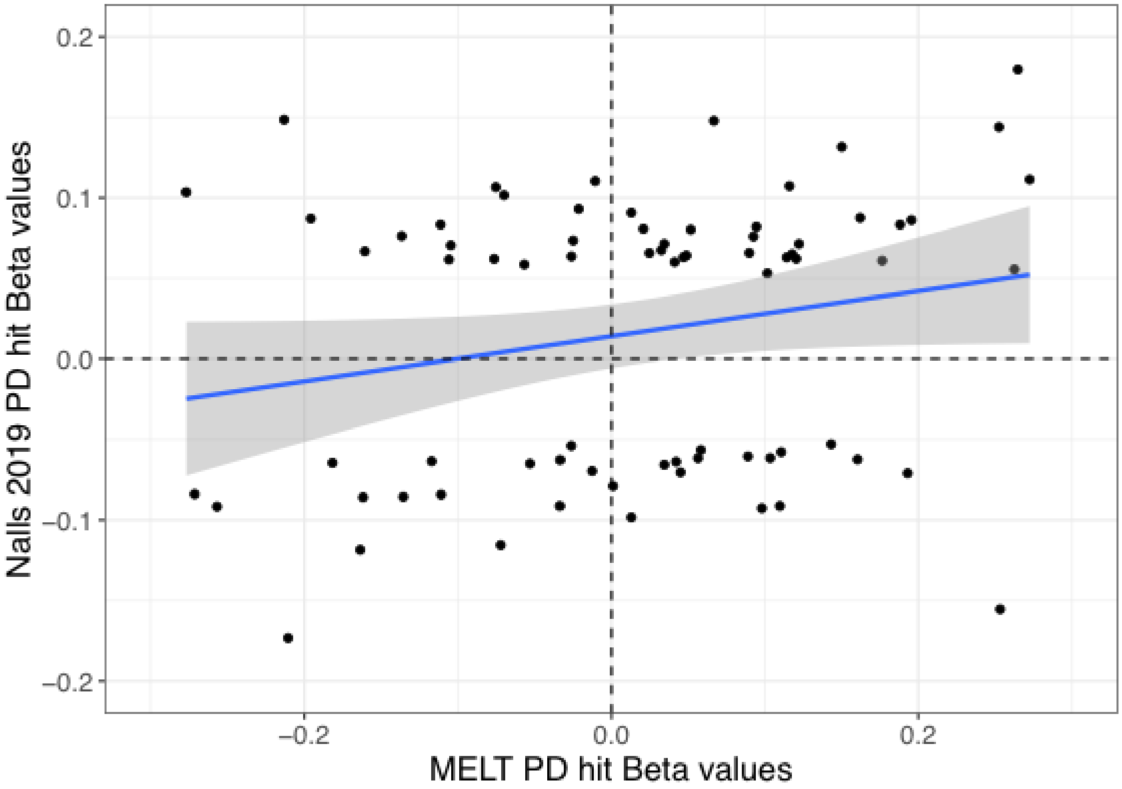


Figure 5.10. Beta coefficient plot: coefficient values of the top PD risk variants from the most recent meta-analysis were plotted against the corresponding betas from our current GWAS(PPMI/NABEC). The beta-values were positively correlated (p=8.05E-04).

5.5 Discussion

TE insertions are currently known to cause over a hundred human diseases, yet no study has considered the role of these elements in PD. In fact, in order to obtain high confidence calls, the initial step of most genetic analysis is to perform filtering that removes TEs as standard. Meaning that the overall contribution of these elements to PD risk and risk of other complex genetic diseases is completely unknown. Following the recent development of programs that genotype these elements genome-wide, we performed the first characterization of non-reference TEs in PD to date. Using an integrative analysis of non-reference TE and SNP variants we ran association, co-localization and LD analysis. We report that non-reference TE colocalize and are enriched at the currently known PD risk loci. In addition, we report that multiple non-reference TE are in linkage with unexplained PD risk variants. Taken together our results highlight that non-reference TEs are an important and potentially disease relevant form of genetic variation that should be further catalogued at these loci.

Consistent with the literature we identified that the majority of non-reference TEs are observed in low frequencies (i.e. low minor allele frequencies), suggesting that in general new TE insertions are highly disruptive and subject to strong purifying selection. This is in line with previous studies that have generated MELT calls for different HAPMAP populations and observed the same allele frequency distribution[218,222]. In terms of the general non-reference TE abundance , again in accordance of *Rishishwar et al* we show that non-reference TE abundance is similar to what is observed in the reference genome, i.e. *Alu* elements are the most common followed by LINE1 and SVAs [222]. Of the common non-reference TE, through our gene-based annotation we show that although non-reference TE are most frequently located intergenic, nearly half of all the common non-reference TEs are located within or 1kb+- of a gene. This is significant given the fact that non-reference TE can impact on gene expression significantly via a number of mechanisms such as; interrupting coding sequences and splicing and introducing novel promoters [94]. Therefore, one could imagine that this form of genetic variation could have a significantly different functional consequence between individuals for gene expression.

In relation to our ANNOVAR GWAS annotation we did not find any non-reference TE variants in strong LD with known GWAS hits (r2 ≥ 0.8 and D' ≥ 0.8). This is contrary to previous reports that have found TE elements to more generally be in strong LD with GWAS variants. One example is a comprehensive study by Payer *et al*, whereby they extensively catalogued and genotyped *Alu* elements at GWAS loci. Through this they identified that 44 non-reference TE *Alu*s were in LD with GWAS hits that are associated with a range of diseases such as Multiple sclerosis, Crohn's disease and many different types of cancers [94]. A number of reasons could explain why this wasn't replicated in our PPMI/NABEC MELT analysis. First our "strong LD" cut off was more stringent than previous studies (such as the later that instead used a R2 cut off of 0.7). In addition, in our present study we focused only on the common *Alu* elements that were already detected in the 1000 genome project, hence we currently have limited detection power. Due to time restraints caused by licensing issues we were only able to run MELT-Single, meaning that the only genotypes we could call at high confidence were the ones with a prior ID so the majority of the non-reference TE *Alu*s were not called in our analysis.

Future studies will focus on running MELT-group which will produce calls for all non-reference TEs and thus substantially increase our detection power. When discussing the LD, it is also important to note that the r2 and D' values may be inaccurate. This may be due to that fact that the current calling method only calls the presence/absence of the non-reference TE. As we have described throughout, TEs are also commonly polymorphic in their sequence length. Therefore, this uncaptured form of variation within these regions may skew the LD values. In addition, non-reference TE are normally observed in much lower frequencies then their proxy SNPs and this difference in allele frequency can weaken the LD values.

Focusing on the contribution of these elements to risk of PD specifically, we report that many non-reference TE map to PD risk loci. Further we identify that common non-reference TEs are in moderate linkage with unexplained PD risk variants. Remarkably in one case we observe that the present non-reference TE is most commonly inherited with the PD risk allele. As shown in Figure 5.11 a non-reference TE SVA is present upstream of *CRH1*, which the nearest gene to the rs62053943 PD risk variant (p=3.58E-68, β=-0.27). This locus contains genes that have already been extensively studied in neurodegenerative disease such as MAPT (which is the gene that encodes for *tau* that is the predominant component of neurofibrillary tangles that are neuropathological hallmarks of AD) and *KANSL1*[22,197,223]. Despite that fact that this locus has already been heavily studied there is still little known about how it contributes to risk of PD. [94] To note, *de novo* SVA insertions have been shown to cause genetic disease by disrupting gene expression and causing alternative splicing. In addition, the size of the SVA repeat can be a disease modifier, as observed with the *TAF1 de novo* SVA insertion causative of XDP, whereby repeat length correlates with age at onset [103]. Here we report that a common ~2/3kb non-reference SVA insertion (MAF = 0.19) is most commonly inherited with the risk haplotype (T). Therefore, this is suggestive that the SVA insertion could be contributing to disease risk at this locus. Further study and validation are needed to understand the relationship between the SVA and risk of PD.

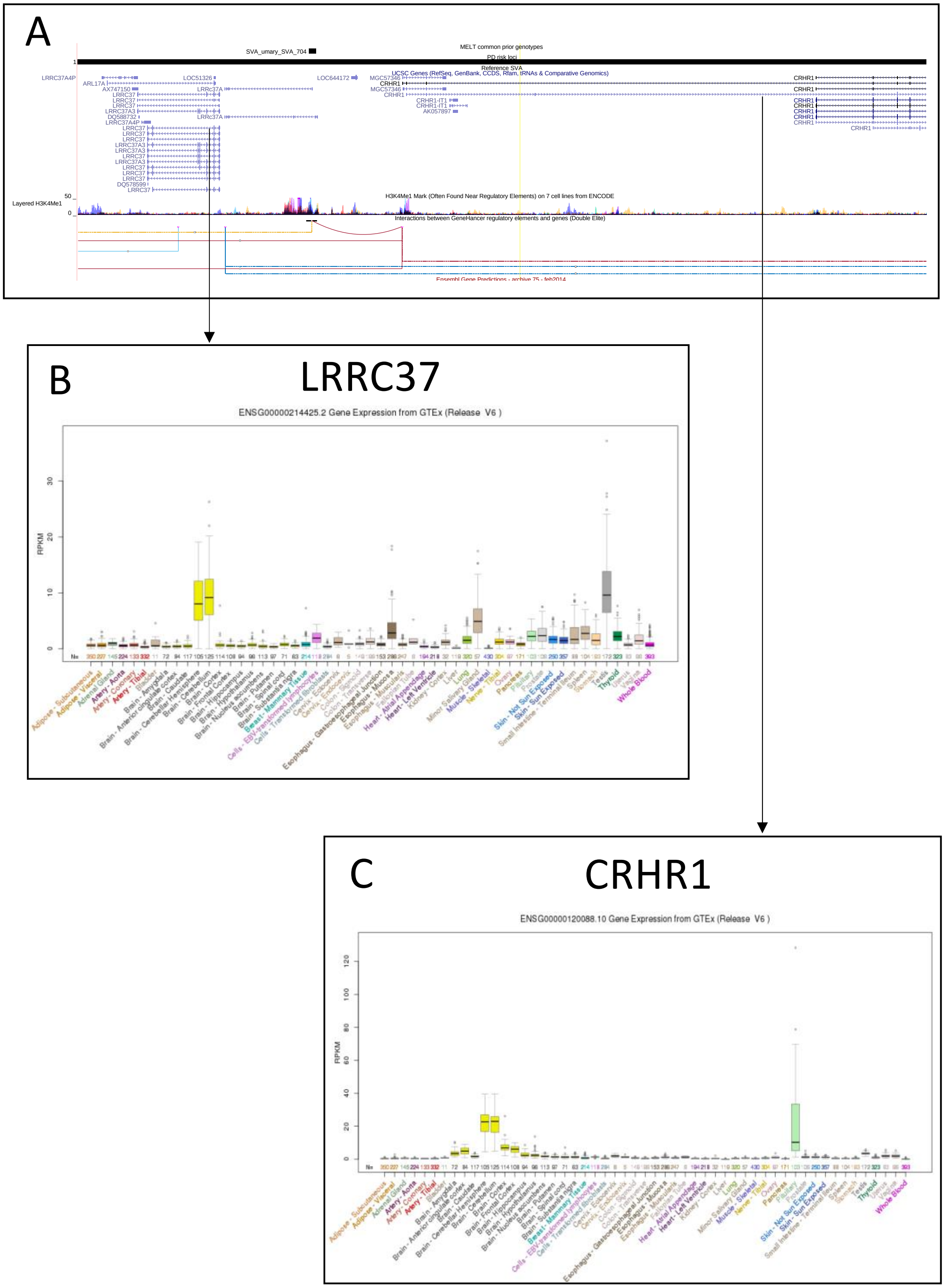


Figure 5.11. A common non-reference SVA that is in LD with the rs62053943 PD risk variant and commonly inherited with the risk allele lies within the promoter of LRRC37A (a PD associated gene)A) The rs62053943 locus, red highlight illustrates the region of the non-reference TE SVA (SVA_umary_SVA_704). The non-reference TE SVA overlaps the promoter of LRRC37A a PD associated gene. Histone marks and geneHancer infer possible functional consequence of the insertions. B) GTEX generated figure showing that the adjacent gene LRRC37 is highly expressed in brain (Cerebellar hemisphere and Cerebellum) and testis. C) GTEX generated figure showing that the

adjacent gene LRRC37 is highly expressed in brain (Cerebellar hemisphere and Cerebellum) and pituitary.

Finally, using the non-reference TE and SNP genotypes we ran an association analysis to identify if any of the non-reference TE were associated with PD risk with genome-wide significance. From this pilot study we did not detect any variant to reach genome-wide significance (Figure 5.9) ($p > 5 \times 10^{-8}$). But this was expected given the number of individuals included. As a control the beta coefficient values for the top PD risk SNPs were plotted from the most recent meta-analysis against the beta coefficient values of this present study, which positively correlated with our values (p=8.05E-04) (Figure 5.10). This highlights that in order to be able to fully establish the role of non-reference TE in PD the GWA should be repeated with a much larger dataset. Further we have shown that multiple non-reference TE are in LD with known PD risk loci, but our study did not have enough power to identify if any variant was significantly associated with risk of PD. To identify if the tagged TE variants are the true risk signal at these associated regions conditional and joint association analysis is needed, which requires significant association signals for testing. Therefore, not only is it imperative to increase sample size to discover new risk loci but it is also essential for further understanding whether the tagged TEs could be strong candidate causal variants for future follow-up study at these known loci.

While this present study marks the first characterization of non-reference TE in PD to date, there are a number of limitations to this analysis. The most significant being that we have not currently validated any of the variants of interest at the bench. It will be crucial to systematically validate the non-reference TE genotypes that have been called by MELT, to be able to take any variant-specific conclusion from this analysis. However, foreseeing this issue, steps were taken in the initial study design to ensure that we obtained robust and high confidence calls. First, we chose to use MELT for non-reference TE detection, which is the gold-standard TE detection tool[217]. MELT has been used by others to explore the role of non-reference TE in disease and was found to have high sensitivity and specificity [224]. It was developed to call non-reference TE in the 1000 genome project and provides a list of insertions that had been found and validated in this initiative. Therefore, because we only focused on these "prior" variants in our initial study it gives confidence that there are corresponding insertions at those defined sites. In addition, we also filtered for common variants (MAF < 0.01). Due to the nature of current non-reference TE detection tools and the samples size used, (similar to SNP calling) calling rare non-reference TE insertions is less reliable. Therefore, we hope that taking these outlined steps has improved the reliability of the non-reference TE calls.

As mentioned above in order to obtain high confidence calls, we focused solely on common and prior detected non-reference TEs. Although this increased the reliability of the calls it does seriously limit the detection power. Therefore, our current analysis is potentially a huge underrepresentation of the contribution of non-reference TE to PD risk. Two main factors contribute to this underrepresentation 1) only calling non-reference TE that are called in the 1000 genome project means we have ignored at least ten times the number of non-reference TE variants and 2) focusing on common variants means that we cannot capture the contribution of rare non-reference TEs to PD. In relation to the later, rare coding SNP variation has already been shown to contribute to

PD risk in existing exome studies [225]. Concerning TE variation, although there isn't the power to address this yet in PD, a recent study addressed the role of rare variation in severe developmental disorders (DD). Using WES from nearly ten thousand DD individuals Gardner *et al* identified 4 *de novo* TEs which were likely causative of the patient's symptoms[226]. While assessing the role of *de novo* TEs in sporadic PD is currently beyond the scope of our analysis, (as our data does not contain a large number of trios), the DD study does highlight that in the future assessing the genetic burden of rare non-reference TEs would be possible and could be informative in PD datasets.

Overall, we report that non-reference TE are a common yet ignored form of genetic variation that are enriched at PD loci and could be contributing to disease risk. Non-reference TEs are not currently catalogued at these loci, in fact these common variants are routinely removed from most genetic analysis as the first step of filtering. In recent years the technology has been developed to confidently call TEs genome-wide. Therefore, it is evident that as we enter the era of WGS and long-read sequencing, to aid in fine-mapping of existing risk loci and the discovery of novel hits, it will be imperative to incorporate non-reference TE detection as part of the standard NGS pipelines.

Chapter 6 Analysis of repetitive element expression in the blood and skin of patients with Parkinson’s disease identifies differential expression of satellite elements

6.1 Introduction

Despite huge successes for the field identifying genetic mutations and risk factors associated with PD, to date there has been little success in developing definitive diagnostic and prognostic biomarkers for the disease. The only definitive diagnosis for PD is performed post-mortem. As onset of the molecular and cellular neuropathology seen in PD likely initiates decades before the manifestation of the motor symptoms, the need for developing a diagnostic marker in readily available tissue is urgent, not only for early intervention but also to monitor progression of therapeutic treatments [227]. Recent efforts have focused on identifying biomarkers of PD in peripheral tissues, with studies identifying molecular alterations in the blood and skin of PD patients. Notably the transcriptional profile from the blood and skin of PD patients demonstrated dysregulation of genes known to be associated with PD [228–230].

Repetitive element (RE) sequence constitutes the majority of the human genome. Recently the field have seen the development of more sophisticated bioinformatic methods, that can now accurately analyse RE expression and this has led to the role of RE in disease aetiology becoming increasingly apparent. Although we have given a detailed description of TEs in Chapter 1, RE's include all repetitive elements including TEs, in brief; RE can be broadly split into five categories each of which, have very distinct functions. The first four minor categories account for ~10% of the genome, and include; simple sequence repeats, segmental duplications, tandem repeats and satellite DNA sequences, and processed pseudogenes. The fifth and most major and abundant group of RE are transposable elements (TE)s[231].

Differential expression of TEs has been associated with several neurological disorders and increased expression is linked with toxicity and genomic instability [232–235]. In response to this, cells have developed various epigenetic mechanisms to ensure TEs are tightly suppressed. However it would appear this mechanism goes awry in pathological state, with increased RTE expression *being* reported in conditions such as schizophrenia, Rett syndrome, Creutzfeldt-Jakob disease (CJD), ataxia telangiectasia and many cancers [236–238]. Specifically, accumulation of RTE transcripts has been described in several neurodegenerative diseases such as Alzheimer's disease (AD) and Amyotrophic lateral sclerosis(ALS)[232–234,239]. A more relevant example of the potential toxicity of RTEs has been highlighted in a recent PD related study, which focussed on assessing the role of LINE1 (a non-LTR, LINE RTE element) in mesencephalic dopaminergic neurons. In an Engrailed-1 heterozygote model Blaudin *de The* show that LINE1 RNA upregulation correlates with increased DNA damage and cell death induced by oxidative stress. Subsequently reduction of LINE1 protects against oxidative stress *in vitro* and *in vivo*[240]. Despite the growing body of evidence that shows that RE expression is associated with many diseases, genome-wide expression of these elements is yet to be characterized in PD. In this chapter, we utilized existing RNA-Seq data from the skin[229] and blood of the same individuals and characterized RE expression in both PD patients and healthy controls. To gain novel insight into the expression of all classes of RE, unlike other methods that focus only on TE, we utilized the well-established RepEnrich pipeline that quantifies all known RE class sequences and report the first characterization of RE expression in PD to date.

6.2 Aims

- Characterize the expression of repetitive elements in the blood and skin
- Identify if repetitive elements are differentially expressed in the blood or skin of patients with Parkinson disease

6.3 Methods

6.3.1. Study participants and ethics

Blood samples were collected from 12 patients with PD (6 men and 6 women, aged 72.2±9.9 years, mean ±SD), and 12 healthy control subjects (blood; 6 men and 6 women, aged 68.9±6.9 years, skin). All patients fulfilled the Queen Square Brain Bank Criteria for idiopathic PD[241,242]. The mean onset age and duration of disease at sample collection were 65.5±8.6 and 7.3±6.3 years, respectively (Table 6.1).

Table 6.1. Characteristics of the PD patients: =Hoehn and Yahr stage; SE-ADL, Schwab and England Activities of Daily Living Scale; MMSE, Mini Mental State Examination. * 1-tremor-dominant; 2-akinetic-rigid; 3-postural instability and gait disorder.

SEX	AGE (YR)	DISEASE ONSET AGE (YR)	DISEASE DURATIO N (YR)	DISEASE SUBTYPE *	HY	SE-ADL	MMS E
M	85	67	18	3	4	40	27
M	76	75	1	1	2.5	90	30
M	73	65	9	2	3	80	30
M	67	50	17	2	4	70	28
M	69	68	2	3	3	80	30
M	73	72	1	1	1	95	29
F	82	74	9	3	4	60	26
F	68	65	4	1	1.5	100	30
F	71	66	6	2	3	70	24
F	48	47	2	2	1.5	90	30
F	69	67	3	1	2.5	80	29
F	85	70	15	1	2.5	60	23

A detailed description of individuals enrolled on the skin study please refer to Planken *et al* [229]. In brief, the median total score of the Movement Disorder Society sponsored revision of the Unified Parkinson Disease Rating Scale (MDS-UPDRS)[243]was 65 (ranging from 22 to 159). The median disease severity assessed by the Hoehn and Yahr scale[244] was 2.75 (ranging from 1 to 4). The median disability score assessed by the Schwab and England Activities of Daily Living Scale (SE-ADL) was 80 % (ranging from 40% to 100%). None of the PD patients were current smokers and two had a history of smoking in the past.

`The study was approved by the Research Ethics Committee of the University of Tartu. Volunteer PD patients and healthy controls were recruited from the Department of Neurology and Neurosurgery at the University Hospital of Tartu. A signed informed consent was acquired from all subjects participating in this study.

6.3.2. Library preparation

The venous blood of all study subjects was collected into Tempus Blood RNA Tubes (Thermo Fisher Scientific Inc, CA, USA). The RNA was extracted applying Tempus Spin RNA Isolation Kit (Thermo Fisher Scientific Inc, CA, USA) combined with DNase treatment (RNase-Free DNase Set, Qiagen, Hilde, Germany), according to the manufacturer's' protocols. The globin mRNA was removed from the extracted total RNA using GLOBIN clear Kit for human (Thermo Fisher Scientific Inc, CA, USA). For the skin, one 4 mm punch-biopsy specimen was taken from non- sun-exposed skin of each subject from both study groups. All biopsy specimens were instantly frozen in liquid nitrogen and stored at -80C° until RNA extraction. Biopsies were homogenized with Precellys 24 homogenizer with the Cryolys system (Bertin Technologies). RNeasy Fibrous Tissue Mini Kit (Qiagen, Hilde, Germany) was used for total RNA extraction, according to the manufacturer’s protocol. During the purification on-column DNase I treatment was performed (Qiagen Hilde, Germany). The RNA quality was assessed using Agilent 2100 Bioanalyzer, with the RNA 6000 Nano kit (Agilent Technologies) and the quantity was evaluated with Qubit fluorometer and Qubit RNA HS Assay kit (Life Technologies). The study samples RIN ranged from 6.7- 9.5 in the blood and skin samples.

6.3.3.RNA sequencing

50 ng of each RNA sample was amplified with Ovation RNA-Seq System V2 Kit (NuGen Technologies Inc, CA, USA) and the output double stranded DNA was used to prepare SOLiD 5500 W System DNA fragment libraries according to manufacturer's protocols (Thermo Fisher Scientific Inc, CA, USA). For library preparation, the barcoding adapters were used, and 12 libraries were pooled prior to sequencing. For sequencing skin samples, the SOLiD 5500 W XL with paired-end chemistry (75 bp in forward and 35 bp in reverse direction) in 6-lane mode was applied. In the case of blood samples SOLiD 5500 W XL platform with fragment sequencing chemistry (75 bp in forward directions) in 3-lane mode was used. In both cases approximately 40 million mappable reads were expected per one sample, which is enough for successful whole transcriptome expression pattern analysis.

6.3.4. Read alignment and quantification

Raw colour-space reads were filtered for rRNA, active tRNA, and SOLiD adaptor sequences. The remaining reads were aligned as single end reads to the GRCh37/hg19 reference genome, while allowing multi-mapping to detect reads aligning to possible repeat sequences. To ensure secondary alignments were reported in the BAM files no mapping quality cut-off was set. LifeScope software (LifeTechnologies) with recommended settings designed for colour-space read alignment and analysis was used for both mapping steps.

The number of reads aligning to known exonic gene sequences were counted and visualized by plotting the first two principal components in order to exclude the possibility of bias in the RNA-seq datasets. The base Stats package in R was used to conduct the principle component analysis (PCA)[245]. Prior to PCA, the read counts were normalized as z-scored counts per million mapped reads (CPM) values, where the standard deviation and mean were calculated separately for each gene.

To connect colour-space mapping with the RepEnrich pipeline https://github.com/nskvir/RepEnrich[231], the GRCh37/hg19 mapped BAM files were parsed using samtools and in-house perl scripts to separate unique and non-unique mapped reads. For uniquely mapped reads, only alignments with MAPQ ≥ 10 (in Phred scale) were retained, on average ~77% of all reads mapped to the reference genome. For multi-mapping reads, the base-space sequence was inferred from the longest alignment and these reads were converted to FASTQ format. This enabled us to convert the colour-space data into suitable format for downstream analysis. Next, RepEnrich with default parameters was applied to obtain read counts of REs. RepEnrich aligns multi-mapping reads separately to a pseudogenome containing the RE loci lifted from the RepeatMasker GRch37/hg19 Library in order to more accurately infer read counts which estimate the abundance of expressed RE. The RepEnrich pipeline applies different quantification strategies to uniquely and multi-mapping reads in order to more accurately infer read counts, which accurately estimates the abundance of expressed repetitive elements. The alignments were quantified at repeat class, family and subfamily level. Repeat subfamilies are a collection of highly similar sequences representing all known instances of a given repetitive element copies in the hg19 genome build annotated in the RepeatMasker Library."

6.3.5. Analysis of differentially expressed repetitive elements

The R package edgeR (Robinson et al. 2010) was used to identify differentially expressed REs at sub-family, family and class level between the case and control in the blood and skin. The EdgeR package uses a negative binomial model to infer the significance of differential read counts. The RE overall library sizes were used for normalization of read counts for each sample. Prior to differential expression testing, the edgeR normalized RE pseudocounts were visualized by plotting the first two principal components in order to check for potential bias in the RE counts data. The generalized linear model approach of EdgeR was then applied to compare PD to control. The workflow used for the RepEnrich analysis is outlined in (Figure 6.2).

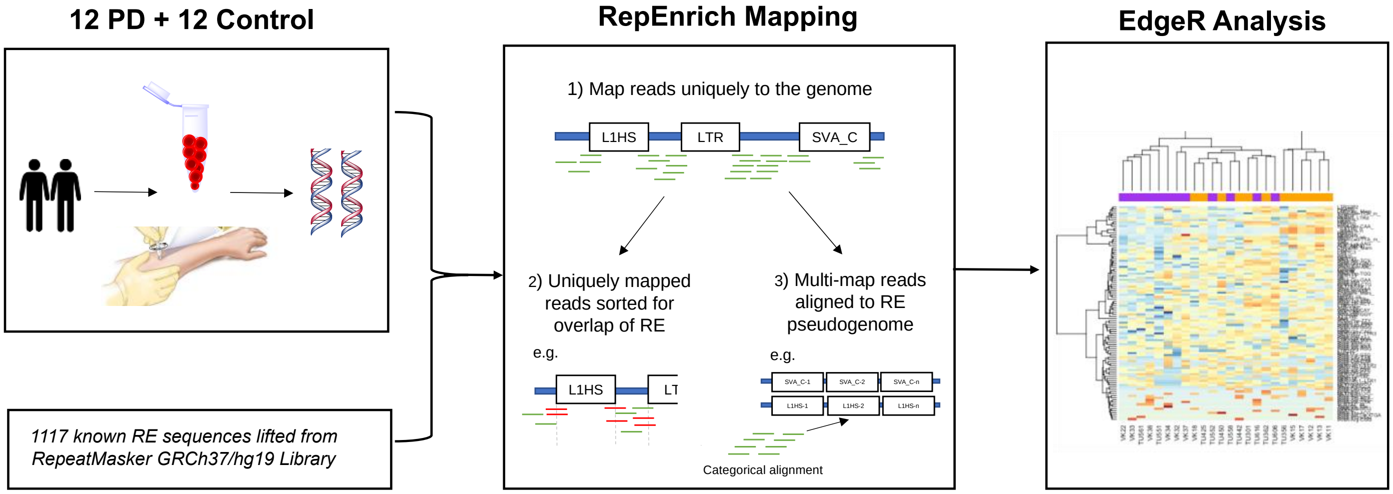

12 PD + 12 Control
RepEnrich Mapping
EdgeR Analysis
1) Map reads uniquely to the genome
L1HS
LTR
SVA_C
2) Uniquely mapped reads sorted for overlap of RE
e.g.
L1HS
LT
3) Multi-map reads aligned to RE pseudogenome
e.g.
SVA_C-1
SVA_C-2
SVA_C-n
L1HS-1
L1HS-2
L1HS-n
Categorical alignment
1117 known RE sequences lifted from RepeatMasker GRCh37/hg19 Library

Figure 6.1. RepEnrich Workflow. RNA-Sequencing data was obtained from blood and skin of 12 PD patients and 12 healthy controls. A RE pseudogenome assembly was constructed by concatenating the genomic sequence for the 1117 RE elements from the ReRCh37/hg19 Library. Reads were mapped using the RepEnrich pipeline and differential RE expression was identified following EdgeR analysis.

6.4 Results

6.4.1 Repetitive elements are widely expressed in the blood and skin

Analysis of RNA-Seq data from 12 PD patients and 12 controls (with an average of 24 (blood) or 31 (skin) million reads per individual) identified that 20.3% (blood) and 23.8% (skin) of reads mapping to the reference genome GRCh37/hg19 aligned to the RE annotated in the RepeatMasker GRch37/hg19 Library (Figure 6.3).

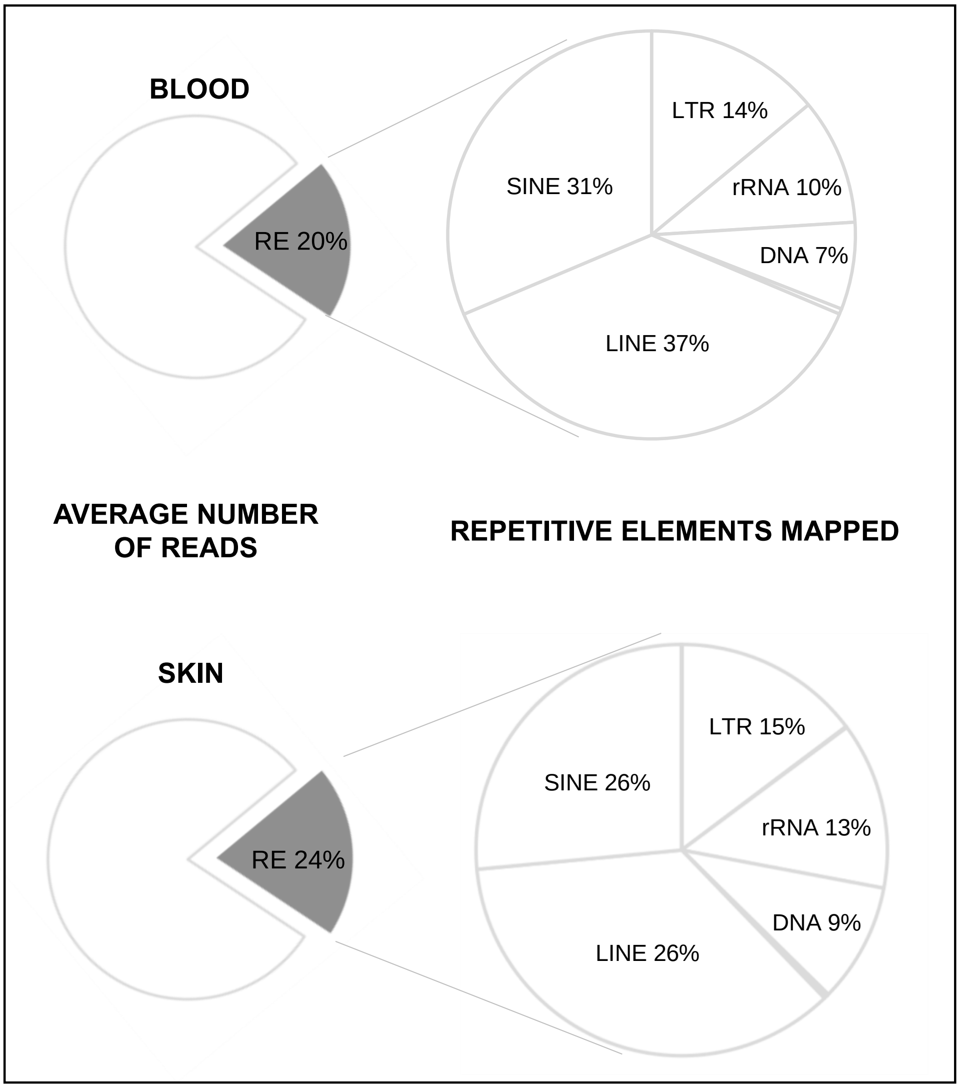


Figure 6.2. Mapped repetitive element expression in the blood and skin. Analysis of RNA-Seq data from the blood of 12 PD patients and 12 controls (with an average of 24 (blood) or 31 (skin) million reads per individual) identified 20 % of reads mapping to the reference genome GRCh37/hg19, aligned to the custom built RE psuedogenome assembly used in RepEnrich. Of the reads that mapped to the RE pseudogenome assembly, in the blood on average 37.10% originated from LINE elements, 31.22% from SINE, 13.93% from LTR, 10.41% from rRNA, 6.90% from DNA and 0.44% other (satellite, snRNA, tRNA,RNA,RC,scRNA). In the skin on average 35.71% originated from LINE elements, 26.47% from SINE, 14.70% from LTR, 13.14% from rRNA, 9.41% from DNA and 0.57% other (satellite snRNA, tRNA,RNA,RC,scRNA).

No significant bias was detected in the RNA-Seq datasets by visualizing the read counts for expressed genes as principal component analysis (PCA) plots. REs were widely expressed, on average of the 1117 REs queried, 1086 were detected in the blood (RPKM≥1) (97%) and 1099 in the skin. Based on RepeatMasker annotations RepEnrich output is quantified into three categories 1) expression of every RE that has a known consensus sequence in RepeatMasker, which is named as 'subfamily' classification (n= 1117) 2) grouping all of the REs by 'family' (n=48) and 3) further sub grouping of the families by 'class' (n= 13). An explanation of the grouping of the RepEnrich output for expression levels is given below (Table 6.2).

Table 6.2. Explanation of RepEnrich output for expression levels. Based on RepeatMasker data the output is quantified in three categories 1) expression of every RE that has a known consensus sequence in RepeatMasker (subfamily) (n= 1117) 2) grouping all of the RE by family (n=48) and 3) grouping the families further into classes (n= 13).

RE Subdfamily (n=1117)	**RE family (n=48)**	**RE class (n=13)**
Every known RE consensus sequence in RepeatMasker	satellite,centr,acro,telo	Satellite
	RNA	RNA
	Helitron	RC
	scRNA	scRNA
	rRNA	rRNA
	tRNA	tRNA
	srpRNA	srpRNA
	ERVL,ERVL-MaLR,ERV1,Gypsy,LTR,ERVK,ERV1?,Deu,Gypsy?,ERVL?	LTR
	Other	Other
	Dong-R4,L1,CR1,RTE-BovB,L2,L1?,RTE-X	LINE
	Alu,MIR	SINE
	TcMar-Mariner,TcMar?,hAT-Tip100,DNA,hAT-Charlie,hAT-Tip100?,hAT-Blackjack,PiggyBac?,hAT?,TcMar-Tc2,TcMar-Pogo,TcMar,PiggyBac,TcMar-Tigger,hAT,Merlin,MULE-MuDR	DNA
	snRNA	snRNA

No significant difference for relative abundance of reads originating from each of the different RE classes was observed between PD and control in the blood or in the skin. Of the reads that mapped to the RE pseudogenome assembly, in the blood on average 37.10% originated from LINE elements, 31.22% from SINE, 13.93% from LTR, 10.41% from rRNA, 6.90% from DNA and 0.44% other (satellite, snRNA, tRNA, RNA, RC ,scRNA).In the skin on average 35.71% originated from LINE elements, 26.47% from SINE, 14.70%

from LTR, 13.14% from rRNA, 9.41% from DNA and 0.57% other (satellite, snRNA, tRNA, RNA, RC, scRNA) (Figure 6.3).

However when the abundance of individual members of each class were compared between the two tissues analysed, there were significant differences in relative abundance of RE between blood and skin for the majority of the classes, highlighting previously reported tissue-specific nature of RE expression[246].

6.4.2. Satellite elements are significantly upregulated in the blood of PD patients

EdgeR analysis was applied to compare RE expression between PD patients and healthy control subjects in both the blood and skin. The edgeR -normalized pseudocounts of REs were visualized on PCA plots and no possible biases affecting the analysis were observed (Supplementary Figure 3-6).. No significant differences in RE expression were observed comparing PD patients and healthy controls in the skin (FDR ≤0.01). However, upregulation of satellite REs at the class level with a $\log_2$FC increase of 1.93 (FDR = 7.7E-06) was identified in the blood from PD patients compared to the blood from healthy control subjects**.**

At the family level, 3 RE families were significantly differentially expressed at FDR ≤ 0.01; satellite, centr and acro (FDR = 7.88E-06, 7.88E-06, 6.39E-04 respectively), which were upregulated with a logFC increase of 2.05 for satellite, 1.84 for centr and 1.30 for acro in the blood from the PD patients**)**.

At the subfamily level four specific satellite class REs were significantly differentially expressed at FDR ≤ 0.01, two simple satellite IIIs (repName= CATTC_n and _GAATG_n) a high-copy satellite II (repName= HSATII) and a centromeric satellite (repName= ALR_Alpha) all of which were upregulated in the blood of PD patients (Table 6.3).

Table 6.3. Differentially expressed repetitive elements identified in blood from PD patients. Showing characteristics of each differentially expressed element, log FC, log CPM, p-value and FDR (≤0.01 cut off).

Class	**Family**	**RE**	**Description**	**Log FC**	**Log CPM**	**P-value**	**FDR**
Satellite	Satellite	_CATTC_n	Simple satellite III	4.4	7.74	2.27E-12	1.56E-09
Satellite	-	HSATII	High-copy satellite II	4.12	5.69	2.79E-12	1.56E-09
Satellite	Satellite	_GAATG_n	Simple satellite III	4.23	7.32	1.68E-11	6.25E-09
Satellite	Centromeric	ALR_Alpha	171bp satellite associate with human centromeres	2.02	8.69	5.20E-08	1.45E-05

Moreover, the expression levels of these specific satellite elements displayed little inter-variability in the PD patient group when compared to healthy controls (Figure 6.3).

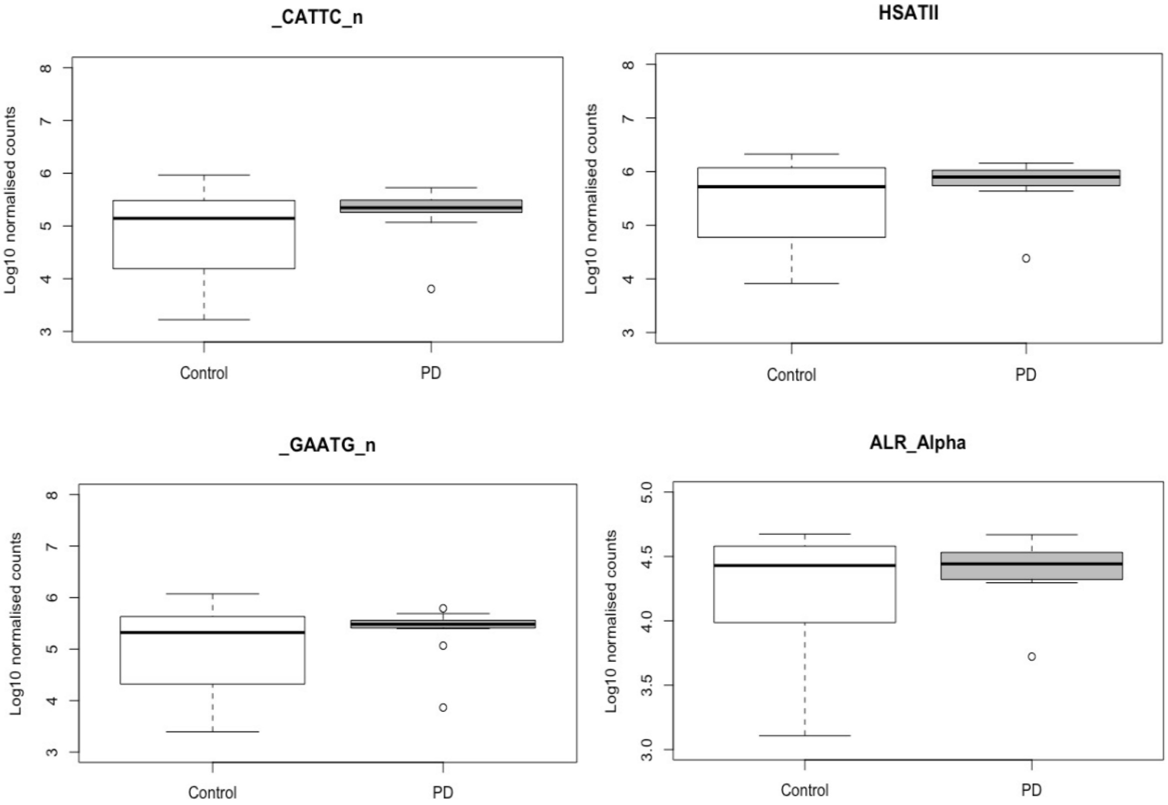


Figure 6.3. Upregulation of Satellite elements in the blood of PD patients. At the subfamily level four satellite class repetitive elements were significantly differentially expressed at FDR ≤ 0.01, two simple satellite III (repName= CATTC_n and _GAATG_n) a high-copy satellite II (repName= HSATII) and a centromeric satellite (repName= ALR_Alpha) all of which were upregulated in the blood of PD patients. Simple satellite III repeat (CATTC)n RNAs were the most significantly upregulated in the blood of PD patients (p-value= 2.27E-12) with a logFC increase of 4.40. Pericentromeric human satellite II (HSATII) repeat derived RNAs were significantly upregulated (p-value=2.79E-12) with a logFC increase of 4.12. Simple satellite III (GAATG)n derived RNAs were upregulated in PD (p-value=1.68E-11) with a logFC increase of 4.23. Finally, human alpha centromeric satellite (ALR_Alpha) derived RNAs were also upregulated in PD (p-value=5.20E-08) with a 2.02 logFC increase.

Simple satellite III repeat (CATTC)n RNAs were the most significantly upregulated in PD blood (p-value= 2.27E-12) with a logFC increase of 4.40. Pericentromeric human satellite II (HSATII) repeat derived RNAs were also significantly upregulated (p-value=2.79E-12) with a logFC increase of 4.12. HSATII derived RNA should be undetectable in normal tissue and dysregulation of these elements has shown to induce genomic instability [247]. Simple satellite III (GAATG)n derived RNAs were also upregulated in the blood of PD patients compared to the blood of healthy control subjects(p-value=1.68E-11) with a logFC increase of 4.23. Finally, human alpha centromeric satellite (ALR_Alpha) derived RNAs were also upregulated in the blood of PD patients compared to the blood of healthy control patients (p-value=5.20E-08) with a 2.02 logFC increase (Figure. 6.3).

6.5 Discussion

This analysis included a detailed characterization of the expression of REs in the blood and skin. It and also presents the first genome-wide analysis of RE expression in PD and highlights the previously reported tissue-specific nature of RE expression [246,248]. Using a stringent FDR cut-off of 0.01 we found that there was no differential expression of REs in the skin when PD patients and healthy controls were compared. However in the blood we identified that satellite elements are upregulated in PD patients and a group of satellite elements, (repName= CATTC_n, HSATII, ALR_Alpha) which are a group of elements that have been collectively associated with genome instability [247,249], are significantly differentially expressed.

Characterization of RNA-Seq data in the 24 subjects studied identified that REs were widely expressed and constituted 20% of all expressed transcripts in the blood and 24 % of all expressed transcripts in the skin. We report no significant difference in relative abundance of global RE expression between patients versus controls for either tissue. Our data is in agreement with that reported by Faulkner *et al* who performed a comprehensive RE analysis using cap analysis gene expression (CAGE) sequencing data and determined that in human tissue, on average around 20% of all CAGE tags detected were mapped to REs and overall RE expression varied significantly between tissue[246]. Our data is also supported by a recent study from our group that analysed skin RNA-Seq data from 12 individuals with psoriasis and 12 healthy controls with the same RepEnrich pipeline and found that on average 27.5% of reads aligned to REs[248].

In light of the recent associations between TE dysregulation and neurodegenerative disease, more specifically the identification of LINE1 overexpression inducing death of mesencephalic dopaminergic neurons, we set out to identify if these elements were differentially expressed in the peripheral tissues of PD patients. Our analysis used a method that not only determined differential expression of the different classes of TEs but all class of RE. Although we did not observe differentially expressed TEs in the skin or blood of PD patients, we did identify upregulation of satellite elements in the blood of PD patients (Figure 6.3).

Overall aberrant overexpression of satellite repeats has been associated with genomic instability [250,251]. Collectively three of the upregulated satellite elements identified in this study (repName= CATTC_n and _HSATII and ALR_Alpha) have been named as a group of REs involved in genomic instability and considered to be transcriptionally silent in the genome[247]. Although the simple satellite III RE (repName=_GAATG_n) was also significantly upregulated, like many of the REs, there is a paucity of information available for the function of this RE in the literature. Interestingly, in addition to increased expression levels, a considerably small degree of variability was observed in the PD patients compared to the healthy control subjects. This could hint at a yet undetermined regulatory process that can be associated with PD pathophysiology.

Literature have shown that inducing dysregulation of genes known to silence satellite elements can promote genomic instability, which consequently can result in; growth arrest, impaired homologous recombination and spontaneous DNA breaks through this pathway. An example of this is apparent with the gene *SIRT1* (Silent information regulator-1) which is a known repressor of repetitive DNA. It has been shown in mouse ES cells that when *SIRT1* is inhibited one of the major consequences is activation of major satellite repeats and thus an increase in expressed satellite transcripts is observed[252]. This process has been associated with mitochondrial dysfunction and oxidative stress, which are common pathways that are affected during ageing and that have been implicated in PD [253]. However, it is unknown if the observed upregulation of satellite elements is a mechanistically important factor in the aetiology of PD or if it is simply an indicator of pathophysiological state; thus, it will of interest in the future whether such pathways and genes are modulated in the CNS or in immune cells in PD. As shown in Figure 6.3 there is a striking loss of variability in the expression of the group of satellite elements upregulated in PD, which is pattern that is not observed in other RE elements (Figure 6.4). It is also important to mention here that there is a possibility that the white cell count of the blood used in this present study may be contributing to the observed differences of satellites element expression in the blood of individuals with PD. The precise function of the immune system in PD aetiology is currently a controversial topic of debate. Despite this it has been repeatedly shown than white blood cells are dysregulated in PD270. Of interest, it has been recently shown that HSATII RNA is also highly expressed in human cells infects with two herpes viruses. Further this study suggests that HSATII RNA synthesis post infection has important functional consequences for viral replication271. Therefore, this could suggest that

HSATII dysregulation is not a disease-specific observation but an indicator or cell-stress following disease or infection.

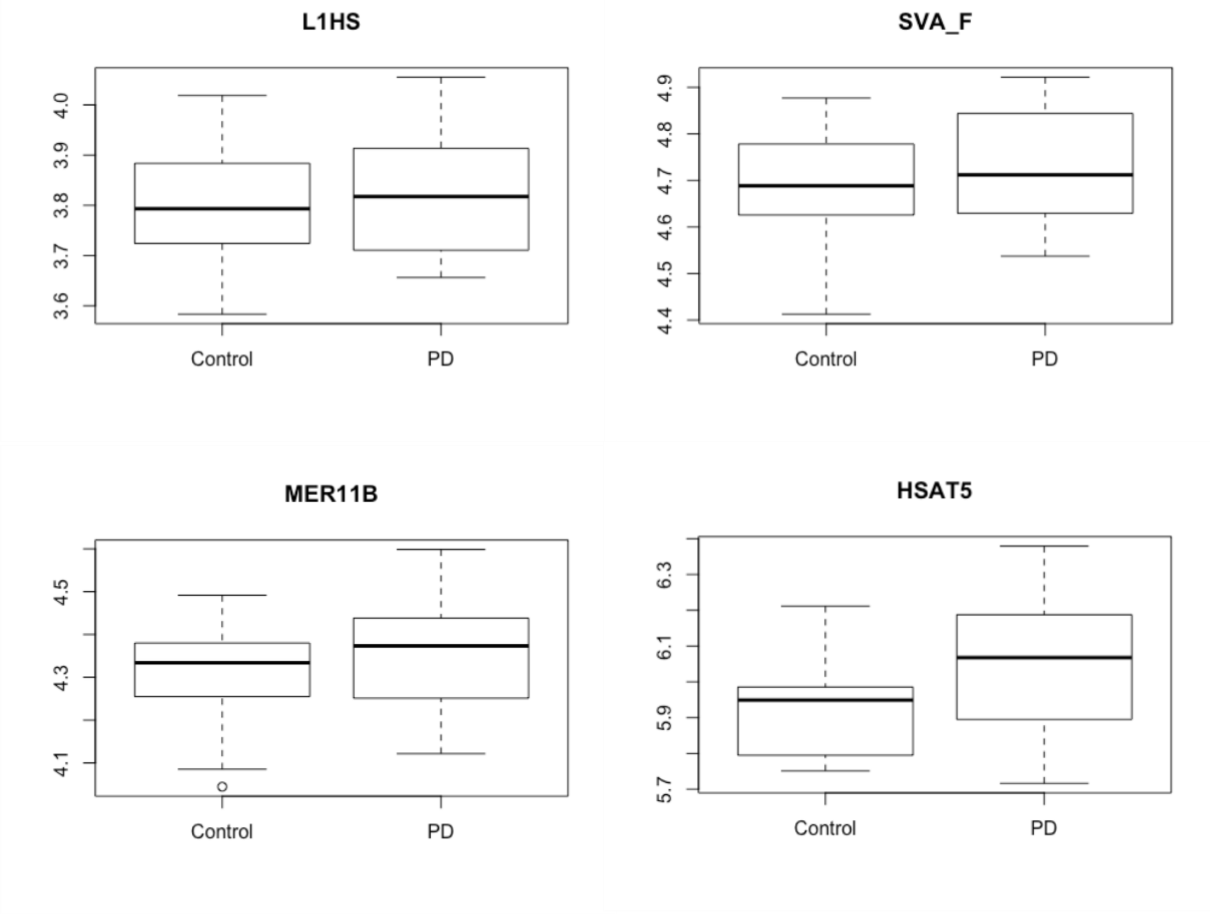


Figure 6.4 Following the lack of observed variability in the upregulated satellite elements in PD individuals, four non-significantly expressed , randomly selected REs were plotted. This included 3 other non-significant REs (Repname =L1HS, SVA_F, MER11B) and another satellite element (Repname = HSAT5). We did not observe this lack of variability in PD individuals in the non-significant individuals, further strengthening the point that this is a disease-specific signature only observed with this particular set of upregulated elements.

Therefore our data indicates that overexpression of the specific satellite elements is a potential disease-specific signature in the blood for PD and highlights the need for further characterization of our model in a larger, better clinically defined, data set to also address if there is possible association with progression of the disease. The PPMI cohort has blood RNA-Seq available and therefore TE dysregulation will be further addressed in this cohort during my post-doc position.

Overall in this chapter we set out to characterise RE expression with previously published PD RNA-Seq data in response to recent studies that have associated TE expression with neurodegenerative disease. This type of analysis has not been explored previously, mainly due to the lack of technology to do so. Using a tool that not only quantifies TE expression but all class of REs we report that a group of satellite elements are differentially expressed and appear to have a PD specific expression signature in the blood. Although we report the first genome-wide analysis of RE expression in PD to date, this analysis does have a number of limitations. First, we acknowledge that the cohort size (n=24) is particularly small. After demonstrating that there are significant disease-specific signatures of RE expression, I will now be addressing this in a larger PD expression dataset during my post-doctoral position at NIH. Another limitation is that we show through of our analysis (of the blood and skin of the same individuals) that RE expression occurs in a tissue specific manner. Thus, we cannot make any presumptions that this reflects what could be happening in the brain. Although a very interesting concept, addressing RE expression in the brains of PD individuals is particularly problematic as the areas of interest will have occurred a lot of cell-death at time of post-mortem. Despite the literature suggesting a possible role of RTE in the brain and neurodegeneration in particular, the purpose of our present study was to look for possible differences in RE expression in readily available tissue (such as blood and skin) as a potential biomarker.

In summary, the field is still far from establishing the molecular mechanisms underpinning PD and much progress is needed to develop an objective biomarker for the disease. In this chapter we utilized previously existing RNA-Seq data and characterized RE expression in the blood and skin of PD patients, our rationale being, 1) recent studies have shown that disease pathology can be detected in the peripheral tissue 2) more sophisticated bioinformatic methods have provided a growing body of evidence that RTE dysregulation is associated with neurodegenerative diseases such as AD,ALS and PD[232,233,239,240]. We identified firstly, overall tissue-specific differences in RE expression as supported by the literature and secondly that a specific group of satellite elements, that have been strongly associated with epigenetic instability, displayed altered expression in the blood of patients with PD. Further characterization is required to determine the consequence of upregulation of satellite elements in PD, however our data supplies a potential novel non-invasive biomarker of the disease and its progression.

Chapter 7

Thesis Summary

7.1. Conclusions:

The ultimate goal of the research presented in this thesis was to characterize the role of TEs in PD and present the first genome-wide analysis of TE expression and variation in PD to date. Previous to this, due to lack of technology to do so, TEs have not been included in any PD genetic or expression analyses and so are a completely overlooked source of genetic variation in the genome. In fact, TE's are routinely removed from WGS as an initial filtering step in analyses. Repetitive elements are largely completely "masked" in the genome, in fact for genome-wide analyses that generate "whole genome" datasets this data is routinely limited to a concentrated analysis of ~20-50% of the human genome corresponding to the non-repetitive portion defined by RepeatMasker[232]. This is because the majority of genetic data has been generated with illumina short-read sequencing, which has limited read length (50-300bp). Therefore, there is a fundamental superior performance for single-copy genic regions compared to repetitive DNA and due to this ambiguity of mapping the latter they are routinely removed in QC. Within recent years genome-enabled technologies have rapidly advanced which now allow for accurate and scalable TE detection. Following this development, not only is it evident that TEs are associated with many eQTLs and disease risk variants, but specific TE variants are now identified as being causative of disease (such as the XDP causing SVA at TAF1). In support of this, taken together the data from this thesis suggests that TE variation could be involved in PD aetiology. Further the data we present suggests that not only could integrating TE variants be a valuable and critical step forward for furthering our understanding of existing risk loci, but it could also be important for establishing new risk hits. At present, only one third of the heritable component of PD can be explained by the known risk variants, therefore incorporating TE variant detection into routine genetic analysis to address the "missing heritability" of this complex genetic disease is beneficial.

7.1.1. Reference SVA variation is an imputable and common source of uncaptured genetic variation in the genome

Expanding on the previous characterization of a reference SVA upstream of the *PARK7* gene by Savage et al, the initial aim of the research presented in this thesis was to address whether the reference SVA was an eQTL or associated with risk of PD. At the time of the original *PARK7* SVA analysis only five genes had been associated with PD. These five genes were "PD associated genes" as they harboured mutations that were causative of Mendelian forms of the disease[126]. Since then, following extensive GWAS and meta-analysis initiatives, ninety risk loci have been identified and through this it is now clear that the *PARK7* locus is not pleiotropic, i.e. not associated with risk of sporadic PD. Therefore, it is not surprising that we found no significant association between the *PARK7* SVA genotype and risk of PD in our present analysis (Chapter 2.). In addition, reference SVAs have been shown to direct gene expression in an allele-like manner [5], so we extensively addressed whether the SVA was acting as an eQTL for the *PARK7* transcripts (RefSeq and a longer, brain-specific transcript (that originated from another transcriptional start site adjacent to the *PARK7* SVA)) in the normal brain. We did not find a significant association between *PARK7* expression and SVA genotype, but our analysis is currently incomplete as we were unable to test for allele 4. Now that WGS is available it could be possible to generate proxy SNPs for the rarer allele and reanalyse whether the SVA is an eQTL at this locus. At present pursuing the *PARK7 l*ocus to further investigate sporadic PD would not be appropriate given that it is not an identified risk region. However, *PARK7* has also been associated with cancer, specifically tumour development and progression in non-small cell lung carcinoma[254,255]and breast cancer. Therefore, further work to generate proxy SNPs for all of the *PARK7* alleles would be beneficial to allow for bioinformatic analysis in other disease GWAS and expression datasets.

Most significantly through the work demonstrated in (Chapter 2) we have shown for the first time that common reference SVA variation can be imputed. Although we could not test allele 4 of the SVA using the present proxy SNPs, if WGS data was used rather than the original HAPMAP GWAS data, then the rarer allele will most likely be imputable. This is substantial as it is evident from (Chapter 3) that because of their genic nature, PD risk loci contain many reference SVAs. As SVAs are known to modulate gene expression, lie within potentially functional domains at PD loci and are variable in primary sequence, this is another layer of genetic variation that is currently unknown at these regions but could be disease related.

An example that highlights the potential advantage and importance of imputing TE variation and integrating this with SNP data is shown by a recent study from *Payer et al.* This study focused on *Alu* variants around GWAS loci and PCR genotyped over one hundred *Alu* insertions at these regions. After further bioinformatic analysis it was identified that several *Alu*s were in LD with GWAS risk variants, many of which were in LD with well-established cancer associations. Using individual-level cancer GWAS datasets and the associated *Alu*s proxy SNP's, *Payer et al* imputed the *Alu* variants and validated that *Alu* variants were on the risk haplotypes and associated with disease. This suggested that the associated *Alu* elements were good causal variant candidates and therefore they were nominated as good candidates for future functional follow-up studies at these loci[160]. This is an example of how integrating TE analysis can be shed light on possible causal variants at a risk locus. The majority of complex genetic diseases that are polygenic in nature have multiple risk loci spread across the genome. A major challenge to GWAS is understanding how these signals contribute to disease. In most cases GWAS signals still do not 1) identify the causal variant or 2) successfully nominate the target gene within these loci that are involved in disease mechanism. Hence, integrating TE genotypes with SNP datasets is not only beneficial for the study of the genetic contribution to PD but potentially for all complex genetic diseases.

7.1.2. Non-reference transposable elements are potentially important variants at PD risk loci

Moving beyond simple enrichment analysis we leveraged new TE detection tools (MELT) and individual level PD GWAS datasets (PPMI WGS) and performed the first characterization of non-reference TE in PD to date. It is only possible to detect presence/absence of TEs with the current TE detection tools. However, there is also variation within the “present” TE between individuals, but this cannot be addressed with the current technology. In addition, as highlighted in the previous section it would be extremely beneficial to address the variation within reference TEs, but the current technology does not exist to genotype reference TE variation on a genome-wide scale using short-read sequencing data (as reference TE’s are very repetitive and hence map to multiple regions of the genome).

The aim of our MELT pilot study was to identify whether running TE detection on this scale (~1000 genome) was feasible and informative for the study of PD. We hypothesized that TE variation could represent new risk loci and also explain existing PD risk variants. Although we did not have the power to discover new hits in our initial MELT association analysis, we could explore whether the TE variants were in LD with the known PD risk variants. The analysis was designed to pursue variants that were being called with high confidence in MELT so that detailed PCR analysis was not necessary to validate general conclusions from the initial pilot-data that would support future analysis. Like most SV detection tools MELT TE detection is more accurate if regions of interest are pre-defined. Therefore, variants were filtered that were common (MAF > 0.01) and that had been detected previously in the 1000 genome project (priors). Despite the fact that filtering for high confidence calls caused a significant decrease in the number of TEs that were included in the analysis, we report that even when only 2657 non-reference TE variants were analysed, remarkably we identified that two were in linkage with known PD risk variants and 165 mapped to known PD risk loci. Alu, LINE-1 and SVA insertions can all impact on gene expression networks through many different mechanisms such as alternative splicing and exonisation. Thus, TE's are another layer of genetic variation that could be modulating regulatory networks at PD loci that are not currently catalogued at these regions despite potentially being involved in disease aetiology.

Most significantly we identified that a non-reference SVA was in moderate linkage with a known PD risk variant and is commonly inherited with the risk haplotype at the *MAPT* locus. A non-reference TEs SVA is present upstream of *CRH1*, which the nearest gene to the rs62053943 PD risk variant (p=3.58E-68, β=-0.27). This locus contains genes that have already been extensively studied in neurodegenerative disease such as *MAPT* (which encodes for *tau* that is the predominant component of neurofibrillary tangles that are neuropathological hallmarks of AD) and *KANSL1*[22,197,223]. Despite that fact that this locus has already been heavily studied there is still little known about how it contributes to risk of PD. [94]Here we report that a common ~2/3kb non-reference SVA insertion (MAF = 0.19) is most commonly inherited with the risk haplotype (T). Therefore, this is suggestive that the SVA insertion is a strong causal variant candidate which should be validated with follow-up functional studies.

An example of a common non-reference TE that is associated with disease and affects the function of the gene that it inserted into is the non-reference SVA-E insertion in the *CASP8* gene. The SVA-E insertion within intron 8 of the *CASP8* gene is associated with transcript abnormalities and an increased risk of breast cancer but a decreased risk of prostate cancer[256]. As mentioned above we have identified that a non-reference common SVA is in moderate LD with the PD risk variant at the *MAPT* locus, which is also a locus associated with AD. As with PD, the role of the *MAPT* locus in influencing AD risk is still largely unknown. A recent study that explored the overlap between AD and PD at this locus found and replicated association of both AD and PD with the A allele of rs393152 within the extended *MAPT* region on chromosome 17 (meta-analysis p-value across 5 independent AD cohorts = 1.65 × 10−7). To the note, the rs393152 variant is in LD with the rs62053943 PD risk variant (CEU population D' = 1 R^2=0.80[257]). Given the pleiotropy at this locus, and in light of the *CASP8* example, it is a feasible that the non-reference SVA we have detected could be involved in both diseases, which requires further follow-up study.

The genetic analysis presented in this thesis began with the overall aspiration of exploring TE as common genetic risk factors that could identify new PD hits or further explain existing risk loci. However, using this common genetic risk data, we were not able to address other key questions regarding how TE could be contributing to PD aetiology, such as whether somatic TE insertions or rare TE are associated with disease. The later could be addressed in future studies with burden analyses to identify if rare TE are collectively contributing to PD risk. However, addressing somatic TE insertion in PD is a much more difficult question.

Non-LTR TE can affect cellular function through insertions in the germline (as described throughout this thesis) but also via their transposition in adult tissues, or mosaic during development, which includes neuronal cells. The evidence for the later in neuronal cells has come from a combination of studies using cell lines, animal models and human tissue. By applying engineered LINE-1 transposition reporter elements in cell culture assays, it was identified that rat neuronal progenitor cells, human fetal brain neuronal progenitor cells, neuronal progenitor cells derived from human embryonic stem cells and mature non-dividing neurons can support human LINE-1 transposition *in vitro* [258–260]. In addition, it has also been shown that an enhanced green fluorescent protein marked human LINE-1 transposition reporter transgene in mice led to somatic mosaicism in the brain[260]. Another study by *Coufal et al* used a quantitative multiplexed PCR assay to determine the endogenous LINE-1 copy number in a given genome. From this they identified that there is an increase in LINE-1 copies in several brain regions compared with the heart and liver from the same human individual, with the highest number found in the hippocampus[259,260]. In support of this, using a technique termed retrotransposon capture sequencing (RC-Seq), endogenous somatic transposition events have been identified in the human brain[166,261]. RC-Seq was used to generate libraries of TE insertions in genomic DNA from the hippocampus and caudate nucleus of three donors of advanced age (average 92 years old). Subsequent next-generation sequencing identified 7743 LINE-1, 13 692 *Alu* and 1350 SVA putative somatic de novo insertions in total in the three individuals which were present in one brain region but absent in the other and not previously identified as a germline variant[166]. Thirty-three of these potential de novo insertions were chosen for validation by genotyping PCR and capillary sequencing of the resulting PCR products, successfully validating twenty eight of them as somatic de novo insertions that were absent from the second brain region[166]. The number of neurons affected by an individual somatic de novo TE insertion would be dependent on the point in time of the transposition event, i.e. whether it occurred in a single mature post mitotic neuron, during neurogenesis or early in embryonic development affecting neuronal lineages. Therefore, characterizing the extent of neuronal mosaicism is challenging[262]. Further to identify how TE contribute to neuronal mosaicism in PD specifically, would be even more problematic. This is because it is not currently informative, from a disease perspective, to work with brain tissues of patients who died from neurodegenerative diseases such as PD. For neurodegenerative disease there is considerable cell loss in the brain, so assaying in these disease tissues is not informative for understanding the disease process.

Despite the challenges, techniques have been developed to more easily detect somatic mosaicism and CNV. The evidence for mosaicism in healthy and diseased brain is increasing rapidly, with somatic copy number gains of APP reported in the brain of individuals with AD. For PD specifically, *Proukakis et al* originally hypothesized that somatic *SNCA* CNV could lead to mosaicism in the brain and this could have a role in synucleinopathies[263]. Further the group recently reported evidence of somatic SNCA gains in brain, which was more commonly observed in nigral dopaminergic neurons of sporadic PD than controls and negatively correlated with AAO. They suggest from this data that somatic SNCA gains may be a risk factor for sporadic synucleinopathies, such as PD[263,264]. To note, when the SNCA CNV is inherited it causes autosomal dominant forms of PD and in these cases an enrichment of TE has been reported at the CNV breakpoints[264,265]. It has been demonstrated that TE enrichment encourages DNA damage which can cause CNV events. Specifically, Alu/Alu-mediated rearrangement

(AAMRs) is a mechanism that has been shown to be causative of CNV in the genome. A recent study by *Song et al* developed a tool (AluAluCNVpredictor)for predicting hot spots for CNV events based on *Alu* positions in the genome. From this analysis Song et al noted that the younger *Alu* elements were more likely to cause CNV. In light of this it is feasible to imagine a scenario whereby an individual could inherit an enrichment of non-reference “new” *Alu*, which in response to a stressor could give rise to the observed somatic SNV, which could contribute to PD risk. Unfortunately, WGS is not available for the individuals included in the PD somatic CNV study described above. But further WGS and MELT locus-specific analysis would be an exciting future study to further explore the possibility somatic CNV in the brain can be caused by TE enrichment at CNV breakpoints.

7.1.3 Future work

TE detection is now possible in a scalable and affordable manner, yet there are still many hurdles to overcome to be able to comprehensively assess the role of TE, not only in PD but all complex genetic diseases. Many of the limitations of the work described throughout this thesis exist because current genomic tools are still not adapted for the analysis of repetitive sequences such as TEs. One example is that annotation packages such as ANNOVA run annotation based on a single bp rather than the region spanning the TE. In addition, TE detection tools are very computer intensive to run, especially when genome-wide discovery is run with no pre-defined regions, which makes running on cloud-based system at scale problematic. Consequently, there is still room for improvement in many areas of TE bioinformatic analysis.

Another point to note that could halt the progression of TE detection in the human genome is lack of a unified TE online repository which contains detailed variant information, such as primary sequence and allele frequencies. For example, SNP variation can easily be compared between populations and disease states due to resources such as dbgap. At present dbRip is an online repository which includes instances of any active TE organized by genomic loci, is not maintained and contains detailed information for around 4.5k non-reference TE variants from 28 different populations[266]. It is important that TE discovery and characterization is compiled in dedicated online public repositories. Despite the fact that dbRip and others are available, there is a lack of unified TE repositories that have a long-term sustainability plan. The need for this is especially important for TE-associated human variants and mutations, particularly in the context of the millions of genomes currently being sequenced. Further, as TE detection tools are more accurate when regions of known TE insertion are predefined this also highlights the need for a unified resource so that more detailed "priors" files are available for subsequent studies.

Moving forward, I am incredibly lucky that the research described in this thesis will be expanded during my postdoctoral position at NIH. We have now shown that TE collectively are transcriptionally active, over-represented at PD risk loci and in moderate LD with known PD risk hits. But this is potentially a huge underestimation of the contribution of TE to PD. My future work will focus on two main lines of analysis 1) Understanding the role of reference TE variation in the genome, (such as the contribution of reference SVA variants) and 2) Understanding the role of non-reference

TE variation (such as the contribution from the variants outlined in our pilot MELT analysis).

7.1.3.1. Future studies including reference SVA variation:

MELT which characterises presence/absence of non-reference TE variation, is possible to run because there is WGS and TE detection tools available. However as shown in Chapter 3 reference SVA at these loci could be contributing to disease risk at these regions but are completely uncatalogued in the genome and at present there is no technology available to detect this variation genome-wide. Therefore, to begin to address reference SVA variation (so that this information can be integrated into reference panels) target-sequencing for the ~2700 reference SVA in the genome would be necessary. We have identified that common variation within these elements is imputable (*see Chapter 3*) and in addition the NABEC resource has expression and WGS data available. Therefore, utilizing the NABEC cohort to deeply characterize reference SVAs and integrate this data into existing WGS would be incredibly informative on the contribution of SVA variation to disease and gene expression. Although it should also be noted that gaining accurate information on the variation within reference SVA will be difficult, mainly due to the fact that SVA's contain multiple repeat domains that are difficult to uniquely map with short reads. This problem could in part be elevated with the use of long-read sequencing. Software tools have been developed to characterize SV in long-read sequencing, such as SMRT-SV and SNIFFLES. Compared with short-read tools, long-read methods are better adapted to return the full sequence of the TE insertions, which enables better functional downstream analysis and bioinformatic validation of the insertions. Although long-read technologies are improving rapidly, the broader application of long-read technologies is currently limited by a lower throughput, higher error rate and higher cost per base relative to short read sequencing.

7.1.3.2. Future studies including non-reference TE variation

Following our non-reference TE MELT pilot study, it is evident that non-reference TE variation is a source of genetic variation that could be involved in PD aetiology. Not only do many non-reference TE's lie within PD risk regions but even from our initial analysis of ~2.6k variants, two variants are in moderate LD with known PD risk variants and therefore are good causal variant candidates for future follow-up study. We did not have the power to detect genome-wide significant hits in the pilot study, nor did we have the ability to assess non-reference TE that weren't discovered in the 1000 genome project. In light of this, a very exciting extension of the current work will be expanding the MELT analysis into a large-scale analysis including more genomes and covering all detectable non-reference common TE variants.

In addition, the MELT pilot study was specifically designed to include the NABEC cohort as these individuals already have existing expression data available. The NABEC datasets include methylation, RNA-seq and alternative splicing data, which can now be correlated with the MELT TE variants, to identify TE QTLs in the brain. As TE insertions are known to cause alternative splicing and effect gene expression this will be a very informative analysis. To be able to utilize future large-scale transcriptomic and epigenetic datasets, we also included individuals from the PPMI cohort that are part of the FounDIn initiative in our MELT analysis:

FOUNDIN-PD – Foundational Data Initiative for Parkinson Diseasehttps://www.foundinpd.org/wp/

FounDIn is a $6 million two year program funded by the Michael J.Fox Foundation that is focused on further understanding how known risk loci and causal factors are contributing to PD onset and progression. Around one hundred iPS cell lines from the PPMI will be cultured and differentiated into dopaminergic neurons. Further, advanced "omics" techniques will be used (e.g., genomics, proteomics, metabolomics) to map how various genetic changes lead to cellular and molecular changes associated with PD. Through this ATAC-sequencing, Hi-C and RNA PacBio seq will be available for a subset of the PPMI individuals that have been included in our MELT analysis. Hence, we will be able to utilize this data to further correlate the non-reference TE variants, so that we can infer functional consequence of these insertions in control and PD. Although this will give incredible insight into what could be the possible consequence of non-reference TE insertions, they could be causal variants, which will need extensive functional follow-up with such as CRISPR studies in iPSs. A detailed workflow of the future MELT TE analyses is shown in (Figure 7.1)

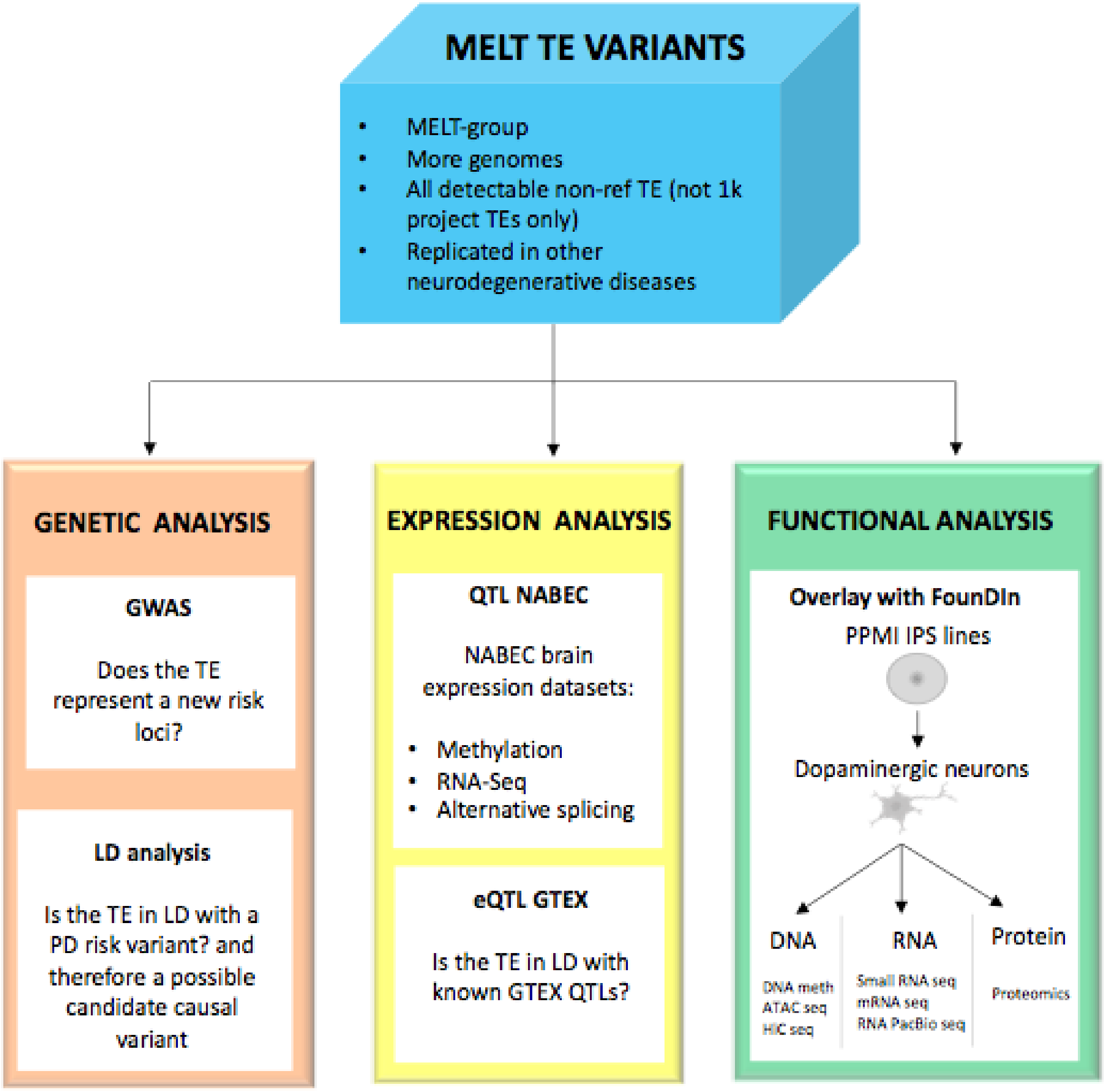


Figure 7.1. General workflow of the next phase of the TE in PD analyses

The main aim of our ongoing TE analysis is to aid in dissecting existing genetic risk and identifying new loci that could contribute to PD risk and progression. Therefore finally, it would be extremely informative to add datasets from non-European populations in future studies. Our phasing of the TE insertions with surrounding SNPs to dissect haplotypes, currently focuses on common TE variants in the European population. Inclusion of more diverse human populations for variant discovery and targeted discoveries in patient populations will likely increase the number of candidate functional variants.

Abbreviations:

AAO	Age at onset
AD	Alzheimer Disease
ALS	Amyotrophic Lateral Sclerosis
CAGE	Cap analyses gene expression
CDCV	Common disease common variant
CDRV	Common disease rare variant
CJD	Creutzfeldt-Jakob disease
CNV	Copy number variants
CPM	Counts per million mapped
DD	Developmental disorders
eQTL	Expression quantitative trait loci
FE	Functional equivalence
GCTA	Genome-wide complex trait analysis
GTEX	Genotype-Tissue Expression
GWAS	Genome wide association studies
IPDGC	International Parkinson’s Disease; Genomic Consortium
IPSC	Induced pluripotent stem cell
LD	Linkage Disequilibrium
LINE	Long interspersed nuclear elements
LOOMA	Leave one out meta-analyse
LTD	Long terminal repeats
LTR	Long terminal repeats
MAF	Minor allele frequency
MELT	Mobile elements location tool
MR	Mendelian Randomisation
MS	Muscular Sclerosis
NABEC	North American Brain Expression Consortium

NGS	Next generation sequencing
Non-LTR	Non-Long terminal repeats
ORF	Open Read Frame
PC	Principal components
PCA	Principle component analysis
PD	Parkinson’s Disease
PPG	Pseudogene-gene
PPMI	Parkinson’s Progression markers initiative
PRS	Polygenic risk scores
QTL	Quantitative trait loci
RE	Repetitive element
RTE	Retrotransposable elements
SINE	Short interspersed nuclear elements
SNP	Single nucleotide polymorphism
SV	Structural variants
TE	Transposable Elements
TPM	Transcript per million
TPRT	Target primed reverse transcription
TSD	Target site duplications
UTR	Untranslated region
VCF	Variant call format
VNTR	Variable number tandem repeat
VQSR	Variant Quality Score Recalibration
WES	Whole exome sequencing
WGS	Whole genome sequencing data
XDP	X-linked dystonia Parkinsonism

References

1. Nalls, M. A. *et al.* Expanding Parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk. doi:10.1101/388165
2. Book, A. *et al.* A Meta-Analysis of α-Synuclein Multiplication in Familial Parkinsonism. *Front. Neurol.* **9**, (2018).
3. Aneichyk T, E. al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/29474918. (Accessed: 30th April 2019)
4. Savage, A. L., Bubb, V. J., Breen, G. & Quinn, J. P. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol. Biol.* **13**, 101 (2013).
5. 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* **9**, e90833 (2014).
6. 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.* **13**, 828–838 (2017).
7. de Lau, L. M. L. & Breteler, M. M. B. Epidemiology of Parkinson's disease. *Lancet Neurol.* **5**, 525–535 (2006).
8. Gasser, T. Genetics of parkinson's disease. *Annals of Neurology* **44**, S53–S57 (1998).
9. Calabrese, V. P. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* **69**, 223–4; author reply 224 (2007).
10. Dorsey, E. R., Ray Dorsey, E. & Bloem, B. R. The Parkinson Pandemic—A Call to Action. *JAMA Neurology* **75**, 9 (2018).
11. Ward, C. D. *et al.* Parkinson's disease in 65 pairs of twins and in a set of quadruplets. *Neurology* **33**, 815–815 (1983).
12. Duvoisin, R. C., Eldridge, R., Williams, A., Nutt, J. & Calne, D. Twin study of Parkinson disease. *Neurology* **31**, 77–77 (1981).
13. Parkinson's Disease in a Chemist Working with 1-Methyl-4-Phenyl-L,2,5,6-Tetrahydropyridine. *N. Engl. J. Med.* **309**, 310–310 (1983).
14. Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).
15. Singleton, A. & Hardy, J. A generalizable hypothesis for the genetic architecture of disease:

pleomorphic risk loci. *Hum. Mol. Genet.* **20**, R158–62 (2011).

16. Lander, E. S. The New Genomics: Global Views of Biology. *Science* **274**, 536–539 (1996).
17. Spillantini, M. G. *et al.* Alpha-synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
18. Chartier-Harlin, M.-C. *et al.* Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364**, 1167–1169 (2004).
19. Singleton, A. B. -Synuclein Locus Triplication Causes Parkinson's Disease. *Science* **302**, 841–841 (2003).
20. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).
21. Maraganore, D. M. *et al.* Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease. *JAMA* **296**, 661–670 (2006).
22. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).
23. UK Parkinson's Disease Consortium *et al.* Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.* **20**, 345–353 (2011).
24. Lill, C. M. *et al.* Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. *PLoS Genet.* **8**, e1002548 (2012).
25. 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.* **46**, 989–993 (2014).
26. Spatola, M. & Wider, C. Genetics of Parkinson's disease: the yield. *Parkinsonism & Related Disorders* **20**, S35–S38 (2014).
27. Healy, D. G., Wood, N. W. & Schapira, A. H. V. Test for LRRK2 mutations in patients with Parkinson's disease. *Pract. Neurol.* **8**, 381–385 (2008).
28. Lesage, S. *et al.* G2019S LRRK2 mutation in French and North African families with Parkinson's disease. *Annals of Neurology* **58**, 784–787 (2005).
29. Ross, O. A. *et al.* Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: a case-control study. *Lancet Neurol.* **10**, 898–908 (2011).
30. Lesage, S. *et al.* LRRK2 haplotype analyses in European and North African families with Parkinson disease: a common founder for the G2019S mutation dating from the 13th century. *Am. J. Hum. Genet.* **77**, 330–332 (2005).
31. Di Fonzo, A. *et al.* A common missense variant in the LRRK2 gene, Gly2385Arg, associated

with Parkinson’s disease risk in Taiwan. *Neurogenetics* **7**, 133–138 (2006).

32. Farrer, M. J. *et al.* Lrrk2 G2385R is an ancestral risk factor for Parkinson’s disease in Asia. *Parkinsonism Relat. Disord.* **13**, 89–92 (2007).
33. Tan, E. K. & Schapira, A. H. Uniting Chinese across Asia: the LRRK2 Gly2385Arg risk variant. *European journal of neurology: the official journal of the European Federation of Neurological Societies* **15**, 203–204 (2008).
34. Neudorfer, O. *et al.* Occurrence of Parkinson’s syndrome in type 1 Gaucher disease. *QJM* **89**, 691–694 (1996).
35. Halperin, A., Elstein, D. & Zimran, A. Increased incidence of Parkinson disease among relatives of patients with Gaucher disease. *Blood Cells, Molecules, and Diseases* **36**, 426–428 (2006).
36. Sidransky, E. *et al.* Multicenter analysis of glucocerebrosidase mutations in Parkinson’s disease. *N. Engl. J. Med.* **361**, 1651–1661 (2009).
37. Beutler, E. Gaucher disease: new molecular approaches to diagnosis and treatment. *Science* **256**, 794–799 (1992).
38. Gegg, M. E. *et al.* Glucocerebrosidase deficiency in substantia nigra of parkinson disease brains. *Ann. Neurol.* **72**, 455–463 (2012).
39. Anheim, M. *et al.* Penetrance of Parkinson disease in glucocerebrosidase gene mutation carriers. *Neurology* **78**, 417–420 (2012).
40. Mitsui, J. *et al.* Mutations for Gaucher disease confer high susceptibility to Parkinson disease. *Arch. Neurol.* **66**, 571–576 (2009).
41. Hutton, M. *et al.* Association of missense and 5’-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393**, 702–705 (1998).
42. Baker, M. *et al.* Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum. Mol. Genet.* **8**, 711–715 (1999).
43. Evans, W. *et al.* The tau H2 haplotype is almost exclusively Caucasian in origin. *Neurosci. Lett.* **369**, 183–185 (2004).
44. Pittman, A. M. *et al.* Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. *J. Med. Genet.* **42**, 837–846 (2005).
45. Skipper, L. *et al.* Linkage Disequilibrium and Association of MAPT H1 in Parkinson Disease. *The American Journal of Human Genetics* **75**, 669–677 (2004).

46. Edwards, T. L. *et al.* Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann. Hum. Genet.* **74**, 97–109 (2010).

47. Hernandez, D. G. *et al.* Genome Wide Assessment of Young Onset Parkinson's Disease from Finland. *PLoS ONE* **7**, e41859 (2012).

48. Pihlstrøm, L. *et al.* Supportive evidence for 11 loci from genome-wide association studies in Parkinson's disease. *Neurobiol. Aging* **34**, 1708.e7–1708.e13 (2013).

49. International Parkinson Disease Genomics Consortium *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377**, 641–649 (2011).

50. Satake, W. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41**, 1303–1307 (2009).

51. Vinkhuyzen, A. A. E., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods. *Annual Review of Genetics* **47**, 75–95 (2013).

52. Rijsdijk, F. V. & Sham, P. C. Analytic approaches to twin data using structural equation models. *Brief. Bioinform.* **3**, 119–133 (2002).

53. Wirdefeldt, K., Gatz, M., Schalling, M. & Pedersen, N. L. No evidence for heritability of Parkinson disease in Swedish twins. *Neurology* **63**, 305–311 (2004).

54. Thacker, E. L. & Ascherio, A. Familial aggregation of Parkinson's disease: a meta-analysis. *Mov. Disord.* **23**, 1174–1183 (2008).

55. Do, C. B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet.* **7**, e1002141 (2011).

56. Tanner, C. M. *et al.* Parkinson disease in twins: an etiologic study. *JAMA* **281**, 341–346 (1999).

57. Keller, M. F. *et al.* Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum. Mol. Genet.* **21**, 4996–5009 (2012).

58. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).

59. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).

60. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).

61. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).

62. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

63. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).

64. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).

65. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).

66. Vries, B. B. A. de *et al.* Diagnostic Genome Profiling in Mental Retardation. *The American Journal of Human Genetics* **77**, 606–616 (2005).

67. Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).

68. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* **18**, 74–82 (2002).

69. Fellermann, K. *et al.* A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon. *The American Journal of Human Genetics* **79**, 439–448 (2006).

70. Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).

71. Gamazon, E. R. & Stranger, B. E. The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* **14**, 352–357 (2015).

72. Nishioka, K. *et al.* Expanding the clinical phenotype of SNCA duplication carriers. *Mov. Disord.* **24**, 1811–1819 (2009).

73. Bradley, W. E. C. *et al.* Hotspots of large rare deletions in the human genome. *PLoS One* **5**, e9401 (2010).

74. Hedrich, K. *et al.* Distribution, type, and origin ofParkin mutations: Review and case studies. *Movement Disorders* **19**, 1146–1157 (2004).

75. Marongiu, R. *et al.* Whole gene deletion and splicing mutations expand the PINK1 genotypic spectrum. *Hum. Mutat.* **28**, 98 (2007).

76. Macedo, M. G. *et al.* Genotypic and phenotypic characteristics of Dutch patients with early

onset Parkinson’s disease. *Movement Disorders* **24**, 196–203 (2009).

77. Bose, P., Hermetz, K. E., Conneely, K. N. & Katharine Rudd, M. Tandem Repeats and G-Rich Sequences Are Enriched at Human CNV Breakpoints. *PLoS One* **9**, e101607 (2014).

78. Ross, O. A. *et al.* Genomic investigation of α-synuclein multiplication and parkinsonism. *Annals of Neurology* **63**, 743–750 (2008).

79. Bonifati, V. *et al.* DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurol. Sci.* **24**, 159–160 (2003).

80. Morais, S., Bastos-Ferreira, R., Sequeiros, J. & Alonso, I. Genomic mechanisms underlyingPARK2large deletions identified in a cohort of patients with PD. *Neurology Genetics* **2**, e73 (2016).

81. McCLINTOCK, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 344–355 (1950).

82. Consortium, I. H. G. S. & International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

83. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5280–5285 (2003).

84. 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* **332**, 164–166 (1988).

85. Batzer, M. A. & Deininger, P. L. A human-specific subfamily of Alu sequences. *Genomics* **9**, 481–487 (1991).

86. Batzer, M. A. *et al.* Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res.* **19**, 3619–3623 (1991).

87. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H., Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* **73**, 1444–1451 (2003).

88. Wang, H. *et al.* SVA Elements: A Hominid-specific Retroposon Family. *J. Mol. Biol.* **354**, 994–1007 (2005).

89. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).

90. Salem, A.-H. & -H. Salem, A. Recently Integrated Alu Elements and Human Genomic Diversity. *Molecular Biology and Evolution* **20**, 1349–1361 (2003).

91. Zhang, Z. Millions of Years of Evolution Preserved: A Comprehensive Catalog of the

Processed Pseudogenes in the Human Genome. *Genome Research* **13**, 2541–2558 (2003).

92. Hancks, D. C. & Kazazian, H. H., Jr. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).

93. Scott, A. F. *et al.* Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**, 113–125 (1987).

94. Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).

95. Babushok, D. V. & Kazazian, H. H. Progress in understanding the biology of the human mutagen LINE-1. *Human Mutation* **28**, 527–539 (2007).

96. Minakami, R. *et al.* Identification of an internal cis-element essential for the human Li transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Research* **20**, 3139–3145 (1992).

97. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).

98. Hedges, D. J. *et al.* Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14**, 1068–1075 (2004).

99. Mighell, A. J., Markham, A. F. & Robinson, P. A. Alu sequences. *FEBS Lett.* **417**, 1–5 (1997).

100. Payer, L. M. *et al.* Alu insertion variants alter mRNA splicing. *Nucleic Acids Res.* **47**, 421–431 (2019).

101. Hancks, D. C. & Kazazian, H. H., Jr. Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* **22**, 191–203 (2012).

102. Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. *Nucleic Acids Research* **40**, 1666–1683 (2012).

103. Levy, O., Knisbacher, B. A., Levanon, E. Y. & Havlin, S. Integrating networks and comparative genomics reveals retroelement proliferation dynamics in hominid genomes. *Science Advances* **3**, e1701256 (2017).

104. Bantysh, O. B. & Buzdin, A. A. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochemistry (Moscow)* **74**, 1393–1399 (2009).

105. Muotri, A. R., Marchetto, M. C. N., Coufal, N. G. & Gage, F. H. The necessary junk: new functions for transposable elements. *Hum. Mol. Genet.* **16 Spec No. 2**, R159–67 (2007).

106. Goodier, J. L. & Kazazian, H. H., Jr. Retrotransposons revisited: the restraint and

rehabilitation of parasites. *Cell* **135**, 23–35 (2008).

107. Gianfrancesco, O., Bubb, V. J. & Quinn, J. P. SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides* **64**, 3–7 (2017).
108. Hancks, D. C., Mandal, P. K., Cheung, L. E. & Kazazian, H. H. The Minimal Active Human SVA Retrotransposon Requires Only the 5'-Hexamer and Alu-Like Domains. *Molecular and Cellular Biology* **32**, 4718–4726 (2012).
109. 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.* **114**, E11020–E11028 (2017).
110. Makino, S. *et al.* Reduced Neuron-Specific Expression of the TAF1 Gene Is Associated with X-Linked Dystonia-Parkinsonism. *The American Journal of Human Genetics* **80**, 393–406 (2007).
111. Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *bioRxiv* 149872 (2017). doi:10.1101/149872
112. Chiò, A. *et al.* A de novo missense mutation of the FUS gene in a 'true' sporadic ALS case. *Neurobiology of Aging* **32**, 553.e23–553.e26 (2011).
113. Lai, S.-L. *et al.* FUS mutations in sporadic amyotrophic lateral sclerosis. *Neurobiology of Aging* **32**, 550.e1–550.e4 (2011).
114. Rademakers, R. *et al.* Fus gene mutations in familial and sporadic amyotrophic lateral sclerosis. *Muscle Nerve* **42**, 170–176 (2010).
115. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**, 397–405 (2008).
116. Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* **46**, 21–42 (2012).
117. Conley, A. B., Piriyapongsa, J. & Jordan, I. K. Retroviral promoters in the human genome. *Bioinformatics* **24**, 1563–1567 (2008).
118. Jordan, I. K., King Jordan, I., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* **19**, 68–72 (2003).
119. Mariño-Ramírez, L., Lewis, K. C., Landsman, D. & Jordan, I. K. Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* **110**, 333–

341 (2005).

120. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).

121. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).

122. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.* **45**, 325–329 (2013).

123. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).

124. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).

125. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).

126. Piriyapongsa, J., Mariño-Ramírez, L. & Jordan, I. K. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**, 1323–1337 (2007).

127. Weber, M. J. Mammalian Small Nucleolar RNAs Are Mobile Genetic Elements. *PLoS Genetics* **2**, e205 (2006).

128. Conley, A. B. & Jordan, I. Cell type-specific termination of transcription by transposable element sequences. *Mobile DNA* **3**, 15 (2012).

129. Raviram, R. *et al.* Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol.* **19**, 216 (2018).

130. Glinsky, G. V. Contribution of transposable elements and distal enhancers to evolution of human-specific features of interphase chromatin architecture in embryonic stem cells. *Chromosome Res.* **26**, 61–84 (2018).

131. Cournac, A., Koszul, R. & Mozziconacci, J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.* **44**, 245–255 (2016).

132. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

133. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).

134. Darrow, E. M. *et al.* Deletion of DXZ4 on the human inactive X chromosome alters higher-

order genome architecture. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4504–12 (2016).

135. Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575–579 (2016).

136. Rennie, S., Dalby, M., van Duin, L. & Andersson, R. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat. Commun.* **9**, 487 (2018).

137. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

138. Dixon, J. R., Jung, I., Selvaraj, S. & Shen, Y. Antosiewicz--Bour get JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B. *Chromatin architecture reorganization during stem cell differentiation. Nature* **518**, 331–336 (2015).

139. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* **62**, 668–680 (2016).

140. Bennett, E. A. Natural Genetic Variation Caused by Transposable Elements in Humans. *Genetics* **168**, 933–951 (2004).

141. Savage, A. L., Bubb, V. J., Breen, G. & Quinn, J. P. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol. Biol.* **13**, 101 (2013).

142. Quinn, J. & Bubb, V. Hominid retrotransposons as a modulator of genomic function. 264 (2000).

143. Alasdair MacKenzie, J. Q. A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer-like properties in the mouse embryo. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 15251 (1999).

144. Website. Available at: https://www.genome.gov/27563570/. (Accessed: 28th January 2019)

145. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).

146. Website. Available at: https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md. (Accessed: 15th April 2019)

147. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).

148. Website. Available at: https://github.com/gatk-workflows/broad-prod-wgs-germline-snps-

indels. (Accessed: 15th April 2019)

149. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

150. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

151. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).

152. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).

153. van der Brug, M. P. *et al.* RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10244–10249 (2008).

154. Workman, C. *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* **3**, research0048 (2002).

155. Schadt, E. E., Li, C., Ellis, B. & Wong, W. H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem. Suppl.* **Suppl 37**, 120–125 (2001).

156. Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557 (2001).

157. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

158. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).

159. Parts, L., Stegle, O., Winn, J. & Durbin, R. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet.* **7**, e1001276 (2011).

160. Blauwendraat, C. *et al.* Comprehensive promoter level expression quantitative trait loci analysis of the human frontal lobe. *Genome Med.* **8**, 65 (2016).

161. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).

162. Pardo, L. M. *et al.* Regional differences in gene expression and promoter usage in aged human brains. *Neurobiol. Aging* **34**, 1825–1836 (2013).

163. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

164. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* **13**, 307–308 (2015).

165. Bonifati, V. Mutations in the DJ-1 Gene Associated with Autosomal Recessive Early-Onset Parkinsonism. *Science* **299**, 256–259 (2003).

166. Nalls, M. A. *et al.* Parkinson's disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk. (2018). doi:10.1101/388165

167. Ariga, H. *et al.* Neuroprotective Function of DJ-1 in Parkinson's Disease. *Oxid. Med. Cell. Longev.* **2013**, (2013).

168. Hancks, D. C. & Kazazian, H. H. SVA retrotransposons: Evolution and genetic instability. *Seminars in Cancer Biology* **20**, 234–245 (2010).

169. Wang H, E. al. SVA elements: a hominid-specific retroposon family. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16288912. (Accessed: 17th April 2019)

170. Glinsky, G. V. Mechanistically Distinct Pathways of Divergent Regulatory DNA Creation Contribute to Evolution of Human-Specific Genomic Regulatory Networks Driving Phenotypic Divergence ofHomo sapiens. *Genome Biology and Evolution* **8**, 2774–2788 (2016).

171. Robbez-Masson, L. & Rowe, H. M. Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology* **12**, 45 (2015).

172. Ewing, A. D. Transposable element detection from whole genome sequence data. *Mob. DNA* **6**, 24 (2015).

173. Rishishwar, L. *et al.* Population and clinical genetics of human transposable elements in the (post) genomic era. *Mob. Genet. Elements* **7**, 1 (2017).

174. Wang L, E. al. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27998931. (Accessed: 17th April 2019)

175. Payer LM, E. al. Structural variants caused by Alu insertions are associated with risks for many human diseases. - PubMed - NCBI. Available at:

https://www.ncbi.nlm.nih.gov/pubmed/28465436. (Accessed: 17th April 2019)

176. Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. doi:10.1101/149872

177. Abigail L.Savage, V. J. B. A. J. P. Q. What role do human specific retrotransposons play in mental health and behaviour. *Curr. Trends Neurol.* **7**, 57–68 (2013).

178. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

179. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

180. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

181. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).

182. Larsen, P. A., Hunnicutt, K. E., Larsen, R. J., Yoder, A. D. & Saunders, A. M. Warning SINEs: Alu elements, evolution of the human brain, and the spectrum of neurological disease. *Chromosome Res.* **26**, 93–111 (2018).

183. Liu, X. *et al.* The association between TOMM40 gene polymorphism and spontaneous brain activity in amnestic mild cognitive impairment. *J. Neurol.* **261**, 1499–1507 (2014).

184. Yu, L. *et al.* TOMM40 ‘523 VARIANT AND COGNITIVE DECLINE IN COMMUNITY BASED OLDER PERSONS WITH APOE E3/3 GENOTYPE. *Alzheimer’s & Dementia* **12**, P1146–P1147 (2016).

185. Payton, A. *et al.* A TOMM40 poly-T variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging. *Neurobiol. Aging* **39**, 217.e1–7 (2016).

186. Singleton, A. & Hardy, J. The Evolution of Genetics: Alzheimer’s and Parkinson's Diseases. *Neuron* **90**, 1154–1163 (2016).

187. Nalls, M. A. *et al.* Parkinson’s disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk. (2018). doi:10.1101/388165

188. Robak, L. A. *et al.* Excessive burden of lysosomal storage disorder gene variants in Parkinson’s disease. *Brain* **140**, 3191–3203 (2017).

189. Hardy, J. Genetic Analysis of Pathways to Parkinson Disease. *Neuron* **68**, 201–206 (2010).

190. Billingsley, K. J., Bandres-Ciga, S., Saez-Atienzar, S. & Singleton, A. B. Genetic risk factors in Parkinson’s disease. *Cell Tissue Res.* **373**, 9–20 (2018).

191. Langston, J. W., Ballard, P., Tetrud, J. W. & Irwin, I. Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis. *Science* **219**, 979–980 (1983).

192. Schapira, A. *Mitochondrial Function and Dysfunction*. (Elsevier, 2003).

193. Canet-Avilés, R. M. *et al.* The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9103–9108 (2004).

194. Funayama, M. *et al.* CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease: a genome-wide linkage and sequencing study. *Lancet Neurol.* **14**, 274–282 (2015).

195. Burchell, V. S. *et al.* The Parkinson's disease-linked proteins Fbxo7 and Parkin interact to mediate mitophagy. *Nat. Neurosci.* **16**, 1257–1265 (2013).

196. Lesage, S. *et al.* Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. *Am. J. Hum. Genet.* **98**, 500–513 (2016).

197. Gorman, G. S. *et al.* Mitochondrial diseases. *Nature Reviews Disease Primers* **2**, 16080 (2016).

198. Pedro J. Garcia-Ruiz, A. J. E. Parkinson Disease: An Evolutionary Perspective. *Front. Neurol.* **8**, (2017).

199. Blauwendraat, C. *et al.* Parkinson disease age of onset GWAS: defining heritability, genetic loci and a-synuclein mechanisms. (2018). doi:10.1101/424010

200. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

201. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

202. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

203. Wirdefeldt, K., Gatz, M., Reynolds, C. A., Prescott, C. A. & Pedersen, N. L. Heritability of Parkinson disease in Swedish twins: a longitudinal study. *Neurobiol. Aging* **32**, 1923.e1–8 (2011).

204. Gasser, T. Genetics of Parkinson's disease. *Curr. Opin. Neurol.* **18**, 363–369 (2005).

205. Wickremaratchi, M. M. *et al.* Prevalence and age of onset of Parkinson's disease in Cardiff: a community based cross sectional study and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **80**, 805–807 (2009).

206. Porter, B., Macfarlane, R., Unwin, N. & Walker, R. The Prevalence of Parkinson's Disease in an Area of North Tyneside in the North-East of England. *Neuroepidemiology* **26**, 156–161 (2006).

207. Viechtbauer, W. Conducting Meta-Analyses inRwith themetaforPackage. *J. Stat. Softw.* **36**, (2010).

208. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2014).

209. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).

210. International Parkinson Disease Genomics Consortium *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377**, 641–649 (2011).

211. Satake, W. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41**, 1303–1307 (2009).

212. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

213. 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.* **46**, 989–993 (2014).

214. Guffanti, G. *et al.* LINE1 insertions as a genomic risk factor for schizophrenia: Preliminary evidence from an affected family. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **171**, 534–545 (2016).

215. Goebel, H. H., Heipertz, R., Scholz, W., Iqbal, K. & Tellez-Nagel, I. Juvenile Huntington chorea: Clinical, ultrastructural, and biochemical studies. *Neurology* **28**, 23–23 (1978).

216. Carmo, C., Naia, L., Lopes, C. & Cristina Rego, A. Mitochondrial Dysfunction in Huntington's Disease. in *Advances in Experimental Medicine and Biology* 59–83 (2018).

217. Atsumi, T. The ultrastructure of intramuscular nerves in amyotrophic lateral sclerosis. *Acta Neuropathol.* **55**, 193–198 (1981).

218. Sasaki, S. & Iwata, M. Mitochondrial alterations in the spinal cord of patients with sporadic amyotrophic lateral sclerosis. *J. Neuropathol. Exp. Neurol.* **66**, 10–16 (2007).

219. Moreira, P. I., Cardoso, S. M., Santos, M. S. & Oliveira, C. R. The key role of mitochondria in Alzheimer's disease. *J. Alzheimers. Dis.* **9**, 101–110 (2006).

220. Nunomura, A. *et al.* Oxidative Damage Is the Earliest Event in Alzheimer Disease. *J. Neuropathol. Exp. Neurol.* **60**, 759–767 (2001).

221. Moreira, P. I., Duarte, A. I., Santos, M. S., Cristina Rego, A. & Oliveira, C. R. An Integrative View of the Role of Oxidative Stress, Mitochondria and Insulin in Alzheimer's Disease. *J. Alzheimers. Dis.* **16**, 741–761 (2009).

222. Perry, G., Zhu, X. & Smith†, M. A. *Alzheimer's Disease: Advances for a New Century*. (IOS Press, 2013).

223. Nalls, M. A. *et al.* Genetic risk and age in Parkinson's disease: Continuum not stratum. *Mov. Disord.* **30**, 850–854 (2015).

224. Escott-Price, V. *et al.* Polygenic risk of Parkinson disease is correlated with disease age at onset. *Ann. Neurol.* **77**, 582–591 (2015).

225. Lill, C. M. *et al.* Impact of Parkinson's disease risk loci on age at onset. *Mov. Disord.* **30**, 847–850 (2015).

226. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).

227. Pihlstrøm, L. *et al.* Supportive evidence for 11 loci from genome-wide association studies in Parkinson's disease. *Neurobiol. Aging* **34**, 1708.e7–1708.e13 (2013).

228. Lill, C. M. *et al.* Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. *PLoS Genet.* **8**, e1002548 (2012).

229. Gaare, J. J. *et al.* Rare genetic variation in mitochondrial pathways influences the risk for Parkinson's disease. *Mov. Disord.* **33**, 1591–1600 (2018).

230. Lin, M. T. *et al.* Somatic mitochondrial DNA mutations in early parkinson and incidental lewy body disease. *Annals of Neurology* **71**, 850–854 (2012).

231. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).

232. Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable elements. *Nat. Rev. Genet.* **19**, 688–704 (2018).

233. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

234. Wang, L., Rishishwar, L., Mariño-Ramírez, L. & King Jordan, I. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res.* **45**, 2318 (2017).

235. Website. Available at: https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md. (Accessed: 28th January 2019)

236. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

237. Rishishwar, L., Tellez Villa, C. E. & Jordan, I. K. Transposable element polymorphisms recapitulate human evolution. *Mob. DNA* **6**, 21 (2015).

238. Krüger, J., Moilanen, V., Majamaa, K. & Remes, A. M. Molecular Genetic Analysis of the APP, PSEN1, and PSEN2 Genes in Finnish Patients With Early-onset Alzheimer Disease and Frontotemporal Lobar Degeneration. *Alzheimer Disease & Associated Disorders* **26**, 272–276 (2012).

239. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).

240. Jansen, I. E. *et al.* Discovery and functional prioritization of Parkinson's disease candidate genes from large-scale whole exome sequencing. *Genome Biology* **18**, (2017).

241. Gardner, E. J. *et al.* Contribution of Retrotransposition to Developmental Disorders. doi:10.1101/471375

242. Chen-Plotkin, A. S. Unbiased Approaches to Biomarker Discovery in Neurodegenerative Diseases. *Neuron* **84**, 594–607 (2014).

243. Santiago, J. A. & Potashkin, J. A. Blood Transcriptomic Meta-analysis Identifies Dysregulation of Hemoglobin and Iron Metabolism in Parkinson' Disease. *Front. Aging Neurosci.* **9**, 73 (2017).

244. Planken, A. *et al.* Looking beyond the brain to improve the pathogenic understanding of Parkinson's disease: implications of whole transcriptome profiling of Patients' skin. *BMC Neurol.* **17**, 6 (2017).

245. Shamir, R. *et al.* Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676–1683 (2017).

246. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, (2014).

247. Douville, R., Liu, J., Rothstein, J. & Nath, A. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. *Ann. Neurol.* **69**, 141–151 (2011).

248. Guo, C. *et al.* Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Rep.* **23**, 2874–2880 (2018).

249. Krug, L. *et al.* Retrotransposon Activation Contributes to Neurodegeneration in a Drosophila TDP-43 Model of ALS. (2016). doi:10.1101/090175

250. Maxwell, P. H., Burhans, W. C. & Curcio, M. J. Retrotransposition is associated with genome instability during chronological aging. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20376–20381 (2011).

251. Jeong, B.-H., Lee, Y.-J., Carp, R. I. & Kim, Y.-S. The prevalence of human endogenous retroviruses in cerebrospinal fluids from patients with sporadic Creutzfeldt–Jakob disease. *J. Clin. Virol.* **47**, 136–142 (2010).

252. Tan, H. *et al.* Retrotransposon activation contributes to fragile X premutation rCGG-mediated neurodegeneration. *Hum. Mol. Genet.* **21**, 57–65 (2012).

253. Bundo, M. *et al.* Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* **81**, 306–313 (2014).

254. Sun, W., Samimi, H., Gamez, M., Zare, H. & Frost, B. Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nat. Neurosci.* **21**, 1038–1048 (2018).

255. Blaudin de Thé, F. *et al.* Engrailed homeoprotein blocks degeneration in adult dopaminergic neurons through LINE-1 repression. *EMBO J.* e97374 (2018).

256. Gibb, W. R. & Lees, A. J. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **51**, 745–752 (1988).

257. Lees, A. J., Hardy, J. & Revesz, T. Parkinson's disease. *Lancet* **373**, 2055–2066 (2009).

258. Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord.* **23**, 2129–2170 (2008).

259. Hoehn, M. M. & Yahr, M. D. Parkinsonism: onset, progression, and mortality. *Neurology* **17**, 427–427 (1967).

260. Website. Available at: R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. (Accessed: 6th December 2018)

261. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).

262. Hall, L. L. *et al.* Demethylated HSATII DNA and HSATII RNA Foci Sequester PRC1 and MeCP2 into Cancer-Specific Nuclear Bodies. *Cell Rep.* **18**, 2943–2956 (2017).

263. Lättekivi, F. *et al.* Transcriptional landscape of human endogenous retroviruses (HERVs) and other repetitive elements in psoriatic skin. *Sci. Rep.* **8**, 4358 (2018).

264. Bersani, F. *et al.* Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15148–15153 (2015).

265. Bouzinba-Segard, H., Guais, A. & Francastel, C. Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8709–8714 (2006).

266. Valgardsdottir, R. & Chiodi - Nucleic Acids Res, I. Giordano Metal (2008) Transcription of satellite III non-coding RNAs is a general stress response in human cells.

267. Oberdoerffer, P. *et al.* SIRT1 redistribution on chromatin promotes genomic stability but alters gene expression during aging. *Cell* **135**, 907–918 (2008).

268. Kubben, N. & Misteli, T. Shared molecular and cellular mechanisms of premature ageing and ageing-associated diseases. *Nat. Rev. Mol. Cell Biol.* **18**, 595–609 (2017).

269. Sulzer, D. *et al.* T cells from patients with Parkinson's disease recognize α-synuclein peptides. *Nature* **546**, 656–661 (2017).

270. Nogalski, M. T. *et al.* A tumor-specific endogenous repetitive element is induced by herpesviruses. *Nat. Commun.* **10**, 90 (2019).

271. Le Naour, F. *et al.* Proteomics-based identification of RS/DJ-1 as a novel circulating tumor antigen in breast cancer. *Clin. Cancer Res.* **7**, 3328–3335 (2001).

272. MacKeigan, J. P. *et al.* Proteomic profiling drug-induced apoptosis in non-small cell lung carcinoma: identification of RS/DJ-1 and RhoGDIalpha. *Cancer Res.* **63**, 6928–6934 (2003).

273. Stacey, S. N. *et al.* Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. *Hum. Mol. Genet.* **25**, 1008 (2016).

274. Desikan, R. S. *et al.* Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. *Mol. Psychiatry* **20**, 1588 (2015).

275. Macia, A. *et al.* Engineered LINE-1 retrotransposition in nondividing human neurons. *Genome Res.* **27**, 335–348 (2017).

276. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).

277. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1

retrotransposition. *Nature* **435**, 903–910 (2005).

278. Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228–239 (2015).

279. Faulkner, G. J. & Garcia-Perez, J. L. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. *Trends Genet.* **33**, 802–816 (2017).

280. Proukakis, C., Houlden, H. & Schapira, A. H. Somatic alpha-synuclein mutations in Parkinson's disease: Hypothesis and preliminary data. *Movement Disorders* **28**, 705–712 (2013).

281. Mokretar K, E. al. Somatic copy number gains of α-synuclein (SNCA) in Parkinson's disease and multiple system atrophy brains. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/29917054. (Accessed: 1st May 2019)

282. Mathias Toft, O. A. R. Copy number variation in Parkinson's disease. *Genome Med.* **2**, 62 (2010).

283. Wang, J. *et al.* dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation* **27**, 323–329 (2006).