The role of non-coding tandem repeat DNA and non-LTR retrotransposons in Amyotrophic Lateral Sclerosis risk loci

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

Jack Marshall MBiolSci

March 2021

Acknowledgments

Firstly, I would like to express my utmost appreciation to my PhD supervisors, Prof. John Quinn and Dr. Jill Bubb. Thank you for all your support and guidance, both on a personal and professional level. Thank you for always pushing me and believing in me. I am so grateful for all the opportunities you have given me, allowing me to network and collaborate with so many amazing people.

I would also like to extend my deepest gratitude to my collaborators and everyone who has helped with my training. I am deeply indebted to Dr. Johnathan Cooper-Knock, who invited me to Project MinE working group 6. This opportunity has been a real turning point in my PhD journey and I am extremely grateful and honoured to have become part of this worldwide consortium. Thank you to all other members of working group 6, particularly for your advice and support in all meetings; I look forward to working with you all in the future. I would also like to thank Prof. Ammar Al-Chalabi and his team at King's College London for allowing me to visit and work alongside them, with a special thanks to Dr. Alfredo Iacoangeli and Dr. Ashley Jones for all the bioinformatics training.

Thank you to all members of Quinn lab, past and present, especially Ben Middlehurst, Kimberley Billingsley, Olympia Gianfrancesco, Emma Price, Maurizio Manca, Ashley Hall, Ana Illera, Veri Pessoa, John Mears, Josh Sullivan and Sarah Doran. Ben, it has been a pleasure to sit next to you in the office and work alongside you for the past 4 years; I will really miss it. When I started this PhD I expected to gain some great friends, but I did not expect to gain a brother; thank you for all the laughs and support, laddy. Kim and Oly, you introduced me to Quinn lab and made me so welcome and for that I am so grateful. You are both amazing scientists and I hope to work alongside you again in the future. Quinn lab, you are a wonderful group of people and have been like a family to me; I wish you all the best for the future.

I would also like to thank my GCSE and A-level Biology teacher, Dr Dando: you gave me the passion and drive to pursue a career in science. I would not be where I am now without your support, guidance, and encouragement.

I wish to show my utmost gratitude to my amazing Mum and Dad, who have always encouraged and supported me throughout my studies. You have always been there for me, listened and guided me and I am extremely grateful. To the rest of my wonderful family, thank you for all your love and support.

Finally, I would like to thank the MRC who funded this research and made this PhD possible.

Abstract

Genome-wide association studies and functional data have shown that there is a genetic basis contributing to sporadic and familial forms of amyotrophic lateral sclerosis (ALS), with a vast plethora of genes being associated with the disease. Since the discovery of the *C9orf72* intronic repeat expansion there has been a growing awareness of non-coding repetitive DNA being associated with ALS risk. Originally labelled "junk DNA", non-coding repetitive DNA is now known to be vital in regulating and shaping the human genome and has become of particular interest in the context of disease risk and pathogenesis. Repetitive DNA can be found in two forms: static and mobile, both of which were addressed in this PhD project, with an emphasis on variable number tandem repeats (VNTRs) and non-long terminal repeat retrotransposons.

Previous studies have demonstrated that polymorphic VNTRs can be associated with predisposition to disease and this often correlated with the differential transcriptional regulatory properties of the VNTR based upon the copy number of the repeat element. Analysis of the ALS risk gene *CFAP410* led to the identification of an intronic VNTR, with the discovery and characterisation of novel genetic variants present only in ALS patients. This VNTR was also shown to be functional in two reporter gene assays: both driving expression in the absence of a minimal promoter and also acting as a transcriptional regulatory domain, modulating expression in an allelic dependent manner on the basis of repeat copy number.

Loss of function variants of *NEK1* have been found to be associated with ALS risk. Current studies have focused on the coding regions of this locus, therefore our work aimed to address non-coding variation at this region, with the hypothesis that such genetic variation could impact *NEK1* gene regulation. This analysis led to the discovery of an intronic SINE-VNTR-*Alu* (SVA) retrotransposon within *NEK1*, which was found to be polymorphic within three domains: the 5' CT element, central VNTR and 3' poly A tail. Of note, two CT element variants were only found in ALS patients from a UK MNDA cohort. This region was then assessed using Isaac variant caller, facilitating high-throughput characterisation of SVA genetic variation within cohorts of Project MinE, which still agreed with the previous trend in genotype frequency but one particular cohort did not conform to that observed in the UK MNDA cohort. The *NEK1* SVA was tested functionally and demonstrated regulatory function *in vitro*, inducing significant repression within two reporter gene constructs, inducing both cell line and orientation specific expression profiles. By utilising the CRISPR Cas 9 system it was possible to excise the SVA element in HEK293 cells, but under basal conditions this deletion did not induce a significant change in *NEK1* gene expression. Ultimately this work aimed to raise the profile of VNTRs and SVA elements as potential sources of missing heritability in complex disease and to better understand their function in gene regulation.

Abbreviations

ABCA7 - ATP binding casette subfamily A member 7

ACTB - Actin Beta

AD - Alzheimer's disease

ALS - Amyotrophic lateral sclerosis

ALT - Alternate allele

ApE - A plasmid editor

APOBEC - Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like

ASD - Autism spectrum disorder

Axial SMD - Axial spondylometphyseal dyslasia

Aβ - Amyloid beta

BASH - Bourne again shell

BCC - Cutaneous basal-cell carcinoma

CAT - Chloramphenicol acetyltransferase

CC - Coiled-coil

cDNA - Complimentary DNA

CFAP410 - Cilia and flagella associated protein 410

ChIP - Chromatin immunoprecipitation

Chk - Checkpoint kinase

CHROM - Chromosome

CLIPseq - Crosslinking immunoprecipitation sequencing

CNS - Central nervous system

CRISPR - Clustered regularly interspered short palindromic repeat

crRNA - CRISPR RNA

Ct - Cycle threshold

CTCF - CCCTC binding-protein

DDR - DNA damage response

DLB - Dementia with Lewy bodies

DMEM - Dulbecco's modified eagle media

DSB - Double strand break

ECR - Evolutionary conserved region

EDTA - Ethylenediaminetetraacetic acid

EN - Endonuclease

ENCODE - Encyclopaedia of DNA Elements

Env - Envelope

ESC - Embryonic stem cell

FA - Friedrich ataxia

FALS - Familial amyotrophic lateral sclerosis

FBS - Fetal bovine serum

FEV1 - FEV transcription factor

FMRI - Fragile X mental retardation

FTD - Frontotemporal dementia

FUS - Fused in sarcoma

FXS - Fragile X syndrome

FXTAS - Fragile X-associated tremor/ataxia syndrome

Gag - group specific antigen

GAPDH - Glyceraldehyde-3-phosphate dehydrogenase

gRNA - Guide RNA

GSEA - Gene set enrichment analysis

GTEx - Genotype-tissues expression portal project

GWAS - Genome-wide Association Studies

HD - Huntington's disease

HDAC - Histone deacetylase

HDL2 - Huntington disease-like 2

HEK293 - Human embryonic kidney cell line

HeLa - Henrietta Lacks clonal cell line

HERV - Human endogenous retrovirus

HPC - High performing computer cluster

HR - Homologous recombination

IME - Intron-mediated enhancement

iN - Induced cortical neurons

Indel - Insertion-deletion

iPSC - Induced pluripotent stem cell

IR - Ionising radiation

IVC - Isaac variant caller

KCL - King's college London

KO - Knockout

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

LAR II - Luciferase assay reagent

LCL - Lymphoblastoid cell line

LINE - Long interspersed nuclear element

lncRNA - Long non-coding RNA

LRR - Leucine rich repeat

LTR - Long terminal repeat

MAF - Minor allele frequency

MAOA - Monoamine oxidase A

MC - Motor cortex

MD - Myotonic dystrophy

MeCP2 - Methyl-CpG binding protein 2

miR-137 - MicroRNA-137

miRNA - MicroRNA

MND - Motor neurone disease

MNDA - Motor neurone disease association

mRNA - Messenger RNA

MYA - Million years ago

NCBI - National center for biotechnology information

NEK1 - NIMA (Never in mitosis gene A)-related kinase 1

NGS - Next generation sequencing

NIID - Neuronal intranuclear inclusion disease

NIPA1 - NIPA magnesium transporter 1

NMDARS - N-methyl-D aspartate receptors

NRSF - Neuron restrictive silencing factor

NSC - Neural stem cell

NT - Non-target guides

NYGC - New York genome center

ORF - Open Reading Frame

OS - Oxidative stress

PAM - Protospacer adjacent motif

PBP - Progressive bulbar palsy

PBS - Phosphate-buffered saline

PCR - Polymerase chain reaction

PD - Parkinson's disease

PHAST - Phylogenic analysis with space/time models

PLB - Passive lysis buffer

PLS - Primary lateral sclerosis

PMA - Progressive musclar atrophy

Pol - Polymerase

POS - Position

Pro - Protease

QTL - Quantitative trait loci

QX - QIAxcel

RAN - Repeat-associated non-AUG

RB - Retinoblastoma

RC-Seq - Retrotransposon capture sequencing

RE - Restriction enzyme

REF - Reference allele

REST - RE1-silencing transcription factor

RIP - Retrotransposon insertion polymorphism

RNP - Ribonucleoprotein

rRNA - Ribosomal RNA

RT - Reverse transcriptase

RT-PCR - Reverse transcription polymerase chain reaction

SAG - Stop & Glo

SALS - Sporadic amyotrophic lateral sclerosis

SAM - Sequencing alignment map format

SBMA - Spinobulbar muscular atrophy

SCA - Spinocerebellar ataxia

SINE - Short interspersed nuclear element

siRNA - Short interfering RNA

SLC6A4 - Serotonin transporter/Solute Carrier Family 6 Member 4

SNP - Single nucleotide polymorphism

SOD1 - Superoxide dismutase 1

SPN - Spiny projection neurons

STR - Short tandem repeat

SV - Structural variant

SVA - SINE-VNTR-Alu

SZ - Schizophrenia

TAD - Topological associated domains

TAF1 - TATA-box binding protein associated factor 1

TARDBP - TAR DNA-binding protein

TDP-43 - TAR DNA-binding protein of 43 kDa

TE - Transposable element

TF - Transcription factor

TFIIIC - RNA polymerase transcription factor C

TIR - Terminal inverted repeats

TPI - Triosephosphate isomerase

TPMT - thiopurine s-methyltransferase

TPRT - Target primed reverse transcription

tracrRNA - Trans-activating RNA

TRD - Tandem repeat disorder

TSD - Target site duplication

TSS - Transcriptional start site

UBC - Ubiquitin C

UCSC - University of Santa Cruz California

UCSD - University of California San Diego

UGT1A1 - UDP-glycosyltransferase 1 family polypeptide A1

ULD/EPM1 - Unverricht-Lundborg disease

Un - Untransfected

UTR - Untranslated region

VCF - Variant call format

VDAC1 - Voltage dependent anion channel 1

VNTR - Variable number tandem repeat

WDR7 - WD repeat domain 7

WGS - Whole genome sequencing

WT - Wildtype

XDP - X-linked dystonia Parkinsonism

YB1 - Y-box binding protein

ZNF - Zinc finger protein

Contents

List of figures

List of tables

Chapter 1: General Introduction

Thesis overview

The aim of this PhD was to characterise the genetic variation within and investigate the role of non-coding repetitive DNA in amyotrophic lateral sclerosis (ALS) risk loci. The genomic regions in question took the form of both variable number tandem repeats (VNTRs) and transposable (mobile) elements, specifically non-LTR retrotransposons: both of which have been previously shown that they may have profound regulatory impact on host gene expression. However, repetitive and mobile DNA remain difficult to characterise, particularly larger, GC-rich and more complex (imperfect) tandem repeats and structural variants. Several tandem repeat mutations confer risk for ALS including one which is a well characterised cause of the disease: an intronic GGGGCC repeat expansion within *C9orf72*, the most common cause of ALS in European populations. Such breakthroughs have helped to subvert and disprove the once dogmatic view that non-coding DNA is "junk DNA". The work presented here outlines optimisation of several methods (both wet lab and bioinformatic) used to characterise genetic variation within VNTRs and retrotransposons, followed by *in vitro* studies to test the functional capacity of such regions. Ultimately this PhD thesis constitutes a programme of work looking at repetitive DNA as potential sources of missing heritability and regulatory elements involved mechanistically in disease pathogenesis. This led us to discover novel rare genetic variants in ALS patients, raising the profile of non-coding tandem repeat and transposon variation in ALS.

Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disorder and the most common form of motor neurone disease (MND)[1,2]. The term ALS was first devised in the 1800's by Jean Martin Charcot to describe muscular atrophy (amyotrophic), as well as tissue hardening and scarring within the lateral spinal cord (lateral sclerosis)[3]. Clinically ALS is characterised by the degeneration and loss of the upper motor neurons in the brainstem and the lower motor neurons in the anterior horns of the spinal cord; resulting in weakness, atrophy, paralysis and death due to respiratory failure, with sufferers dying between 3 and 5 years following symptom onset[3,4]. Approximately 50% of ALS patients suffer from cognitive impairment, and 15% of all patients manifest symptoms of frontotemporal dementia (FTD)[5,6]. A 2019 review by Longinetti and Fang on the epidemiology of ALS reported that the worldwide rate of incidence ranged between 0.6 to 3.8 per 100,000 person-years, with prevalence reported to be between 4.1 and 8.4 per 100,000 individuals[7]. Unfortunately, there is no cure for this progressive and fatal disease and it the cause of more than 1 in 500 deaths in both the UK and the USA[8]. Four other forms of MND are known: primary lateral sclerosis (PLS), progressive bulbar palsy (PBP), progressive muscular atrophy (PMA) and pseudobulbar palsy[8]. Although MND is technically the umbrella term for this group of diseases, ALS is often used to refer to the collective and is routinely used in scientific literature[9]; from this point on this thesis will adopt the same nomenclature.

ALS Genetics – the big four

Approximately 5-10% of ALS cases have a family history of the disease, known as familial ALS (FALS), which is usually inherited and transmitted in a highly penetrant and autosomal dominant fashion[8,10]. The remaining cases are classified as sporadic ALS (SALS), where patients have no family history[10]. Although SALS patients lack a known family history of the disease, candidate gene and genome-wide association studies (GWAS) have now shown that there is a genetic basis contributing to both sporadic and familial forms of ALS and that it is a complex disease with a polygenic architecture[8,11].

The first ALS associated gene discovered was superoxide dismutase 1 (*SOD1*), initially uncovered back in 1993 when Rosen *et al*. identified 11 missense mutations of *SOD1* in 13 FALS families[12]. This gene encodes a Cu/Zn superoxide dismutase, a metalloprotein which is mainly localised in the cytoplasm but also present in the nucleus, mitochondria, and lysosomes[13]. SOD1 functions as an antioxidant enzyme, involved in the protection against oxidative stress and toxicity induced by reactive oxygen species such as superoxide radicals and hydrogen peroxide[13,14]. To date there are over 185 known ALS associated mutations of *SOD1*, accounting for approximately 1.2% of SALS and 14.8% FALS in European populations and 1.5% of SALS and 30% FALS in Asian populations[10,15]; meaning this gene is the second most commonly mutated gene in European ALS cases and the most commonly mutated gene in Asian ALS cases[15].

In 2006, TAR DNA-binding protein of 43 kDa (TDP-43) was discovered to be a major component of ubiquitin-positive tau-negative cytoplasmic inclusions in both ALS and FTD patients[16,17]; a pathological hallmark which is observed in ~97%

of ALS cases and ~45% of FTD cases[18]. TDP-43 is an RNA binding protein, involved in regulating of mRNA transcription, splicing, transport and translation, regulation and promotion of microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) and the formation of stress granules[18,19]. Following the breakthrough in 2006, Kabashi *et al*. in 2008 discovered eight missense mutations in *TARDBP* (which encodes TDP-43) in three FALS and six SALS patients. Many of the mutations occurred in the carboxy-terminus of the protein, postulated to disrupt TDP-43 transport to the nuclear pore and splicing activity; these mutants also showed increased TDP-43 aggregation, highlighting the important interplay between genetic predisposition and pathogenesis in ALS[20]. There are over fifty mutations of *TARDBP* associated with ALS[21], accounting for approximately 0.8% of SALS cases and 4.2% of FALS cases in European populations and 0.2% of SALS cases and 1.5% of FALS cases in Asian populations[15].

Fused in sarcoma (*FUS*), a major ALS associated risk gene, encodes a DNA and RNA binding protein involved in regulation and stimulation of transcription[22,23], DNA damage response[24,25], and RNA metabolism pathways[26]: pre-mRNA splicing[27,28], mRNA transport, translation and stability[29-31]. This protein is primarily localised in the nucleus of neurons, but due to its RNA binding capacity it can move to the cytoplasm and facilitate nucleocytoplasmic transport of RNA[32]. *FUS* was first associated with ALS in 2009, when Kwiatkowski *et al*. identified 13 coding mutations within the *FUS* gene in 17 different families of FALS cases. They also found that these mutations led to aberrant localisation, cytoplasmic retention and aggregation of the mutant FUS protein[33]. An additional study in 2009 by Vance *et al*. also identified three FALS cases harbouring *FUS* missense mutations, which

also were positive for cytoplasmic inclusions of the mutated protein[34]. It was later confirmed that *FUS* mutations were also present in SALS cases[35,36]. There are currently over 50 identified mutations of *FUS* associated with ALS[37], accounting for 0.3% of SALS cases and 2.8% of FALS cases in European populations and 0.9% of SALS cases and 6.4% of FALS cases in Asian populations[15].

The chromosome 9p21 locus was first identified as an ALS and FTD risk locus through genome-wide linkage studies in 2006[38,39], with two ALS GWAS SNPs in 9p21 later being discovered[40]. In 2011 two research groups discovered the key mutation behind the risk association of the 9p21 locus: the *C9orf72* intronic GGGGCC (G_4C_2) repeat expansion[41,42], now known to be the most common cause of ALS and FTD in European populations[43]. The C9orf72 protein remains to be extensively characterised, but has been found to localise in the cytoplasm and lysosome and elicits roles in membrane trafficking, endocytosis and autophagy[44-46]. Healthy individuals have up to 23 copies of the *C9orf72* intronic GGGGCC repeat, but expansions beyond this have been found to be associated with ALS, with some patients having more than 1000 repeats[47,48]. Overall, the GGGGCC repeat expansion of *C9orf72* accounts for 5.1% of SALS cases and 33.7% of FALS cases in European populations and 0.3% of SALS cases and 2.3% of FALS cases in Asian populations: constituting the most commonly reported mutation in European cases[15]. The molecular mechanisms exhibited by tandem repeats and the functional consequence of repeat expansions will be discussed in detail later in this chapter.

Since the discovery of the 'big four' risk variants (*SOD1*, *TARDBP*, *FUS* and *C9orf72*)[49], the rise of next generation sequencing (NGS) and whole genome

sequencing, there are now over 100 genes associated with ALS (https://alsod.ac.uk/[50]) (many of which have been extensively reviewed by Theunissen *et al*.[51] and Shatunov and Al-Chalabi[52]), indicating the significant heterogeneity of the disease.

Candidate gene studies, GWA studies and rare variant burden analysis have elucidated novel ALS risk loci, which is paramount to defining and understanding the molecular mechanisms and pathways which contribute to motor neuron degeneration. Genetic variation in ALS associated genes and several key pathological mechanisms have been hypothesised to be involved in ALS progression, including protein misfolding and aggregation, impaired RNA processing and trafficking, stress granule formation, axonal defects, oxidative stress, apoptosis, impairment of autophagy and the proteasome, mitochondrial dysfunction and inflammation[6,53]. Better understanding of the interplay between genetics and biochemistry is an integral part of developing novel therapies for ALS. At present, uncovering the extent of the genetic basis (heritability) of ALS remains a key focus, with the aim of delineating the heterogeneity of the disease.

Table 1.1. List of ALS associated genes

Genes found to contain ALS-associated genetic variation within coding sequence, function and pathways of the gene and references to ALS studies. Gene list is non-exhaustive and adapted from Mejzini *et al*, 2019[10], Theunissen *et al*, 2020[51], Shatunov and Al-Chalabi, 2021[52], with the addition of other more recently discovered ALS associated genes.

ALS Gene	Function/pathway	Reference
SOD1	Oxidative stress, mitochondrial dysfunction	Rosen *et al.*, 1993[12]
CHCHD10	Oxidative stress, mitochondrial dysfunction	Bannwarth *et al.*, 2014[54]
TARDBP	RNA processing, transport and splicing	Kabashi *et al.*, 2008[20]; Gitcho *et al.*, 2008[55]
FUS	RNA processing, transport and splicing	Kwiatkowski *et al.*, 2009[33]; Vance *et al.*, 2009[34]
C9orf72	RNA processing and metabolism, autophagy, splicing	Morita *et al.*, 2006[38]; Vance *et al.*, 2006[39]; DeJesus-Hernandez *et al.*, 2011[41]; Renton *et al.*, 2011[42]
MATR3	RNA processing and metabolism	Johnson *et al.*, 2014[56]
ANG	RNA processing and metabolism	Greenway *et al.*, 2006[57]
TAF15	RNA processing, transcription initiation	Ticozzi *et al.*, 2011[58]
TIA1	RNA metabolism, stress granule formation	Mackenzie *et al.*, 2017[59]
HNRNPA1	RNA processing and metabolism	Kim *et al.*, 2013[60]
HNRNPA2/B1	RNA processing and metabolism	Kim *et al.*, 2013[60]
EWSR1	RNA processing and metabolism	Couthouis *et al.*, 2012[61]
NEK1	DNA damage response, cell-cycle regulation, mitochondrial dysfunction	Cirulli *et al.*, 2015[62]; Kenna *et al.*, 2016[63]; Brenner *et al.*, 2016[64]; Nguyen *et al.*, 2018[65]
CFAP410 (C21orf2)	DNA damage response, cilia regulation	Van Rheenen *et al.*, 2016[11]
SPG11	DNA damage	Orlacchio *et al.*, 2010[66]
ELP3	Transcript elongation	Simpson *et al.*, 2009[67]
SMN1	RNA binding protein interaction	Corcia *et al.*, 2002[68]
SETX	DNA/RNA processing	Chen *et al.*, 2004[69]
VCP	Protein quality control, trafficking and degradation	Johnson *et al.*, 2010[70]
VAPB	Protein quality control, trafficking and degradation	Nishimura *et al.*, 2004[71]
OPTN	Protein quality control, trafficking and degradation	Maruyama *et al.*, 2010[72]
TBK1	Protein quality control, trafficking and degradation	Cirulli *et al.*, 2015[62]
CCNF	Protein quality control, trafficking and degradation	Williams *et al.*, 2016[73]

SQSTM1	Protein quality control, trafficking and degradation	Fecto *et al*., 2011[74]
APEX1	DNA repair and oxidative stress	Greenway *et al*., 2004[75]
KIF5A	Cytoskeletal and trafficking, organelle transport	Nicolas *et al*., 2018[76]
DCTN1	Cytoskeletal and trafficking	Munch *et al*., 2004[77]
TUBA4A	Cytoskeletal and trafficking, axonal transport	Smith *et al*., 2014[78]
PFN1	Cytoskeletal and trafficking	Wu *et al*., 2012[79]
ANXA11	Cytoskeletal and trafficking	Smith *et al*., 2017[80]
PRPH	Cytoskeletal protein	Leung *et al*., 2004[81]
NEFH	Cytoskeletal and trafficking	Figlewicz *et al*., 1993[82]
FIG4	Cytoskeletal organisation and vesicle trafficking	Chow *et al*., 2009[83]
KIFAP3	Anterograde transport and chromosomal cytokinesis	Landers *et al*., 2009[84]
ALS2	Endosomal dynamics	Hadano *et al*., 2001[85]
ATXN1	RNA processing, chromatin binding	Lattante *et al*., 2018[86]
ATXN2	RNA processing, cell survival, endocytosis	Elden *et al*., 2010[87]
SIGMAR1	Endoplasmic reticulum chaperone	Al-Saif *et al*., 2011[88]
UNC13A	Neurotransmitter release	Van Es *et al*., 2009[89]
GRN	Cell growth regulator	Schymick *et al*., 2007[90]
CHMP2B	Cell receptor recycling	Parkinson et al., 2006[91]
DPP6	Calcium gated channel modification	Van Es *et al*., 2008[92]
DAO	Detoxifying agent	Mitchell *et al*., 2010[93]
VEGFA	Migration of endothelial cells, angiogenesis	Lambrechts *et al*., 2003[94]
ITPR2	Neurotransmission, apoptosis	Van Es *et al*., 2007[95]
ARHGEF28	Nucleotide exchange factor	Droppelmann *et al*., 2013[96]
PON1	Organophosphate hydrolysis	Slowik *et al*., 2006[97]
HFE	Iron absorption	Wang *et al*., 2004[98]
DNAJC7	Heat shock protein, autophagy	Farhan *et al*., 2019[99]
GLT8D1	Glycosyltransferase	Cooper-Knock *et al*., 2019[100]
NIPA1	Synapse and axon development	Blauw *et al*., 2012[101]

ALS is a complex genetic disease

Heritability describes the phenotypic variance which can be explained by genetic factors[102]. In other words, it indicates how much variability in a trait (such as variation in risk of developing a disease) can be attributed to genetic variation[103,104]. Al-Chalabi *et al*. have predicted ALS heritability by performing two twin studies and estimate SALS heritability to be 0.61[105]: meaning that 61% of the variability in ALS susceptibility/risk can be explained by genetic variation[106]. This highlights that there is a significant genetic contribution to ALS risk variance. The genetic basis of ALS also works in concert with a number of environmental factors[107], such as age and gender[108-110], as well as exposure to heavy metals and organophosphates-containing chemicals such as pesticides[111,112], indicating the complexity of the disorder[113]. It is now hypothesised that ALS pathogenesis is a multistep process, with the contribution of both genetic and environmental risk factors via gene-environment interaction[110,114-116].

Although monogenic forms of ALS exist with single mutations inherited in a Mendelian fashion which are known to cause the disease, such as single high effect and pathogenic mutations in *SOD1*, this does not often explain disease segregation across families[5,49]. This has been found within four European ALS pedigrees, highlighting that some ALS patients within a family pedigree do carry *SOD1* mutations but some affected individuals do not: unaffected family members have also been found to harbour *SOD1* mutations[117]. Due to such complexities, an oligogenic susceptibility profile of ALS has now been proposed[113], with numerous studies highlighting that ALS patients can carry mutations in more than one gene,

with multiple variants influencing disease pathogenesis[15,65,118-122]. Interestingly, a study by Beck *et al*. identified 11 unaffected individuals in the UK with large *C9orf72* repeat expansions (>400 repeats)[123], again signifying incomplete penetrance and that a combinatorial effect with mutations in other ALS genes may be required to determine presentation of the disease[5]. It is now also known that genetic variation and alterations in expression of ALS genes can correlate with clinical profiles, influencing phenotypes such as site of onset, age of onset and duration of disease[124]. Further complication also lies with the pleiotropic effect of some gene variants, inducing multiple clinical phenotypes: including the *C9orf72* repeat expansion being implicated in both ALS and FTD and the *ATXN2* repeat expansion being associated with ALS, FTD and spinocerebellar ataxia[41,42,125,126]. Nguyen *et al*. have recently reviewed the reported ALS and ALS-FTD patients who are carriers of multiple gene variants (n = 74). They found that 51 (69%) of these patients harboured the *C9orf72* repeat expansion, with the next most highly observed double mutations being found with either *TARDBP* or the more recently discovered ALS gene, *NEK1*[5]. Two recent studies have also identified the co-occurrence of an ALS risk associated *ATXN1* repeat expansion with the GGGGCC *C9orf72* repeat expansion, indicating the presence of multiple tandem repeat expansions in the same patient[86,127]. Although both low-frequency (with high effect) and common-frequency (with low effect) mutations contribute to ALS, most of the identified ALS variants are hypothesised to have an intermediate effect size and confer moderate risk[115]. It was previously postulated that rare variants are likely to be population specific and therefore will make replication studies difficult[115].

While GWA studies have helped identify numerous novel loci associated with ALS, such studies almost always only focuses on common variation (minor allele frequency>1%) and therefore only captures a fraction of the genetic contribution[128]. A recent GWA study in 2016 by van Rheenen *at el*. analysed 12,577 ALS cases and 23,475 controls and identified a nonsynonymous variant within *CFAP410* (previously known as *C21orf2*) (rs75087725, P=8.7 x 10^{-11}). They identified a genome-wide association for this variant by imputing genotypes (estimating missing genotypes using haplotype data) from a merged reference panel constructed from high coverage (43x) WGS data of Dutch ALS cases (n=1246) and controls (615) and the 1000 Genomes Project Phase I reference panel. Merging the reference panels led to the *CFAP410* variant (rs75087725) being identified in more haplotypes: 62 instead of the 10 found by using the 1000 Genomes Project Phase I panel alone, allowing more samples to pass quality control as they were over the 1% allele frequency threshold and therefore leading to genome-wide association[11]. Overall, in this study they estimated SNP-based heritability to be 8.5% but the identified GWAS loci only accounted for 0.2%, meaning that the remaining associations were not genome-wide significant. This finding led van Rheenen *et al*. to estimate heritability but partition it by minor allele frequency (MAF), to try and determine where the undiscovered risk loci lay in terms of allele frequency. They found that the majority (~50%) of ALS risk heritability could be attributed to low frequency variation (SNPs with MAF between 0.001-0.1). Ultimately this finding showed that ALS had a polygenic architecture with rare variation making a large contribution to ALS heritability[11].

Analysing and discovering rare genetic variation associated with ALS is a major research focus through Project MinE: an international collaboration which aims to investigate whole genome sequencing (WGS) of 15,000 ALS cases and 7,500 controls[128]. This global initiative seeks to discover and analyse the rare genetic variation often missed by GWAS, thus helping to build a more complete picture of the genetic contribution to ALS risk. Through large-scale collaboration Project MinE has generated large cohorts in order to detect rare variation using techniques such as rare genic burden testing, for example helping to discover *NEK1* loss of function (LOF) variants which confer risk for ALS[63]. Other breakthroughs include the discovery of *CFAP410*, *NIPA1* and *ATXN1* as novel ALS risk loci. The ongoing work within Project MinE has led to the formation of several working groups, each focusing on particular aspects of ALS research such as phenotyping, association testing, epigenetics, structural variation and non-coding genomic variation[129]. Two of these recently discovered ALS genes will be discussed in detail within the data chapters of this thesis: *CFAP410* (previously known as *C21orf2*) in Chapter 3 and *NEK1* in Chapter 4.

Theunissen *et al*. argue that structural variants (SV) could be an important source of missing heritability within ALS[51]. Two main sources of structural variation are tandem repeats and transposable elements; both forms of repetitive DNA. Due to the limitations of short read sequencing and wet lab assessment being difficult, repetitive DNA characterisation remains a complicated avenue of genetics. The latter half of this chapter will highlight the importance of repetitive DNA (both tandem repeats and transposable elements) in the context of genomic regulation and pathogenesis of neurodegenerative disease.

Repetitive DNA

Over 50% of human DNA is repetitive in nature: DNA which is identical or similar in sequence to other regions within the genome[130]. This variation can occur in tandem but also in a dispersed manner in the form of transposable (mobile) elements leading to large structural variation across the human genome[131], often referred to as the "repeatome"[132]. This section will focus on these two forms of repetitive elements which make up the "repeatome", making the argument for their importance in shaping and regulating the human genome.

Variable Number Tandem Repeats (VNTRs)

Tandem repeats (TRs) are DNA sequence motifs which are repeated numerous times and appear contiguously at a locus. Constituting approximately 3%, these elements are found throughout the human genome[132,133]. It is hypothesised that this variation has arisen due to slippage errors in DNA replication, such as slipped-strand mispairing which results in the misalignment of DNA strands and thus expansion or contraction of the copy number of the DNA motif[134]; thus they can be variable in sequence length in the human population, differing in repeat size across individuals, leading to the term variable number tandem repeat (VNTR). TRs are often referred to as short tandem repeats (STRs), simple sequence repeats as well as micro and minisatellites: microsatellites are defined as elements containing repeats of 1-6 bp in length, while minisatellites contain approximately 6-500 bp motifs[135]. For simplicity, these repeat elements will be referred to as variable number tandem repeats (VNTRs) throughout this thesis.

VNTRs – modulators of gene expression

Originally it was thought that VNTRs lacked function and were inactive in the human genome, often being termed "junk" DNA[136]. However, it is now known that this is not the case and it has been extensively shown that VNTRs have a profound impact on genetic architecture and gene regulation. VNTRs are found within exons, introns and intergenic space, each exhibiting particular regulatory and functional consequences (Figure 1.1).

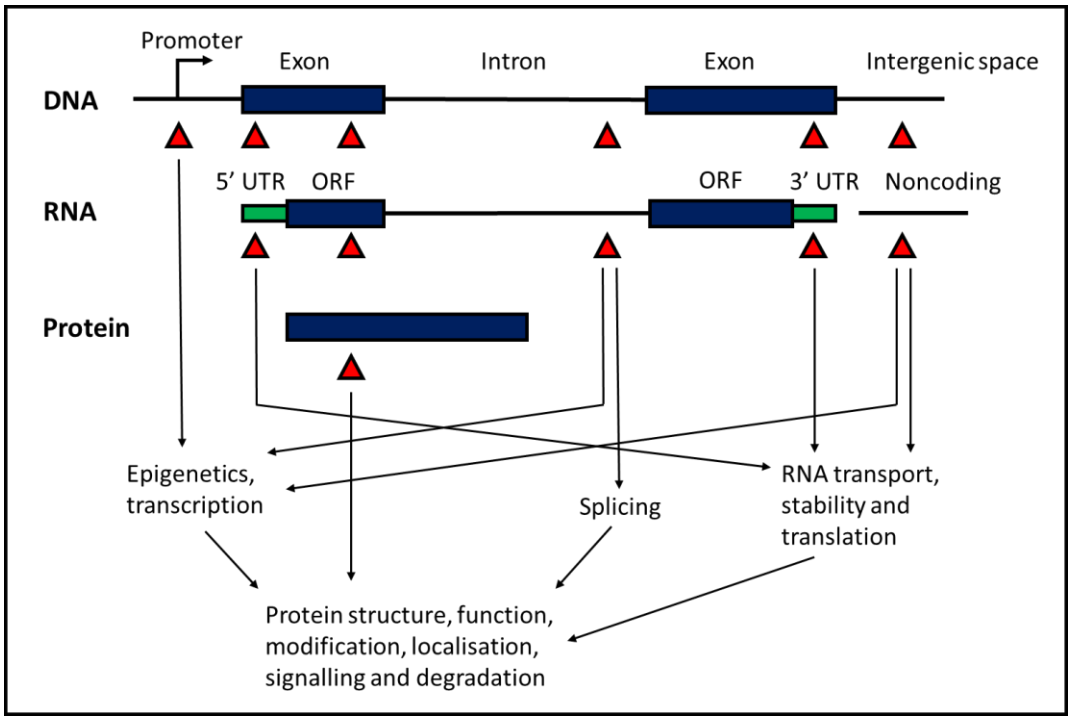


Figure 1.1. VNTR polymorphisms and their downstream effects

Schematic representing a gene locus and the possible functional effects of polymorphic VNTRs at specific locations. Polymorphic VNTR are shown as red triangles and can be present in promoters, untranslated regions (UTRs), exons, introns and the intergenic spaces. Sequence length polymorphisms of VNTRs can alter regulation of transcription, splicing and translation, thus altering DNA, RNA and protein levels. Adapted from Hannan, 2010[137].

There is a wealth of literature demonstrating that VNTRs can act in a transcriptional capacity, which has been previously reviewed[138-140]. VNTRs can serve as transcription factor (TF) binding sites, where repeat number can modulate gene expression profiles by altering TF affinity at gene promoters[141,142]. For example, Hsieh *et al*. have shown that an extra TA insertion within a TATA-like box sequence found in the promoter of *UDP-glycosyltransferase 1 family polypeptide A1* (*UGT1A1*) causes a decrease in gene expression by lowering the binding affinity for TATA-binding protein: this affinity was shown to decrease as the number of TA insertions increased[143]. Zukic *et al*. identified a GC-rich VNTR within the promoter of thiopurine s-methyltransferase (*TPMT*) gene and showed in K562 cells that the VNTR can influence gene expression based on tandem repeat size and also mediate *TMPT* transcription through the binding of GC-rich box transcription factors Sp1 and Sp3[144]. Two VNTRs within the serotonin transporter (*SLC6A4*) gene which are present in the linked polymorphic region (LPR) of the 5' promoter and intron 2 (Stin2) have been assessed. Ali *et al*. using renilla luciferase reporter gene assays in rat prefrontal cortical cells have shown that the 9, 10 and 12-copy Stin2 VNTR variants were all regulated by CCCTC-binding factor (CTCF), each showing significantly reduced reporter gene expression: CTCF did not regulate activity of the LPR VNTR[145]. An additional study by Haddley *et al*. showed that both *SLC6A4* VNTRs bind several transcription factors, including CTCF, y-box binding protein (YB1) and methyl-CpG binding protein 2 (MeCP2), but exhibited allele specific binding of these factors in response to cocaine, highlighting that VNTRs can regulate gene expression in an allele-specific and stimulus inducible manner[146].

VNTRs have been shown to drive gene expression in an allele dependent manner. Warburton *et al*. identified an internal VNTR within the promoter of microRNA-137 (miR-137). Through cloning the promoter region encompassing the 4-copy and 12-copy variants of VNTR into pGL3-B reporter gene construct (lmir137(4) and lm137(12)) they found that the internal promoter containing the identified VNTR supported luciferase reporter gene expression in SH-SY5Y cells. The 4-copy and 12-copy VNTR alone were and cloned into pGL3-P reporter gene construct (VNTRmir137(4) and VNTRmir137(12)): when compared to the empty pGL3-P vector, the 4-copy VNTR drove an increase in luciferase activity while the 12-copy VNTR led to a decrease in reporter gene expression, highlighting the capacity for VNTR repeat number to drive differential gene expression profiles[147]. The previously mentioned study by Ali *et al*. found that the 9 and 10-copy variants of the Stin2 VNTR within the *SLC6A4* gene exhibited differential reporter gene activity in rat prefrontal cortical cells, further highlighting that VNTRs can exhibit allele-specific expression profiles[145].

VNTR polymorphisms also work in concert with other genetic variants to exhibit distinct and combinatorial regulatory mechanisms. An additional study from Warburton *et al*. using haplotyping analysis at the *MIR137* locus discovered a proxy SNP (rs2660304) within the internal promoter region of Mir137 which was in linkage with the rs1625579 schizophrenia GWAS SNP. It was found that the VNTR from the aforementioned study worked together with rs2660304 risk SNP to drive differential promoter activity. Expression constructs containing the internal Mir137 promoter (with the 4-copy VNTR) were generated, either containing the rs2660304 SNP major allele (A) or minor allele (C): lmir137(4)+A and lmir137(4)+C.

The lmir137(4)+A construct led to a significant reduction (1.32 fold decrease) in luciferase expression, while the lmir137(4)+C did not drive any changes in reporter gene activity[148]. It has also been shown that both the LPR and Stin2 VNTRs of the *SLC6A4* gene display a combinatorial regulatory effect *in vitro*, with dual VNTR constructs (LPR and Stin2 in conjunction) producing increased reporter gene expression when compared to single VNTR constructs[145]. Recent work by Manca *et al*. has also shown that VNTRs can exhibit isoform specific modes of regulation and multiple VNTRs can have an additive effect on gene expression profiles. They assessed two VNTRs (dVNTR and uVNTR) found upstream of the monoamine oxidase A (*MAOA*) gene and discovered that a single knock out (KO) of the dVNTR and a double KO of the dVNTR and uVNTR both led to a significant decrease in total *MAOA* expression. They also measured expression of a minor alterative isoform of *MAOA* and found that single KOs of each VNTR led to a significant increase in expression of this isoform and a double KO induced a further increase in expression, highlighting isoform specific regulatory effects and an additive effect of the two VNTRs at this locus[149].

Intronic VNTR polymorphisms have also been shown to alter splicing. De Roeck *et al*. identified a 592 bp VNTR within intron 18 of *ATP binding cassette subfamily A member 7* (*ABCA7*), built of 23.7 units of a 25 bp repeat. Genotyping analysis of Alzheimer's disease (AD) patients (n=275) and controls (n=177) by Southern Blotting led to the discovery of several VNTR variants, ranging from 298 bp (12 repeats) to 10678 bp (427 repeats). Interestingly, expanded VNTR allele (>5720 bp) frequency was significantly higher in AD (7.3%) compared to controls (1.7%). Expression analysis also showed that in both AD and controls, increasing

VNTR size correlated with a decrease in *ABCA7* gene expression. This VNTR was shown to regulate splicing of *ABCA7*, specifically inducing exon 19 skipping, leading to loss of 44 amino acids. Furthermore, a strong positive correlation was observed between ratio of exon 19 skipping and VNTR length, with an expansion of the VNTR increasing skipping of exon 19[150]. Due to the repetitive nature of VNTRs they can form secondary structures which can induce genomic instability[151,152]. Formation of RNA secondary structures exhibited by VNTRs has been shown to alter splicing efficiency[153]. Non-canonical structures formed by VNTRs include triplexes, hairpins, cruciforms, R-loops and G4 quadruplexes, which have roles in genome organisation and the regulation of transcription and DNA replication[154,155].

It has previously been shown that VNTRs contribute to gene expression change of nearby genes on a genome-wide level, while also contributing the clinical phenotypes. Gymrek *et al*. assessed 311 individuals with lymphoblastoid cell line RNA-seq data and short tandem repeat (STR) genotyping data to identify VNTR genetic variation which associated with nearby gene expression changes (termed expression STRs; eSTRs). Of the 15,000 protein coding genes within the RNA-seq dataset, a total of 2,060 protein coding genes had significantly associated eSTRs. Furthermore, these eSTRs were found to be enriched in regulatory regions and transcriptional start sites, as they co-localised with peaks of histone modifications associated with those regions. Using the TwinsUK cohort of the UK10K project they were able to test association between eSTRs and 38 phenotypes. A total of 12 associations between eSTRs and clinical phenotypes were found to statistically significant, including FEV transcription factor (FEV1) lung function, diastolic blood pressure and changes in levels of uric acid, albumin, C-

reactive protein, and haemoglobin[156]. A study by Quilez *et al*. used targeted sequencing to genotype promoter associated VNTRs. VNTR variants were then correlated with gene expression changes and CpG methylation levels to identify expression quantitative loci (eQTLs) and methylation quantitative loci (mQTLs). These were then overlaid with regulatory regions on the basis of transcription factor binding sites and DNaseI hypersensitive regions. They found that VNTRs which significantly altered expression and methylation levels of nearby genes (significant eQTLs and mQTLs) were more frequently found in regulatory regions, highlighting a preferential overlap and enrichment of functional VNTRs in regulatory regions of the genome. Quilez *et al*. also performed linkage disequilibrium analysis and found that VNTRs, especially those with multiple variants, were poorly tagged by nearby (≤250 kb) SNPs: an r^2 value above 0.6 was not attained by any VNTR, indicating the limitations of using SNP mapping to genotype VNTRs in the human genome[157].

VNTRs and disease

Polymorphic VNTRs are associated with a number of neuropathological diseases, often referred to as tandem repeat disorders (TRDs). There are two mechanisms by which VNTRs can contribute to the aetiology of disease: expansion mutations which cause disease and mutations which confer risk for disease. To date there are over 20 known TRPs which cause disease, most of which are CAG polyglutamine tract expansions (Table 1.2). Many of the TRPs cause conditions of the nervous system and some have been known for over 20 years, including those

involved in fragile X syndrome (FXS) myotonic dystrophy (MD), Huntington's disease (HD) and spinocerebellar ataxias (SCA)[132,158-168].

Table 1.2. List of human diseases caused by polymorphic VNTRs

Disease, gene, region of gene and tandem repeat mutation are shown. Table is non-exhaustive and adapted from Hannan, 2018[169], with the addition of NIID[170] and ULD/EPM1[171].

Disease	Gene	Region of gene	Tandem repeat
Huntington's disease (HD)	*Huntingtin (HTT)*	Exon	CAG
Huntington disease-like 2 (HDL2)	*Junctophilin 3 (JPH3)*	Exon	CTG
Amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD)	*C9orf72*	Intron	GGGGCC
Myotonic dystrophy 1 (DM)	*DM1 protein kinase (DMPK)*	3' UTR	CTG
Myotonic dystrophy 2	*CCHC-type zinc finger nucleic acid binding protein (CNBP)*	Intron	CCTG
Fragile X syndrome (FXS)	*FMRP translational regulator (FMR1)*	5' UTR	CGG
Fragile X tremor-ataxia syndrome (FXTAS)	*FMRP translational regulator (FMR1)*	5' UTR	CGG
Spinocerebellar ataxia 1 (SCA1)	*Ataxin 1 (ATXN1)*	Exon	CAG
Spinocerebellar ataxia 2 (SCA2)	*Ataxin 2 (ATXN2), ataxin 2 anti-sense RNA (ATXN2-AS)*	Exon	CAG
Spinocerebellar ataxia 3 (SCA3)	*Ataxin 3 (ATXN3)*	Exon	CAG
Spinocerebellar ataxia 6 (SCA6)	*Calcium voltage-gated channel subunit alpha 1A (CACNA1A)*	Exon	CAG
Spinocerebellar ataxia 7 (SCA7)	*Ataxin 7 (ATXN7)*	Exon	CAG
Spinocerebellar ataxia 8 (SCA8)	*Ataxin 8 (ATXN8), ataxin 8 opposite strand lncRNA (ATXN8-OS)*	Exon	CTG
Spinocerebellar ataxia 10 (SCA10)	*Ataxin 10 (ATXN10)*	Intron	ATTGT
Spinocerebellar ataxia 12 (SCA12)	*Protein phosphatase 2 regulatory subunit B beta (PPP2R2B)*	5' UTR	CAG
Spinocerebellar ataxia 17 (SCA17)	*TATA-box binding protein (TBP)*	Exon	CAG
Spinobulbar muscular atrophy (SBMA)	*Androgen receptor (AR)*	Exon	CAG
Dentatorubral-pallidolusian atrophy (DRPLA)	*Atrophin 1 (ATN1)*	Exon	CAG
Friedreich ataxia (FA)	*Frataxin (FXN)*	Intron	GAA
Neuronal intranuclear inclusion disease (NIID)	*Notch 2 N-terminal like C (NOTCH2NLC)*	5' UTR	GGC
Unverricht-Lundborg disease (ULD/EPM1)	*Cystatin B (CSTB)*	5' flank/promoter	CCCCGCCCCGCG

Recently the discovery of the GGGGCC repeat expansion in *C9orf72* has led to an improved understanding of the neuropathological mechanisms that TPRs can promote[41,42]. One of the proposed mechanisms induced by the GGGGCC repeat expansion is the loss of function of the C9orf72 protein due to haploinsufficiency, as previous studies have shown that the intronic repeat expansion induces a

decrease in C9orf72 protein levels[41,42,47]. A second mechanism of action is toxic gain of function via RNA foci. These sense and anti-sense foci can accumulate in the nucleus of neurons and sequester RNA-binding proteins leading to cellular dysfunction and toxicity[172]. It has also been previously shown that there is a negative correlation between age of onset of disease and RNA foci burden in the frontal cortex of ALS patients[173]. The final pathological mechanism of action by tandem repeats is through a process known as repeat-associated non-AUG (RAN) translation. This is a noncanonical form of translation that occurs in the absence of an AUG start codon and was discovered through CAG expansion constructs expressing homopolymeric proteins[174]. Similarly, RAN translation of CAG expansions in human spinocerebellar ataxia type 8 (SCA8) and myotonic dystrophy type 1 (DM1) can generate polyalanine and polyglutamine protein tracts respectively[175]. Another result of RAN translation is the formation of dipeptide repeat expansions, with poly glycine-alanine, poly glycine-arginine and poly glycine-proline dipeptide repeats all being associated with the *C9orf72* GGGGCC repeat expansion found in ALS and FTD[176]. Formation of homopolymeric proteins or dipeptide repeats can lead to cellular toxicity through a number of mechanisms, such as loss of function of the protein or toxic gain of function at either the RNA or protein level, causing a number of downstream effects such as endoplasmic reticulum stress, protein sequestration, proteasome inhibition and reduction of dendritic branching[47,177,178].

Not all VNTR mutations are directly causative of disease, but have been identified as associated risk factors, increasing susceptibility to disorders[150,179,180]. One recent example in ALS is *NIPA magnesium transporter 1* (*NIPA1*) CGC

polyalanine repeat length. Blauw *et al*. genotyped the polyalanine repeat found within exon 1 of *NIPA1* in 2292 ALS cases and 2777 controls from three European populations (Dutch, Belgian and German) and discovered that long repeat length of the *NIPA1* polyalanine tract (>8 GCG repeats) not only led to a significant increase in susceptibility for ALS, but also modulated age of onset and median survival. The genotyping results showed that 7 and 8 copies of GCG were the most frequent and defined as 'normal', while any repeat length larger than 8 repeats was defined as 'long'. It was found that there was a statistically significant difference in long repeat frequency between ALS cases and controls and that there was a significant increase in ALS susceptibility in all three populations. When combining survival data from Dutch and Belgian cohorts there was a significant decrease in median survival of long repeat carriers (shorter by 3 months). The combined age of onset of long repeat carriers was also significantly lower (by 3.6 years)[101]. This association has also been replicated within an international cohort by Tazelaar *et al*. in 2019. They genotyped *NIPA1* CGC repeats in 3955 ALS cases and 2276 controls and found that there was an increased ALS risk in cases with greater than 8 CGC repeats. These results were then combined with previous studies and this meta-analysis (6245 ALS cases and 5051 controls) also showed an increased risk of ALS for carriers of long *NIPA1* CGC repeats[181].

As mentioned previously in Table 1.2, a CAG repeat expansion in *ATXN2* is a known cause of spinocerebellar ataxia type 2 (SCA2), specifically due to an expansion of >34 CAG repeats[126]. Yet this repeat expansion has also been found to be associated with increased risk for ALS. From assessing 2802 ALS cases and 1258 controls from UK and Dutch cohorts, Sproviero *et al*. found that there was an

overall increased risk of ALS in people harbouring intermediate repeat expansions (between 24 and 34 CAG repeats). They also revealed that relative risk of ALS increased as the size of the intermediate repeat increased (between 29-32 repeats) but the risk then dropped above 33 repeats: the range associated with causing SCA2[125].

The Project MinE ALS Sequencing Consortium and associated collaborators identified a tandem repeat expansion within *ATXN1* conferring risk for ALS. Tazelaar *et al*. analysed samples from four different cohorts (Ireland, France, Belgium, and The Netherlands), encompassing 2672 ALS patients and 2416 controls. *ATXN1* CAG/CAT intermediate expansions (≥33 repeats) were found in 12.2% of ALS patients and 10.1% of controls: a significant association between the presence of at least *ATXN1* intermediate repeat and ALS status was observed. Tazelaar *et al*. also assessed the *ATXN1* trinucleotide repeat length using the STR calling tool, Expansion Hunter[182], in a separate cohort of 2048 ALS patients and 891 controls. It was identified that a subset of this cohort (n=1129) showed a 98% agreement across PCR and Expansion Hunter genotyping calls, indicating high accuracy and confidence in the generated calls from Expansion Hunter. Furthermore, they observed intermediate expansions in 12.0% of ALS patients and 8.8% of controls. From this they performed a largescale meta-analysis with 11,700 individuals and found a statistically significant association between *ATXN1* intermediate repeat (≥33) expansions and ALS risk ($p=3.33x10^{-7}$). However, they observed that the presence of *ATXN1* intermediate repeat had no significant effect on either survival or age of onset of ALS. Aside from association with ALS risk, Tazelaar *et al*. also tested the effect of the *ATXN1* repeat expansion on a

neurodegeneration associated phenotype within *Drosophila*: rough eye, causing depigmentation and necrotic spot formation. Interestingly they found that overexpression of an expanded ATXN1 poly Q (poly glutamine) repeat (82Q compared to 2Q) increased this phenotype slightly, but when co-expressed with human TDP-43 led to a severe phenotype (increased necrotic spot formation). This severe phenotype was also observed when expanded 82Q polyQ length repeats were expressed in a *Drosophila* model of *C9orf72* (which expresses toxic glycine-arginine dipeptides), again proving that the ATXN1 polyQ can modify disease phenotype. Overall, they propose that the *ATXN1* CAG repeat expansion is not pathogenic but could contribute to the multistep process of disease manifestation[127].

Course *et al*. have used long read sequencing to identify a 69 bp VNTR in the final intron of WD repeat domain 7 (*WDR7*), ranging from 1 to 86 copies in length. They genotyped this VNTR in SALS (n=376), PD (n=531) and controls (n=639) and found that per individual the median copy number of the 86-copy variant was significantly higher in SALS patients compared to controls. They also assessed WGS from an ALS Quebec cohort (n=470) and observed a significantly higher repeat number of the VNTR in SALS patients compared to control samples, determining an association between higher repeat number and SALS; however, no association between longer repeat number and age of onset of ALS was observed[183].

Overall, VNTR polymorphisms have been found to not only cause but also increase the risk of developing disease, particularly disorders of the central

nervous system (CNS), including ALS. Over the past ten years, with the discovery of five tandem repeat expansions within genes linked to ALS (*C9orf72*, *ATXN1*, *ATXN2*, *NIPA1* and *WDR7*) it has been made clear that VNTRs are a strong candidate of missing heritability in this disease, highlighting the need to characterise other repetitive regions in ALS patients. Chapter 3 of this thesis summarises work carried out to characterise VNTR polymorphisms in ALS patients, with the aim of identifying potential novel risk factors for ALS.

Transposable Elements

Genomic DNA is not a stable or static entity, with approximately half of the human genome being derived from transposable elements (TEs)[184]; repetitive DNA elements which can mobilise to different locations within the genome[185]. TEs were first discovered in maize in 1950 by Barbara McClintock[186], paving the way for detection of TEs in other organisms and the acceptance of transposition being a widespread phenomenon[187]. TEs are split into two families based on the transposition strategy and intermediate formed during replication: retrotransposons (type I) and DNA transposons (type II) (Figure 1.2)[188]. Retrotransposons replicate through a copy-and-paste mechanism via an RNA intermediate, which is reverse transcribed and re-inserted back into the genome[189,190]. DNA transposons mobilise through a cut and paste mechanism via a DNA intermediate, facilitating insertion into new genomic locations[191]. Although originally branded “junk” DNA, TEs are now known to have a roles in the regulation and evolution of the genome as well as contributing to genetic instability and disease progression[192,193].

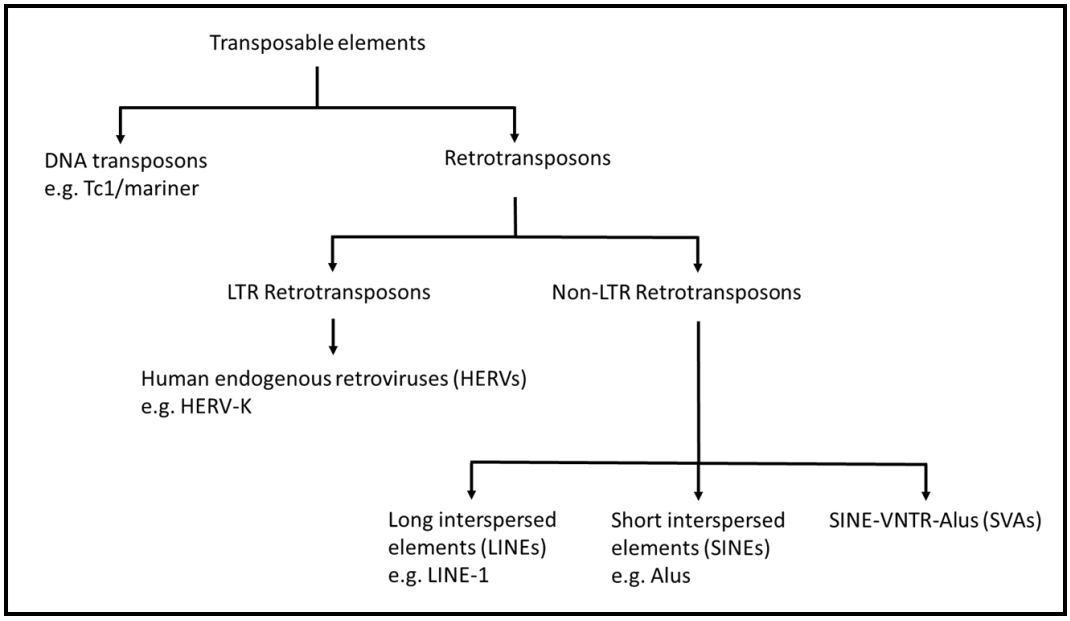


Figure 1.2. Transposable element classification

Transposons are split into two major groups: retrotransposons and DNA transposons. The non-LTR retrotransposons are the only elements which are known to be mobile in the human genome. LINE-1 is autonomous as it encodes the ORF1 and ORF2 machinery necessary for mobilisation: *Alu* and SVA elements are non-autonomous and therefore require the machinery encoded by LINE-1 to transpose. Adapted from Misiak *et al*., 2019[194].

DNA transposons constitute approximately 8% of the human genome, with examples of these including Tc1/mariner, piggyBac, MuDr and hAT elements[194]. These transposons have not been found to be functionally active for mobilisation in the human genome. Retrotransposons account for approximately 42% of the human genome and are characterised into two types: long terminal repeats (LTR) and non-LTRs, with the latter being the only active (mobile) transposable elements in humans[191,195,196]. LTR retrotransposons include the human endogenous retroviruses (HERVs), such as HERV-K, HERV-W. Non-LTRs are divided into three families: Long Interspersed Elements-1 (LINE-1, L1), *Alu* elements (SINEs) and SVA elements (SINE-VNTR-Alu)[191]. Non-LTR retrotransposons are now known to be

mobilisation competent and have been shown to implicated in the aetiology of several diseases and will now be the focus of discussion for the rest of this chapter.

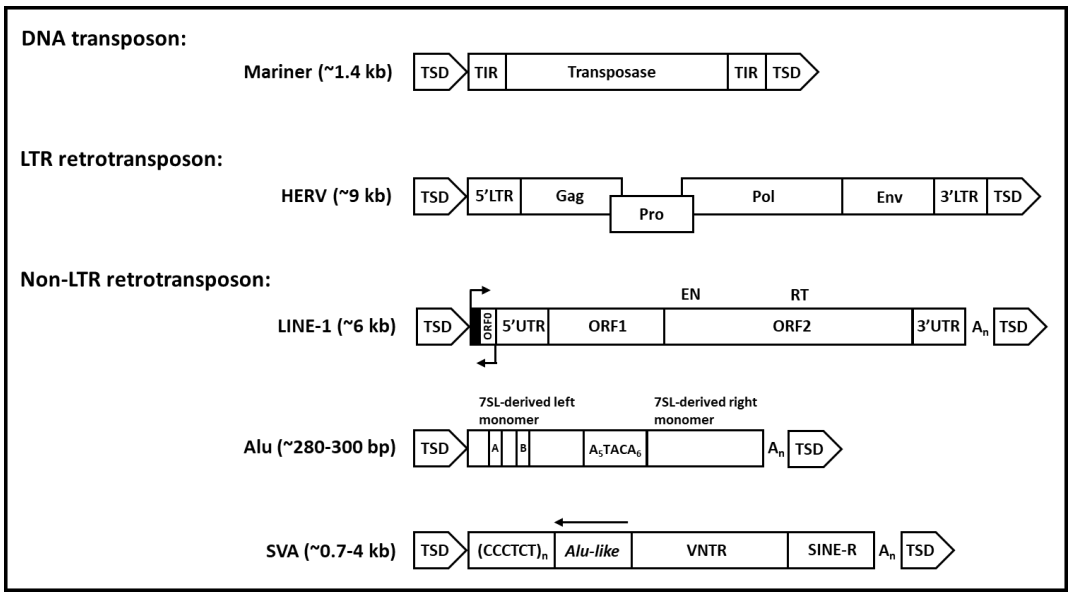


Figure 1.3. Structure of transposable elements

Composite structure of several transposable elements found within the human genome. Example of a Hsmar-1 DNA transposon (part of the mariner-like subfamily), which is approximately 1.4 kb in length: encoding the transposase enzyme which is flanked by terminal inverted repeats (TIRs) and target site duplications (TSDs). HERVs are approximately 9 kb in length and encode the group specific antigen (*gag*), protease (*pro*), polymerase (*pol*) and envelope (*env*) proteins, which are flanked by long terminal repeats (LTRs) and TSDs. A full-length LINE-1 element is approximately 6 kb and encodes three open reading frame (ORF) proteins: ORF0, ORF1 and ORF2. The 5'UTR contains both a sense and anti-sense promoter. The 3' end of the element contains both a 3'UTR and a poly A tail which is a transcriptional termination signal. *Alus* are between 280-300 bp in length and built of two monomers derived from 7SL RNA which are split by an A-rich (A_5TACA_6) connector sequence. SINE-VNTR-Alu (SVA) elements range from approximately 0.7-4 kb in length. The 5' end contains a CT rich ($CCCTCT_n$) domain known as a CT element, followed by an Alu-like sequence which is built of two anti-sense Alu fragments, a VNTR region, a SINE-R domain and finished with a 3' Poly A tail.

Structures not drawn to scale. Adapted from Goodier, 2016[197] and Savage *et al*., 2019[191].

Retrotransposons

HERV

HERVs are sequences derived from ancient and now extinct exogenous retroviruses which infected and integrated into primate genomes approximately 100 million years ago[198]. These provirus remnants have bypassed host defence mechanisms, integrating, and propagating throughout human evolution[199]. Structurally, HERVs are built of 5' and 3' long terminal repeats, encapsulating the viral genes *gag*, *pro*, *pol* and *env*: encoding capsid proteins, viral protease, reverse transcriptase and envelope protein, respectively[200,201]. HERVs account for approximately 8% of the human genome and at least 31 HERV families have been identified in humans, with the most recently acquired being HERV-K[199,202]. HERV-K has now shown to be transcriptionally active during embryogenesis and in certain other circumstances, including during HIV infection and melanoma[203]. Furthermore, these proviruses have been implicated in a number of neurological disorders, including multiple sclerosis, schizophrenia, bipolar disorder and ALS[204-207].

LINE-1

Long interspersed nuclear element-1 (LINE-1, L1) constitutes approximately 17% of the human genome, with over 500,000 copies present[208]; although only 80-100 of these are active[209]. Of the 80-100 active elements it has been found that the majority of mobilisation events have arisen from a small number (approximately 5-20) of these elements, which are referred to as "hot"

(highly active) L1s[208,210,211]. These 6kb elements are retrotransposition competent, encoding for three open reading frames: ORF0, ORF1 and ORF2. ORF0 is an anti-sense domain which through overexpression studies has been shown to increase L1 mobility, indicating a role in influencing and enhancing L1 mobilisation[212]. ORF1 is a 40 kDA protein which has RNA binding and chaperone activity, while ORF2 is a 150 kDa protein with both reverse transcriptase and endonuclease activity; it is this machinery that is necessary for retrotransposition to occur via a process called target primed reverse transcription[213] (Figure 1.4). *Alu* and SVA are non-autonomous elements and possess the ability to hijack this L1 machinery and thus can also translocate and re-insert into the genome[214].

Alu

Alu elements are primate specific short interspersed elements (SINEs) which emerged approximately 65 million years ago. These retrotransposons are one of the most successful mobile elements, with over 1 million copies present within humans, comprising approximately 11% of the human genome[215,216]. *Alu* elements are dimeric in structure and usually around 300 bp in length: built of left and right monomers originating from the 7SL RNA gene (Figure 1.3), a part of the signal recognition particle[217]. Three major subfamilies of *Alu* are known: *Alu*J, *Alu*S and *Alu*Y, with Y being the evolutionary oldest and J being the youngest[216].

SVA

SINE-VNTR-*Alu* (SVA) elements are hominid specific and the youngest of the non-LTRs, tracing back approximately 18-25 million years ago (mya)[218]. SVAs are a composite structure, approximately 0.7-4 kb in length, consisting of a 5' CT-

rich domain (CT element), an anti-sense Alu-like region, a central GC-rich VNTR, SINE-R domain and 3' poly A termination signal (poly A tail) (Figure 1.3). These elements are of retroviral origin, with their SINE-R domain sharing homology with the 3' *env* gene of HERV-K10[219]. There are approximately 2700-3000 copies present in the human genome, which are split into subclasses A-F on the basis of evolutionary age and the composition of the SINE-R domain. SVA-A is the oldest subclass in evolution (dating back approximately 13.6 million years), with the SVA-F subclass being the youngest (approximately 3.2 million years)[220]. An SVA-F1 subclass has also been identified, which resulted from an alternative splicing event at the *MAST2* locus in which exon 1 of the *MAST2* gene spliced into an intronic SVA element which then mobilised[221]: evolutionary age of this subclass has not been predicted (at the time of writing) (Table 1.3).

Gianfrancesco *et al*. calculated genome-wide composition of SVA elements in the human reference genome and discovered that 34.44% of all SVAs in humans were of the older subclasses (A-C), with 65.54% being part of the younger subclasses (D-F1). It was also found that the SVA-Ds were the largest subclass, constituting 44.39% of the total SVA content in the human genome[222]. Currently, little is known about how SVA elements are transcribed or if they contain an internal promoter, but using 5' RACE it has been discovered that SVA elements contain many transcriptional start sites (TSS): both upstream and internally[221]. SVA elements have a high GC-content (usually a minimum of 60%) and can therefore be considered mobile CpG islands and they have been shown to preferentially insert into genic regions of the human genome[220,223].

Additionally, SVA elements, have been shown to have the ability to function as transcriptional regulatory domains and thus modulate gene expression profiles[223,224]. Due to their composite and high GC repetitive structure they are very difficult to characterise and we hypothesise they could harbour disease specific genetic variation and thus be a source of missing heritability in complex disease.

Table 1.3. SVA subclasses

Subclasses, evolutionary age (million years ago), conservation and genome-wide composition (%) of each subclass shown. Statistics taken from Wang *et al*., 2005[225] and Gianfrancesco *et al*., 2019[222].

SVA subclass	Evolutionary age (mya)	Conservation	Human genome composition (%)
A	~13.56	Multiple primates (gorillas, bonobos, chimpanzees and humans)	7.82
B	~11.56	Multiple primates (gorillas, bonobos, chimpanzees and humans)	16.45
C	~10.88	Multiple primates (gorillas, bonobos, chimpanzees and humans)	10.17
D	~9.55	Multiple primates (gorillas, bonobos, chimpanzees and humans). Some are human specific	44.39
E	~3.46	Human specific	4.94
F	~3.18	Human specific	13.43
F1	N/A	Human specific	2.80

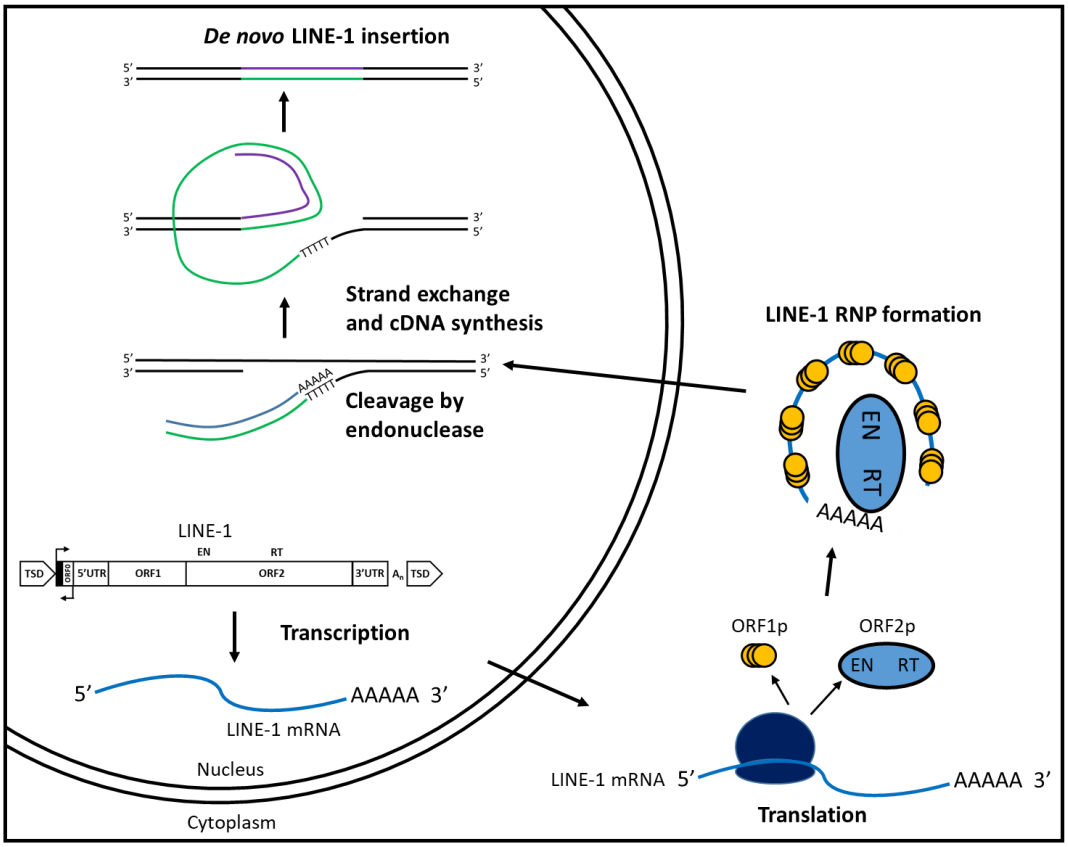


Figure 1.4. Retrotransposon mobilisation

Schematic of retrotransposon mobilisation via a process known as target primed reverse transcription (TPRT). Initially, LINE-1 is transcribed by RNA Polymerase II in the nucleus and the resulting LINE-1 mRNA is then exported to the cytoplasm and translation of open reading frame (ORF) proteins takes place. ORF1p (yellow trimer) and ORF2p (blue oval) can then bind their LINE-1 mRNA, inducing formation of a LINE-1 ribonucleoprotein (RNP) complex. The RNP particle then moves to the nucleus, where TPRT occurs and a *de novo* LINE-1 insertion is integrated into the genome. The ORF1p encodes endonuclease (EN), facilitating a single stranded nick at a LINE-1 consensus site (5'-TTTT/A-3'), then allowing ORF2p encoded reverse transcriptase (RT) to begin first strand LINE-1 cDNA synthesis at an exposed a 3' hydroxyl residue. Second strand cleavage, second strand LINE-1 cDNA synthesis and DNA repair then occurs, but these processes remain poorly understood[191,226]. LINE-1 ORF1p and ORF2p can also bind both *Alu* and SVA mRNA, facilitating trans-mobilisation of these elements via TPRT[191]. Adapted from Martin, 2010[227], Bock and Schumann, 2016[213] and Savage *et al*., 2019[191].

Retrotransposons – driving genome diversity

Non-LTR retrotransposon insertions are a large source of structural variation within the human genome: they can mobilise within the germline and thus such insertions are passed on to the next generation, but *de novo* somatic insertions have also been discovered[228-230]. Non-LTR retrotransposons exhibit two forms of polymorphism: variation within the components that make up these elements and as present or absent with respect to the reference human genome[231], the latter of which are referred to as retrotransposon insertion polymorphisms (RIPs). It is now known that non-LTR retrotransposons are mobilised in the brain and central nervous system (CNS)[232], with this retrotransposition driving somatic mosaicism and neuronal diversity and plasticity[233-235]. Mobilisation of L1 occurs early during embryo development, CNS formation and during adult neurogenesis[234]; it is now known that retrotransposition can occur in non-dividing cells and mature neurons[236,237]. Detection of retrotransposon insertions has been made possible due to two strategies: reporter assays and next generation sequencing. The LINE-1 reporter assay was first adapted in 1996 thus helping to engineer transposition *in vitro*[214,238], but since the emergence of next generation sequencing it has been possible to detect retrotransposon insertion polymorphisms (RIPs) on a global scale, utilising technology such as high-throughput retrotransposon capture sequencing (RC-Seq)[234,239]. Numerous computational tools have now been developed to detect and annotate RIPs within next-generation and whole genome sequencing (WGS) data, which have been reviewed elsewhere[240,241]. Thanks to these advancements, it has been possible to perform genome comparisons to

estimate retrotransposition rate within humans. Previously, it has been postulated that the rate of LINE-1 insertions was between 1/100 and 1/200 births[242], with the rate of Alu and SVA insertions being approximately 1/20 and 1/900 births, respectively[243]. However, more recent work utilising WGS data from three-generation human pedigrees has estimated retrotransposition rates of 1/63 for LINE-1, 1/40 for Alu and 1/63 for SVA elements[244].

While mobilisation is a well-documented phenomenon, retrotransposition is restricted and regulated by numerous host factors[197,245]. It has been previously documented that DNA methylation is one mechanism in which retrotransposons are restricted[246,247], however LINE-1 elements have been shown to become unmethylated in some cancers[248-250]. Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) have been shown to bind and repress retrotransposons, through the recruitment of KAP1 and subsequent assembly of a repressor complex which induces epigenetic alterations (DNA and histone methylation), leading to formation of heterochromatin[251,252]. Another line of defence includes the apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (ABOBEC) proteins which has been shown to restrict LINE-1 retrotransposition[253,254]. Repression of retrotransposons is also possible through the PIWI/piRNA pathway, inducing transcriptional silencing by facilitating epigenetic changes such as repressive chromatin modifications (histone methylation) and also post-transcriptional silencing via PIWI protein-mediated slicing of TE transcripts[255-257]. This dynamic has now been termed an evolutionary arms race, a recurring state of antagonistic co-evolution between retrotransposons and host defence factors: the constant selection for retrotransposons to elude host defences so they can remain

active and mobilise, while host factors also co-evolve to restrict this replication of retrotransposons[258]. Antagonistic co-evolution of host and pathogen through constant cycles of adaptation and counter adaptation is a dynamic of the Red Queen Hypothesis proposed by Leigh Van Valen[259,260], inspired by Lewis Carroll's Alice Through The Looking Glass[258]: "It takes all the running you can do, to keep in the same place"[261].

It is now known that transposable elements are not randomly distributed across the genome[185], preferentially inserting into certain genomic locations due to factors such chromatin accessibility and target site sequences[262]. Natural selection also acts upon TE distribution, as insertions which pose detrimental effects on host fitness are removed from the population: a balancing act is struck between maintaining host fitness and sustaining the ability to propagate[185]. Although often termed "selfish" DNA, some TEs have undergone exaptation, co-opting a beneficial cellular function within the host genome[263,264]. While an arms race between TEs and host defence factors is supported, zinc finger proteins (ZNFs) have been shown to facilitate domestication of TEs in the human genome, contributing to species specific epigenetic and transcriptional regulation[265,266]. Furthermore, TEs are now known to exhibit regulatory properties at the transcriptional level, influencing and shaping the evolution of gene expression profiles in the human genome[267,268].

Retrotransposons – impacting and regulating the human genome

Transposable elements (TEs) have been shown to have several impacts on mammalian genomes. Retrotransposon insertion mutations account for approximately of 0.3% of all mutations within the human genome and can elicit a variety of regulatory effects[193](Figure 1.5).

TE-derived sequences contain transcription factor (TF) binding sites and thus can act as transcriptional regulatory domains which can modify gene expression profiles[269,270](Figure 1.5). Previous studies have shown that TEs, including HERVs, MER elements, LINE-1, *Alus* and SVAs can bind various TFs, including TP53, ESR1, STAT1, NANOG, OCT4 and CTCF[271-277]. TEs have been shown to introduce novel TF binding sites in a species-specific manner, contributing to mammalian evolution of gene regulation. Sundaram *et al*. used ChIP-seq to profile genome-wide TF binding sites in human (K562) and mouse (MEL) cell lines. A total of 26 orthologous pairs of TFs were assessed and 695,042 TF binding peaks were defined, which were then overlaid with annotated TE locations from RepeatMasker. In humans, a total of 135,442 (19%) TF binding sites were found within TEs: in mice, 140,058 (20%) TF binding sites were derived from TEs. Furthermore, they found that the majority of TE derived TF binding sites were species specific: 132,197 (98%) in humans and 138,649 (99%) in mice[278]. Similarly, Kellner and Makalowski collected data from the ENCODE (encyclopedia of DNA elements) project and assessed transcription factor binding sites found within proximal promoters of genes and discovered that 215,964 (6.8%) of 3,173,045 active TF binding sites in proximal promoters were found in TE-derived sequences,

highlighting a significant impact from TEs in shaping and rewiring gene expression networks[268].

TEs have also been shown to influence and shape gene expression profiles in a tissue specific manner. Trizzino *et al*. have assessed histone modification data from the Roadmap Epigenomics Project and gene expression data from 24 primary tissues and cell types within the GTEx Portal and found that particular subfamilies of TE were enriched in regions of active chromatin in a tissue dependent manner. The SINE subfamily was the most significantly enriched across all tissues, accounting for 43-66% of TEs in active chromatin, while LINE elements and long terminal repeat (LTR) retrotransposons were the most significantly depleted across all tissues in regions of active chromatin. Interestingly, SVA elements were found to be significantly enriched in regions of active chromatin in 13/25 tissues. They also found that SVA elements displayed tissue specific gene regulation profiles, binding master regulators in those respective tissues. In adipose tissue, SVAs found in active chromatin were enriched for SOX6 and ZEB1 TF binding sites, key regulators of adipogenesis and adipocyte differentiation, respectively. Furthermore, SVA associated genes in adipose tissue exhibited a significant increase in expression when compared to other tissues, indicating a potential role for SVAs in the activation of transcription in adipose tissue. However, SVAs found within active chromatin in the liver were enriched for STAT3 and CPEB1 TF binding sites, which regulate liver regeneration and insulin signalling, respectively. SVA associated genes in the liver exhibited a significant decrease in expression when compared to other tissues, suggesting a role for SVAs as transcriptional repressors

in the liver[267]. Ultimately, TEs can function as transcriptional regulatory domains, but act in a tissue specific fashion.

TEs insertions can drive deleterious effects on host genes[279]. Insertional mutagenesis created by TEs can lead to nonsense mutations or deletions within exons, generating premature transcription termination signals: insertions within introns can introduce novel alternative splice sites and premature polyadenylation signals[193]. TE insertions can impact gene function. Insertions within genes can lead to exonisation or production of novel gene transcripts, disrupting or altering gene structure and function (Figure 1.5). A study by Rodriguez-Martin *et al*. identified a familial retinoblastoma (RB) case with a *de novo* full length LINE-1 insertion within intron 14 of *RB1*. This insertion was found to disrupt the intron 14/exon 15 boundary, introducing three non-canonical splice acceptor sites, resulting in exonisation of the LINE-1 element and production of aberrant *RB1* mRNA transcripts[280]. Payer *et al*. have shown that *Alu* insertions can alter splicing efficiency, acting as splicing quantitative loci (sQTLs). They initially identified *Alu* insertions within 100 bp of alternatively spliced exons and then went on to test if these insertions affected splicing efficiency. A total of 23 different loci (with and without the Alu insertion) were cloned into an intron within a minigene reporter vector and then transcript expression was measured using RT-PCR. One example was an *Alu* element found 41 bp upstream of exon 33 of *NUP160*. They found that when the Alu insertion was present, skipping of exon 33 was significantly increased (42.5% as opposed to 20%). A similar effect was also found with Alu insertions in *BPFIC*, *SLC2A9* and *CD58*, with the insertions all leading to a significant increase in exon skipping. The alternative effect was also observed, with an Alu insertion in

CCDC110, leading to a significant increase in exon 5 inclusion (from 43.2% to 79%). Overall, this study showed that *Alu* elements can effect splicing in a locus dependent manner[281].

The polymorphic presence or absence of TEs within human populations also adds another layer of complexity to genomic regulation by altering gene expression dynamics and regulatory networks within host loci. Wang *et al*. assessed the relationship between retrotransposon genetic variation (presence or absence polymorphisms known as retrotransposon insertion polymorphisms, RIPs) and gene expression profiles using expression quantitative trait loci (eQTL) analysis. The loci evaluated were regions where retrotransposon insertions (RIPs) had occurred and these were correlated with eQTLs of genes at these loci. Overall they collected RNA-seq data from 445 healthy individuals from 5 human populations (4 European and 1 African) and observed a total of 10,106 retrotransposon insertions (RIPs). Genotyping data (LINE-1, Alu and SVA RIPs) of the 445 healthy individuals was taken from the phase 3 variant release of 1000 Genomes Project[229,282]: RNA-seq data for the same samples was part of the Geuvadis RNA sequencing project[283], both generated from Epstein-Barr virus transformed B-lymphocyte cells (LCL) from these individuals. From these datasets they determined statistically significant associations between transposable element (TE) genotypes and individual gene expression profiles at given loci: referred to as polyTE-gene expression associations (TE-eQTLs) within this study.

They found that the *Alu*-5788 locus was associated with *REL* expression levels, with the insertion correlating with increased gene expression. The *Alu*-

108441 locus was found to be associated with *PSD4* expression levels, with the insertion correlating with reduced gene expression. It is interesting to note that these TE-eQTLs were population specific, with the first being found primarily in the African population and the latter being more prevalent in European populations. The Alu-1870 locus contained a common insertion in both populations, but only associated with a decrease in *PRDM2* gene expression in the African population; which the authors attribute to potential interaction with other population specific variants. In contrast to this, the Alu-8559 locus shared an insertion in both populations and the associated decrease in *HSD17B12* gene expression was observed in each population. Interestingly, the *Alu*-7481 locus had an insertion which led to an association with expression levels of multiple genes. They detected an association between the Alu insertion and increased expression of *PAX5* and three of its target genes *PIK3AP1*, *REL* and *ZSCAN23*, indicating that the insertion altered expression of the transcription factor encoded by *PAX5* which in turn affected regulation of its downstream targets[284]. This study highlights the importance of TE transposition altering levels of transcription of individual genes but also driving gene regulatory networks.

Spirito *et al*. have also assessed LINE-1, *Alu* and SVA RIPs within the same 445 healthy individuals, utilised previously by Wang *et al*.[284]. In this study they were also able to identify *cis*-expression quantitative trait loci (*cis*-eQTLs): constituting an association between a RIP and a gene (within 1Mbp), where there was a significant difference in expression of that gene in individuals with the RIP compared to those without the RIP (referred to as a TE structural variant-*cis*-eQTL in this study). In total, 8551 retrotransposon (presence or absence) polymorphisms

were assessed (7208 *Alus*, 937 LINE-1s and 406 SVAs) and a total of 323 significant TE structural variant-*cis*-eQTL associations were discovered. Interestingly, Spirito *et al*. found that SVAs constituted a statistically significant higher relative proportion of TE structural variant-*cis*-eQTLs (when compared against *Alus* and LINE-1s), indicating that SVAs were generating the greatest effect on gene expression: this significantly higher proportion of associations from SVAs was also observed when they examined the cis-eQTLs from Wang *et al*.[284]. Furthermore, they also found that SVAs were significantly enriched within regulatory regions, whereas *Alus* and LINE-1s were depleted in these sequences, also inferring that SVAs were having the strongest impact on gene regulation[285]. A more recent study from Goubert *et al*., involved TE-eQTL analysis on lymphoblastoid cell lines (LCL) (n = 444) and iPSCs (n = 289), identifying a total of 211 and 176 *cis*-TE-eQTLs, respectively. They found an enrichment of TE-eQTLs within or close to genes, with 92.4% (in LCL) and 83% (in iPSC) of all implicated TEs within 250 kb of the target gene; they were found to be enriched within introns and regions which were within 10 kb upstream or downstream of the target genes. Furthermore, 18% (n =57) of all TE-eQTLs were found in both cell types, indicating that most TE-eQTLs were cell type-specific[286].

The human genome is a complex 3D entity built of multiple layers of organised structures: chromatin folds and loops into regions known as topological associated domains (TADs), sub-megabase regions which preferentially form intradomain interactions, forming loops between regulatory regions (enhancers and promoters)[287,288]; on the megabase-scale chromatin is demarcated into long-range compartments of open (active) and closed (repressive) chromatin (known as

A/B compartments), which in turn form larger chromosome territories within the nucleus[287,289]. The looping of chromatin therefore leads to the formation of organised neighbourhoods; long range interactions across chromosomes which in turn help regulate key processes such as DNA replication and transcription[287,288,290]. TEs (including LINEs, SINEs, and ERVs) can modulate this chromatin architecture by recruiting CTCF (CCCTC-binding factor), an 11 zinc finger protein which has a well-established role in chromatin looping and TAD formation[288,291]. Diehl *et al*. used ChIP-seq to assess CTCF binding sites and Hi-C data to identify chromatin loop interactions with TEs in the human and mouse genome, with >85% of all CTCF sites being species specific. Moreover, in the mouse >47% of CTCF sites were derived from TEs: in humans >30% of CTCF sites were derived from TEs (LINEs, SINEs and ERVs). This study highlights that TEs have contributed to species-specific chromosomal organisation and form functional chromatin loops[292]. Ferrari *et al*. has shown that *Alu* elements can recruit RNA polymerase III transcription factor C (TFIIIC), which facilitates histone acetylation and subsequent binding of CTCF to induce chromatin looping and 3D folding[293].

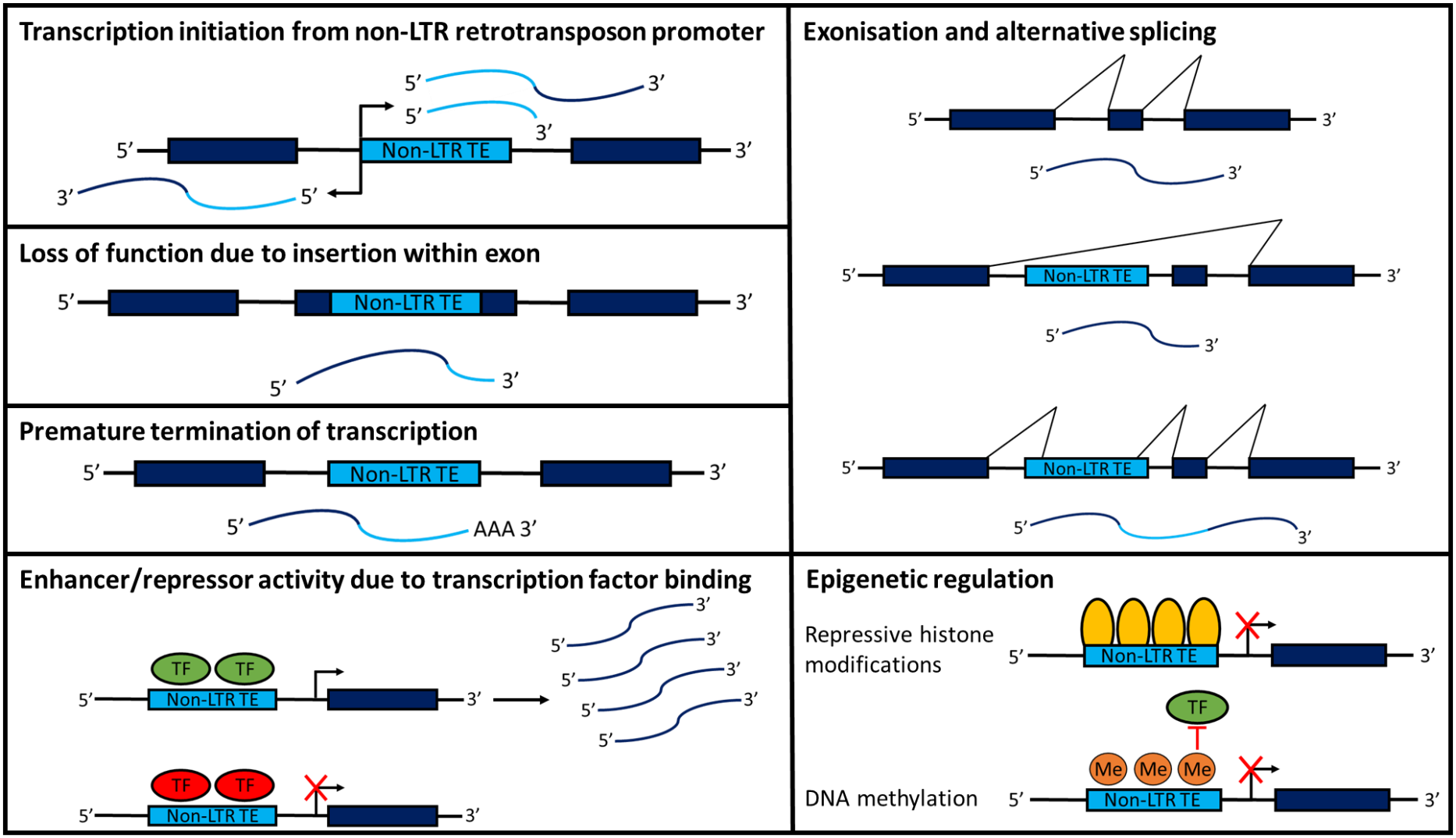

Transcription initiation from non-LTR retrotransposon promoter
5′
3′
5′
3′
5′
Non-LTR TE
3′
3′
5′
Loss of function due to insertion within exon
5′
Non-LTR TE
3′
5′
3′
Premature termination of transcription
5′
Non-LTR TE
3′
5′
AAA 3′
Exonisation and alternative splicing
5′
3′
5′
3′
5′
Non-LTR TE
3′
5′
3′
5′
Non-LTR TE
3′
5′
3′
Enhancer/repressor activity due to transcription factor binding
TF
TF
5′
Non-LTR TE
3′
3′
5′
3′
5′
3′
5′
3′
TF
TF
5′
Non-LTR TE
3′
Epigenetic regulation
Repressive histone modifications
5′
Non-LTR TE
3′
TF
Me
Me
Me
DNA methylation
5′
Non-LTR TE
3′

Figure 1.5. How non-LTR retrotransposons regulate gene expression

Mechanisms of how non-LTR retrotransposons regulate gene expression profiles in the human genome. Non-LTR retrotransposons can exhibit promoter activity, inducing novel transcriptional effects on host gene expression[263,279]. Transcription can be initiated by both sense and anti-sense promoters found within TEs, leading to altered host gene activity[191,263]. Non-LTR retrotransposons contain transcription factors binding sites which can elicit positive (enhancive) or negative (repressive) effects on gene expression profiles in a tissue specific manner[267]. Retrotransposon insertions within introns can alter splicing efficiency and induce exonisation and alternative splicing of host genes[280,281]. Retrotransposons may also influence epigenetic parameters of host loci, including changes in DNA methylation and histone modification. Non-LTR = non long terminal repeat, TE = transposable element, TF = transcription factor. Adapted from Savage *et al*. 2019[191].

Retrotransposons and disease

Non-LTR retrotransposons can have a profound effect on the stability and evolution of the genome, but also drive disease states. Retrotransposon insertions were first discovered as drivers of disease approximately 30 years ago, with the detection of a LINE-1 insertion causing haemophilia A[294]. There are now a plethora of studies which have implicated retrotransposons in the involvement of multiple diseases[209,295].

To date there are 124 recorded L1-mediated disease causing insertions, 13 of which are SVA insertions (Table 1.4)[295]. An example of this is an SVA-F insertion found in intron 32 of the *TAF1* gene which causes X-linked dystonia Parkinsonism (XDP)[296]. Furthermore, CT-element repeat length of this SVA element has been found to inversely correlate with both age of onset and *TAF1* gene expression[297,298]. An SVA insertion has also been found to be associated with increased risk for cutaneous basal-cell carcinoma (BCC) and breast cancer but is protective against prostate cancer. Stacey *et al*. discovered that rs70036, a germline SNP associated with increased risk for BCC and breast cancer, is inherited and associated with an SVA insertion within intron 8 of *CASP8*. Furthermore, carriers of this risk SNP and SVA insertion were found to have preferential intron 8 retention, with an increase in intron 8 retained transcripts being observed when the SVA was present, thus indicating a potential role for this insertion to alter splicing of *CASP8*[299].

Table 1.4. SVA insertions associated with human disease

(Adapted from Hancks and Kazazian, 2016[295])

Gene	Chromosome	Disease	Subfamily
BTK	X	X-linked agammaglobulinemia	N/A
TAF1	X	X-linked dystonia parkinsonism	F
FIX	X	Hemophilia B	F
LDRAP1	1	Autosomal recessive hypercholesterolemia	E
SPTA1	1	Hereditary elliptocytosis and hereditary pyropoikilocytosis	E
CASP8	2	Breast cancer susceptibility	E
A4GNT	3	Chromothripsis	E
HLA-A	6	Leukemia	F1
PMS2	7	Lynch Syndrome	F
FKTN	9	Fukuyama-type congenital muscular dystrophy	E
PNPLA2	11	Neutral lipid storage disease with subclinical myopathy	E
SUZ1P	17	Neurofibromatosis Type I	F1
SUZ1P	17	Neurofibromatosis Type I	F

Although no specific SVA element insertions have been found to be associated with ALS, several types of retrotransposon have been implicated in ALS and are now hypothesised to contribute to disease progression. Douville *et al*. measured HERV-K *pol* transcript levels using quantitative real-time PCR and found that there was a statistically significant increase in HERV-K pol transcripts in the brains of ALS patients (n=28) compared to non-ALS controls (age-matched patients with chronic disease, n=12). This result highlighted an ALS specific pattern of HERV-K pol expression as HERV-K *pol* transcripts were not detected in both patients who suffered accidental death who had no pre-existing conditions and Parkinson's disease. Immunostaining analysis also showed that there was a significant increase in HERV-K reverse transcriptase (RT) protein levels in ALS cortical brain tissue

compared to cortical tissue derived from patients with other systemic disease. Overall, HERV-K RT was detected in 10 of 13 ALS patients compared to only 3 of 10 patients with systemic disease, indicating a significantly higher frequency of RT detection in ALS. As TDP-43 overexpression is a well characterised hallmark of ALS and FTD, Douville *et al*. measured TDP-43 transcript levels in cortical brain tissue of ALS patients and found a significant increase in TDP-43 mRNA compared to non-ALS controls, furthermore there was a positive correlation between HERV-K *pol* (RT RNA) and TDP-43 transcript levels and HERV-K RT was found to be colocalised with TDP-43 protein in motor neurons of these ALS patients[206]. This study identified the potential for HERV-K to serve as a novel biomarker for ALS.

A more recent investigation by Li *et al*. discovered that HERV-K *env*, *gag* and *pol* genes are all expressed in post mortem ALS brain tissue, with a statistically significant increase in expression of all three genes in ALS patients (n=11) compared to healthy controls (n=16). Interestingly, this study also showed that expression of HERV-K can induce toxicity in human neurons. Both the HERV-K genome and *env* gene were transfected separately into induced pluripotent stem cell (iPSC) derived human neurons, which led a significant decrease in total cell count and neurite length (measured 24 hours post transfection). Endogenous HERV-K was also expressed through targeted transcriptional activation using the CRISPR Cas9 system, also leading to a statistically significant decrease in total cell number and neurite length. Additionally, this study also proved that HERV-K is regulated by TDP-43, because transfection of TDP-43 into human stem cell derived neurons induced HERV-K expression, leading to a fold increase of *env* and *pol* transcripts. They also generated a HERV-K reporter construct (HERV-K LTR-

MetLuc), which once co-transfected with TDP-43 led to a significant increase in luciferase activity compared to co-transfection with chloramphenicol acetyltransferase (CAT) (control). complementary to this result, knockdown of endogenous TDP-43 using small interfering RNA (siRNA) led to a decrease in HERV-K expression[207]. Ultimately, this study strengthened the previous proposal citing HERV-K as a potential biomarker for ALS and provided evidence of HERV-K mediated neurotoxicity and regulation by TDP-43.

Li *et al*. have shown that TDP-43 can bind to transposable elements (TEs) and thus regulate mobile element expression. By analysing crosslinking-immunoprecipitation sequencing (CLIPseq) datasets they found that TPD-43 targets a number of TEs, including LINEs, SINEs and LTR transposons. Interestingly, they also discovered that TDP-43 binding to TEs was depleted in FTD patients but not in controls, suggesting that TDP-43 pathology may lead to a deregulation of TEs. This hypothesis was supported by analysis of repetitive element sequencing reads from mRNA-seq data of two mouse models with TDP-43 pathology (overexpression of human TDP-43 and the depletion of striatal TDP-43 respectively. Overexpression of human TDP-43 led to an increase in expression of 86 repetitive elements and TDP-43 depletion (loss of function) induced an increase in expression of 223 repetitive elements (including LINE, SINE, LTR and DNA TEs). Overall, this highlights that dysregulation of TDP-43 results in overexpression of TEs, indicating that this RNA binding protein could be an important regulator of transposon expression and that accumulation of TE transcripts could contribute to neurodegenerative disorders which are mediated by TDP-43 dysfunction[300].

A study by Krug *et al*. further supports a link between TDP-43 dysfunction and TE expression. They found that human TDP-43 (hTDP-43) transgene expression in neurons and glial cells of *Drosophila* led to an increase in expression of several subtypes of retrotransposon, including LTR and LINE elements. Some of these responses were found to be cell specific, with a significant increase in *Gypsy* (ERV) expression observed in glial cells only, confirmed through both RNA-seq and RT-PCR. They also found that hTDP-43 expression in the neurons and glial cells of *Drosophila* led to significant impairment of locomotor function and reduced life span, both of which were more severe in glial cells. To test if the increase in *Gypsy* expression was contributing to these degenerative phenotypes, RNAi of *Gypsy* ORF2 was performed to induce a 50% reduction in *Gypsy* expression. The co-expression of the *Gypsy* RNAi construct with hTDP-43 in glial cells led to a substantial improvement in survival rate compared to co-expression of a control RNAi GFP construct with hTDP-43; indicating that increased *Gypsy* expression contributed to the TDP-43-mediated neurological phenotypes observed in this study[301].

Transcript levels of multiple repetitive elements have been found to be altered in the brains of ALS patients. Prudencio *et al*. observed a significant increase in transcript levels of several repetitive elements in the frontal cortex of *C9orf72*-positive ALS patients. Quantitative RT-PCR analysis confirmed that transcript levels a number of LTRs (LTR2, LTR70, MER21B, MER51C), SINEs (*AluYk12*, *AluYa5*, FRAM) and a LINE-1 element (L1MA9) were all significantly increased in *C9orf72*-positive ALS patients (n=56) compared to controls (n=9): modest increases in transcript levels were observed (for all repetitive elements

mentioned above) in *C9orf72*-negative ALS patients (n=46) compared to controls, but statistical significance was not reached[302].

Differential expression of several retrotransposons has recently been observed in a subset of ALS patients. Using frontal cortex transcriptomes from 77 ALS patients and 18 non-neurological controls from the New York Genome Center (NYGC) ALS Consortium, Tam *et al*., were able to stratify patients into molecular subtypes based on gene expression signatures: 148 ALS and 28 control transcriptomes were used (n = 176), due to multiple regions of the frontal cortex (such as motor cortex) being available for some samples. Overall, a total of three distinct subgroups within these patients were identified. The first group (ALS-Ox) (61%; 91/148) showed elevated expression of stress response genes associated with oxidative and proteotoxic stress, including *SOD1*. The second group (ALS-Glia) (19%; 28/148) had increased expression of markers for glial cell types including astrocytes and oligodendrocytes and microglia. In the final subgroup (ALS-TE) (20%; 29/148) it was found that transposon expression was the most significant hit from Gene Set Enrichment Analysis (GSEA) pathway analysis when compared to control samples, including hits for LINE, SINE and LTR TEs. Furthermore, when compared to ALS-Ox and AL-Glia, the ALS-TE subset of patients displayed increased expression levels of several retrotransposons, including the human specific LINE-1 subfamily (L1HS), L1PA6 subfamily and SVA (SVA-A) elements; a reduction in *TARDBP* expression was also observed in this subgroup. Further validation was performed using fresh frozen motor cortex samples of 13 ALS patients and 6 non-neurological controls from University of California San Diego (UCSD). Seven of these ALS patients were part of the ALS-TE subgroup and increased expression

levels of three active retrotransposons (L1HS, *AluYk12* and *AluYa5*) was validated via qPCR. Using immunostaining it was also found that the UCSD ALS samples within the ALS-TE subgroup displayed staining of phosphorylated TDP-43 (indicative of TDP-43 pathology[16]), which was not found in ALS-Ox, ALS-Glia or control samples, supporting the hypothesis of TE re-activation in response to TDP-43 dysfunction. Tam *et al*, also performed enhanced crosslinking and immunoprecipitation sequencing (eCLIP-seq) on SH-SY5Y cells; sequencing RNAs which were bound to TDP-43, resulting in a total of 36,716 peaks, with 31% mapping to TEs. TDP-43 CLIP reads were found over a number of retrotransposon subclasses, including LINE-1 (L1PA6), SINEs (*AluY*), SVAs (SVA-D) and HERVs (HERV3). Subsequent knockdown of TDP-43 (using short hairpin RNA) also led to a significant upregulation of retrotransposons transcripts (including L1PA6, AluY and SVA-D elements), further supporting that TDP-43 functions by silencing retrotransposons[303].

Although most of the current literature surrounding the involvement of retrotransposons in ALS has focussed on HERV-K and LINE-1, the 2019 study by Tam *et al*. highlighted that SVA-D transcript levels were increased in a subset of ALS patients and bound TDP-43, highlighting potential for these elements to contribute to the disease. Chapters 4 and 5 of this thesis will focus on characterising genetic variation and functional capacity of an SVA element located within an ALS risk locus (*NEK1*). The ultimate goal of this PhD project was to characterise both VNTR and SVA genetic variation and function in ALS risk loci and to strengthen the argument of their contribution to ALS, with the potential for such regions to be missing sources of heritability.

Chapter 2: Materials and Methods

2.1 Materials

2.1.1 Commonly used buffers and reagents

2.1.1.1 TBE buffer

TBE Buffer (5x) was made using 108 g tris base (Sigma), 55 g boric acid (Sigma) 5.84 g EDTA (Sigma) and made up to 2 L with distilled water. TBE was then diluted to a working concentration (0.5x) with distilled water. TBE was used to make agarose gels which were then used for gel agarose electrophoresis.

2.1.1.2 LB Broth

LB Broth (Fluka Analytical) was made by adding 10 g (25 g/L) into 400 ml distilled water (Sigma), autoclaved and then stored at room temperature. LB broth was used as culture medium for bacteria.

2.1.2.2 LB Agar

LB Agar (Fluka Analytical) was made by adding 16 g (40 g/L) into 400 ml distilled water (Sigma), autoclaved and then stored at room temperature. All bacteria were plated on agar plates containing antibiotic (kanamycin for pCR®-Blunt intermediate vectors and ampicillin for pGL3 and pSHM06 vectors). Kanamycin antibiotic stock solutions were prepared by dissolving kanamycin sulphate salt (Sigma) in nuclease free water to a final concentration of 50 mg/ml, filter sterilised and stored at -20 °C. To generate LB agar plates with kanamycin, a total of 400 ml of cooled liquid LB agar was then mixed with 400 µl of kanamycin solution (to generate a final concentration of 50 µg/ml). Ampicillin stock solutions were generated by dissolving ampicillin sodium salt (Sigma) in nuclease free water to a final concentration 100 mg/ml, filter sterilised and stored at -20 °C. To generate LB

agar plates with ampicillin, a total of 400 ml of cooled liquid LB agar was then mixed with 400 µl of ampicillin solution (to generate a final concentration of 100 µg/ml).

2.1.2 Human DNA samples

2.1.2.1 MNDA UK

Genomic DNA purified from blood for MND cases and controls was obtained from the UK MND Collections DNA and Cell Bank (https://www.mndassociation.org/research/for-researchers/resources-for-researchers/ukmndcollections/dna-bank/). A total of 500 MND patients were provided: 456 were diagnosed with ALS, 21 were diagnosed with PMA, 12 were diagnosed with PBP and 11 were diagnosed with PLS. A total of 333 MND cases were male and 167 were female, with an age range of 24-91 years old and a disease age of onset range of 23-88 years old. A panel of 499 controls was also obtained: 188 were male and 311 were female with an age range of 27-84 years old (Appendix 1).

2.1.2.2 UK and Dutch samples from Project MinE

WGS data was obtained from the UK dataset of Project MinE, encompassing a total of 1284 SALS cases and 500 controls. DNA from three Dutch control samples (ALS24457, ALS26061 and ALS26656) were also obtained for genotyping analysis of the *NEK1* SVA.

2.1.3 Plasmids

2.1.3.1 pCR®-Blunt vector

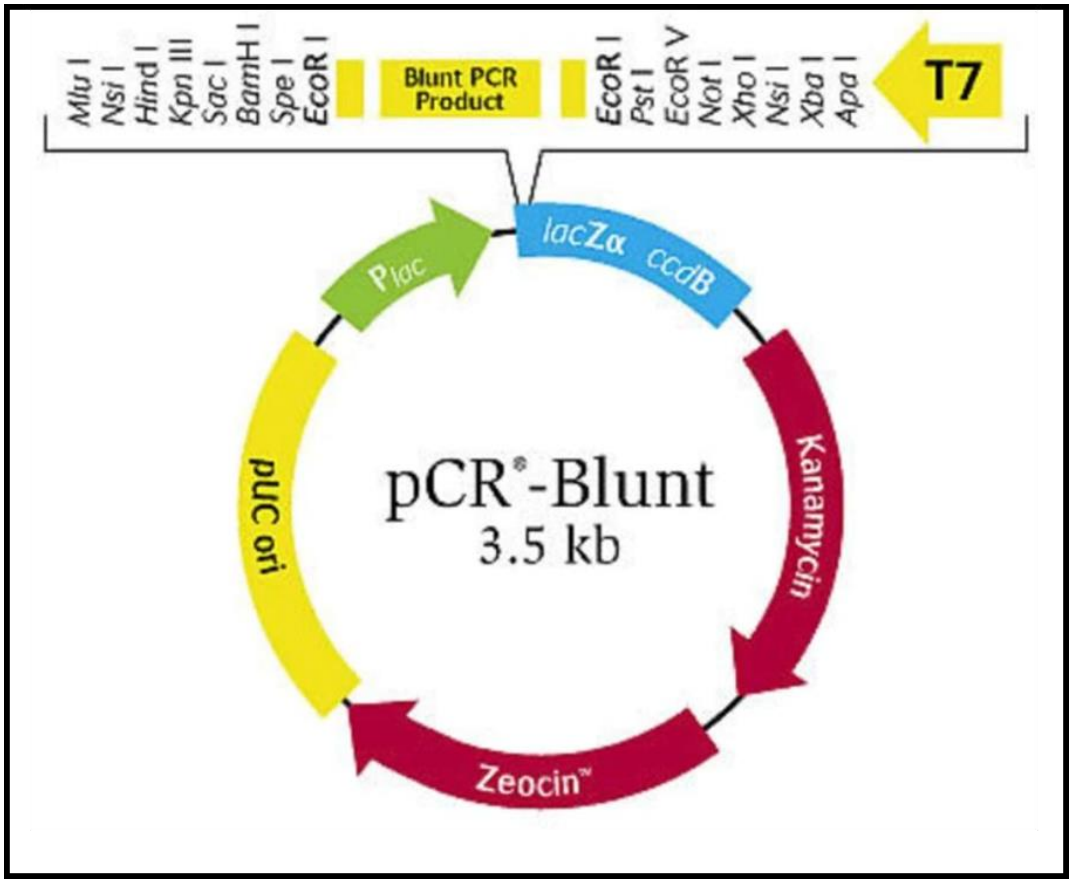


Figure 2.1. pCR®-Blunt vector map

The pCR®-Blunt intermediate vector (ThermoFisher) contains a *Lac*Zα-*ccd*B fusion gene which is used for positive selection of recombinant vector (function of the lethal *ccd*B gene is disrupted and inactivated by a DNA insert). A *lac* promoter is present to facilitate expression of the *Lac*Zα-*ccd*B fusion gene. The T7 promoter (yellow arrow) facilitates transcription and translation both *in vitro* and *in vivo*. The pUC origin of replication is also within this vector, allowing for high copy replication in *E.coli*. The kanamycin resistance gene is present, allowing for selection in *E.coli*. Multiple cloning site and restriction enzyme cut-sites are shown and sites for the M13 forward and M13 reverse primers are also present to allow for sequencing of the vector.

2.1.3.2 pGL3 vectors

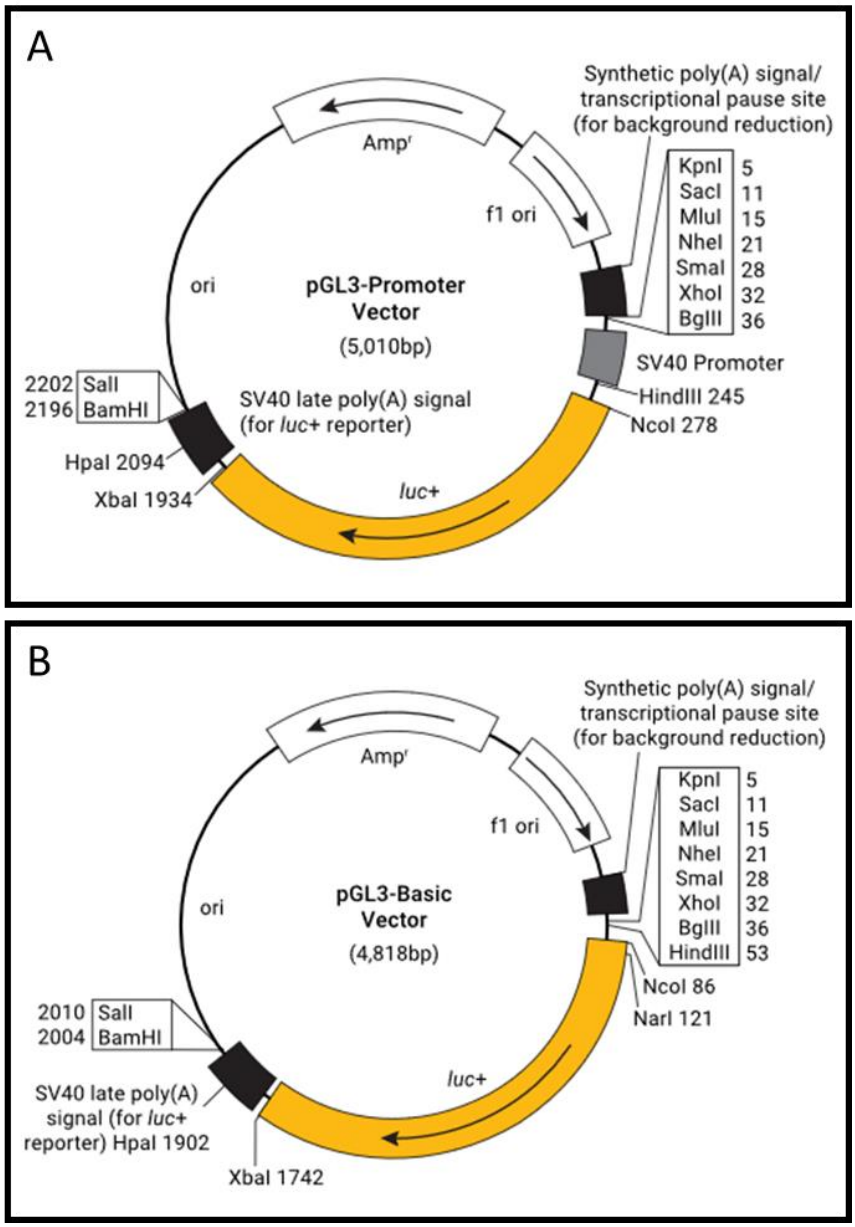


Figure 2.2. pGL3 vectors

A: pGL3-Promoter (pGL3-P) contains an SV40 promoter (grey box) which induces transcription and translation both *in vitro* and *in vivo*. The Firefly (*Photinus pyralis*) luciferase gene (yellow) is found downstream of the SV40 promoter and is used as a reporter gene for luciferase assays. **B**: pGL3-Basic (pGL3-B) has an identical backbone to pGL3-P, also contains the Firefly (*Photinus pyralis*) luciferase gene (yellow) but lacks the SV40 promoter. Multiple cloning site is present in both vectors and restriction enzyme cut-sites are shown. An origin of replication is found in both vectors, allowing for high copy replication in *E.coli*. The ampicillin resistance

gene is also present in both vectors, allowing for selection in *E.coli*. Both vectors are commercially available from Promega.

2.1.3.3 pSHM06 vector

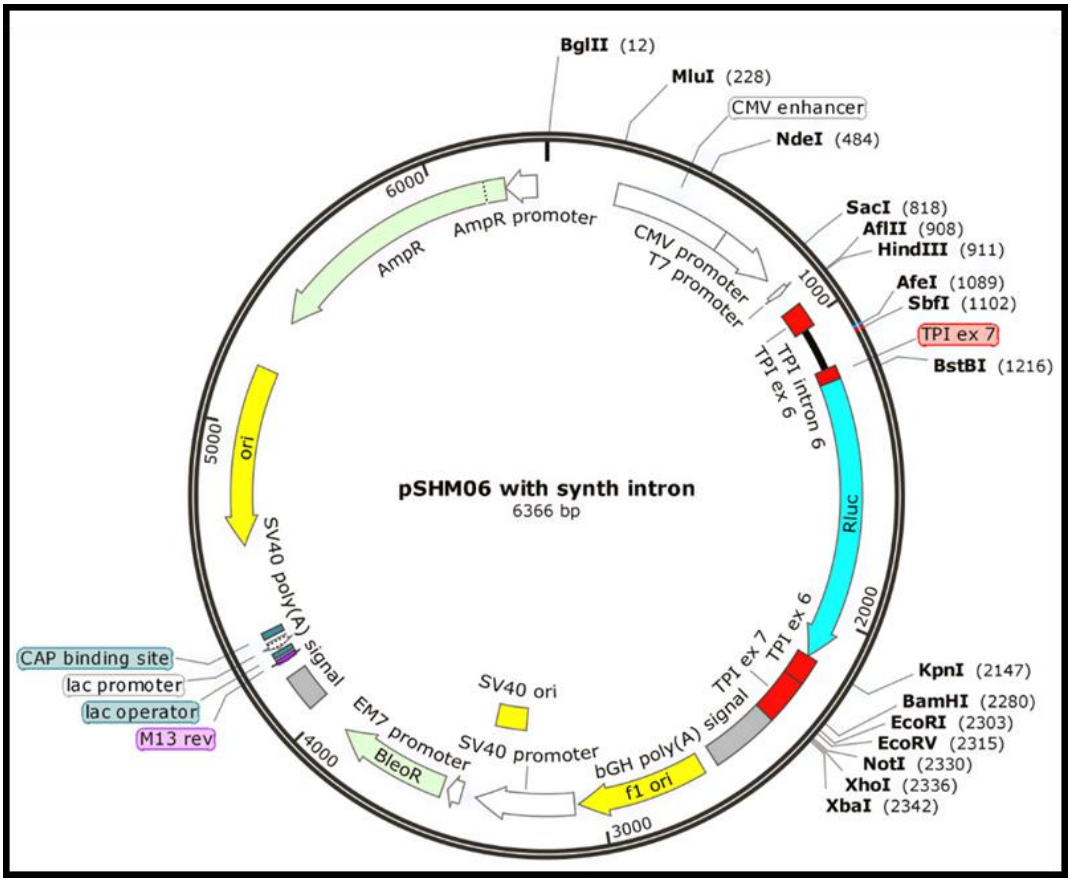


Figure 2.3. pSHM06 vector

The pSHM06 vector contains a high expression CMV promotor (white arrow) to facilitate transcription and translation *in vitro* and *in vivo*. The Renilla (*Renilla reniformis*) luciferase gene (light blue) is found downstream of the CMV promoter and is used as a reporter gene for luciferase assays. reporter gene. Exon 6 and 7 of the *triosephosphate isomerase* (*TPI*) gene (red boxes) are present at both the 5' and 3' end of the Renilla luciferase gene. Intron 6 of *TPI* is also present between exons 6 and 7 at the 5' end (shown as a black line), which indicates the site of insertion used for this plasmid. This vector also contains an origin of replication to facilitate high copy replication and the ampicillin resistance gene to allow for selection in *E.coli.* The pSHM06 vector is not commercially available and was generated by Nott *et al.*[304].

2.1.3.4 pSpCas9(BB)-2A-GFP vector

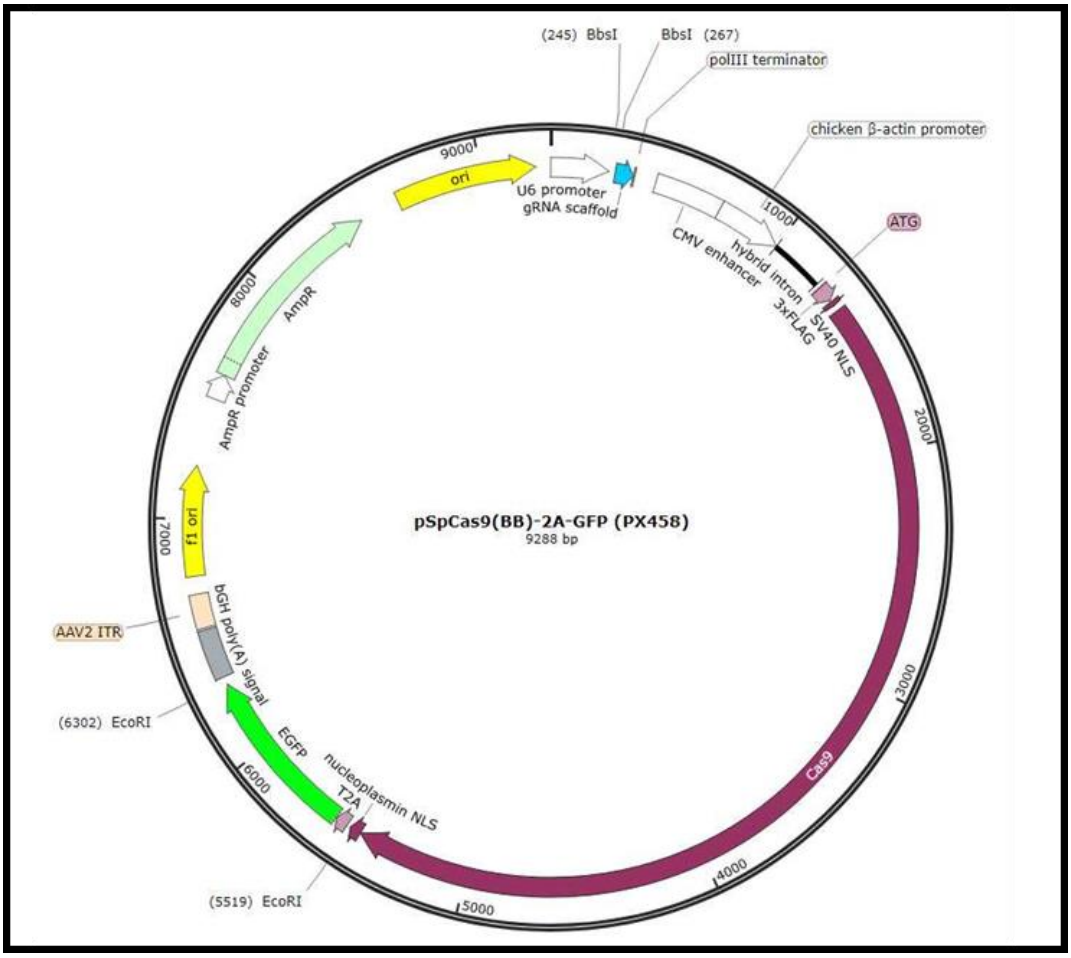


Figure 2.4. pSpCas9(BB)-2A-GFP vector

pSpCas9(BB)-2A-GFP is a Cas9 expression plasmid, containing a U6 promoter, guide RNA scaffold (light blue) and Cas9 coding sequence (purple). Between the *BbsI* cut-sites marks the site of integration for the guide RNA sequence. This vector also contains an origin of replication to facilitate high copy replication and an ampicillin resistance gene to allow for selection in *E.coli*. pSpCas9(BB)-2A-GFP (PX458) was a gift from Feng Zhang (Addgene plasmid # 48138 ; http://n2t.net/addgene:48138 ; RRID:Addgene_48138)[305].

2.1.4 Human cell lines and culture media

2.1.4.1 SH-SY5Y

SH-SY5Y (ATCC® CRL-2266™) is a neuroblastoma cell line derived from SK-N-SH, which was established from a metastatic bone tumour. SH-SY5Y were cultured in a 50:50 mix of nutrient mixture F-12 Ham (Sigma) and minimal essential medium eagle (Sigma), supplemented with 10 % foetal bovine serum (Gibco), 1 % penicillin/streptomycin (100 U/ml, 100 mg/ml; Sigma), 1 % (*v/v*) 200 mM L-glutamine (Sigma), and 1 % (*v/v*) 100 mM sodium pyruvate (Sigma). Cells were incubated at 37 °C in 5 % CO_2.

2.1.4.2 HEK293

HEK293 (ATCC® CRL-1573™) is a hypotriploid cell line derived from human embryonic kidney cells that were transformed with sheared human adenovirus type 5 DNA. HEK293 were cultured in Dulbecco's modified eagle media (DMEM) containing 4.5 g/L D-Glucose and 200 mM L-glutamine (Gibco), supplemented with 10 % foetal bovine serum (Gibco), 1 % penicillin/streptomycin (100 U/ml, 100 mg/ml; Sigma), and 1 % (*v/v*) 100 mM sodium pyruvate (Sigma). Cells were incubated at 37 °C in 5 % CO_2.

2.1.3.3 SKNAS

SKNAS (ATCC® CRL-2137™) is a human cell line derived from bone marrow metastasis from a child with embryonal neuroblastoma. These cells were cultured in Dulbecco's modified eagle media (DMEM) containing 4.5 g/L D-Glucose and 200 mM L-glutamine (Gibco), supplemented with 10 % foetal bovine serum

(Gibco), 1 % penicillin/streptomycin (100 U/ml, 100 mg/ml; Sigma), and 1 % (*v/v*) non-essential amino-acids (Gibco). Cells were incubated at 37 °C in 5 % CO_2.

2.1.4.4 Human cell culture reagents

Table 2.1. Reagents in cell culture

Reagent	Supplier
PBS (1x, pH 7.2)	Gibco
Fetal Bovine Serum (heat inactivated)	Gibco
L-glutamine (200 mM)	Gibco
Penicillin-Streptomycin solution	Sigma
Sodium Pyruvate (100 mM)	Sigma
Trypsin-EDTA solution (0.25%, sterile filtered)	Sigma
DMSO	Sigma

2.1.4.5 Freezing media

All mammalian cell lines were frozen in liquid nitrogen for long term storage, specifically stored in freezing media made of 90% foetal bovine serum (Sigma) and 10% DMSO (Sigma).

All cell lines were regularly tested for mycoplasma and they were authenticated at the beginning of the study.

2.2 Methods

2.2.1 Primer design for PCR

The primary DNA sequence for the region of interest was downloaded from University of California Santa Cruz (UCSC) Genome Browser (https://genome.ucsc.edu/), plus additional flanking sequence to accommodate optimal primer design. Sequences between 18-25 bp were selected and their thermodynamics were inspected using OligoAnalyzer, a tool hosted by Integrated DNA Technologies (https://eu.idtdna.com/calc/analyzer). Primers were designed to have a melting temperature between 55 -65°C and a GC content between 40-60%. OligoAnalyzer was also used to check for hairpin formation and likelihood of homo- and hetero- dimers in a PCR reaction. Potential primer pairs were submitted to BLAT on UCSC to confirm sequence similarity across the genome and were tested using the In-Silico PCR tool to validate primer specificity. All primers were ordered as lyophilised and upon arrival were dissolved in nuclease free water (Gibco) to a final concentration of 100 µM. All primers were then diluted to 20 µM for use in PCR.

2.2.2 Polymerase Chain Reaction

2.2.2.1 DNA Polymerase selection

All VNTRs and full length SVAs were amplified using KOD Hot Start DNA Polymerase (Merck); a high fidelity DNA polymerase, qualified for long, repetitive and GC rich amplicons[306,307]. As this enzyme contains a proofreading mechanism it is less prone to mutation errors and thus downstream processes such as sequencing should be more accurate due to the presence of less PCR generated mutations. This also

minimised mutations within reporter gene constructs, which could alter gene expression experiments. Hot Start DNA Polymerase enzymes have an antibody conjugated to them which renders them inactive, ensuring no reactivity at room temperature and thus reducing nonspecific amplification. Reactions using Hot Start DNA Polymerases therefore included an initial heat activation step to cleave the antibody (95 °C for 5 minutes). All PCR reactions were performed in the SimpliAmp™ Thermal Cycle (Applied Biosystems) and stored at 4 °C for short term storage (~2 weeks) or at -20 °C for long term storage. The extension temperature for all KOD Hot Start DNA Polymerase (Merck) reactions was set to 68 °C as that is the optimal temperature for proof reading activity and therefore ensured the highest DNA sequence fidelity. Please refer to Table 2.2 for specific PCR conditions and thermal cycles.

2.2.2.2 Genotyping of VNTRs and SVAs

Table 2.2. PCR primers and master mix setup

Target Region	Primers (5'-3')	Application	Product size (template)	Master Mix	Thermal cycler conditions
NEK1 SVA (full length)	For: AACTGGGATATCGTCAGCAGCT Rev: ACCGCAACCAGGCTAAGCT	Genotyping	1877 bp (gDNA)	DNA template: 5 ng Buffer for KOD Hot Start (10X): 2 µl dNTPs (2 mM each): 2 µl (0.2 mM each) $MgSO_4$ (25 mM): 1.2 µl (1.5 mM) For (20 µM): 0.3 µl (0.3 µM) Rev (20 µM): 0.3 µl (0.3 µM) KOD Hot Start DNA Polymerase (1 U/µl): 0.4 µl (0.02 U/µl) Betaine (5 M): 4 µl (1 M) DMSO: 1 µl Nuclease free H_20: x µl	1 cycle: 95 °C – 2 mins 30 cycles: 95 °C – 20 secs 61 °C – 10 secs 68 °C – 30 secs 4 °C - ∞
NEK1 SVA CT element	For: AACTGGGATATCGTCAGCAGCT Rev: GCAGTACAGTCCAGCTTCGG	Genotyping	189 bp (gDNA)	DNA template: 5 ng Reddy Mix (2X): 10 µl For (20 µM): 0.3 µl (0.3 µM) Rev (20 µM): 0.3 µl (0.3 µM) Nuclease free H_20: x µl	1 cycle: 95 °C – 2 mins 30 cycles: 95 °C – 30 secs 61.3 °C – 30 secs 72 °C – 15 secs 1 cycle: 72 °C – 2 mins 4 °C - ∞
NEK1 SVA VNTR Nested PCR: (Do full length SVA PCR first but only 20 cycles and then use this as template for the PCR)	For: TACAACCTCCACCTCCCAGC Rev: CCACAAAACCGCCACTGTCATC	Genotyping	888 bp (gDNA)	DNA template: 1 µl product from full length PCR Buffer for KOD Hot Start (10X): 2 µl dNTPs (2 mM each): 2 µl (0.2 mM each) $MgSO_4$ (25 mM): 1.2 µl (1.5 mM) For (20 µM): 0.3 µl (0.3 µM) Rev (20 µM): 0.3 µl (0.3 µM) KOD Hot Start DNA Polymerase (1 U/µl): 0.4 µl (0.02 U/µl) Betaine (5 M): 4 µl (1 M) DMSO: 1 µl Nuclease free H_20: x µl	1 cycle: 95 °C – 2 mins 30 cycles: 95 °C – 20 secs 61 °C – 10 secs 68 °C – 30 secs 4 °C - ∞
NEK1 SVA poly A tail	For: CCTATGACCCTGCCAAATCCCC Rev: TCCCTCTACGCCATTCCCAC	Genotyping	366 bp (gDNA)	DNA template: 5 ng GoTaq® Flexi buffer (5X): 4 µl dNTPs (10 mM each): 0.4 µl (0.2 mM each) $MgCl_2$ (25 mM): 1.6 µl (2 mM) For (20 µM): 0.3 µl (0.3 µM) Rev (20 µM): 0.3 µl (0.3 µM) GoTaq® Hot Start DNA Polymerase (5 U/µl): 0.1 µl (0.025 U/µl) Nuclease free H_20: x µl	1 cycle: 95 °C – 2 mins 35 cycles: 95 °C – 30 secs 60.6 °C – 30 secs 72 °C – 15 secs 4 °C - ∞
NEK1 SVA CRISPR KO region	For: CACTCCCAACTTCCCATCTC Rev: GTACAGCCTTGCCAAACCTG	Confirmation of CRISPR modification	4660 bp (gDNA)	DNA template: 20 ng Xtreme buffer (2X): 10 µl (1X) dNTPs (2 mM each): 4 µl (0.4 mM each) For (20 µM): 0.32 µl (0.32 µM) Rev (20 µM): 0.32 µl (0.32 µM) KOD Xtreme™ Hot Start DNA Polymerase (1 U/µl): 0.48 µl (0.024 U/µl) Nuclease free H_20: x µl	1 cycle: 94 °C – 2 mins 30 cycles: 98 °C – 10 secs 62 °C – 30 secs 68 °C – 3 mins 4 °C - ∞
NEK1 (all transcripts/total expression)	For: AAGTGTGGGAAGAGCATTTG Rev: CTACTGTCCACGCCAACTTC	mRNA expression levels (RT-PCR)	157 bp (cDNA) *1:10 dilution	DNA template*: 1 µl GoTaq® Flexi buffer (5X): 4 µl (1X) dNTPs (10 mM each): 0.4 µl (0.2 mM each) $MgCl_2$ (25 mM): 1.6 µl (2 mM) For: 0.3 µl (0.3 µM) Rev: 0.3 µl (0.3 µM) GoTaq® Hot Start DNA Polymerase (5 U/µl): 0.1 µl (0.025 U/µl) Nuclease free H_20: 12.3 µl	1 cycle: 95 °C – 2 mins 30 cycles: 95 °C – 30 secs 60 °C – 30 secs 72 °C – 30 secs 72 °C – 2 mins 4 °C - ∞
CLCN3 (all transcripts/total expression)	For: CTGATCGTCCAGCAGGCATTGG Rev: AGCCTGATGGAACCTTGATGCCA	mRNA expression levels (RT-PCR)	114 bp (cDNA) *1:10 dilution	DNA template*: 1 µl GoTaq® Flexi buffer (5X): 4 µl (1X) dNTPs (10 mM each): 0.4 µl (0.2 mM each) $MgCl_2$ (25 mM): 1.6 µl (2 mM) For (20 µM): 0.3 µl (0.15 µM) Rev (20 µM): 0.3 µl (0.15 µM) GoTaq® Hot Start DNA Polymerase (5 U/µl): 0.1 µl (0.025 U/µl) Nuclease free H_20: 12.3 µl	1 cycle: 95 °C – 2 mins 30 cycles: 95 °C – 30 secs 60 °C – 30 secs 72 °C – 30 secs 72 °C – 2 mins 4 °C - ∞
CFAP410 (*C21orf2*) VNTR	For: AACCCCAGACAACAGACCC Rev: CTGACGCGGAAGATGGTTC	Genotyping	584 bp (gDNA)	DNA template: 2 ng Buffer for KOD Hot Start (10X): 2 µl dNTPs (2 mM each): 2 µl (0.2 mM each) $MgSO_4$ (25 mM): 1.2 µl (1.5 mM) For (20 µM): 0.3 µl (0.15 µM) Rev (20 µM): 0.3 µl (0.15 µM) KOD Hot Start DNA Polymerase (1 U/µl): 0.4 µl (0.02 U/µl) Betaine (5 M): 4 µl (1 M) DMSO: 1 µl Nuclease free H_20: x µl	1 cycle: 95 °C – 2 mins 30 cycles: 95 °C – 20 secs 63 °C – 10 secs 68 °C – 40 secs 4 °C - ∞
REST VNTR	For: GGCACTCCTTGCTTGGTAGAGG Rev: GCCGCACATTCCAACACAGGAC	Genotyping	178 bp (gDNA)	DNA template: 5 ng Buffer for KOD Hot Start (10X): 2 µl dNTPs (2 mM each): 2 µl (0.2 mM each) $MgSO_4$ (25 mM): 1.2 µl (1.5 mM) Fw (20 µM): 0.3 µl (0.3 µM) Rv (20 µM): 0.3 µl (0.3 µM) KOD Hot Start DNA Polymerase (1 U/µl): 0.4 µl (0.02 U/µl) Betaine (5 M): 4 µl (1 M) DMSO: 1 µl Nuclease free H_20: x µl	1 cycle: 5 °C – 2 mins 30 cycles: 95 °C – 20 secs 58.5 °C – 10 secs 68 °C – 30 secs 4 °C - ∞

2.2.2.3 Gel agarose electrophoresis

Both PCR products and restriction digests were run on agarose gels and visualised using a UV transilluminator (BioDoc-it system). Percentage of agarose ranged between 0.8-3% based on the amplicon size; smaller bands (≤200 bp) were run on higher percentage gels to allow for higher resolution images. Powdered UltraPure™ agarose (Invitrogen) was added to 0.5X TBE buffer (Section 2.1.1.1) and heated up to boiling temperature to allow the agarose to dissolve. Ethidium bromide (EtBr, 500 µg/ml, Sigma), an intercalating dye used to visualise nucleic acids, was added to the solution (at a final concentration of 50 µg per 100 ml). The liquid agarose was then poured into heat-resistant plastic gel trays and allowed to cool at room temperature for approximately 20 minutes. Before cooling, gel combs were added to the tank to create wells to load each PCR sample. Once set, agarose gels were then placed into horizontal gel tanks containing 0.5X TBE buffer, connected to a power supply and run between 100-120 V/cm to separate DNA fragments using an electric field. PCR amplicon size was compared against DNA marker ladders, either 100 bp ladder (Promega) or 1 kb ladder (Promega) depending on predicted amplicon size.

2.2.2.4 QIAxcel Advanced System – gel capillary electrophoresis

Gel capillary electrophoresis using the QIAxcel (QX) advanced system (QIAGEN) was adopted for high throughput genotyping of polymorphic DNA regions. This system was specifically adopted for small amplicons (approximately ≤200 bp) to facilitate high resolution characterisation of VNTR and SVA CT-element variants. All QIAxcel cartridges were calibrated with QX Intensity Calibration Marker prior to use to normalise signal intensity across all 12 gel cartridge channels. Unused

cartridges were placed upright in the packaging blister at 2-8 ˚C for long term storage. Before use, all cartridges were left to equilibrate to room temperature for 20 minutes in the covered cartridge stand containing QX DNA wash buffer. The wash buffer tray was cleaned with 70% ethanol and rinsed thoroughly with nuclease water and DNA separation and wash buffers were regularly changed to maintain a clean signal output. Amplicons ranging from 100-500 bp were run on the High Resolution cartridge (QIAGEN) using the OM800 method as this gave the best resolution possible (3-5 bp). QX DNA Size Marker 25-500 bp v2.0 was run alongside samples to generate a reference marker table and determine size of sample amplicons. QX Alignment Marker 15/600 bp was used in every run to minimise migration time variation across capillaries. L (low) methods were used for DNA samples with concentrations of <10 ng/µl, M (medium) methods were used for DNA samples ranging from 10-100 ng/µl and H (high) methods were used for DNA samples with concentrations of >100 ng/µl. Both the *REST* VNTR and *NEK1* SVA-D CT element (amplicon sizes of approximately ≤200 bp) were screened in the MNDA UK cohort using the QIAxcel advanced system and data was assessed using the QIAxcel ScreenGel software, displaying genotyping results in both an electropherogram and digital gel image format. All DNA products were also verified afterwards on 3% UltraPure™ agarose (Invitrogen) gels and were run for approximately 4 hours at 100 V/cm (Section 2.2.2.2).

2.2.3 Cloning methods

2.2.3.1 Amplification of fragments for subcloning using PCR

DNA fragments of interest were initially subcloned into the pCR®-Blunt vector from the Zero Blunt® PCR Cloning Kit (ThermoFisher) and then cloned into reporter gene constructs. DNA fragments were amplified by PCR using KOD Hot Start DNA Polymerase (Merck), which generated a blunt product and thus was compatible with the Zero Blunt® PCR Cloning Kit (ThermoFisher). Please see Section 2.2.2 for all PCR amplicons and thermal cycler conditions.

2.2.3.2 Extraction of DNA fragments from agarose gel

PCR amplicons were isolated and purified using the Wizard® SV Gel and PCR Clean-Up System (Promega) according to manufacturer's instructions. This membrane column based system allows for extraction of dsDNA between 100 bp and 10 kb and purification of DNA products immediately following amplification by PCR. DNA fragments were run on an agarose gel (1%) and then isolated and cut out using a scalpel. The gel slice was then weighed and dissolved in membrane binding solution at 60°C (10 µl solution for every 10 mg of gel slice up to a total of 350 mg) and then passed through the DNA binding column. DNA was eluted in 30 µl of nuclease free water (Invitrogen) to achieve high yield/concentration without reducing elution efficiency. Purified DNA was then subject to quantity and quality assessment (outlined in Section 2.2.6) and then taken forward for cloning experiments.

2.2.3.3 Ligation of DNA fragments into pCR®-Blunt intermediate vector

DNA fragments of interest were ligated and cloned into the pCR®-Blunt vector (Figure 2.1) using the Zero Blunt® PCR Cloning Kit (ThermoFisher) according to manufacturer's instructions. This kit is designed to clone blunt PCR fragments into the pCR®-Blunt vector using an ExpressLink™ T4 DNA ligase. The pCR®-Blunt vector is supplied linearised and contains a lethal *LacZ-ccdB* gene fusion which is disrupted upon ligation of a blunt product, permitting propagation of only positive recombinants. A 10:1 molar ratio of insert:vector was used as this was the manufacturer's recommended ratio for optimal blunt-ended PCR ligation. The amount of PCR product required for optimal ligation was calculated using the following equation:

$$\text{x ng insert} = \frac{(10)(\text{y bp PCR product})(25 \text{ ng linearised pCR®-Blunt})}{(3500 \text{ bp pCR®-Blunt})}$$

The ligation reaction used is outlined in Table 2.3.

Table 2.3. Ligation reaction for pCR®-Blunt vector cloning

pCR®-Blunt (25 ng)	1 µL
Blunt PCR product	1–5 µL
5X ExpressLink™ T4 DNA Ligase Buffer	2 µL
Nuclease Free Water	to a total of 9 µL
ExpressLinkTM T4 DNA Ligase (5 U/µL)	1 µL
Total Volume	**10 µL**

All ligations were performed overnight at room temperature.

2.2.3.4 Ligation of DNA inserts into pSHM06 vector and pGL3P/B vectors

The *NEK1* SVA-D was non-directionally cloned into the pSHM06 vector and used for luciferase reporter gene assays. Between 1-4 µg of pCR®-Blunt vector

containing the SVA was cut with 10 U/µg of *NsiI-HF* (NEB) restriction endonuclease at 37 °C for 2 hours, which left 4 bp overhangs at both the 5' and 3' ends; the enzyme was then heat inactivated at 80 °C for 20 minutes. Approximately 2 µg of pSHM06 vector was digested with *Sbf-I-HF* (NEB), leaving 4bp overhangs which were complementary to *NsiI* overhangs of the SVA insert. To ensure the linearised pSH0M6 vector did not self-ligate, the vector was dephosphorylated with Antarctic Phosphatase (NEB): 5 units of enzyme per 1 pmol of DNA end were incubated with 2 µl of 10X reaction buffer and nuclease free water (20 µl reaction volume) for 30 minutes at 37 °C and then the phosphatase was heat inactivated for 2 mins at 80 °C.

The *NEK1* SVA-D was cloned into pGL3-P (Promega) and used for luciferase reporter gene assays. Approximately 1 µg of pGL3-P vector was linearised with *SmaI* (NEB). Following this, between 2-3 µg of pSHM06 construct containing the *NEK1* SVA was digested with *EcoRV* (NEB) to cleave out the SVA insert, which was then ligated into the linearised pGL3-P vector as previously described above. The *CFAP410* VNTR was cloned into both pGL3-P and pGL3-B and also used for luciferase reporter gene assays. Approximately 1 µg each of pGL3-P and pGL3-B was digested with *NheI-HF* (NEB) at 37 °C; the enzyme was then heat inactivated at 80 °C for 20 minutes. Following this, approximately 2 µg of pCR®-Blunt vector containing the *CFAP410* VNTR was digested (to cleave out the VNTR insert) with *SpeI* and *XbaI* (NEB) at 37 °C; both enzymes were also heat inactivated at 80 °C for 20 minutes. All ligation reactions were performed with 10:1 molar ratio of insert:vector and left overnight at room temperature and set up as previously

described in Section 2.2.3.3, but with their respective vectors and inserts of interest. Please refer to Table 2.4 for a full list of all constructs.

Table 2.4. Table of constructs

Construct	Vector	Insert orientation	Vector RE sites	Insert RE sites	RE used to check for insert	RE used to check orientation of insert	Application
REST VNTR (6-repeat)	pGL3-B	Forward	*HindIII*	N/A	*BglII* and *NcoI*	N/A	Luciferase and variant sequencing
CFAP410 (*C21orf2*) VNTR (allele 4)	pGL3-P	Forward	*NheI*	*SpeI* and *XbaI*	*KpnI*	*DraIII*	Luciferase and variant sequencing
CFAP410 (*C21orf2*) VNTR (allele 4)	pGL3-P	Reverse	*NheI*	*SpeI* and *XbaI*	*KpnI*	*DraIII*	Luciferase and variant sequencing
CFAP410 (*C21orf2*) VNTR (allele 5)	pGL3-P	Forward	*NheI*	*SpeI* and *XbaI*	*KpnI*	*DraIII*	Luciferase
CFAP410 (*C21orf2*) VNTR (allele 5)	pGL3-P	Reverse	*NheI*	*SpeI* and *XbaI*	*KpnI*	*DraIII*	Luciferase
CFAP410 (*C21orf2*) VNTR (allele 5)	pGL3-B	Forward	*NheI*	*SpeI* and *XbaI*	*KpnI*	*DraIII*	Luciferase
CFAP410 (*C21orf2*) VNTR (allele 5)	pGL3-B	Reverse	*NheI*	*SpeI* and *XbaI*	*KpnI*	*DraIII*	Luciferase
NEK1 SVA (full length, reference genome)	pGL3-P	Forward	*SmaI*	*EcoRV*	*BamHI*	*BamHI*	Luciferase
NEK1 SVA (full length, reference genome)	pGL3-P	Reverse	*SmaI*	*EcoRV*	*BamHI*	*BamHI*	Luciferase
NEK1 SVA (full length, reference genome)	pSHM06	Forward	*SbfI*	*NsiI*	*BamHI*	*BamHI*	Luciferase
NEK1 SVA (full length, reference genome)	pSHM06	Reverse	*SbfI*	*NsiI*	*BamHI*	*BamHI*	Luciferase
NEK1 SVA (full length, CT variant 1)	pCR®-Blunt	Forward	N/A (vector already linearised)	N/A (blunt ended PCR product)	*NsiI*	N/A	Variant sequencing
NEK1 SVA (full length, CT variant 3)	pCR®-Blunt	Forward	N/A (vector already linearised)	N/A (blunt ended PCR product)	*NsiI*	N/A	Variant sequencing
NEK1 SVA (full length, CT variant 4)	pCR®-Blunt	Forward	N/A (vector already linearised)	N/A (blunt ended PCR product)	*NsiI*	N/A	Variant sequencing
NEK1 SVA KO Guide 1	pSpCas9(BB)-2A-GFP	Forward	*BbsI*	*BbsI*	*BbsI*	N/A	CRISPR
NEK1 SVA KO Guide 2	pSpCas9(BB)-2A-GFP	Forward	*BbsI*	*BbsI*	*BbsI*	N/A	CRISPR
NEK1 SVA KO Guide 3	pSpCas9(BB)-2A-GFP	Forward	*BbsI*	*BbsI*	*BbsI*	N/A	CRISPR
NEK1 SVA KO Guide 4	pSpCas9(BB)-2A-GFP	Forward	*BbsI*	*BbsI*	*BbsI*	N/A	CRISPR

2.2.3.5 Transformation of DH5a competent E.coli

Ligation mixes were transformed into Subcloning Efficiency™ DH5 Competent *E.coli* cell (ThermoFisher) according to the manufacturer's instructions. Between 1-5 µl of ligation mix was added to a 50 µl aliquot of chemically competent cells and incubated on ice for 30 minutes. Following this, cells were incubated in a pre-warmed water bath (42 °C) for 30 seconds to induce heatshock and aid cell membrane permeability and thus facilitate entry of DNA into the cytosol of bacterial cells[308,309]. After this the mixture was then placed back on ice for 2 minutes and then 950 µl of LB broth was added. This mixture was then placed into a shaking incubator at 37 ºC and 225 rpm for 1 hour. From this, between 50-200 µL of cells were then plated and spread onto an LB agar plate containing either kanamycin (final concentration of 50 µg/ml) (for all pCR®-Blunt intermediate vectors) or ampicillin (final concentration of 100 µg/ml) (for all pGL3 vectors and the pSHM06 vector) and placed into an incubator at 37 ºC overnight. Single colonies were then picked and grown individually to facilitate clonal expansion.

2.2.3.6 Growing up bacterial culture

Individual colonies of *E.coli* were picked using a sterile pipette tip and placed into a universal tube containing 5 mL of LB broth and 5 µl of antibiotic: Kanamycin (50 mg/ml, Sigma) for pCR®-Blunt vector and Ampicillin (100 µg/ml, Sigma) for pGL3P vectors and pSHM06 vector. These colonies were then grown overnight in a shaking incubator, at 37 °C and shaking at 225 rpm. The following day these cultures were either used for minipreps (Section 2.2.5.1) or expanded into larger cultures. The latter was done by adding 20-50 µl of bacterial culture to a canonical flask containing 100 ml of LB broth and 100 µl of antibiotic (kanamycin, 50 mg/ml

or ampicillin, 100 mg/ml). These flasks were then incubated overnight in a shaking incubator at 37 °C shaking at 225 rpm.

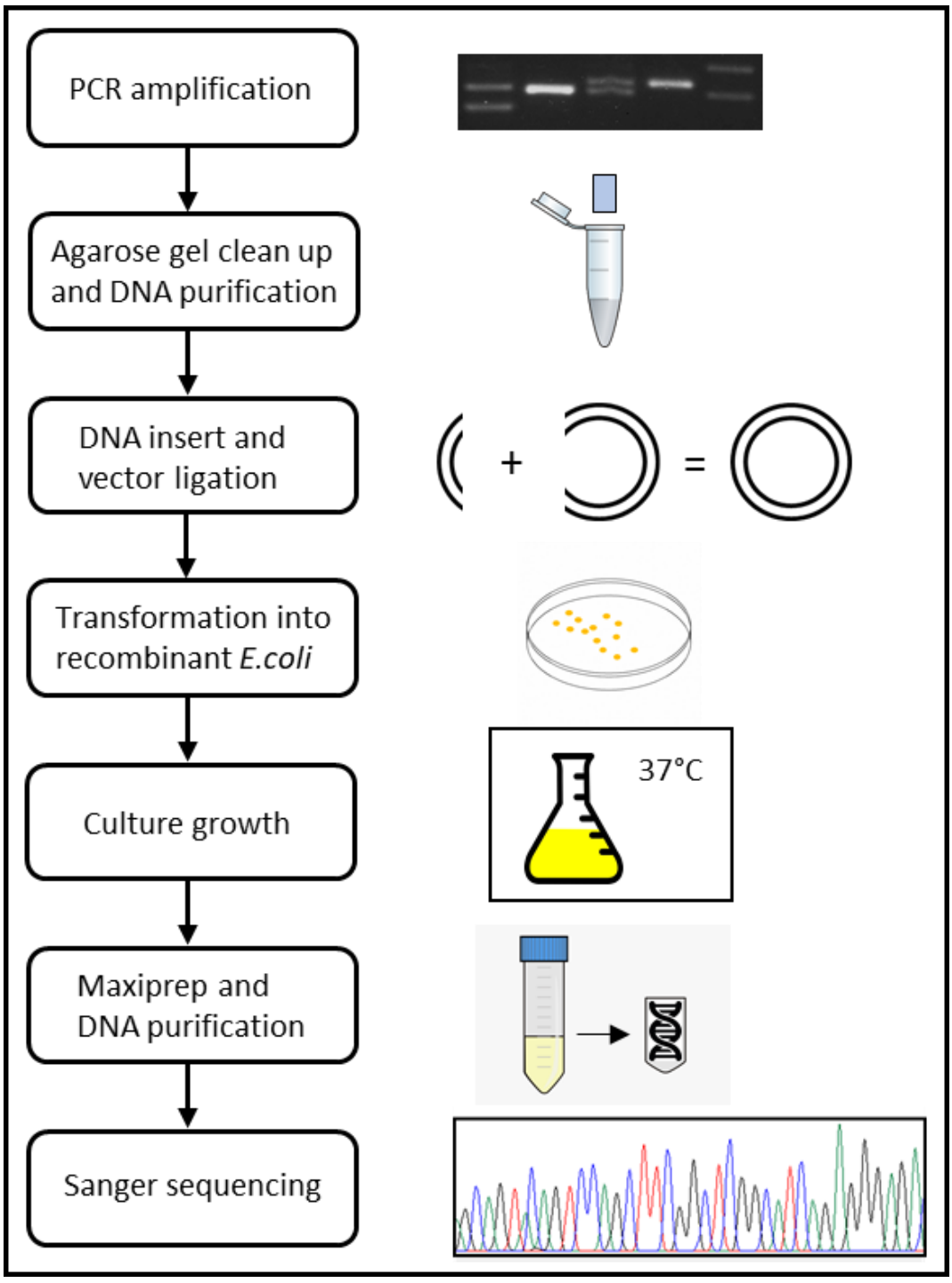


Figure 2.5. Cloning pipeline

An outline of the steps within the cloning pipeline used for all generated constructs.

2.2.3.7 Restriction enzyme digests

To determine presence of a DNA insert within a vector, restriction enzymes (REs) were used to digest the vector. Each vector was digested with restriction enzymes which cut either side of the insert of interest so that the specific insert could be isolated and presence and length confirmed. Primary sequence of vectors were taken from Addgene (https://www.addgene.org/) and uploaded to a plasmid editor (ApE) (https://jorgensen.biology.utah.edu/wayned/ape/) to identify restriction enzyme cut sites. All REs were from NEB or Promega. Restriction enzyme digests were incubated at the temperature appropriate for the enzyme of use in a thermal cycler for approximately 1-3 hours. Enzymes were then heat-inactivated (if possible) by incubating the reaction at 80 °C for 10 minutes. Gel loading dye, purple (6x) (NEB), containing EDTA to stop enzymatic reactions, was added to the digest after the reaction to inactivate the restriction enzyme and to aid sample loading onto an agarose gel. All pSHM06 and pGL3-P/B constructs were digested with *EcoRI* to determine presence of an insert. Each restriction digest was then run on an agarose gel and visualised on a UV transilluminator (Section 2.2.2.3). Following a successful insert check diagnostic digest of all reporter gene constructs, an insert orientation diagnostic digest was performed (a "one-in one-out" digest). This was to check which orientation the insert had cloned into the vector. One restriction cut site within the insert and one within the vector backbone was utilised in this process; this allowed for two distinct fragment patterns depending on the orientation of the insert (Figure 2.6). All *CFAP410* VNTR constructs were digested with *DraIII-HF* (NEB) to determine both presence and orientation of the VNTR (Figure 3.12). All pGL3-P and pSHM06 constructs

containing the *NEK1* SVA were digested with *BamHI* to determine orientation of the SVA (Figure 5.1 and Figure 5.3). All reporter gene constructs were then sequenced verified (Section 2.2.6).

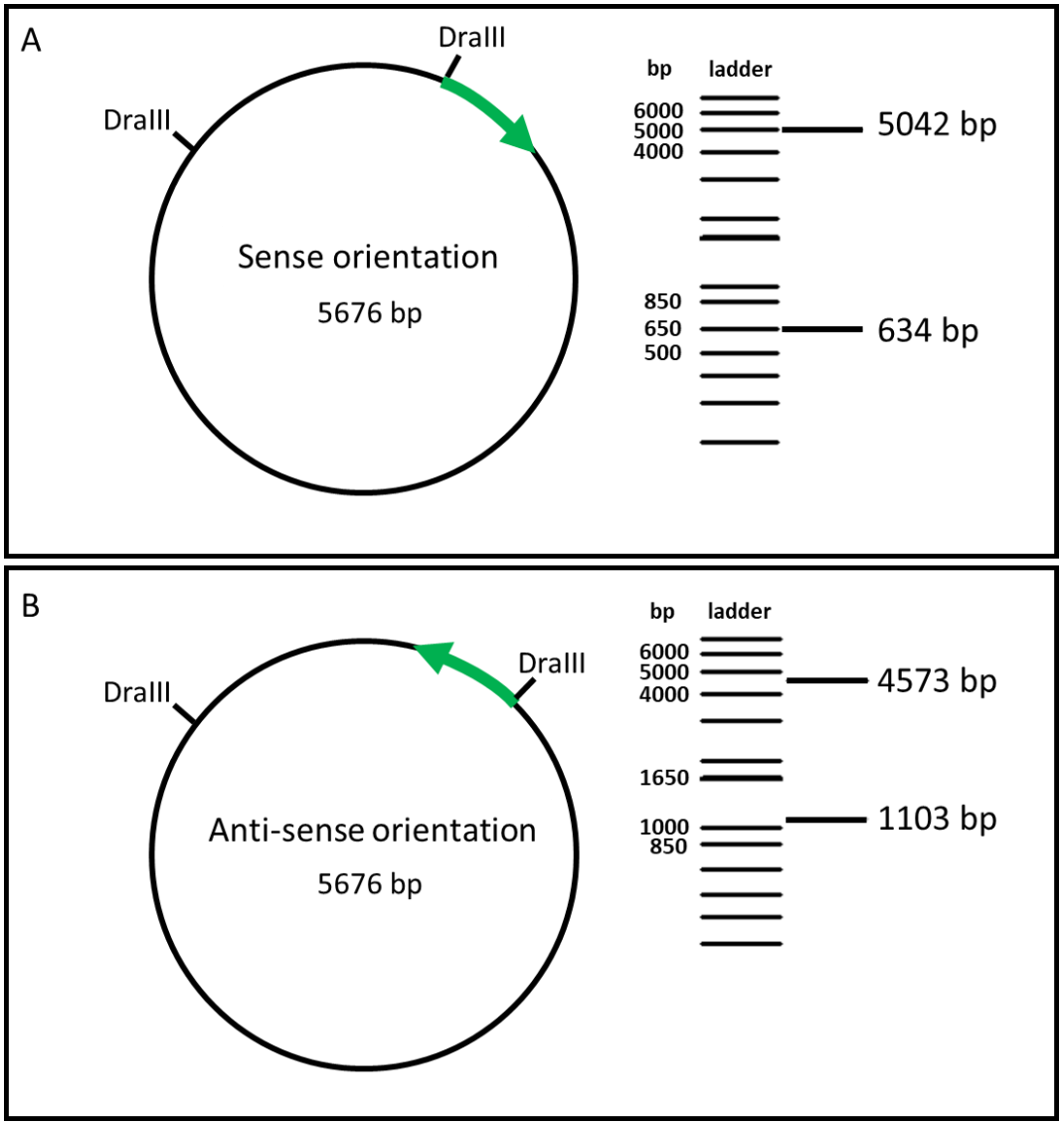


Figure 2.6. Orientation check restriction digest

An example orientation digest schematic from ApE. The enzyme *DraIII* cuts once in the VNTR insert and once in the backbone of the pGL3-P vector. **A**: When the VNTR is in the sense orientation with respect to the promoter of pGL3-P vector this generates a band of 5024 bp and a band of 634 bp. **B**: When the VNTR is in the anti-sense orientation with respect to the promoter of pGL3-P vector this generates a band of 4573 bp and a band of 1103 bp. This difference in banding

pattern allows one to accurately distinguish between VNTR orientation in this construct. All constructs were also verified via Sanger sequencing.

2.2.3.8 Gibson Isothermal Assembly

The *REST* VNTR was cloned into pGL3-B using the Gibson Assembly Master Mix (NEB). This system is an exonuclease-based cloning method which incorporates a T5 exonuclease, *Taq* ligase and Phusion DNA polymerase all within a one-step isothermal reaction[310](Figure 2.7). Firstly, primers were designed for the VNTR insert containing a 16 bp overhang (underlined below) complementary to both sides of the Hind III site used to linearise pGL3-B vector:

Fw: 5' GATCTGCGATCTAAGTGGCACTCCTTGCTTG 3'

Rv: 5' ACAGTACCGGAATGCCGCCGCACATTCCAAC 3'

The *REST* VNTR was amplified by PCR using the standard protocol and primers (Table 2.2). From this, 1 µl of PCR product was used as a template for amplification by PCR using the newly designed Gibson assembly primers (above) (same conditions as the standard *REST* VNTR PCR within Table 2.2). Prior to the Gibson assembly reaction, the pGL3-B vector (Promega) was linearised with *Hind III*. The pmols of VNTR insert required for the Gibson assembly reaction was calculated using the equation below:

pmols = (weight in ng) x 1,000 / (base pairs x 650 daltons)

The Gibson isothermal assembly reaction used is displayed in Table 2.5. Samples were incubated at 50 °C in a thermocycler for 60 minutes to help improve assembly efficiency.

Table 2.5. Gibson assembly reaction

	2-3 Fragment Assembly
Total Amount of Fragments	0.02-0.5 pmols (X µL)
Gibson Assembly Master Mix (2X)	10µL
Deionised H_2O	10-XµL
Total Volume	**20 µL**

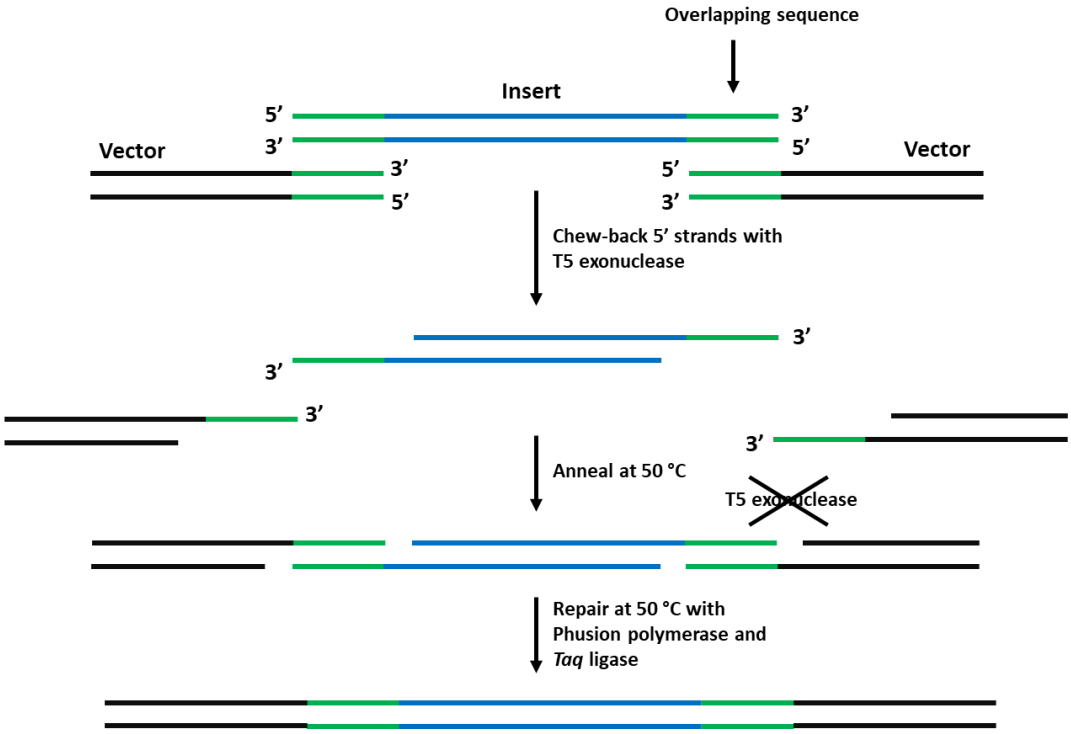


Figure 2.7. Gibson Isothermal Assembly reaction outline

The T5 exonuclease chews back the 5' strand, leaving 3' overhangs on the vector and insert which are complementary and thus induces annealing. Once the complementary strands have annealed, Phusion polymerase fills in gaps of each fragment and *Taq* ligase then seals any nicks in the DNA (Adapted from Gibson et al., 2009[310]).

2.2.4 Isolation of plasmid DNA from bacterial cultures

2.2.4.1 Miniprep

Plasmid DNA from 5 ml cultures was isolated and purified using the Wizard® Plus SV Minipreps DNA Purification System (Promega), a silica membrane minicolumn based system. 5 ml bacterial cultures were placed into 5 ml falcon tubes and centrifuged at 4500 g for 15 minutes. Pelleted cells were then resuspended, lysed and DNA was isolated and purified according to manufacturer's instructions. All purified DNA was eluted in 50 µl of Ultra Pure™ DNase/RNase free distilled water (Invitrogen).

2.2.4.2 Maxiprep

Plasmid DNA from 100 ml cultures was isolated and purified using the QIAGEN Plasmid Maxi kit (QIAGEN). 100 ml bacterial cultures were aliquoted into 50 ml falcons and centrifuged at 4500 g for 15 minutes and the supernatant was then careful discarded. The bacterial cell pellet was then resuspended and lysed; DNA was then isolated and purified according to manufacturer's instructions. DNA pellets were air dried for 10 minutes to ensure evaporation of any residual 70% ethanol (EtOH) and then resuspended in 200 µl of Ultra Pure™ DNase/RNase free distilled water (Invitrogen). All purified DNA was then subject to a quality check and quantified using a NanoDrop™ spectrophotometer (Thermo Scientific) (Section 2.2.5.1).

2.2.5 Nucleic acid quality control

2.2.5.1 DNA and RNA quantification and quality verification using the nanodrop

DNA and RNA quantity and purity were both assessed using a NanoDrop™ spectrophotometer (Thermo Scientific). The NanoDrop instrument was first initialised with nuclease free water and then blanked with the solution that the nucleic acid was in (either nuclease free water or TE buffer). 1 µL of sample was used for each measurement and the nucleic acid concentration, 260/230 ratio and 260/280 ratio were recorded. The 260/230 ratio is used to assess nucleic acid purity, indicating the presence of contaminants such as phenol, TRIzol and carryover of guanidine thiocyanate from nucleic acid purification systems. The 260/280 ratio is another measure of nucleic acid purity, indicating the presence protein contamination. All DNA and RNA used had a 260/280 ratio of approximately 1.8 and 2.0 respectively, which is accepted as pure. Both types of nucleic acid also yielded a 260/230 ratio between 1.8-2.2, which is also accepted as pure.

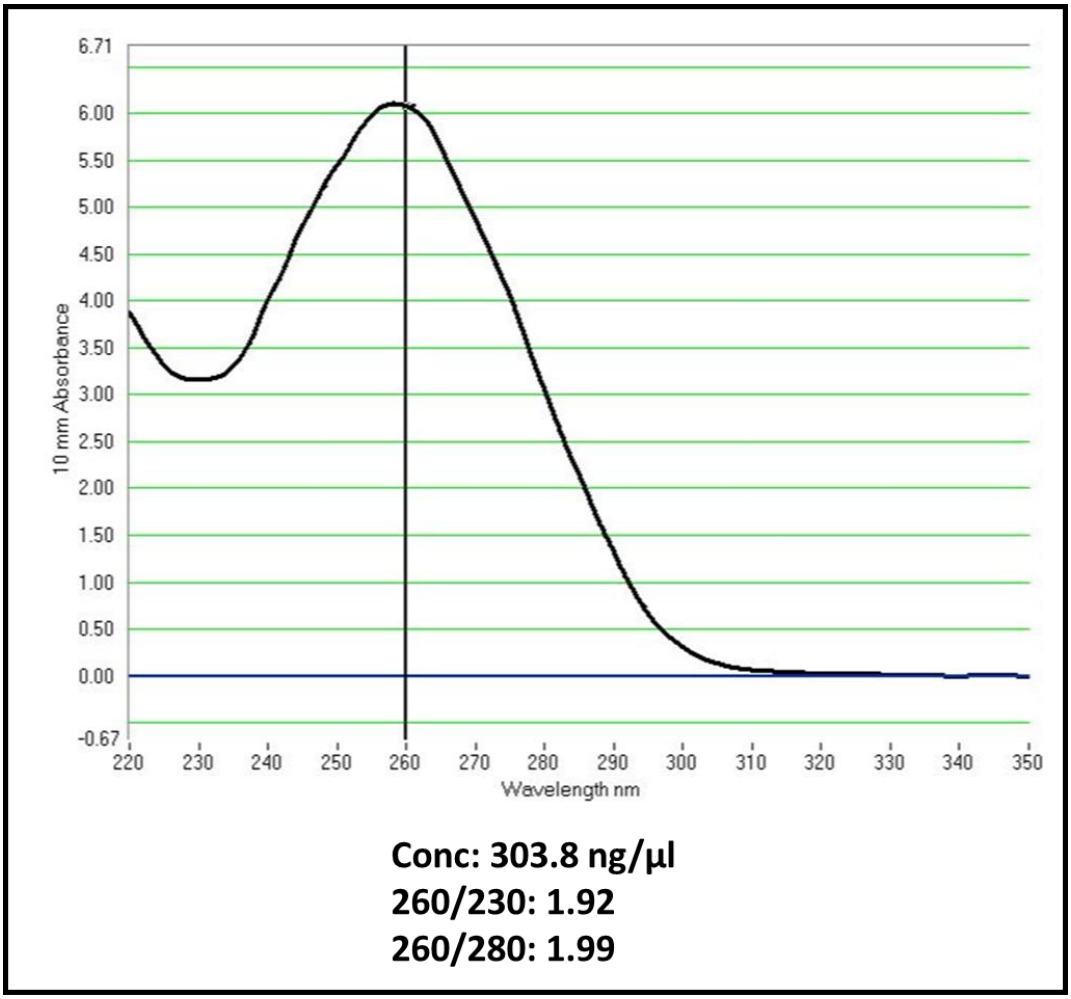


Figure 2.8. Example NanoDrop™ result.

This DNA sample is pure, with a 260/230 ratio between 1.8-2.2 meaning there are no contaminants such as phenol. The 260/280 ratio is between 1.8-2.0 also meaning there are no protein contaminants in this sample.

2.2.5.2 RNA integrity assessment using agarose gels

RNA integrity was visually assessed by running purified and normalised RNA on UltraPure™ agarose (Invitrogen) gels for 1 hour at 100 V/cm, to look for intact mammalian 28S and 18S ribosomal RNA (rRNA). Intact bands (and no smearing) indicated that the RNA was intact and had not degraded from RNase contamination (Figure 2.13).

2.2.6 Sanger Sequencing

Sequencing of DNA was performed externally by Source Bioscience. A total of 5µl of DNA sample normalised to 100 ng/µl and 5 µl of sequencing primer (specific for either pCR®-Blunt, pGL3-P, pGL3-B or pSHM06 vector) normalised to 3.2 pmol/µl was included for each sequencing reaction. All electropherograms were visually assessed for sequencing quality using Chromas (http://technelysium.com.au/wp/chromas/) and FASTA files were aligned to reference genome sequences (UCSC, hg19) to check for presence of SNPs.

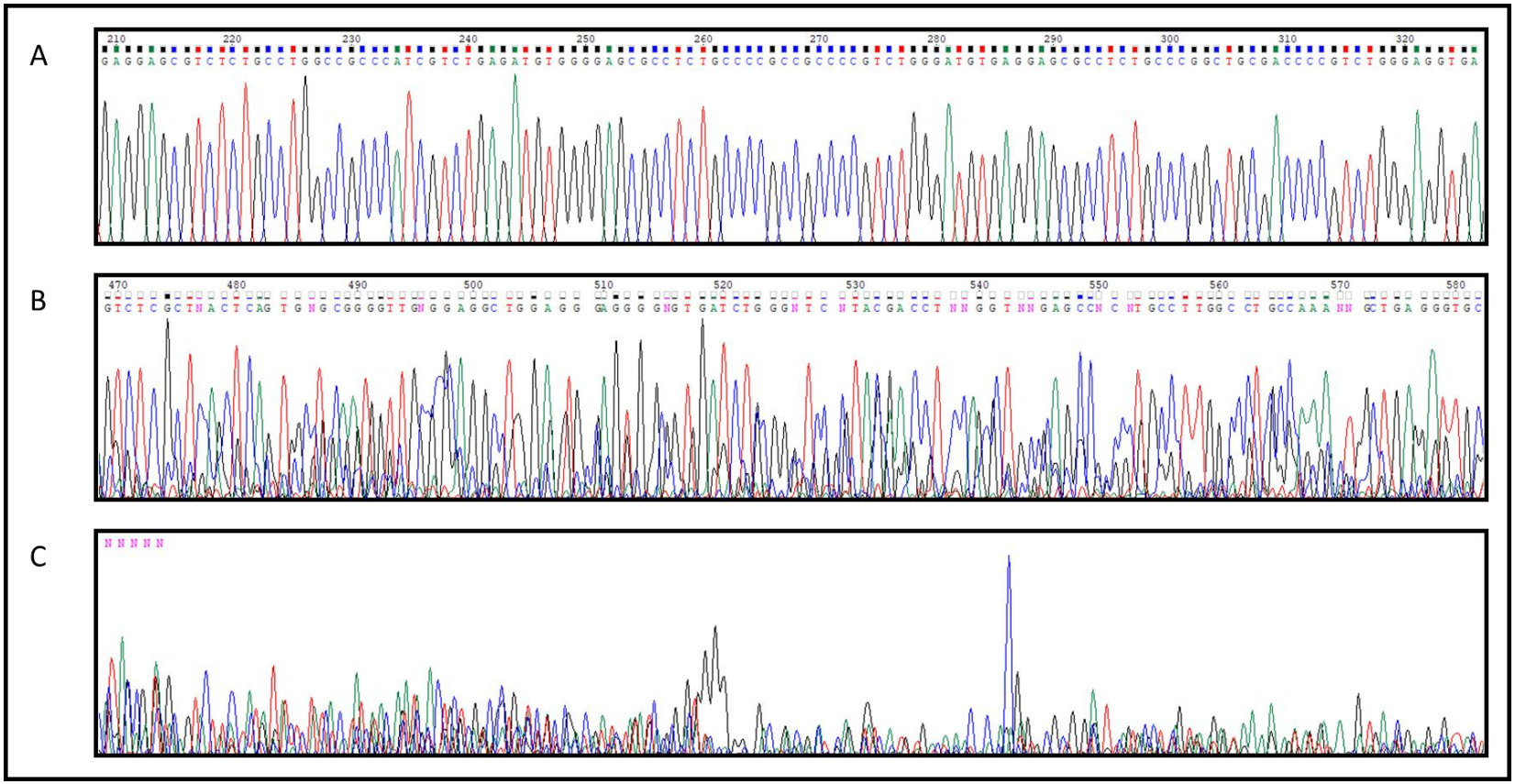


Figure 2.9. Sanger Sequencing – The Good the Bad and the Ugly.

A: High quality sequencing read. **B**: Bad quality sequencing read and noisy background possibly due to clone contamination. **C**: Failed sequencing reaction possibly due to low quality of DNA prep. All DNA sequences were visualised using Chromas (http://technelysium.com.au/wp/chromas/).

2.2.7 Cell Culture

2.2.7.1 Changing media and passaging cells

All adherent mammalian cell lines were cultured in media outlined in Section 2.1.4 in T75 flasks and incubated at 37 °C and 5% CO_2. Cells were passaged when confluency was between 70-90% and transferred into fresh sterile T75 flasks. Culture media, PBS and trypsin were prewarmed to 37 °C before use. The media was gently aspirated from the flask and the cells were washed with 10 ml of sterile phosphate-buffered saline (PBS) (Gibco). After washing, PBS was then gently aspirated away and 1 ml of trypsin (Sigma) was added and the flask was gently rocked to allow the trypsin to cover the entire internal surface of the flask and was then incubated at 37 °C for approximately 2-3 minutes. Once the cells had detached from the internal surface, the trypsin was deactivated through the addition of 10 ml of fresh culture media containing serum. This media was then washed up and down the internal surface of the flask to ensure all cells had detached. The detached mixture of media and cells was then transferred to a sterile 15 ml falcon tube and centrifuged at 130 g for 5 minutes at room temperature. The supernatant was carefully aspirated away and then cell pellet was then resuspended into 10 ml of fresh culture media. From this, 1 ml of the cell mixture was transferred to a fresh and sterile T75 flask containing 19 ml of prewarmed culture media.

2.2.7.2 Freezing cells in liquid nitrogen

All mammalian cell lines used were stored in freezing media outlined in Section 2.1.4.6 and kept in liquid nitrogen for long term storage. Once a T75 flask had

reached 70-90% confluency, the cells were washed and passaged as previously described in Section 2.2.7.1. The cells were then resuspended in 10 ml of freezing media and 1 ml of cells was transferred into each cryovial. These cryovials were then transferred to a room temperature Mr Frosty™ freezing container (Thermo Scientific) and stored at -80 °C for 24-48 hours before being moved to liquid nitrogen for long term storage.

2.2.7.3 Counting cells on a haemocytometer

For individual experiments, prior to being plated, the number of cells per ml was calculated by counting the cells on a haemocytometer. The haemocytometer and cover slip were both cleaned with 70% ethanol before and after use. Once cells reached 70-90% confluency, they were washed and passaged as previously described in Section 2.2.7.1. Following resuspension of cells with 10 ml of fresh media, 10 µl was transferred to a sterile Eppendorf tube and then mixed with 10 ul of Trypan Blue to stain any dead cells: living cells with an intact cell membrane are not stained. 10 µl of cells was then added to the haemocytometer and visualised under a 10x objective of a light microscope. The counting surface of the haemocytometer is composed of 4 sets of 16 squares. The external border consists of two parallel lines and any cells on these borders were consistently included in all counts. The mean of the 4 sets of squares was then calculated, multiplied by 2 for the dilution factors and then multiplied by 10,000 to calculate the number of cells per ml.

2.2.7.4 DNA extraction from cultured cells

Genomic DNA was extracted and isolated from mammalian cell lines using the GenElute™ Mammalian Genomic DNA Miniprep kit (Sigma) according to the manufacturer's instructions. Purified DNA was then subject to quality checks and quantification using the NanoDrop™

2.2.8 Transfection of plasmid DNA into cultured cells

All DNA constructs were transiently transfected into mammalian cell lines using the high efficiency cationic polymer, Turbofect™ Transfection Reagent (Thermo Fisher). Cells were plated 24 hours prior to transfection and the protocol was performed according to manufacturer's instructions for a 24-well plate format (Table 2.6). Due to high cellular toxicity, the media was changed 4 hours after the transfection mixture was added to the cells. All transfected cells were then incubated at 37 °C and 5% CO_2 for 48 hours before any downstream applications were performed.

Table 2.6. Transfection for cells used in luciferase assay

Transfection setup for luciferase assay (24-well format).

Reagent	Volume per well (µl)
Serum free media (media dependent on cell line used)	100 µl
pGL3/pSHM06 construct containing DNA insert (experimental construct) (500 ng/µl)	2 µl (1 µg)
Firefly/Renilla luciferase construct (control construct) (10 ng/µl)	2 µl (20 ng)
Turbofect™ Transfection Reagent	2 µl

2.2.9 Luciferase reporter gene assays

The pGL3-P and pGL3-B constructs contain the Firefly (*Photinus pyralis*) luciferase gene, which expresses the luciferase enzyme (Figure 2.2). This enzyme converts luciferin into oxyluciferin and the reaction generates light which can be detected and measured by a luminometer. The pSHM06 vector contains the Renilla (*Renilla reniformis*) luciferase gene which also produces the luciferase enzyme. In this chemical reaction, coelenterazine is converted to coelenteramide and produces light. Each assay performed in pGL3 vectors (Promega) included pRL-TK vector (Promega) as an internal control (expressing Renilla luciferase), whereas all assays performed in pSHM06 vector included pMLuc-2 vector[145] as an internal control (expressing Firefly luciferase), accounting for cell death and transfection efficiency. All assays were done in triplicate.

2.2.9.1 Cell lysis

All Luciferase reporter gene assays were performed using the Dual-Luciferase® Reporter Assay system (Promega). Cells were plated in 24-well plate format 24 hours before transfection and then left 48 hours before reporter gene expression analysis. All adherent mammalian cell lines were lysed with diluted (1X) passive lysis buffer (PLB) (100 µl per well) and left at room temperature on a rocker for 15 minutes. The lysed cell mixture was then transferred to an opaque 96-well plate and placed into a dual injector GloMax® luminometer (Promega).

2.2.9.2 Measuring reporter gene expression

The two luciferase reporter reagents were prepared during the 15-minute cell lysis stage mentioned previously. Luciferase Assay Reagent II (LAR II) (used to measure

firefly luciferase) was made by resuspending lyophilised Luciferase Assay Substrate in 10 ml of Luciferase Assay Buffer II. A 1X Stop & Glo® (SAG) reagent (used to measure Renilla luciferase) was made by adding 1 part of 50X Stop & Glo® Substrate to 49 parts of Stop & Glo® Buffer: for 100 assays add 200 µl of 50X Stop & Glo® Substrate to 9800 µl of Stop & Glo® Buffer. Following transfer of lysate into 96-well plates the GloMax® luminometer (Promega) dual injectors were cleaned with 70% ethanol and distilled water and then primed with the LAR II and Stop & Glo® reagents to ensure there were no air bubbles present in either injector. The plate was placed within the luminometer and injectors 1 and 2 were set to dispense 100 µL of LARII and SAG reagent respectively into each well, with a 1.5 seconds integration time between each reading. 100 µL of LARII was dispensed by injector 1 and Firefly luciferase activity was measured; after this 100 µL of SAG was dispensed by injector 2 and Renilla luciferase activity was measured. This format was performed for each well containing lysate and results were collated within an Excel spreadsheet.

2.2.10 CRISPR

2.2.10.1 Guide RNA design

CRISPR guide RNAs (gRNA) were designed using the gRNA design tool (http://crispr.mit.edu/). The DNA sequence for chr4:170489609-170493323 was obtained from UCSC (hg19) and uploaded to the http://crispr.mit.edu/ bioinformatic tool, which scanned for 20 bp sequences upstream of a protospacer adjacent motif (PAM) sequence (NGG). These oligos were scored out of 100 based on sequence specificity, with higher scoring oligos having less homology with other

sequences in the genome and therefore being ideal candidates for CRISPR guides, ensuring fewer off target modifications. A total of four guides were designed, two either side of the SVA within *NEK1*. These guides were then ordered from Sigma and then cloned into pSpCas9(BB)-2A-GFP vector by Golden Gate cloning. The queried guide and complement sequence were modified to contain a 5' CACC and 5' CAAA sequence respectively, enabling compatible ends for cloning into the pSpCas9(BB)-2A-GFP vector (Figure 2.10). The top and bottom strands were then resuspended in nuclease free water and annealed together: 6 µl of the sense and anti-sense single stranded sequence was added to 83 µl of nuclease free water and 5 µl of ligase buffer and heated at 95 °C for 5 minutes.

Table 2.7. Example list of guide oligos for CRISPR

Guide	88	ATAGGAAGGCCCGCTGGACG**TGG**	(Fw)
Guide	88	ATATAAATAGGAAGGCCCGC**TGG**	(Fw)
Guide	75	GGAAGGCCCGCTGGACGTGG**TGG**	(Fw)
Guide	69	TCATTGGTATGAAATCAGTC**AGG**	(Rv)

Forward guides

Guide 1

5′ CACCGAAATGCGTGTGTAAATACTG 3′

3′ CTTTACGCACACATTTATGACCAAA 5′ (complement)

Guide 2

5′ CACCGATAGGAAGGCCCGCTGGACG 3′

3′ CTATCCTTCCGGGCGACCTGCCAAA 5′ (complement)

Reverse guides

Guide 3

5′ CACCGTGTTGTTGTTCACTAGTTCC 3′

3′ CACAACAACAAGTGATCAAGGCAAA 5′ (complement)

Guide 4

5′ CACCGGAAACGTGGGAATGGCGTAG 3′

3′ CCTTTGCACCCTTACCGCATCCAAA 5′ (complement)

Figure 2.10. Designed guide oligos with modifications for Golden Gate cloning

2.2.10.2 Golden Gate cloning

Each guide was cloned into pSpCas9(BB)-2A-GFP vector using the Golden Gate cloning strategy. The pSpCas9(BB)-2A-GFP vector was digested with *BbsI*, removing a small insert and leaving overhangs which were compatible with the modifications added to each guide (Figure 2.10), facilitating insertion of each guide into the pSpCas9(BB)-2A-GFP vector. T4 ligase was also added to the same reaction, allowing restriction enzyme digestion and ligation within one reaction (Table 2.8). If the original insert was ligated back into the vector then the *BbsI* site would be

intact and would be cut again in the reaction. However, insertion of the guide destroyed the *BbsI* site and therefore the guide remained inserted and could not be excised out. The reaction was incubated at 37 °C for 5 minutes, followed by 10 cycles at 16 °C for 10 minutes, then 37 °C for 30 minutes and finished with a final heat activation at 80 °C for 20 minutes. A total of 2 µl was taken from this reaction and used for transformation of chemically competent DH5α *E.coli*, previously described in Section 2.2.3.5.

Table 2.8. Golden Gate cloning reaction master mix

Nuclease Free Water	X ul
plasmid - pSpCas9(BB)-2A-GFP	Y ul (150 ng)
Annealed Oligos	3 ul (150 ng)
Ligase buffer 10X	2 µl
Enzyme - Bbs I - HF	1 µl
T4 ligase	1 µl
Total Volume	**20 µL**

2.2.10.3 Screening successful guide cloning

Transformed cells were then picked and expanded to 5 ml cultures and DNA was then extracted as previously described. All samples were then digested with *BbsI*-HF (NEB). If the guide was successfully cloned then the *BbsI* cut site should have been destroyed during the ligation process and therefore the clone should remain uncut (Figure 2.11). All samples were also confirmed through sequencing.

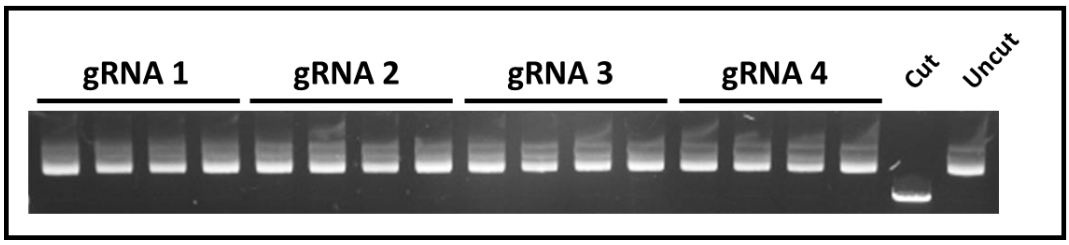


Figure 2.11. Guide RNA cloning verification

Agarose gel electrophoresis of CRISPR gRNA samples which have been successfully cloned into the pSpCas9(BB)-2A-GFP vector using the Golden Gate cloning strategy. Ligation of the guide into the vector destroys the *BbsI* cut site and thus successful cloning is indicated by an uncut sample when digested with *BbsI*.

2.2.10.4 Single cell seeding and clonal expansion

Each combination of forward and reverse gRNAs were co-transfected (gRNA 1 with gRNA 3; gRNA 1 with gRNA 4; gRNA 2 with gRNA 3; gRNA 2 with gRNA 4) into HEK293 cells using Turbofect™ Transfection Reagent (Thermo Fisher) in a 24-well format (100,000 cells per well) (Table 2.9), as previously described in Section 2.2.8. Media was then changed after 4 hours and after 48 hours cells were then counted as previously described in Section 2.2.7.3 and then seeded at 1000 and 2000 cells in sterile TC treated 10 cm petri dishes containing 10 ml of fresh sterile culture media to obtain well-spaced populations of cells. Media changes were done every 2-3 days and cells were left to grow until visible colonies had formed (approximately 14 days). These colonies were then picked with a sterile pipette tip and placed into a 96-well plate.

Table 2.9. Transfection setup for SVA KO CRISPR cells

Reagent	Volume per well (µl)
Serum free media (DMEM, high glucose)	100 µl
CRISPR-Cas9 vector containing 5' gRNA (guide 1 or 2) (250 ng/µl)	2 µl (500 ng)
CRISPR-Cas9 vector containing 3' gRNA (guide 3 or 4) (250 ng/µl)	2 µl (500 ng)
Turbofect™ Transfection Reagent	2 µl

2.2.10.5 CRISPR clone crude lysis

DNA was extracted from clonally expanded cell lines using the DirectPCR® Lysis Reagent (Viagen Biotech), supplemented with Proteinase K (Sigma) (1:100 dilution of reagent to proteinase K). All cells were lysed in 96-well plate format. Firstly, the media was gently aspirated, and cells were washed with 100 µl sterile PBS (Gibco) which was then carefully removed. A total of 50 µl of lysis reagent and proteinase K mixture was added to each well, mixing up and down with the pipette to aid lysis. Plates were then moved to a hybridisation oven and left at 55 °C overnight. The following day, all samples were heat inactivated at 85 °C for 45 minutes. Following lysis, 1 µl of sample was then used as a template for the *NEK1* SVA CRISPR KO region PCR (reaction outlined in Table 2.2).

2.2.10.6 Extraction of gDNA from cell lines

Once lines with successful modifications were identified by PCR, DNA was extracted from these lines using the GenElute™ mammalian genomic DNA miniprep kit (Figure 2.12). Cells were harvested, resuspended, lysed and DNA was then purified on a column according to manufacturer's instructions and eluted in 100 µl of Ultra Pure™ DNase/RNase free distilled water (Invitrogen). This step was performed to confirm that the results of the crude lysis were correct.

2.2.10.7 Genotyping CRISPR clones

Both crude lysis and purified DNA templates were genotyped using KOD Xtreme™ Hot Start DNA Polymerase (Merck): an ultra-high fidelity enzyme suitable for long, GC rich templates and crude sample PCR. Please see Table for details on master mix and thermal cycles.

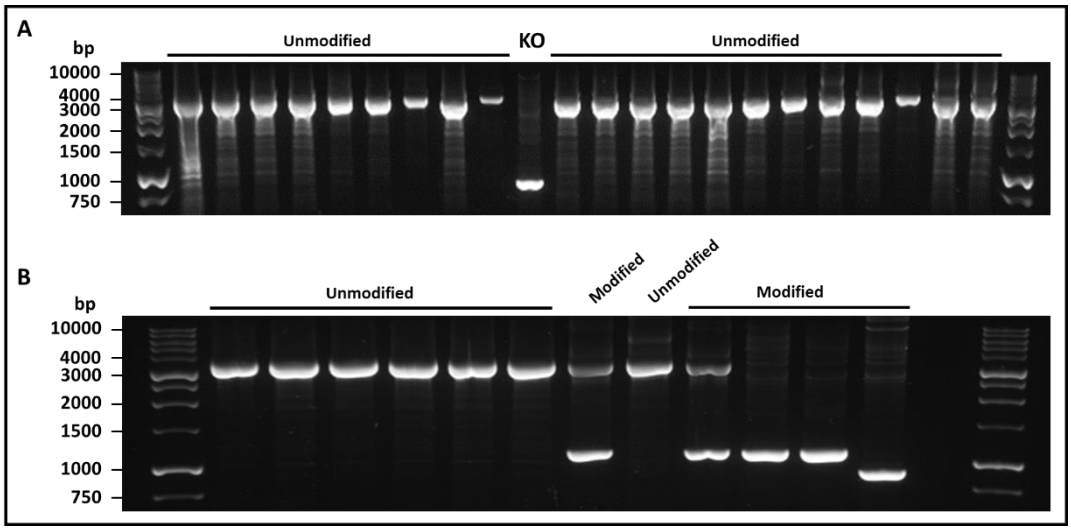


Figure 2.12. Comparison of CRISPR genotyping results

Gel agarose electrophoresis of the *NEK1* SVA-D CRIPSR KO region, validating SVA KO. **A**: PCR template generated using DirectPCR Lysis Reagent (Viagen). **B**: PCR template generated using the GenElute™ Mammalian Genomic DNA Miniprep kit (Sigma). Unmodified amplicon=3432 bp, modified amplicon=1089 bp, KO=knock out.

2.2.11 mRNA expression analysis

2.2.11.1 RNA extraction and quality control

RNA was isolated from mammalian cell lines using the Monarch Total RNA Miniprep kit (NEB) according to manufacturer's instructions. RNA quality and quantity was assessed using a NanoDrop™ spectrophotometer (Thermo Scientific) as previously described in Section 2.2.5. RNA integrity was also assessed by running

each sample on an agarose gel (Figure 2.13), also as previously described in Section 2.2.5.

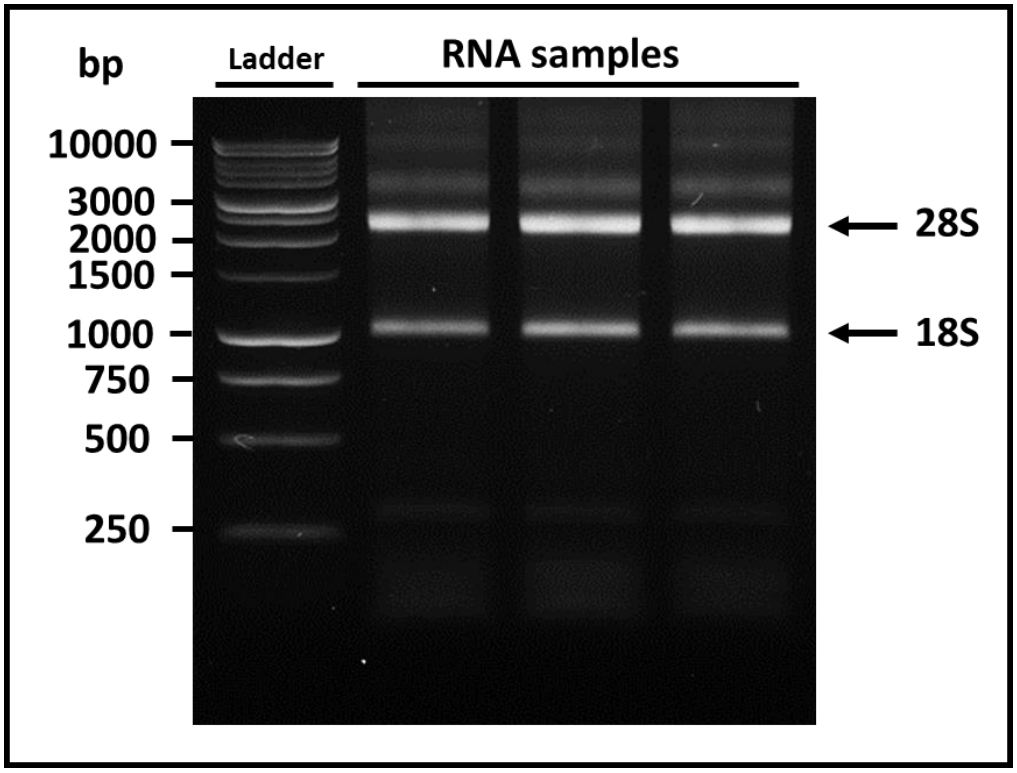


Figure 2.13. RNA integrity analysis

Gel agarose electrophoresis of RNA samples to check for degradation of ribosomal RNA (rRNA) subunits. Clear bands indicate that the 28S and 18S subunits are intact and degradation has not occurred.

2.2.11.2 cDNA synthesis

Once quality control was completed RNA samples were then converted to complementary DNA (cDNA) using the GoScript™ Reverse Transcription System (Promega). All RNA samples were normalised prior to the conversion (all were diluted to 70 ng/µl with nuclease free water): a total of 3 µl of RNA sample (equating to no more than 500 ng) was then incubated with 2 µl of a 50:50 mixture of random hexamers and oligo $(dT)_{15}$ primers. The mixture was used to ensure improved cDNA synthesis: oligo $(dT)_{15}$ primers have specificity for the 3' Poly A tail

while the hexamers will bind randomly, helping to improve full coverage of the transcripts. The RNA and primer mixture (Table 2.10) was briefly mixed and centrifuged. All tubes were then placed in a heat block at 70 °C for 5 minutes and then immediately chilled on ice for 5 minutes to facilitate primer annealing. All tubes were then centrifuged for 10 seconds to collect condensate. All samples were then kept on ice while the reverse transcription reaction (Table 2.11) was set up.

Table 2.10. cDNA synthesis reaction setup

Reagent	Volume (µl)
RNA (up to 5µg/reaction)	3.0
Primer [Oligo(dT)$_{15}$ (0.5µg/reaction) and Random Primer (0.5µg/reaction)	2.0

The 5 µl RNA/primer mix was then combined with 15 µl of reaction mix (Table 2.11) and placed in a thermal cycler at 25 °C for 5 minutes (annealing step). Following this, the samples were incubated at 42 for an hour (extension step). After this the samples were incubated at 70 °C for 15 minutes (reverse transcriptase heat inactivation step).

Table 2.11. Reverse transcription reaction

Reagent	Volume (µl)
GoScript™ 5X Reaction Buffer	4.0
$MgCl_2$ (25 mM)	3.0
PCR Nucleotide Mix (10 mM)	3.0
Recombinant RNasin® Ribonuclease Inhibitor	0.5
GoScript™ Reverse Transcriptase	1.0
Ultra Pure™ DNase/RNase Free Distilled Water	3.5

2.2.11.3 RT-PCR

GoTaq® Hot Start DNA Polymerase (Promega) was used for Reverse Transcription Polymerase Chain Reaction (RT-PCR) of the cDNA targets of interest for gene expression analysis. Please refer to Table 2.2 for master mix and thermal cycle details of all RT-PCR experiments.

2.2.11.4 qPCR

2.2.11.4.1 Assay setup

All qPCR experiments were performed using the GoTaq® qPCR system (Promega) using the master mix outlined in Table 2.12. All reactions were set up in triplicate, run on the Mx3005P Real-Time PCR System (Stratagene) and amplification plots, dissociation curves and text reports containing cycle threshold (Ct) values were generated for each experiment. Thermal cycler conditions for qPCR experiments were the same as the RT-PCR experiments (*NEK1* and *CLCN3*) and can be found in Table 2.2.

Table 2.12. qPCR reaction mix

Reagent	Supplier	Volume (µl)
cDNA (1:20)	N/A	2.0
GoTaq® qPCR Master Mix (2X)	Promega	5.0
Fw primer (10 µM)	N/A	0.1
Rv primer (10 µM)	N/A	0.1
CRX reference dye	Promega	0.1
Ultra Pure™ DNase/RNase Free Distilled Water	Invitrogen	2.7

2.2.11.4.2 Testing primer efficiencies

Amplification efficiency of *NEK1*, *CLCN3* and *ACTB* primers was obtained through serial dilution of template cDNA. All obtained Ct values were plotted on a log scale of the corresponding concentrations and the slope of the trend line was calculated.

The primer efficiency was then calculated from the slope of the trend line using the following equation: $\text{Efficiency}\,(\%) = \left(10^{\frac{-1}{slope}} - 1\right) x100$. All primer efficiencies for *NEK1*, *CLCN3* and *ACTB* can be found in Figure 5.13.

2.2.11.4.3 Relative quantification of gene expression

Relative gene expression was measured by qPCR. All gene expression experiments were performed using the delta delta Ct (ΔΔCt) method. The following steps were performed in Microsoft Excel:

1. Calculate mean Ct for both target and reference gene from technical replicates.
2. Calculate delta Ct (ΔCt): ΔCt = Ct (target gene) – Ct (reference gene).
3. Calculate mean of ΔCt control (non-target guide) from technical replicates.
4. Calculate delta delta Ct (ΔΔCt): ΔCt (sample) – ΔCt (control average).
5. Calculate fold expression using: 2^-(ΔΔCt).

2.2.12 Bioinformatic Analysis

2.2.12.1 UCSC Genome Browser

All VNTRs and SVAs studied in this thesis were first examined using the UCSC Genome Browser (https://genome.ucsc.edu/). This is a genome browser containing a comprehensive set of tools and tracks which can be overlaid within a graphical interface, facilitating visualisation of multiple datasets within a defined genomic region. In particular we used RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org), a program used to screen and identify simple tandem and low complexity DNA repeats, which annotates

transposons and tandem repeat DNA within the reference genome. RepeatMasker annotations were overlaid with conservation data, specifically the vertebrate multiz alignment & conservation of 100 vertebrate species from Phylogenetic Analysis with Space/Time models (PHAST program)[311], allowing us to determine if the retrotransposons and tandem repeat DNA of interest was human specific or present in other primates. Routinely we also addressed tracks containing information from the Encyclopedia of DNA Elements (ENCODE)[312], used to identify DNA with regulatory function. Specifically, we looked at layered H3K4Me1, H3K4Me3 and H3K27Ac histone marks: which are found near regulatory regions, promoters, and active regulatory elements, respectively. Transcription factor ChIP-seq clusters (Txn Factr ChIP E3) was also used to identify transcription factor binding sites around regions of interest. Predicted transcripts from Ensembl were also assessed for *CFAP410*, which were then queried on Genotype-Tissue Expression (GTEx) project portal: a publicly available database and tissue bank resource which can be used to assess specific gene expression levels in 54 non-disease tissues[313].

2.2.12.2 ECR Browser

To determine conserved non-coding regions of *NEK1*/*CLCN3* locus was visualised on ECR browser[314]. This is a genome alignment tool designed to identify evolutionary conserved regions (ECRs) across multiple species genomes, allowing the user to automatically align a genomic sequence of choice to multiple vertebrate genomes to determine conservation. The coordinates of the human *NEK1* and *CLCN3* genes (hg19) were submitted to ECR browser and this was aligned to multiple vertebrate genomes, including chimp, macaque, mouse, rat, xenopus,

chicken, and zebrafish. Conservation peaks found within introns, repetitive elements and intergenic space of other vertebrates were assessed. This was used to identify non-coding regions within *NEK1* or *CLCN3* which were conserved across multiple species, to identify non-coding regions which could be regulatory in function (Supplementary Figure 3).

2.2.12.3 Rosalind HPC cluster and cloud server

All bioinformatic data was stored and analysed on the High Performing Computing (HPC) cluster, Rosalind (https://rosalind.kcl.ac.uk), which uses a Linux based command line. This research computing infrastructure was used to generate and assess all Isaac Variant Caller data (please refer to Chapter 4 Sections 4.3.7-4.3.12).

2.2.12.4 Isaac Variant Caller data analysis and manipulation

Isaac Variant Caller (IVC) is a genetic polymorphism/variant caller program developed by Illumina and is used as part of their Isaac Whole Genome Sequencing (WGS) v2 pipeline for calling of single nucleotide polymorphisms (SNPs) and small indels (insertions and deletions)[315]. IVC is a multistep process, first of which involves read filtration, removing reads which fail quality checks. Secondly, candidate indels are identified by utilising a multiple sequence aligner: indels are found as gaps in the alignment and are examined based on number of reads containing it. SNPs are then called, and a probability of each genotype is computed on the basis of the aligned reads and prior genome distribution. Indels are then also genotyped, assigning a probability of all possible genotypes based on the read data[315,316]. The output from IVC is a genome variant call format (VCF) file: a text file format which stores DNA sequencing data, specifically genetic

polymorphisms/variants, including SNPs, indels and structural variants[317]. All IVC data was already generated by Illumina, via the Isaac WGS pipeline[316], as part of Project MinE[128]. Please refer to Chapter 4 for pipelines of the optimisation process used in the extraction and analysis of IVC results from Project MinE whole genome sequencing data (Figure 4.9 and Figure 4.10).

All IVC data was analysed via PuTTY, an open source emulator which was used to remotely access the Rosalind server. All BASH (Bourne Again Shell) scripts were run through command line and were generated over a period of 3 weeks, thanks to training from Dr Alfredo Iacoangeli and Dr Ashley Jones (Kings College London). All data generated from IVC was analysed and manipulated using several programs and commands which are part of SAMtools: a software package which can be used to manipulate Sequencing Alignment Map/Format (SAM) files, which are used to store aligned next generation sequencing reads[318]. Firstly, the *NEK1* genomic region was extracted from the genome VCF files of all ALS patients and controls within the UK dataset of Project MinE (n=1784) using bcftools (a set of commands used to manipulate VCF files and a component of SAMtools[319]). All files were then compressed using bgzip and indexed using tabix[320].

Once all files were available, the *NEK1* SVA CT element region was specifically extracted using the same process described above. All VCF files were then merged and opened within Excel, and samples were filtered down to those with alternative (ALT) calls matching the size of both allele 1 and 4 of the *NEK1* SVA CT element (n=13). These 13 VCFs were visually inspected on Rosalind using the zless command to confirm the presence of allele 1 or 4 (Figure 4.13). Once the rare CT

element variants were validated in these 13 ALS patients, we then assessed whether these patients contained any coding mutations within *NEK1*. To achieve this, the previously extracted *NEK1* (entire locus) VCF files were inspected for the presence of missense, stop codon gained, stop codon lost, splice variant and frameshift variants (using the view -i command within bcftools). All BASH scripts generated in this study can be found in Appendix 2.

2.2.12.5 dbSNP

To determine minor allele frequency (MAF) of *NEK1* coding variants in ALS cases each SNP was queried on dbSNP[321](https://www.ncbi.nlm.nih.gov/snp/). This is a publicly available database of nucleotide variant data from the National Center for Biotechnology Information (NCBI), compromising data from multiple cohorts including TOPMED, 1000 Genomes project, TWINSUK and GnomAD. This database was used to identify MAF of *NEK1* coding mutations which were present in the 13 ALS patients within UK dataset of Project MinE which contain rare SVA CT element variants (Table 4.6). These 13 ALS patients were then assessed for the presence of known *NEK1* mutations which confer risk for ALS (Supplementary Table 2) to therefore determine if the rare SVA CT elements were inherited with previously discovered ALS risk variants.

Chapter 3: Determining genetic variation and transcriptional activity of variable number tandem repeats (VNTRs)

3.1 Introduction

Tandem repeats are a large source of genetic variation, with variable number tandem repeats (VNTRs) accounting for approximately 3% of the human genome[133,322]. Yet these polymorphisms are often disregarded as they are very difficult to characterise and map using short read sequencing and are missed by genetic association studies which focus solely on single nucleotide variants[140]. Tandem repeat polymorphisms have been implicated in the increased susceptibility and risk of several diseases, such as Alzheimer's disease (AD), hepatocellular carcinoma, depression and schizophrenia[150,179,180,323]. Likewise, expansions of tandem repeats can also cause disease, including Fragile X Tremor Ataxia Syndrome, Myotonic Dystrophy, Huntington's disease, Spinocerebellar Ataxia, and ALS[41,42,158-167]. As a lab we are interested in VNTR polymorphism detection and testing VNTR function *in vitro* and *in vivo*, assessing their capacity to serve as fine tuners of transcription. VNTRs have previously been identified as risk factors for disease, but are particularly difficult to characterise by PCR and cloning (particularly GC-rich repeats). This chapter highlights the optimisation of such techniques to characterise polymorphic variants that are potentially risk factors for ALS.

This chapter is split into two projects which are related to neurodegeneration, constituting assessment of two distinct VNTRs within two different genes. The first project focussed on work initiated by a previous member of our lab, Dr Maurizio Manca, investigating a VNTR found within the promotor

region of the Repressor Element 1 Silencing Transcription Factor (*REST)* gene: the second focussed on a VNTR within the *CFAP410* gene.

REST, also referred to as neuron-restrictive silencing factor (NRSF), is a zinc finger transcription factor (TF) which regulates gene expression via chromatin remodelling[324,325]. REST contains two key repressor domains (RDs); the N-terminal RD1 which interacts with corepressor mSin3 and the C-terminal RD2 which associates with the REST/NRSF corepressor, CoRest[325]. Mechanistically this TF binds a 21 bp repressor element/neuron restrictive silencer element (RE1/NRSE) within neuronal genes, recruiting the corepressors CoRest and mSin3A, leading to the recruitment of chromatin remodelling factors such as histone deacetylases (HDACs), methyltransferases and demethylases which facilitate chromatin plasticity and ultimately induce repression and silencing of neuron-specific genes[326,327]. Initially, *REST* was known as a master transcriptional regulator of neurodevelopment[326]. REST has been shown to repress neuronal genes in non-neuronal cells and during embryogenesis, also modulating neurogenesis related genes within stem cells and neural progenitors and playing an important role in regulating neural stem cell renewal capacity, neuronal precursor differentiation and dictating neuronal cell phenotype specificity[326,328-331]. However, more recently it has been shown that *REST* is also expressed in differentiated neurons and is involved in modulating synaptic plasticity of N-methyl-D aspartate receptors (NMDARS), which are important in synaptogenesis, neural communication and high cognitive functions such as memory and learning[332,333]. Ultimately, this transcriptional regulator is an important fine tuner of gene expression, acting in a

context dependent manner as both an activator and repressor of genes involved in neuronal development, differentiation and survival[325].

Dysregulation of *REST* has been associated with a number of neurological conditions, including Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD), Down's syndrome, epilepsy and schizophrenia (SZ)[326,328]. Mouse models have helped to elucidate phenotypic changes in response to the dysregulation of *REST*, serving to better understand its potential role in disease. Lu *et al*., have shown that overexpression of human *REST* (*hREST*) in adult mice leads to a significant reduction in spontaneous locomotor activity, specifically reducing speed and total distance travelled[334]. Conditional knockout of *REST* in mice has previously been shown to regulate seizure acceleration and activity[335,336], with another study observing that knockout of *REST* increases cell death induced by the dopaminergic neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) in a PD mouse model[337]. Recent *in vitro* expression analysis by Lu *et al*. has shown that *REST* is significantly increased at both the mRNA and protein level in the prefrontal cortex of aged individuals (73-106 years) when compared to young adults (20-35 years). However, they have also shown through immunocytochemistry and western blotting that nuclear REST protein levels were significantly reduced in neurons of affected brain regions in AD patients, including the prefrontal cortex and CA1, CA2 and CA3 regions of the hippocampus; however this reduction was not seen in dentate gyrus granule cells and cerebellar Purkinje cell neurons. Furthermore, this study assessed *REST* targets by performing ChIP-seq in SH-SY5Y cells and found a significant enrichment of REST binding sites in genes associated with AD and cell death pathways. Regulation of these genes was

assessed in normal ageing brain and AD; within the AD population it was found that REST binding of these AD pathology and cell death pathway genes was reduced and mRNA of these genes was elevated. Loss of nuclear REST in cortical and hippocampal neurons of AD cases was also found to be accompanied by the localisation of cytoplasmic structures containing autophagosome markers including microtubule-associated protein 1A/1B-light chain 3 (LC3). Moreover, autophagy activation in SH-SY5Y cells led to a reduction of nuclear REST. Analysis of cortical sections from cases with FTD or dementia with Lewy bodies (DLB) also showed depleted levels of nuclear REST protein and immunofluorescence microscopy in these sections identified that REST co-localised with pathogenic misfolded proteins. In cortical and hippocampal neurons, REST was found to localise with amyloid beta (Aβ) in LC3-positive autophagosomes of AD cases, with TDP-43 in TDP-43-positive FTD patients, with phosphorylated tau in tau-positive FTD patients and with α-synuclein-positive autophagosomes of DLB cases. Overall, this study by Lu *et al*. suggested a neuroprotective role of REST and its dysfunction could be a common mechanism contributing to the pathogenesis of AD, FTD and DLB[338].

REST has not been extensively studied in the context of ALS, however Rockowitz and Zheng in 2015 showed that human specific REST binding sites were enriched in ALS and oxidative stress (OS) genes. Furthermore, through ChIP-seq they assessed REST occupancy in both mouse and human embryonic stem cells (ESCs), identifying approximately 1200 core synteny regions with REST binding capacity across the two species, referred to as REST/NRSF cistromes. However, the majority of REST binding sites were species specific, as most REST bound genes in

human ESCs were not targeted in mouse ESCs (2995 of 4480 genes). This human ESC REST bound gene cistrome was found to be enriched for the processes of memory and learning as well as pathways involved in axon guidance. Rockowitz and Zheng also assessed REST binding sites in genes associated with idiopathic AD, PD, HD, SZ, ALS, OS, autism spectrum disorder (ASD) and intellectual disorders (ID). Aside from ID, all other disorders were enriched for REST binding sites which they postulate is due to the enrichment of neuronal genes in REST target sites. However, human specific REST binding sites were only found to be enriched in OS and ALS, hypothesising that newly emerged human specific REST binding sites could play a role in these disorders. Furthermore, they also found four REST binding peaks upstream of ALS genes and two peaks upstream of OS genes which were overlapped by TEs, which they hypothesise could induce expansion of RE1 sites and REST binding and thus promote human-specific regulation of *REST*[324].

Cilia and flagella associated protein 410 (*CFAP410*; previously known as *C21orf2*), is a gene recently associated with ALS risk through GWAS[11], encoding a leucine rich repeat (LRR) protein that is localised within the primary cilium[339,340]. The *CFAP410* gene was discovered in 1998 through exon trapping and cDNA library sequencing, yet the protein encoded by this gene still remains to be extensively characterised[341]. A study by Lai *et el*. showed that small interfering RNA (siRNA) knockdown (RNA interference) of *CFAP410* led to cilium defects and thus suggested a role for this protein in cilia maintenance and formation[342]. Previous work has hypothesised that primary cilia dysfunction in motor neurons could be involved in ALS pathogenesis. Ma *et al*., assessed co-staining of SMI32 (a motor neuron marker) and adenylyl cyclase 3

(a primary cilia marker[343]) to quantify the number of ciliated motor neurons in both wildtype (WT) and G93A SOD1 mice. They discovered that G93A mice had reduced number of motor neurons with primary cilia compared to WT, which they hypothesised was inducing a reduction in sonic hedgehog (Shh) signalling[344]. A later study by Ma *et al*., confirmed that Shh treatment (250 and 500 ng/ml) in WT and G93A SOD1 mouse motor neurons led to trophic effects, leading to a significant increase in total number of motor neurons compared to untreated cultures. This treatment also led to significant increases in neuronal progenitor proliferation, motor neuron survival and cell differentiation into motor neurons, suggesting that Shh could be of therapeutic use in ALS. Furthermore they found that treatment with Shh led to significant increase in both total motor neuron number and ciliated motor neuron number: the percentage increase in ciliated motor neurons was higher than the increase in total motor neurons, suggesting that primary cilia are integral to efficient Shh signalling[345].

CFAP410 is also now known to be involved in DNA repair. Fang *et al*. have found that *CFAP410* depleted HeLa cells are more sensitive to DNA damage and are less efficient at repairing double strand breaks compared to control cells, specifically affecting homologous recombination repair. Furthermore, overexpression of *NEK1* in *CFAP410* depleted cells led to a significant rescue of defects in DNA damage repair, indicating that these proteins potentially work together within the NR DNA repair pathway[346]. This LRR protein has proven to be an essential player in ciliogenesis and an important modulator of DNA damage repair.

The subsequent focus of *CFAP410* research has been its implications in disease, with mutations in this gene being linked to a number of disorders[11,339,340,347-349]. *CFAP410* has previously been identified as a candidate disease gene for a number of ciliopathies such as retinal dystrophy[347]. Arif *et al*. performed genetic testing on three unrelated patients with early-onset retinal dystrophy with macular staphyloma and found that all three harboured homozygous *CFAP410* mutations: two with a frameshift deletion c.436_466del (p.Glu146Serfs*6) and one with a missense mutation c.182G>A (p.Cys61Tyr). Furthermore, they found that these recessive mutations were the cause of the retinal dystrophy phenotype, causing similar retinal phenotype changes and macular staphyloma in all three patients[339]. Suga *et al*. identified missense *CFAP410* mutations in Japanese patients with autosomal recessive cone-rod dystrophy (arCRD) and autosomal recessive retinitis pigmentosa (arRP)[340]. Moreover, two of these mutations, c.319T>C (p.Tyr107His) and c.331G>A (p.Val111Met) were tested *in vitro* (HEK293T cells) and when compared to wild type led to significantly reduced CFAP410 protein levels, inducing altered CFAP410 protein localisation and enhanced degradation. The mutations were present in the short leucine-rich repeat C-terminal (LCCRT) domain, which is essential for structural integrity and correct folding of LRR proteins[340,350]. Recessive mutations in CFAP410 have also been reported in other ciliopathies, including Joubert and Jeune syndrome and axial spondylometphyseal dysplasia (axial SMD)[348,349].

The scope for *CFAP410* driving disease expanded when it was identified as a novel risk locus for ALS. Across two cohorts, this study identified a low-frequency non-synonymous variant (rs75087725) (missense variant: V58L) which reached

genome-wide significance for ALS risk (OR = 1.65, $P = 3.08x10^{-10}$). Additionally, rare variant burden analysis also showed a significant excess of LOF and nonsynonymous variants of *CFAP410* in ALS cases compared to controls[11]. More recently, Watanabe *et al*., have shown that CFAP410 and NEK1 interact and stabilise each other at the protein level, leading to an accumulation of both proteins which in turn leads to a significant reduction in neurite length of mouse motor neurons. They discovered that CFAP410 is negatively regulated by FBXO3, which induces ubiquitylation and subsequent degradation of CFAP410. However, NEK1 can phosphorylate CFAP410 and impair this interaction with FBXO3 and thus inhibit ubiquitylation. It was also shown that knockdown of *CFAP410* in HEK293T cells leads to a significant reduction in NEK1 protein levels, suggesting that CFAP410 can stabilise NEK1. The V58L mutant of CFAP410 was found to be more stable than wildtype as it does not interact with FBXO3 and therefore is not ubiquitylated. They concluded that increased stability of the V58L mutant in turn enhances stability of NEK1, causing an accumulation of NEK1 and CFAP410, ultimately leading to an aberrant phenotype in mouse motor neurons[351].

Dysregulation of *REST* and *CFAP410* facilitating disease pathogenesis is known, yet there has been a focus on the coding regions of these genes. The study presented here investigates non-coding VNTRs within *REST* and *CFAP410*, which we hypothesis could both be risk factors of disease and potential modulators of transcription and therefore be important regulators of these loci. The aim of this work was two fold: firstly to assess the functionality of these VNTRs and to understand how they might regulate and even drive transcription. Secondly, to investigate and characterise the repeat variation within these elements, to assess

if there was any association with ALS, which ultimately led to the discovery of novel genetic variants in these loci.

3.2 Hypothesis and aims

Hypothesis:

Both the *REST* and *CFAP410* VNTRs have the potential to drive and regulate transcription of their respective loci and genetic variation within these repetitive domains could be potential risk factors for ALS.

Aims:

Genotype the *REST* promoter VNTR in an MNDA cohort of ALS cases (n=175) and control (n=127) to identify potential risk factors for ALS.

Genotype the *REST* promoter VNTR in a NABEC sample set. Use the available SNP data from this cohort to generate tagging SNPs for the three common VNTR variants.

Genotype the *CFAP410* VNTR in an MNDA cohort of ALS cases (n=199) and controls (n=180) to identify potential risk factors for ALS.

Sequence validate the polymorphisms of the *REST* and *CFAP410* VNTRs.

Test the functionality of the VNTR *in vitro* using a reporter gene assay, both as a transcriptional regulator and a promoter. Determine if copy number of the VNTR drives differential expression profiles.

3.3 Results

3.3.1 The promoter VNTR of *REST*

Dr Maurizio Manca had previously identified the *REST* VNTR, determined it was variable in the general population and characterised three common variants: a 7, 9 and 12 copy number repeat of "GGC" found within the promoter region of the main transcript (protein coding isoform 1) of *REST*; the VNTR overlaps the 5' UTR by 36 bp (Figure 3.1). Due to the implications of *REST* in AD as previously discussed, he genotyped this VNTR in a cohort of AD cases and controls to try and identify potential risk variants for this disease. Overall, he found no significant difference in either allele or genotype frequency of VNTR between AD cases and controls. Furthermore, he also screened this region in both SZ and FTD cases, again finding no significant difference in allele or genotype frequency of the *REST* VNTR across case and control[352].

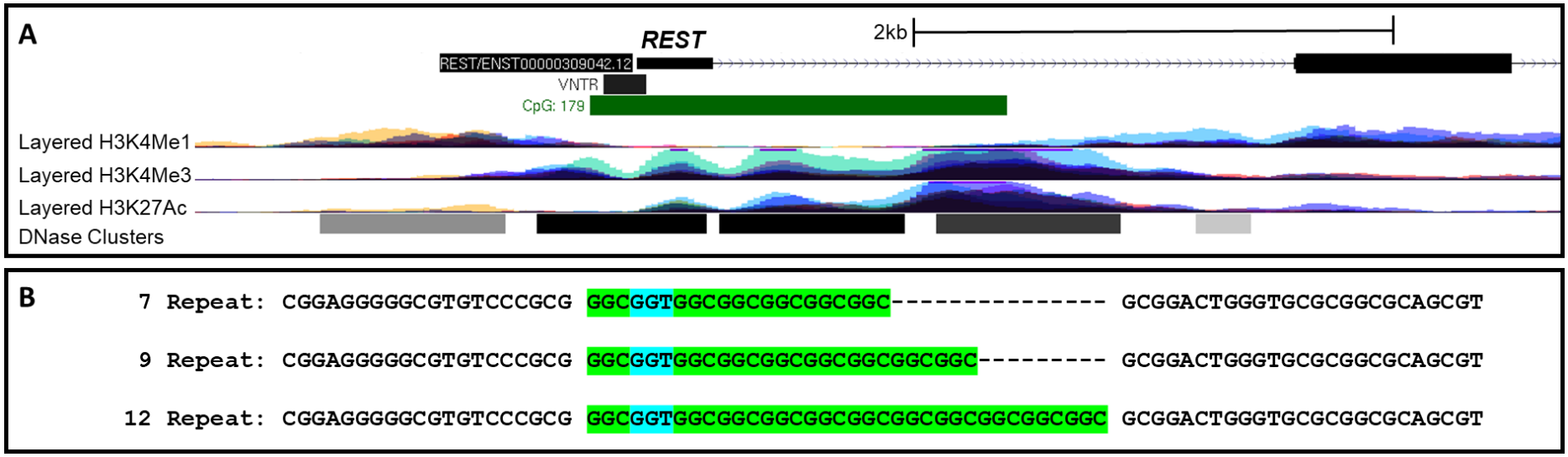


Figure 3.1. There is a VNTR within the promoter region of *REST*.

A: The *REST* locus (UCSC, hg38). There is a VNTR which overlaps the 5' UTR of protein coding isoform 1 of *REST*. **B**: Sequencing of the previously characterised common variants (MAF>5%). The VNTR is built of 3 bp repeats (highlighted in green and blue, with dashes indicating deletions), commonly found in 7, 9 and 12 copies in the general population. The 7 repeat VNTR is present within the human reference genome.

3.3.2 Characterising genetic variation of the *REST* VNTR in ALS

It has been previously shown that there is a significant genetic correlation between ALS and schizophrenia; there is a statistically significant number of overlapping risk loci that share identical risk alleles. Using LD score regression it has been estimated that genetic correlation between these two diseases equates to approximately 14.3%, meaning there is a distinct amount of polygenic overlap between these conditions[353]. Furthermore, there is strong genetic commonality between ALS and FTD, such as the *C9orf72* repeat expansion being uncovered in both conditions, and the presence of shared *C9orf72* and *UNC13A* GWAS SNPs[354,355]. As the *REST* VNTR had been previously characterised in Schizophrenia and FTD cases, this study aimed to assess variation of this region in ALS patients also (Figure 3.2, Table 3.1 and Table 3.2). The *REST* VNTR was genotyped in a Motor Neuron Disease Association (MNDA) cohort of ALS cases and controls.

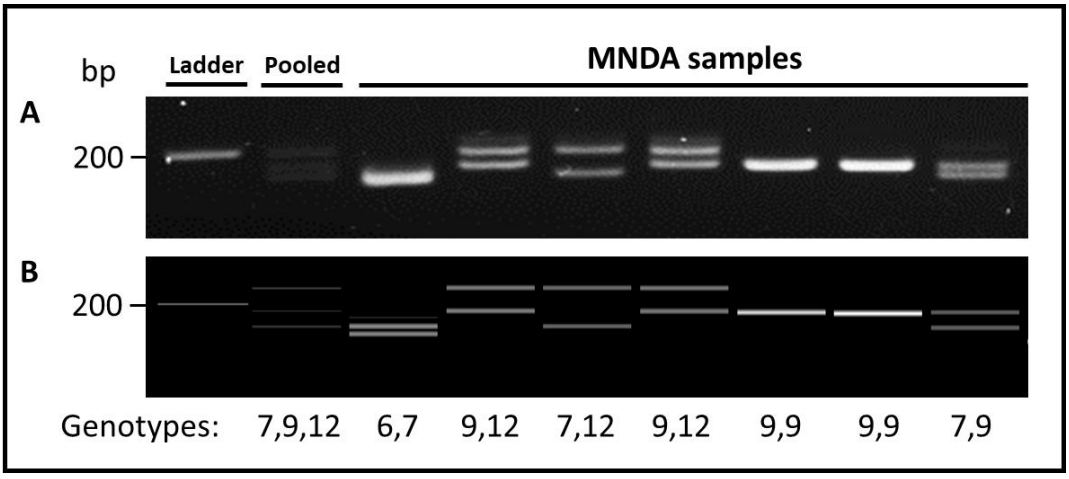


Figure 3.2. *REST* VNTR genotyping in an MNDA cohort.

PCR amplification and gel electrophoresis of the *REST* VNTR in and MNDA cohort (ALS and matched control) (n=302). **A**: agarose gel electrophoresis performed; samples run on 3% agarose at 100V for 4 hours. **B**: gel capillary electrophoresis performed on the same samples using the QIAxcel advanced system and electronic

gel image generated using the QIAxcel ScreenGel software. All three common variants were identified in this cohort: 7, 9 and 12 repeat VNTR. A rare 6 repeat variant was also identified. A pooled genomic DNA was also used as a positive control for all three common variants.

3.3.3 Resolving the *REST* VNTR repeat number polymorphisms

As illustrated in Figure 3.2A it was difficult to resolve the 3 bp difference between the 6 repeat and 7 repeat VNTR using agarose gels. To resolve such small sequence changes, the QIAxcel advanced system (Qiagen) was utilised. This technology uses gel capillary electrophoresis to separate and accurately size DNA fragments, digitally displaying the output as a digital gel image (Figure 3.2B) and an electropherogram (Figure 3.3). Using the high-resolution cartridge available for this system it was possible to resolve 200bp DNA fragments within a 1-3 bp resolution, allowing us to accurately size the difference between the 6 repeat and 7 repeat VNTR (Figure 3.3B) (please refer to Chapter 2 Section 2.2.2.4 for a detailed overview of the QIAxcel methodology). Due to the success of screening the *REST* VNTR using the QIAxcel advanced system it was genotyped in the MNDA cohort using this method, rather than physical agarose gels.

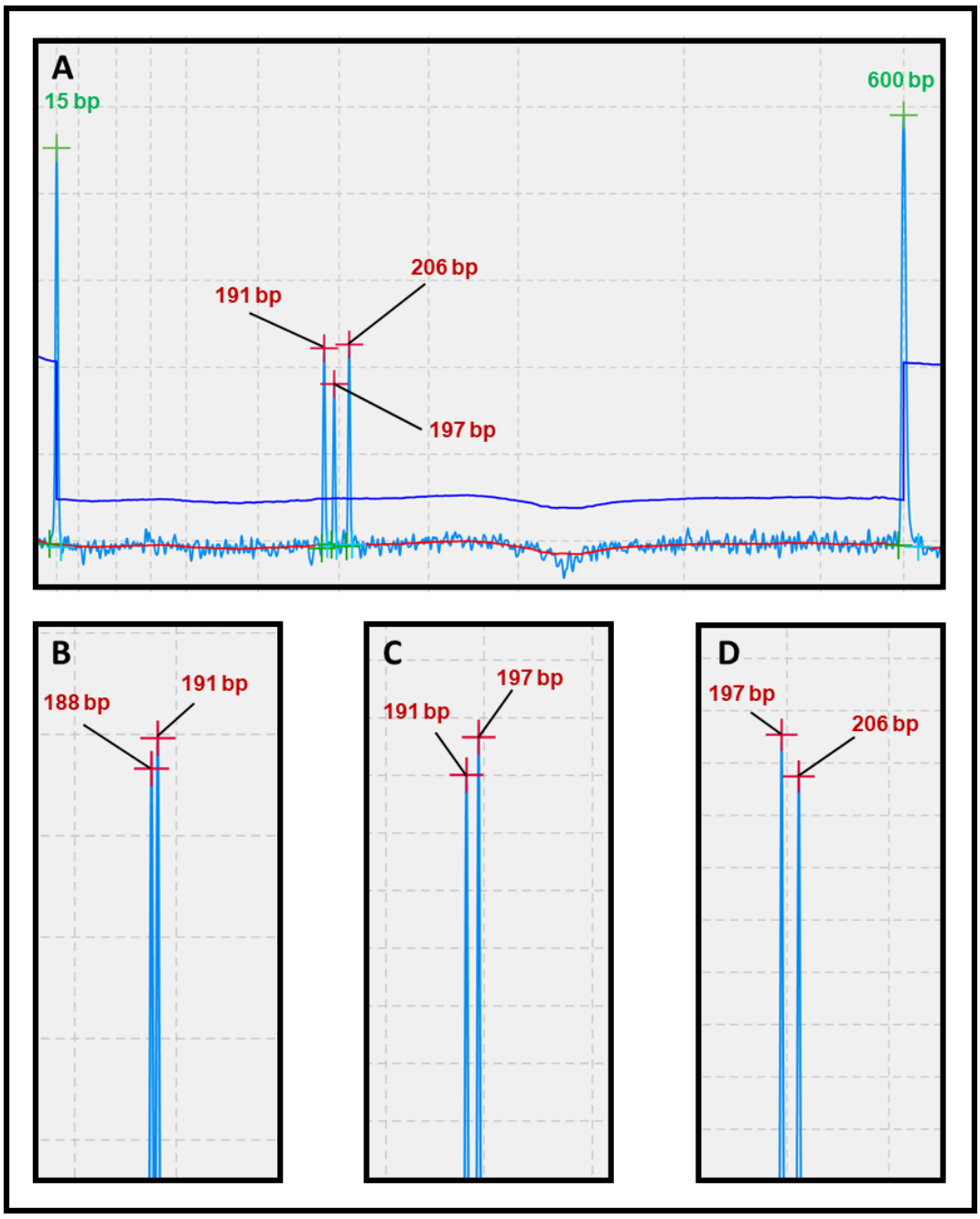


Figure 3.3 Resolving the *REST* VNTR repeat number with high accuracy.

Gel capillary electrophoresis using the QIAxcel advanced system. All samples were run on the high resolution cartridge using the OM800 method. Alignment markers are shown in green and PCR fragment peaks are shown in red. **A**: pooled genomic DNA sample run as a reference point for all three common variants of the *REST* VNTR: 7, 9 and 12 repeat. 191 bp = 7 repeat, 197 bp = 9 repeat, 206 bp = 12 repeat. **B**: sample with the 6/7 genotype displaying a 3 bp smaller variant to the 7 repeat (191 bp). The 188 bp variant = 6 repeat. **C**: example of a sample with 7/9 genotype, displaying an exact 6 bp (two repeats of the VNTR) difference. **D**: example of a

sample with 9/12 genotype, displaying an exact 9 bp (three repeats of the VNTR) difference.

Table 3.1. Allele frequencies of *REST* VNTR in ALS cohort and matched controls.

The four identified alleles of the *REST* VNTR in an ALS cohort (n = 350) and matched controls (n = 254), labelled on the basis of repeat number (R). The 6 repeat (6R) VNTR was only found in ALS cases (n = 1). There was a significant difference in 7R frequency between cases and controls (Fisher's exact test; p-value = 0.03). There was no significant difference in frequency of any other alleles between the ALS cohort and matched controls (Fisher's exact test).

Cohort / Allele	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
6R	1	0.29	0	0.00	1	0.29	1.00
7R	134	38.29	120	47.24	254	-8.96	0.03
9R	132	37.71	78	30.71	210	7.01	0.08
12R	83	23.71	56	22.05	139	1.67	0.70
Total	350	100.00	254	100.00	604	0.00	N/A

Overall, 4 variants of the VNTR were identified in the MNDA cohort. The 7 repeat (7R) was the most common variant, being identified in 38.29% of cases and 47% of controls: this difference was statistically significant (Fisher's exact test; p-value = 0.03). The second most common variant was the 9R, found in 37.71% of cases and 30.71% of controls. While there was a 7.01% difference in 9R frequency across the two cohorts, this result did not reach significance (Fisher's exact test; p-value = 0.08). The 12R was identified in 23.71% of cases and 22.05% of controls; there was no significant difference between the frequency of this variants across cases and controls. Interestingly, the 6R was only present in ALS (n = 1). Overall, there was a significant difference in 7R frequency between ALS cases and controls and we identified a variant only in the ALS cohort (6R) (Table 3.1).

Table 3.2. Genotype frequencies of the *REST* VNTR in an ALS cohort and matched controls.

The seven observed genotypes of the *REST* VNTR in and ALS cohort (n = 175) and matched controls (n = 127). There was no significant difference in genotype frequency between the ALS cohort and matched controls (Fisher's exact test).

Cohort / Genotype	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
6,7	1	0.57	0	0.00	1	0.57	1.00
7,7	27	15.43	30	23.62	57	-8.19	0.08
7,9	47	26.86	35	27.56	82	-0.70	0.90
7,12	32	18.29	25	19.69	57	-1.40	0.77
9,9	25	14.29	13	10.24	38	4.05	0.38
9,12	35	20.00	17	13.39	52	6.61	0.12
12,12	8	4.57	7	5.51	15	-0.94	0.79
Total	175	100.00	127	100.00	302	0.00	N/A

An 8.19% difference in frequency of the 7,7 genotype across cases (15.43%) and controls (23.62%) was observed, but this difference was not statistically significant (Fisher's exact test; p-value = 0.08). The 7,9 genotype was the most common genotype, being found in 26.86% of cases and 27.56% of controls; there was no significant difference in frequency across the two populations (Fisher's exact test, p-value = 0.90). Similarly, no significant difference in 7,12 genotype frequency was observed between ALS cases (18.29%) and controls (19.69%) (Fisher's exact test, p-value = 0.77). The 9,9 genotype was present in 14.29% of cases and 10.24% of controls but this difference did also not reach statistical significance (Fisher's exact test, p-value = 0.38). While there was a 6.61% difference in frequency of the 9,12 genotype, there was no statistically significant difference across case (20.00%) and control (13.39%) (Fisher's exact test, p-value = 0.12). Additionally, the 12,12 genotype was identified in 4.57% of ALS cases and 5.51% of controls, also resulting in no significant difference (Fisher's exact test, p-value = 0.79). The 6R variants was identified in one ALS patient with a 6,7 genotype (0.57% frequency) and was the only rare (MAF<5%) genotype identified in this study. Overall, there was no significant differences in any genotype frequency across ALS cases and control (Fisher's exact test) (Table 3.2).

3.3.4 Validating and sequencing the rare 6 repeat VNTR variant

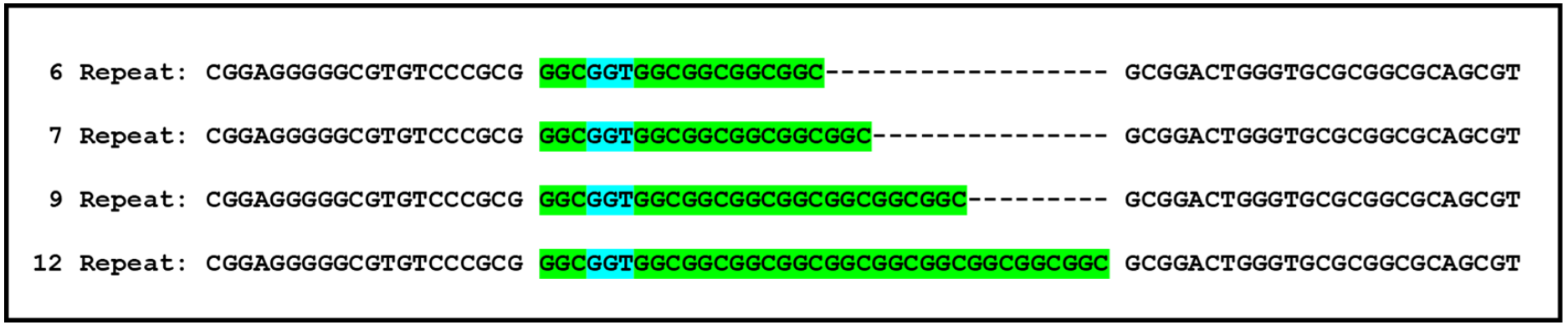


Figure 3.4 *REST* VNTR variant sequencing and alignment.

The rare 6 repeat VNTR aligned against the common variants of the VNTR: 7, 9 and 12 copy number variants. Two repeats of “GGC” and “GGT” were identified and highlighted in green and blue respectively.

Through genotyping the *REST* VNTR using both the QIAxcel advanced system and Sanger sequencing it was confirmed that the 6 repeat is exactly one GGC repeat less than the reference allele (7 repeat VNTR).

3.3.5 *REST* VNTR repeat number variation drives differential gene expression in SH-SY5Y cells

Dr Manca had previously tested the activity of the 7, 9 and 12R variants of the *REST* VNTR in pGL3-B vector (Promega). As this VNTR is within the promoter region of *REST* the pGL3-B vector was chosen; this does not contain a promoter and thus cannot drive luciferase expression. The *REST* VNTR was therefore cloned upstream of the Firefly luciferase reporter gene to test if it could initiate transcription within the repeat. Originally each of the common variants were cloned by Dr Manca using the Gibson Isothermal Assembly technique (please refer to Chapter 2 Section 2.2.4 for a detailed overview of this methodology). It was therefore decided to adopt the same approach to clone the 6R and from this sequence and validate this variant. The luciferase assay was repeated and the 6R variant was included, to test if this rare polymorphism drove a different expression profile to the three common variants previously tested. Each of the VNTR-containing pGL3-B constructs were compared to the empty pGL3-B vector (empty vector) (Figure 3.5).

In HEK293 cells, an 8.83 fold increase in luciferase activity from the 6R construct was observed when compared to the empty vector (8.83±0.46, Mann-Whitney U test, p-value = 3.66E-05). Similarly, there was an 9.35 fold increase in luciferase expression in the 7R construct (9.35±0.45, Mann-Whitney U test, p-value = 3.66E-05). There was no significant difference between the fold expression of the 6R and 7R constructs (Mann-Whitney U test, p-value = 0.47). Interestingly, when compared to the empty vector there was a smaller increase in the 9R construct: a 7 fold increase (7.00±0.62, Mann-Whitney U test, p-value = 3.66E-05). A significant difference in luciferase expression between the 6R and 9R constructs was observed (Mann-Whitney U test, p-value = 3.51E-02). The 12R construct elicited a 8.04 fold increase in expression compared to the empty vector (8.04±0.60, Mann-Whitney U test, p-value = 3.66E-05); there was no significant difference in fold activity between the 6R and 12R (Mann-Whitney U test, p-value = 0.44). Overall, the 6R VNTR-construct did not drive a statistically significant difference in fold activity when compared to any of the common variant (7R, 9R and 12R) constructs. There was statistically significant difference in expression was between the 7R and 9R (Mann-Whitney U test, p-value = 6.10E-03), but no significant difference in reporter gene expression across the 7R and 12R constructs was observed (Mann-Whitney U test, p-value = 0.16). Similarly, no significant difference in luciferase activity was found between 9R and 12R constructs (Mann-Whitney U test, p-value = 0.19) (Figure 3.5).

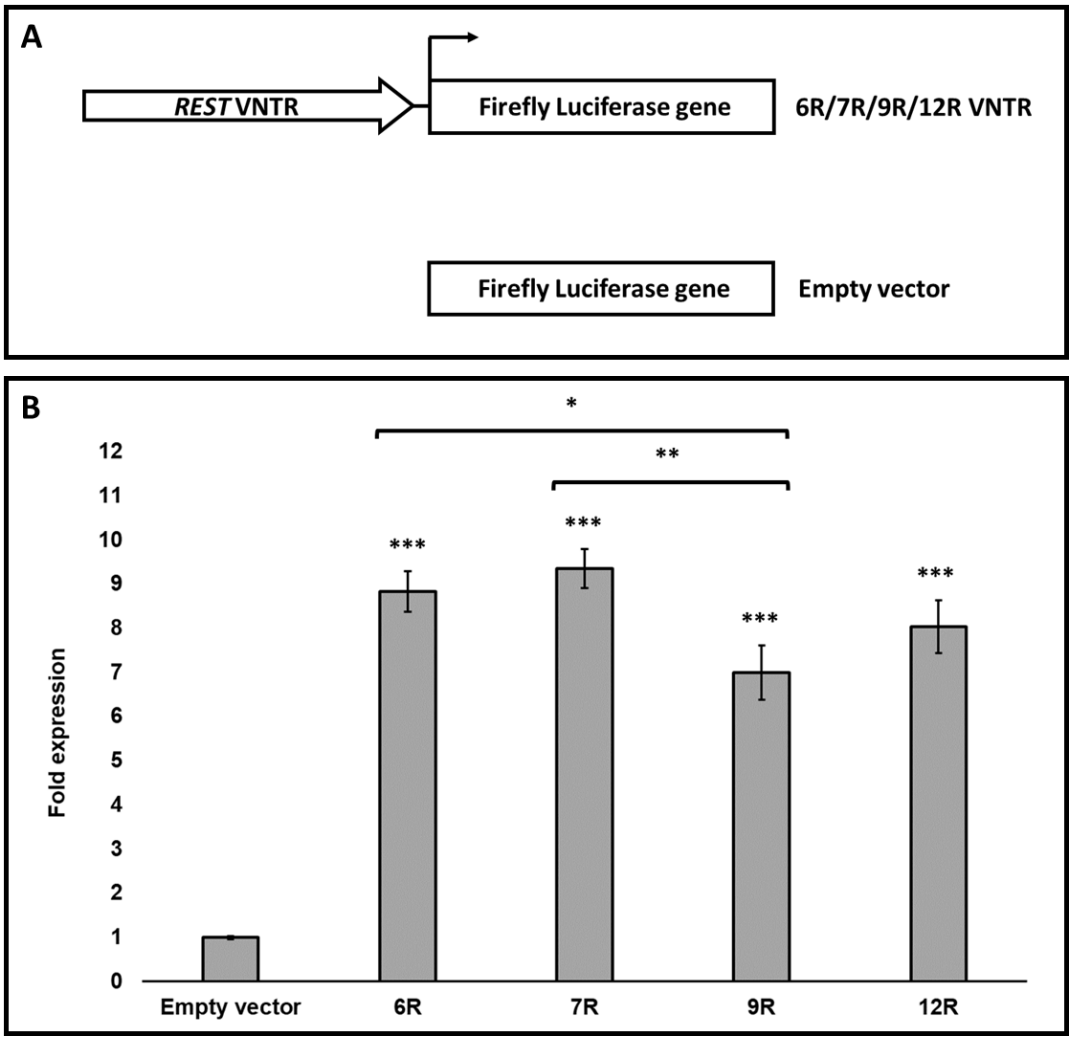


Figure 3.5. *REST* VNTR repeat number variation drives differential gene expression in SH-SY5Y.

A: Schematic of the *REST* VNTR pGL3B constructs. **B**: The fold activity of the *REST* VNTR within the pGL3-B vector normalised to the internal control Renilla Luciferase. SH-SY5Y cell line was transfected with The *REST* VNTR constructs (shown in grey). (biological replicate n = 3, technical replicate per assay n = 4). Mann-Whitney U test was used to compare VNTR containing constructs (6R, 7R, 9R and 12R) to empty vector alone (pGL3-B), then also used to compare between VNTR containing constructs * P<0.05 **P<0.01 ***P<0.001.

3.3.6 Bioinformatic analysis of the *CFAP410* locus

The *CFAP410* locus was analysed using UCSC genome browser (Figure 3.6), specifically evaluating simple tandem repeats, conservation and ENCODE data over this region. The simple repeats track displays results from Tandem Repeat Finder (TRF), a specialised program which locates tandem repeats within DNA sequences[356]. This analysis identified six tandem repeats in the *CFAP410* locus and one was discovered to be a variable number tandem repeat (VNTR) (Figure 3.6A). This VNTR was found within intron 1 of the isoform 1 of *CFAP410* and is approximately 323 bp. This locus was inspected in hg19 (GRCh37/hg19) due this build having the Ensembl Genes track: a database of annotated genes and predicted transcripts from the Ensembl project (http://www.ensembl.org/)[357].

ENCODE data at this region showed that the selected VNTR was within histone marks: layered H3K4Me3 mark which is found by active promoters, layered H3K4Me3 and H3K27Ac which are associated with regulatory regions[312]. The VNTR also falls with a DNaseI hypersensitive cluster which is also indicative of a regulatory region as this is a region of open and accessible chromatin. The vertebrate multiz alignment and conservation track showed that this genomic region was not conserved in chimps, gorillas, orangutans, gibbons, rhesus macaques, rats and mice, indicating that this VNTR is human specific (Figure 3.6B).

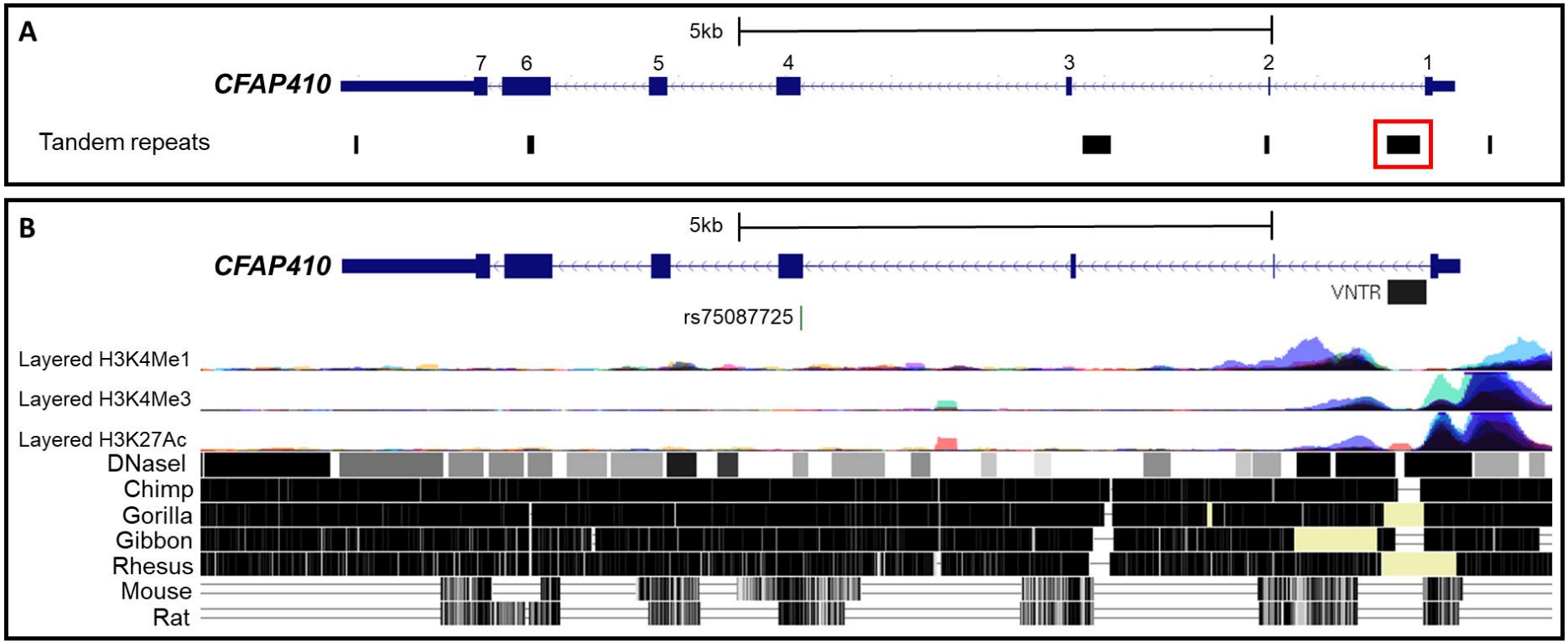


Figure 3.6. There is a VNTR downstream of the main promoter of *CFAP410.*

A: *CFAP410* loci present on chromosome 21 (UCSC, hg19), with exons (blue boxes) numbered; introns indicated as blue lines with arrows. There are six tandem repeats in the *CFAP410* locus, one of which is variable in repeat length (VNTR) (red box). The *CFAP410* gene contains a GWAS SNP (rs75087725) which is associated with Amyotrophic Lateral Sclerosis risk ($P = 3.08 \times 10^{-10}$). **B:** VNTR within intro 1 of transcript 1 *CFAP410* (protein coding isoform). ENCODE data from UCSC shows the levels of enrichment of histone marks within this locus, specifically a signal for H3K4Me1

which is found near regulatory elements, H3K4Me3 which is associated with active promoters and H3K27Ac which is indicative of active regulatory elements. The *CFAP410* locus has been overlaid with a conservation track of several primates and rodents. There is a break in conservation over the VNTR, indicating a human specific expansion of this VNTR.

Figure 3.7. CFAP410 VNTR sequence

Primary sequence of the *CFAP410* VNTR (displayed as 5'-3') from the sense strand, (UCSC hg19, chr21:45758488-45759071). The VNTR is approximately 323 bp and has a GC content of 84%. PCR amplicon size is 584 bp and primers are shown in bold and underlined. 35 bp and 22 bp repeats are boxed and have been aligned by eye and are therefore arbitrary.

The primary sequence of the *CFAP410* VNTR was taken from the simple repeats track on UCSC genome browser and visually inspected in word. The sequence was then broken down into repeat units and aligned (Figure 3.7). It was found that the VNTR consists of a 35 bp and a 22 bp repeat. The reference genome VNTR repeat units were aligned and it was found with the reference allele contains eight 35 bp repeats and two 22 bp repeats (Figure 3.7).

3.3.7 Characterising genetic variation of the *CFAP410* VNTR

CFAP410 is a recently discovered ALS risk locus, we therefore decided to characterise the genetic variation of the VNTR found within this gene and assess its potential association with ALS. The *CFAP410* VNTR was genotyped in an MNDA cohort of ALS patients and controls. Overall, seven variants of the VNTR were identified in the MNDA cohort (n = 379) and two of these were only present in ALS patients (alleles 2 and 7) (Figure 3.8 and Table 3.3).

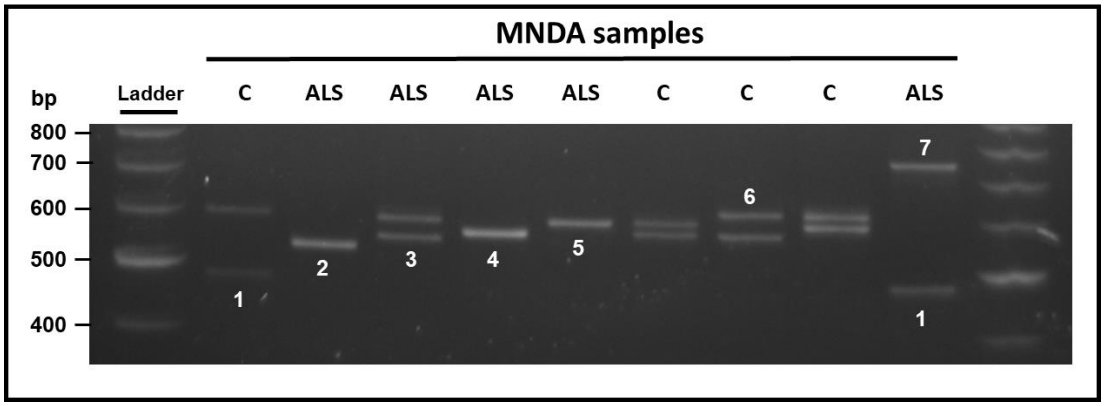


Figure 3.8. *CFAP410* VNTR genotyping.

PCR amplification and gel electrophoresis of the VNTR within the *CFAP410* locus in ALS patients and matched control samples. This region was found to be polymorphic and seven alleles have been identified (1-7) (n = 379). Variants 2 and 7 were found only in ALS patients. All samples were run at 100V on 2% agarose for 3.5 hours. C = control.

Table 3.3. Allele frequencies of *CFAP410* VNTR in ALS cohort and matched controls.

The seven identified alleles of the *CFAP410* VNTR in an ALS cohort (n = 398) and matched controls (n = 360). Alleles 2 and 7 were identified only in the ALS cohort. There is no significant difference in allele frequency between the ALS cohort and matched controls (Fisher's exact test).

Cohort / Allele	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1	2	0.50	3	0.83	5	-0.33	0.67
2	2	0.50	0	0.00	2	0.50	0.50
3	4	1.01	4	1.11	8	-0.11	1.00
4	102	25.63	91	25.28	193	0.35	0.93
5	283	71.11	260	72.22	543	-1.12	0.75
6	3	0.75	2	0.56	5	0.20	1.00
7	2	0.50	0	0.00	2	0.50	0.50
Total	398	100.00	360	100.00	758	0.00	N/A

A total of 7 seven alleles of the *CFAP410* VNTR were identified in the MNDA cohort. Allele 5 was found in the reference genome and was present in 71.11% of cases and 72.22% of controls; no significant difference in frequency was found (Fisher's exact test, p-value = 0.75). Allele 4 was the next most common variant (MAF<1%) and was identified in 25.63% of cases and 25.28% of controls, also not resulting in a significant difference in allele frequency (Fisher's exact test, p-value = 0.93). Variant 3 was low frequency (MAF 1-5%) and was present in both cases (1.01%) and controls (1.11%) (Fisher's exact test, p-value = 1.00). Alleles 1 and 6 were both rare (MAF>1%) and were also both present in both cohorts. Interestingly, alleles 2 and 7 were found only in ALS patients (both 0.5%). Overall, there was no significant difference in frequency of any allele across ALS patients and controls (Fisher's exact test) (Table 3.3) and therefore no relationship with ALS was found.

Table 3.4. Genotype frequencies of *CFAP410* VNTR in ALS cohort and matched controls.

The twelve identified genotypes of the *CFAP410* VNTR in an ALS cohort (n = 199) and matched controls (n = 180). There is no significant difference in genotype frequency between the ALS cohort and matched controls (Fisher's exact test).

Cohort / Genotype	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1,4	1	0.50	1	0.56	2	-0.05	1.00
1,5	0	0.00	2	1.11	2	-1.11	0.22
1,7	1	0.50	0	0.00	1	0.50	1.00
2,2	1	0.50	0	0.00	1	0.50	1.00
3,4	1	0.50	0	0.00	1	0.50	1.00
3,5	3	1.51	4	2.22	7	-0.71	0.71
4,4	14	7.04	16	8.89	30	-1.85	0.57
4,5	70	35.18	58	32.22	128	2.95	0.59
4,6	2	1.01	0	0.00	2	1.01	0.50
5,5	104	52.26	97	53.89	201	-1.63	0.76
5,6	1	0.50	2	1.11	3	-0.61	0.61
5,7	1	0.50	0	0.00	1	0.50	1.00
Total	199	100.00	180	100.00	379	0.00	N/A

Twelve genotypes of the *CFAP410* VNTR were found in the MNDA cohort. Homozygous 5,5 was the most frequent in both cohorts, found in 52.26% of cases and 53.89% of controls (Fisher's exact test, p-value = 0.76). Heterozygous 4,5 was the second most frequent in both populations and was present in 35.18% of cases and 32.22% of controls (Fisher's exact test, p-value = 0.59). Homozygous 4,4 was also common and found in 7.04% of cases and 8.89% of controls (Fisher's exact test, p-value = 0.57). ALS specific alleles 2 and 7 (Table 3.3) were found in 3 patients, identified as genotypes 2,2, 1,7 and 5,7. Overall, there was no significant difference in genotype frequencies across ALS patients and controls (Fisher's exact test) (Table 3.4) and thus no association with ALS was observed.

3.3.8 The *CFAP410* VNTR is stable across brain and blood of the same ALS patient

As previously mentioned in Chapter 1, studies have shown that VNTR repeat size can vary across tissues in the same individual. We therefore decided to access *CFAP410* VNTR genotype in motor cortex and blood DNA from the same ALS patient (n = 7) (please refer to Chapter 4, Section 4.3.3 for the same analysis in the *NEK1* SVA-D CT element). Overall, no expansion or deletion differences across tissues from the same person were observed, confirming in this small sample size that the VNTR genotype was the same in both the motor cortex and blood of the same ALS patient (Figure 3.9).

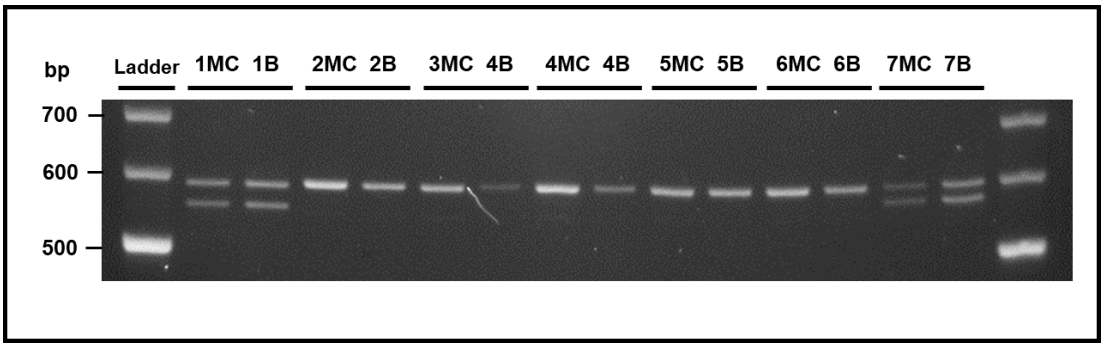


Figure 3.9 *CFAP410* VNTR genotyping in matched brain and blood.

PCR amplification and gel electrophoresis of the *CFAP410* VNTR within matched motor cortex (MC) and blood (B) for 7 ALS patients. There was no variation in genotype across motor cortex and blood of the same ALS patient. All samples were run at 100V on 2% agarose for 3.5 hours.

3.3.9 *CFAP410* VNTR sequencing

To determine the sizes and lengths of the VNTR alleles identified in the MNDA cohort, each variant was cloned and sequence validated. Each of the VNTRs was amplified via PCR for cloning into the pCR®-Blunt vector (ThermoFisher) as previously described in Chapter 2 Section 2.2.3. Unfortunately, after several attempts, only 4 of the 7 variants were successfully cloned (variants 2, 4, 5 and 7) (Figure 3.10). In comparison to allele 5 (the reference genome variant), it was found that allele 4 had a 22 bp repeat deletion. Allele 2, when compared to allele 5, lacked one copy of each repeat. One comparison with the reference allele, allele 7 had an expansion of 149 bp; equating to two more 22 bp repeats and three more 35 bp repeats. Variants 2, 4,

5 and 7 equated to 262 bp, 297 bp, 319 bp, and 468 bp in length, respectively (Figure 3.10).

Thus, overall, allele 2 was built of seven 35 bp repeats and one 22 bp repeat. Allele 4 is formed of eight 35 bp repeats and one 22 bp repeat. Allele 5 is constructed on eight 35 bp repeats and two 22 bp repeats. Allele 7 consists of eleven 35 bp repeats and four 22 bp repeats (Figure 3.10).

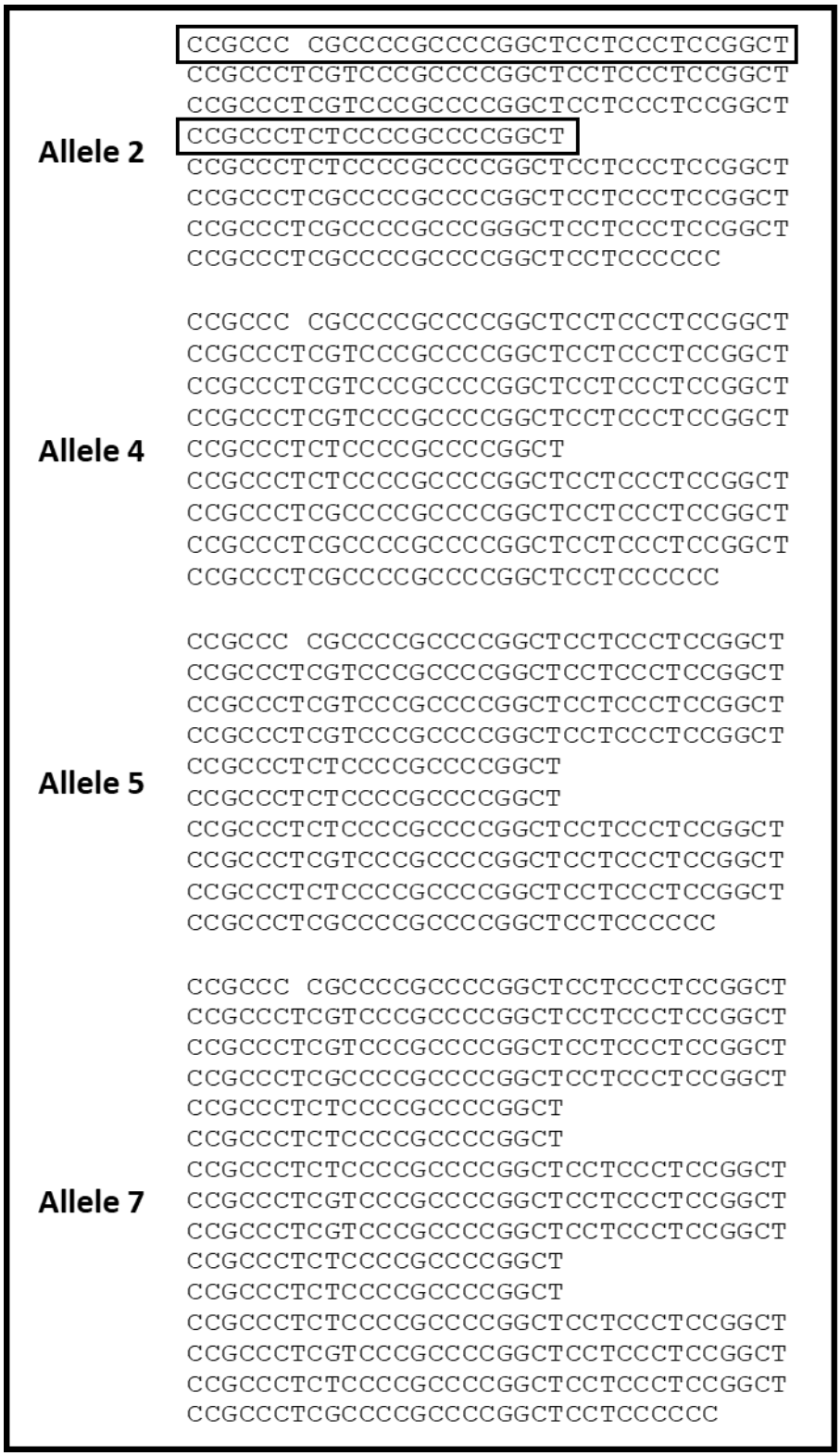


Figure 3.10. *CFAP410* VNTR variant sequences aligned.

Variant 2, 4, 5 and 7 of the *CFAP410* VNTR aligned and split into respective 22 bp and 35 bp repeat units. Allele 2 = 262 bp, allele 4 = 297 bp, allele 5 = 319 bp, allele 7 = 468 bp. 35 bp and 22 bp repeats are boxed. All repeats have been aligned by eye and are therefore arbitrary.

3.3.10 The *CFAP410* VNTR shows functional properties in pGL3-P vector in HEK293 cell line

To determine the potential functionality of the *CFAP410* VNTR it was tested *in vitro* using a reporter gene assay; measuring the effect of the VNTR on luciferase expression in the pGL3-P reporter gene construct (Promega). To test if repeat number variation had the capacity to alter gene expression profiles, both common variants (alleles 4 and 5) were cloned and tested. The endogenous and reverse orientation of each variant was included, to confirm if orientation of the VNTR had any effect on reporter gene expression in this model. The forward construct was termed endogenous due to the VNTR sequence being in the same orientation (and on the same DNA strand) as the promoter of *CFAP410* in the human genome. The reverse VNTR construct was the reverse complement of this sequence (Figure 3.13).

According to UCSC (hg19), there are four RefSeq validated transcripts of *CFAP410* (highlighted in dark blue in Figure 3.11). The dark blue annotation means these particular isoforms have been reviewed and validated by NCBI (either through assessment of sequencing data or literature review). Refseq also provides provisional and predicted transcripts (which are annotated in a lighter blue). Interestingly, the 5' UTR of isoform 4 (short isoform) was found 428 bp downstream of the VNTR. The VNTR also overlaps with a predicted Ensembl transcript (ENST00000462742) (highlighted in red in panel A of Figure 3.11). According to gene expression data from Genotype-Tissue Expression (GTEx) project portal[313], this isoform is ubiquitously expressed (present in 54 non-disease tissue sites) (panel B of Figure 3.11). We hypothesised that the VNTR could initiate transcription of this isoform. To test this

hypothesis the VNTR was cloned into pGL3-B (Promega); a reporter gene construct that is used to test if a region of DNA can act as a promoter *in vitro*. This construct is identical to pGL3-P, other than it lacking the SV40 promoter and therefore the empty vector alone cannot initiate transcription and therefore cannot drive luciferase expression (as indicated in the no promoter lane of Figure 3.13B and empty vector lane of Figure 3.13B). The empty pGL3-P vector (Figure 2.2) was included in this assay as it contains the minimal SV40 early promoter and therefore serves as a suitable positive control and reference point for a low expression promoter in this model for comparison with VNTR constructs (labelled as SV40 promoter in Figure 3.14).

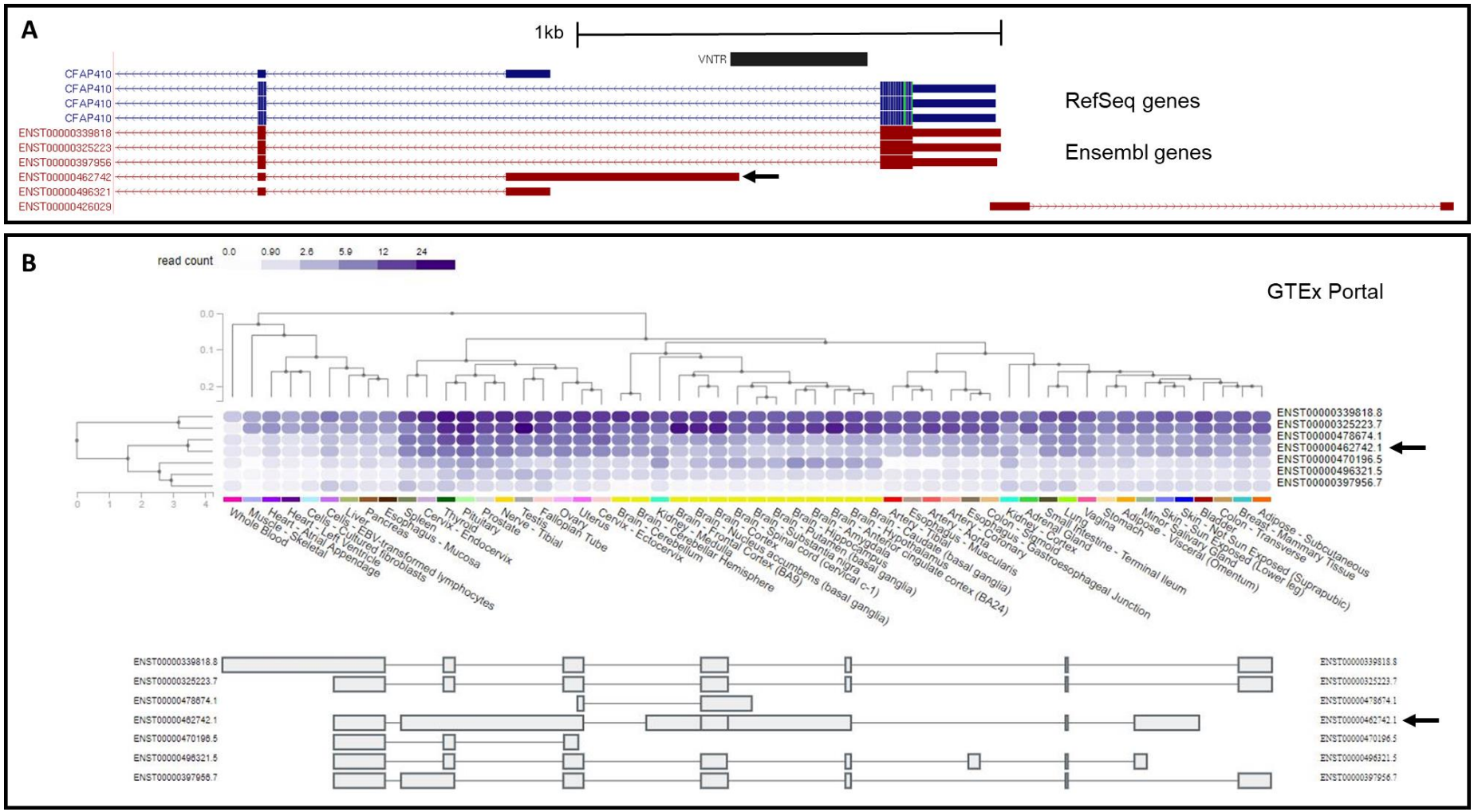


Figure 3.11. There is an expressed *CFAP410* isoform downstream of the VNTR.

A: *CFAP410* loci present on chromosome 21 (UCSC, hg19). There are four curated RefSeq isoforms of *CFAP410* (blue). There are six predicted transcripts from the Ensemble database project (red). **B**: *CFAP410* (*C21orf2*) isoform expression data taken from GTEx Portal (analysis release V8), read count shown (transcripts per million, TPM). Black arrows indicate the expressed isoform (ENST00000462742) which begins downstream of the VNTR.

The *CFAP410* VNTR was amplified using PCR and then ligated into the intermediate pCR®-Blunt vector (ThermoFisher). Variant 4 and 5 were both cloned into pGL3-P, each in both the endogenous and reverse orientation with respect to the SV40 promoter of pGL3-P (four pGL3-P constructs in total). Variant 5 was cloned into pGL3-B in both the endogenous and reverse orientation with respect to the Firefly luciferase reporter gene. The presence of the VNTR in all constructs was confirmed using restriction digest (Figure 3.12) and Sanger sequencing (Supplementary Figure).

Using the Dual-Luciferase® Reporter Assay (Promega) the luminescent signal generated by the luciferase reaction of the VNTR-containing pGL3-P constructs was measured and directly compared to the signal generated by the pGL3-P vector containing the SV40 promoter alone (empty vector) (Figure 3.13); VNTR-containing pGL3-B constructs were compared to the pGL3-B vector alone (empty vector) (Figure 3.14). The luminescent signal of Renilla (*Renilla reinormis*) luciferase was used as an internal control to normalise all samples, accounting for variation in cell number, cell death and transfection efficiency (please refer to Chapter 5 Section 5.3.1 for the same assay procedure for *NEK1* SVA-D constructs).

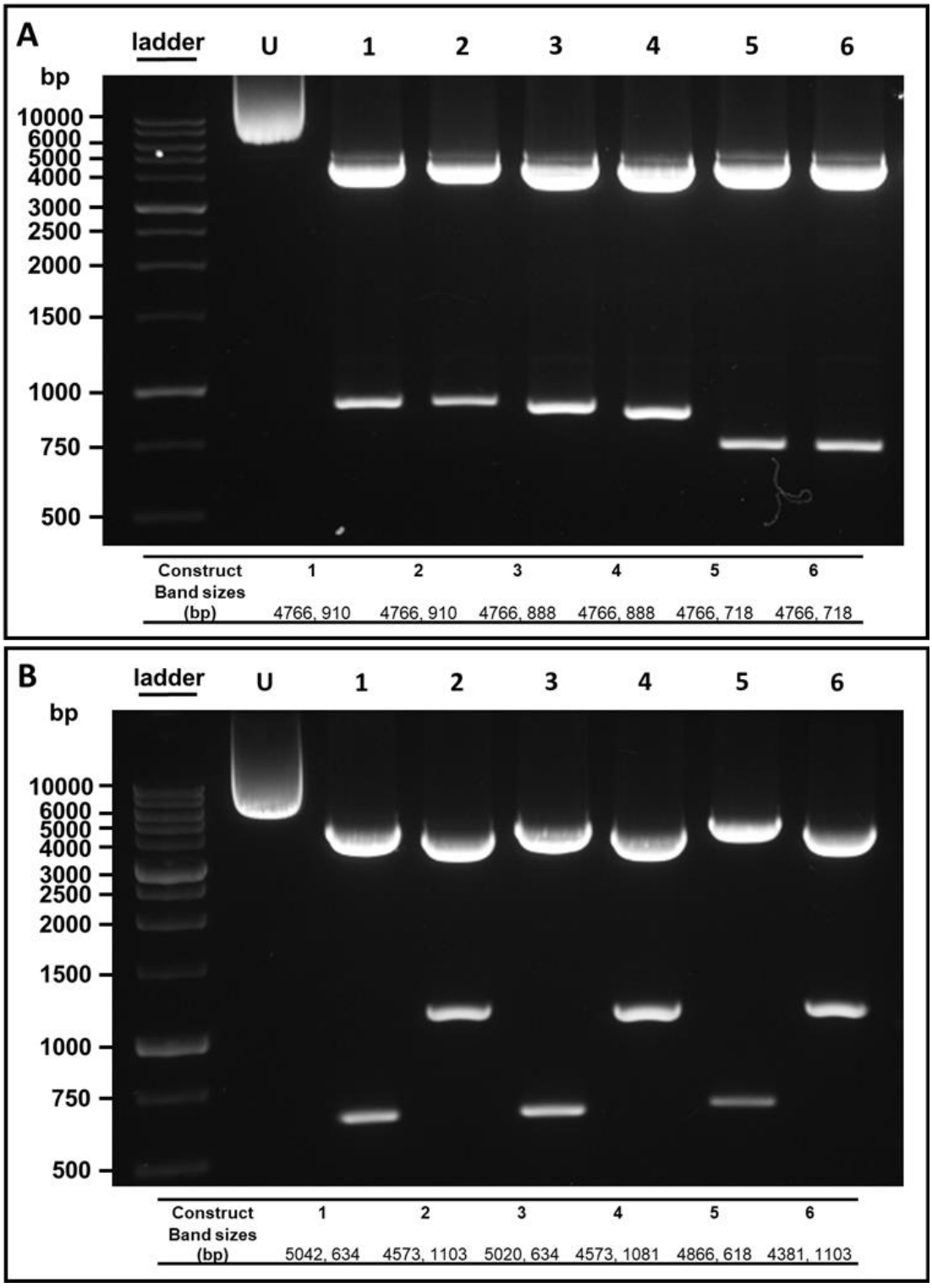


Figure 3.12. *CFAP410* VNTR restriction digests.

A: Restriction enzyme digest of *CFAP410* VNTR reporter gene constructs to test for multimers of the VNTR. Constructs were cut with *KpnI* and *HindIII* and run on a 1% agarose gel for 1.5 hours at 120V. Results from this digest show that there was only one copy of the VNTR present in each construct. **B**: Restriction enzyme digest of *CFAP410* VNTR reporter gene constructs to confirm orientation of the VNTR.

Constructs were cut with *DraIII* and run on a 1% agarose gel for 1 hour at 120V. Expected band sizes for each construct restriction digest are displayed in a table below the gel image. U = uncut, 1 = VNTR allele 5 pGL3P reverse, 2 = VNTR allele 5 pGL3-P endogenous, 3 = VNTR allele 4 pGL3-P reverse, 4 = VNTR allele 4 pGL3-P endogenous, 5 = VNTR allele 5 pGL3-B reverse, 6 = VNTR allele 5 pGL3-B endogenous. Results from this digest showed that constructs 1, 3 and 5 were all in the reverse orientation and constructs 2, 4 and 6 were all in the endogenous orientation.

Due to the strategy adopted for cloning the *CFAP410* VNTR it was possible to clone multiple copies of this region of DNA as the blunt ends can ligate together and form multimers. To check how many copies of the VNTR were present in each construct, a double digest on each reporter gene construct with KpnI and HindIII was performed. These enzymes will only cut the backbone of the pGL3 constructs and thus from size of the insert we could confirm copy number of the VNTR. Overall, only one copy of the VNTR was cloned into each construct (Figure 3.12A). Secondly an orientation confirmation digest by cutting each reporter gene construct with DraIII was also performed (Figure 3.12B). This particular enzyme cut once in the pGL3 backbone and once in the VNTR (adopting the "one in one out" strategy shown in Figure 2.6) (please refer to Chapter 2 Section 2.2.3.7 Restriction enzyme digests for a detailed overview of this process). All restriction enzyme recognition sites were confirmed using the software, A plasmid Editor (ApE) (https://jorgensen.biology.utah.edu/wayned/ape/). Overall, it was confirmed that variant 4 and 5 were cloned in both the endogenous and reverse orientation in pGL3-P. Furthermore it was determined that variant 5 was cloned in pGL3-B in both orientations (Figure 3.12B)

When compared to the empty pGL3-P vector, a 1.71 fold increase in expression of luciferase in the allele 5 endogenous orientation VNTR construct was observed (1.71±0.13, Mann-Whitney U test, p-value = 3.66E-05). Interestingly the opposite effect in the allele 5 reverse orientation VNTR construct occurred; a 1.2 fold decrease in luciferase activity when compared to empty pGL3-P was found (0.82±0.03, Mann-Whitney U test, p-value = 6.01E-05). Similar to the allele 5 endogenous VNTR, a 2.64 fold increase in luciferase expression in the allele 4 endogenous VNTR construct when compared to empty pGL3P was seen (2.64±0.18, Mann-Whitney U test, p-value = 3.66E-05). However, no significant difference in luciferase activity in the allele 4 reverse VNTR construct was observed (1.04±0.06, Mann-Whitney U test, p-value = 0.40). Overall, in the endogenous VNTR constructs it was found that there was a statistically significant increase in fold expression of luciferase when compared to the empty vector. Furthermore, a statistically significant difference in luciferase activity between the endogenous constructs of allele 4 and 5 was also observed (Mann-Whitney U test, p-value = 9.73E-05), with a larger increase found in the allele 4 construct; copy number variation of the VNTR did alter expression levels in this model. The increase in luciferase expression was not seen in either of the reverse constructs, confirming that there were orientation specific expression profiles in this model. There was also a significant difference between the reverse constructs of allele 4 and 5 (Mann-Whitney U test, p-value = 1.56E-04) (Figure 3.13).

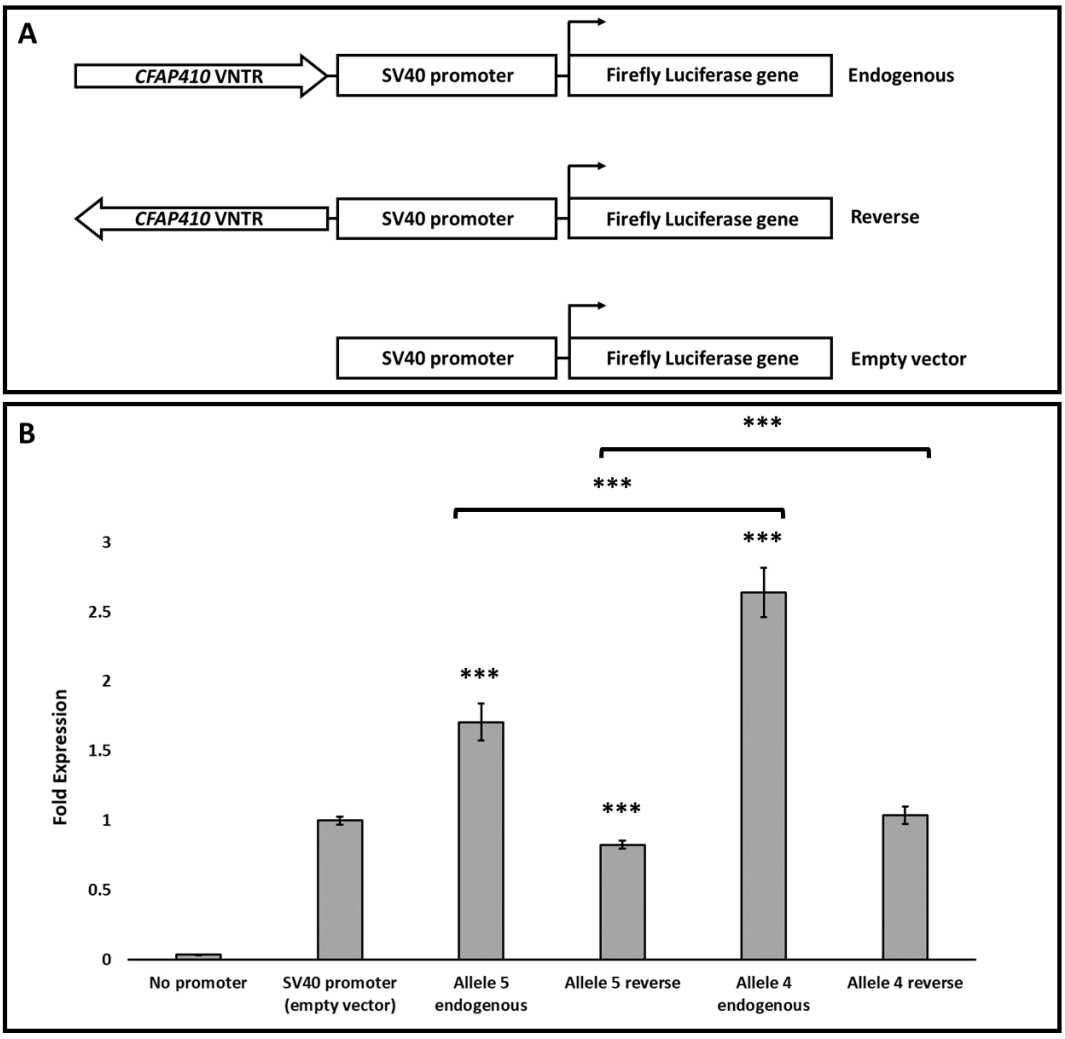


Figure 3.13. The *CFAP410* VNTR shows functional properties in pGL3-P vector in HEK293.

A: Schematic of *CFAP410* VNTR pGL3P constructs. Alleles 4 and 5 of the VNTR were isolated and cloned within the pGL3-Promoter vector. This plasmid contains a reporter gene (firefly luciferase) and an SV40 minimal promoter. Constructs were transfected and after 48 hours Luciferase activity of the VNTR containing constructs was measured on a luminometer and compared to the activity of the empty vector alone. **B:** The fold activity of allele 4 and 5 of the *CFAP410* VNTR in the endogenous and reverse orientation within the pGL3P vector normalised to the internal control Renilla Luciferase. HEK293 cell line was transfected with the CFAP410 VNTR constructs (shown in grey). (Biological replicate n = 3, technical replicate per assay n = 4). The no promoter vector (pGL3-B) was included as a negative control. Mann-Whitney U test was used to compare VNTR containing

constructs to SV40 promoter vector (empty vector) (pGL3-P) and to compare all VNTR containing constructs against each other. ***P<0.001.

3.3.11 The *CFAP410* VNTR shows promoter activity in the pGL3-B vector in HEK293 cell line

Compared to empty pGL3-B, a 7.38 fold increase in luciferase expression in the allele 5 endogenous orientation VNTR construct was found (7.38±0.40, Mann-Whitney U test, p-value = 3.66E-05). Interestingly there was a 71.48 fold increase in luciferase activity in the allele 5 reverse orientation VNTR construct (71.48±6.67, Mann-Whitney U test, p-value = 3.66E-05). The empty pGL3-P vector was included as a positive control and when compared to empty pGL3-B a 29.99 fold increase in luciferase expression was observed (29.99±0.86, Mann-Whitney U test, p-value = 3.66E-05). Overall, a moderate increase in luciferase activity was observed in the endogenous VNTR construct when compared to empty pGL3-B, but this was significantly less than the SV40 minimal promoter activity of pGL3-P (Mann-Whitney U test, p-value = 3.66E-05). In contrast, there was a 2.4 fold increase in luciferase expression in the reverse VNTR construct when compared to empty pGL3-P, showing a large and significant increase in luciferase activity when compared to a known promoter (Mann-Whitney U test, p-value = 3.66E-05) (Figure 3.14).

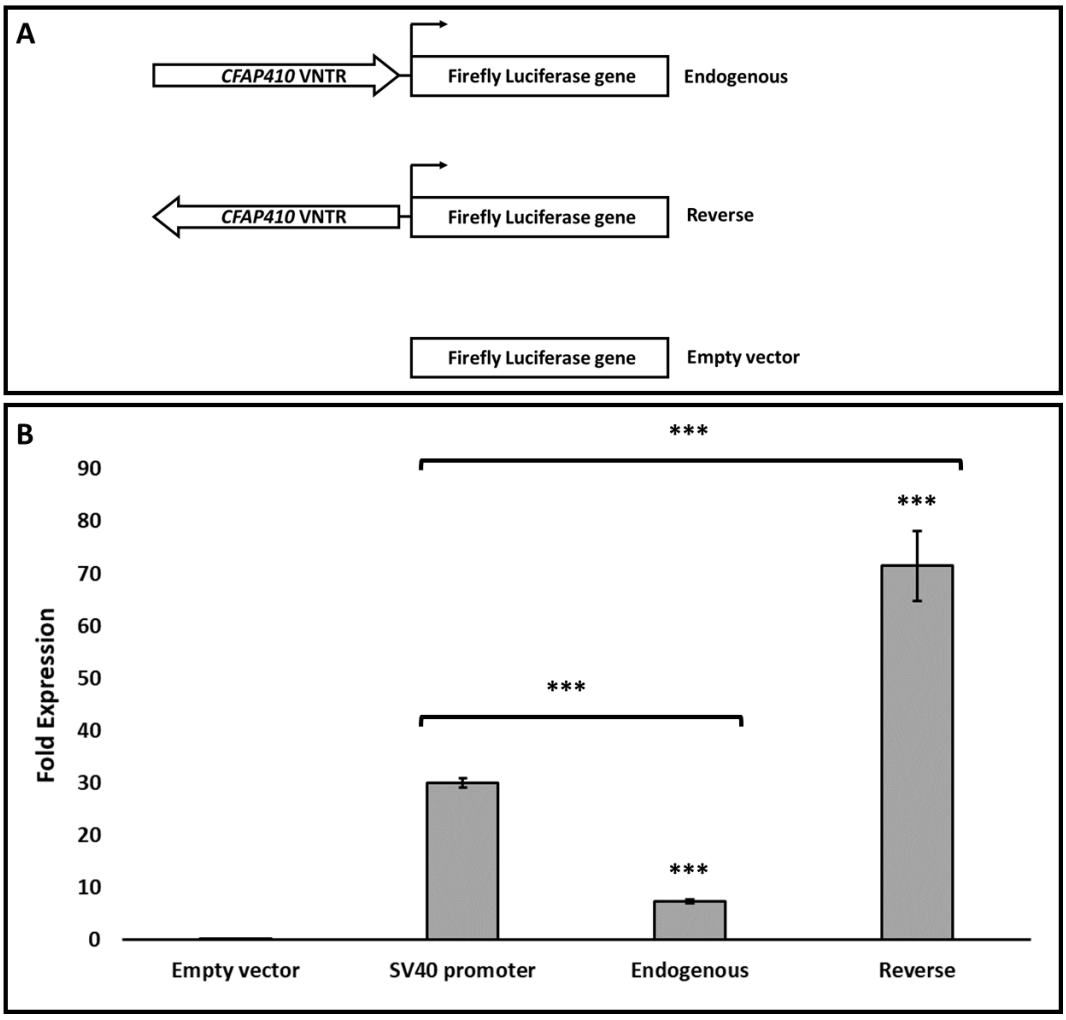


Figure 3.14. The *CFAP410* VNTR shows promoter activity in the pGL3-B vector in HEK293.

A: Schematic of *CFAP410* VNTR pGL3B constructs. Allele 5 of the VNTR was isolated and cloned within the pGL3-Basic vector. This plasmid contains a reporter gene (firefly luciferase) but no promoter. Constructs were transfected and after 48 hours Luciferase activity of the VNTR containing constructs was measured on a luminometer and compared to the activity of the empty vector alone and a known SV40 minimal promoter. **B**: The fold activity of the *CFAP410* VNTR in the endogenous and reverse orientation within the pGL3-B vector normalised to the internal control Renilla Luciferase. HEK293 cell line was transfected with the CFAP410 VNTR pGL3B constructs (shown in grey). (Biological replicate n = 3, technical replicate per assay n = 4). The SV40 promoter vector (pGL3-P) was

included as a positive control. Mann-Whitney U test was used to compare VNTR containing constructs to empty vector alone (pGL3-B). Mann-Whitney U test also used to compare VNTR containing constructs to SV40 promoter vector (pGL3-P) ***P<0.001.

3.4 Discussion

In this study, a novel rare variant in the previously identified promoter VNTR of *REST* has been identified: a 6 tandem repeat variant (6R). The 6R VNTR was successfully cloned using Gibson Assembly (Figure 2.7) and sequence verified (Figure 3.4). Furthermore, using the QIAxcel advanced system, it was possible to resolve each VNTR variant to the exact bp and thus accurately genotype this region. Moreover, it was found that the 6R variant elicits comparable expression profiles to all three of the common VNTR variants: the 7R, 9R and 12R (Figure 3.5).

Using UCSC and tandem repeat finder a VNTR was identified within the ALS risk gene, *CFAP410* (*C21orf2*) (Figure 3.6). It was determined that this VNTR is highly polymorphic, with 7 variants being identified in the MNDA cohort (n = 379) (Figure 3.8). This is the first time this VNTR has been characterised, but the variation identified has no association with ALS (Figure 3.8, Table 3.1 and Table 3.2). Additionally, it was determined that the VNTR genotype is stable across motor cortex and blood of the same ALS patient (Figure 3.9). Finally, it was shown that this VNTR can both regulate and drive transcription of a reporter gene *in vitro* (Figure 3.13 and Figure 3.14).

Conforming with the original hypothesis that both VNTRs could harbour ALS mutations, novel VNTR variants only present in ALS patients of the MNDA cohort were discovered. Overall, one ALS case was identified with the 6R VNTR in the MNDA cohort (allele frequency = 0.29%) (Table 3.1). Dr Manca did identify one American sample from STR browser with the 6 copy repeat, but this variant was not present in the AD, Schizophrenia or FTD cohorts he assessed[352]. While it is likely

that the 6R variant is present in people without ALS, it would still be of interest to screen more ALS patients and matched controls to determine if this variant could be a risk factor for the disease. Similarly, ALS specific variants of the *CFAP410* VNTR were identified in the MNDA cohort, with one case having variant 2 (allele frequency = 0.50%) and two ALS cases harbouring variant 7 (allele frequency = 0.50%) (Table 3.3). The discovery of these variants highlights the importance of assessing not just SNPs in these loci, but larger forms of genetic variation including tandem repeats. Ultimately, this VNTR variation should be screened in a larger sample size to determine if such variants could be risk factors for ALS.

Due to the repetitive nature and often high GC content of VNTRs, these domains are not well characterised and even when assessed are difficult to amplify by PCR, clone and sequence. This study proves that it is possible to characterise such domains and resolve the genetic polymorphisms with high resolution. Utilising the QIAxcel advanced system each *REST* VNTR variant could be distinguished as this system resolved 3 bp differences, constituting a single repeat unit of this VNTR (Figure 3.3). This was possible due to the design of a short PCR amplicon (approximately 200 bp), as the High Resolution cartridge used in this project will allow a 3-5 bp resolution for fragments of 100-500 bp. However, analysis using this system was not as accurate for the *CFAP410* VNTR, leading to variant 7 being uncalled on QIAxcel; due to the length of the expansion (150 bp) falling outside of the 600 bp alignment marker range. Therefore, for high resolution genotyping with the QIAxcel advanced system, it is advised that the PCR amplicon be no larger than 500 bp and to only assess VNTRs with small (<6 bp) repeat units. While it is possible to genotype larger VNTRs by using alignment

markers which span a larger genomic window, this is to the detriment of high resolution and therefore makes distinguishing each variant difficult. Moreover, it is not advised to use this system to determine *de novo* variation; in case large indels remain uncalled due to them being outside the scope of the chosen alignment markers. Overall, this system is useful for high-throughput screening and allows one to genotype many samples in a time efficient manner, but it is best to first assess VNTR genetic variation using physical agarose gels and characterise the region prior to moving to this high resolution system. This optimisation process highlights why repetitive DNA is often not well characterised: it is labour intensive, the variation can be difficult to accurately resolve and is more expensive than techniques such GWAS.

The *REST* VNTR is expanding and contracting as repeats of "GGC" and "GGT". While these repeats are found in the general population and therefore the common variants (MAF>5%) of the *REST* VNTR are not causal disease variants, it is important to understand the potential functional implications of GC-rich repeats and these will now be discussed. The majority of disease causing tandem repeats have been identified as triplets of CAG or GGC, with the latter usually being found within protein coding or 5' UTR regions[358]. Interestingly, long read whole genome sequencing analysis has found that GGC repeats found within coding regions are more stable than those found in 5' UTRs, indicating the latter are more polymorphic in the general population[358]. Furthermore, CGG expansions in the 5' UTR of the *Fragile X Mental Retardation 1* (*FMRI)* gene are a known cause of the neurodegenerative disease, Fragile X-associated tremor/ataxia syndrome (FXTAS)[359,360]. Using long read sequencing, a GGC repeat expansion in the 5' UTR

of the *NOTCH2NLC* gene has been identified in patients with neuronal intranuclear inclusion disease[170]. A further study also discovered this *NOTCH2NLC* GGC repeat expansion in two families affected with AD and three families affected with PD, implying the involvement of this particular repeat expansion in other neurodegenerative disorders[361]. The GGC and CGG trinucleotides code for glycine and thus expansions of these bases can induce poly glycine repeats[362]. It is possible that expansions of GGC repeats upstream of the *REST* promoter could lead to expression of such poly glycine repeats, inducing cellular toxicity either through loss of function or indeed through gain of function of either the RNA or protein expressed.

The *CFAP410* VNTR is a much larger domain, consisting of 22 bp and 35 bp repeats (Figure 3.10). Although trinucleotide repeat expansions are a common cause of disease, larger disease-causing repeat expansions do exist. Examples of these include the pentanucleotide ATTTC repeat in the 5' UTR of *DAB1* which causes spinocerebellar ataxia 37 (SCA37), the hexanucleotide GGGGCC repeat expansion in intron 1 of *C9orf72* which is the most common cause of ALS and FTD and the dodecamer CCCCGCCCCGCG repeat expansion in the promoter region of the *CSTB* gene which causes Unverricht-Lundborg disease (ULD/EPM1)[41,42,171,363]. One mechanism of regulation promoted by GC-rich repeats is hypermethylation, which can induce silencing of genes. Additionally, GC-rich sequences are more likely to form intramolecular folds which can bind RNA binding proteins (such as DM1 and C9orf72) and thus dysregulate splicing[364]. Furthermore, GC rich structures have the potential to form abnormal DNA structures, including triple

helices (H-DNA) and G-quadruplexes (G4) which can have a profound effect on the dynamics of transcription[177].

The *CFAP410 VNTR* is a novel GC-rich tandem repeat expansion observed in ALS (variant 7) and further indicates the importance of such domains in this disease. Not only does this region harbour disease specific mutations, but also has the capacity to modulate and drive transcription, with the latter process being bidirectional. Due to the nature of VNTRs they are inherently dynamic and therefore can alter in size when passed to the next generation[364]. Likewise, the phenotypic range of repeat expansion disorders can differ drastically and is partly down to the expansion size of the tandem repeat[364]. Common features are shared across many repeat expansion diseases, including increasing repeat number positively correlating with severity of the disorder and negatively correlating with age of onset: this phenomenon has been observed in a number of diseases, including HD, DM1, XDP and several forms of SCAs[177]. If more ALS cases are identified to harbour variant 2 or 7 of the *CFAP410* VNTR then it would of interest to assess if there is any correlation between the rare VNTR genotypes and disease severity or age of onset. It is possible that contraction or expansion of this VNTR outside of a particular repeat number threshold could lead to dysregulation of this region and thus alter downstream processes involving *CFAP410*, including DNA damage response and cilliogenesis; but this would need to be functionally validated to confirm this hypothesis.

Previous studies have shown that non-coding VNTRs are functional *in vitro*, acting as fine tuners of transcription[145,146,365]. Haddley *et al*. have shown that

VNTRs can act in a tissue specific and stimulus inducible manner, with repeat number variation also inducing allele specific expression[138]. Functional assays in this chapter have shown that both the *REST* and *CFAP410* VNTR are functional modulators and drivers of transcription, agreeing with our hypothesis that these elements could act as potential regulatory elements. Dr Manca had previously tested the potential for the common variants of the *REST* VNTR to drive transcription of the Firefly Luciferase reporter gene *in vitro*; this experiment was repeated and the rare 6R variant was included and compared to the common variants (Figure 3.5). Our results build on the existing evidence of the REST VNTR acting as a potential promoter in vitro. Compared to Dr Manca's previous experiment, a larger fold activity for all common variants was observed in this study: a 9.35, 7.00 and 8.04 fold activity increase for the 7R, 9R and 12R respectively (Figure 3.5) as opposed to the 6.21, 3.75 and 2.9 fold increase which Dr Manca observed. This project has also shown that the *CFAP410* VNTR can regulate transcription and that VNTR polymorphism can induce allele specific expression profiles *in vitro*, as a statistically significant difference in luciferase gene expression between the allele 4 and 5 endogenous orientation constructs was observed (Mann-Whitney U test, p-value = 9.73E-05) (Figure 3.13). Future work should be done to test allele 7 of the *CFAP410* VNTR in this model, to conclude if this ALS-specific expansion induces alterations to gene expression. Interestingly, it was found that the *CFAP410* VNTR could also drive transcription in both the forward and reverse orientation, suggesting this VNTR could serve as a bidirectional promoter. In the endogenous orientation (driving transcription from the same strand as the *CFAP410* gene), a 7.38 fold increase in luciferase activity

compared to the empty pGL3-B vector was observed (Figure 3.14). While this activity was much lower than the SV40 promoter of pGL3-P (which elicited a 29.99 fold increase in luciferase expression), the activity of the endogenous orientation *CFAP410* VNTR construct was comparable to the expression profiles seen for the *REST* VNTR (Figure 3.5). Furthermore, the *CFAP410* VNTR in the reverse orientation (driving transcription in the opposite direction to the *CFAP410* gene) produced a 71.48 increase in luciferase expression when compared to the empty vector. When compared to pGL3-P vector (which contains an SV40 promoter), this is a 2.4 fold increase in reporter gene expression and highlights the high expression capacity of this novel promoter (Figure 3.14).

Overall, we have identified two potential non-coding regulatory domains which are variable in the population. These VNTR polymorphisms could alter affinity for transcription factors and also modulate methylation status at these regions, leading to allele specific gene expression. Furthermore, contraction and expansion of these GC-rich repetitive regions could induce changes in DNA secondary structure, forming structures such as G4 quadruplexes and R-loops, again regulating the dynamics of transcription (all of which have been previously discussed in Chapter 1). Transcriptional dysregulation and thus alteration of gene expression from non-coding genetic variation could disrupt downstream function and cellular processes, which we argue would be just as detrimental as protein loss of function or toxic gain of function which can result from coding mutations. This chapter highlights not only the functional relevance of non-coding regulatory domains but also argues in favour of such regions being potential sources of

missing heritability in complex disease, strengthened by the discovery of novel ALS variants in this study.

Chapter 4: Evaluating genetic variation of a human specific SVA retrotransposon in the *NEK1* locus and its association with ALS risk.

4.1 Introduction

It is understood that missense, nonsense and frameshift mutations in coding sequence of genes will alter protein structure and function and consequently effect downstream pathways and thus such mutations are associated with disease risk and predisposition. As mentioned previously in Chapter 1, *NEK1* coding mutations have now been discovered which confer risk for ALS[62-65,366].

NIMA (Never in Mitosis Gene A)-Related Kinase 1 (*NEK1*) is part of a family of 11 serine/threonine kinases which regulate the cell cycle[367]. NEK2, 6, 7 and 9 regulate the centrosome cycle and spindle assembly[368], while NEK1, and 8 modulate stability and physiology of primary cilium[367]. NEK1, 8, 10 and 11 all have established roles in DNA damage response (DDR) also[367,369]. NEK1 consists of an N-terminal kinase catalytic domain and several coiled-coil (CC) domains and localises in primary cilia, centrosomes, the cytoplasm, mitochondria and DNA damage sites within the nucleus[367,369,370].

Chen *et al* have shown that *NEK1* is required for early DDR and cell cycle checkpoint activation. Immunofluorescence on human kidney (HK2) cells showed that NEK1 protein localises to the nucleus in response to DNA damage induced by chemotherapeutic drugs (cisplatin) ionising radiation (IR), oxidative injury (H_2O_2) and UV radiation. Furthermore, using RNA interference to silence *NEK1*, it was found that *NEK1* -/- cells (HK2 and murine kat2J) exhibited hypersensitivity to DNA damage caused by IR and that these cells failed to induce arrest of both G1/S and G2/M-phase cell cycle checkpoints. These *NEK1* deficient cells failed to activate

checkpoint kinase 1 (Chk1) and checkpoint kinase 2 (Chk2), highlighting the requirement of *NEK1* to activate these cell cycle checkpoint kinases. By assessing γH2AX nuclear foci loss as an indicator of double strand break repair, it was found that *NEK1* -/- murine cells had persistent DSBs 24 hours after treatment with IR (while all γH2AX nuclear foci were depleted in wildtype cells), showing that NEK1 -/- cells had difficulty repairing DSBs. An excessive number of chromosome breaks per spread was also identified in NEK1 -/- cells. Overall, NEK1 deficient cells fail to arrest cell cycle checkpoints due to lack of Chk1 and Chk2 activation, leading to persistent DNA damage and increased chromosomal instability[371]. An additional study by Chen *et al.* found that *NEK1* -/- katj2 cells become aneuploid, highlighting a role for NEK1 in chromosome segregation. Knockdown of *NEK1* in renal tubular epithelial cells and tail fibroblasts also led to cell immortalisation and loss of contact inhibition[372].

NEK1 has established role in homologous recombination (HR). Spies *et al*. generated *NEK1* depleted HeLa cells and found persistent Rad51 foci 10 hours after treatment with IR[373]; a known marker of DSBs which is removed by Rad54 to facilitate repair through HR[374]. Protein immunoprecipitation and immunoblotting proved that NEK1 directly interacts with Rad54, specifically phosphorylating Ser572 on Rad52 in G2 phase, inducing the removal of Rad51 from chromatin and mediating DNA damage repair through HR[373].

NEK1 is also a known regulator of cell death[370,375]. Voltage dependent anion channel 1 (VDAC1) is an outer mitochondrial membrane protein which serves as a metabolite gatekeeper and receptor for pro and anti-apoptotic proteins,

modulating apoptosis and cell survival[376]. Chen *et al*. through yeast two-hybrid screening, GST pulldown assays and co-immunoprecipitation have shown that VDAC1 interacts with NEK1. They found that under basal conditions and in response to DNA damage that NEK1 phosphorylation can regulate the opening and closing of this channel: phosphorylation of VDAC1 on Ser193 by NEK1 caused the channel to close, preventing an efflux of cytochrome c from initiating pro-apoptotic cascades and thus limiting mitochondrial mediated cell death[370,375].

Prior to implication in ALS, *NEK1* mutations had been discovered in several conditions[377-380]. Thiel *et al*. used homozygosity mapping in two families to identify three autosomal recessive *NEK1* mutations which cause short-rib polydactyly syndrome type majewski (SRPS type Majewski): (c.379>T - p.Arg127X, c.869-2A>G - intronic, c.1640 dup - p.Asn547LysfsX2)[378]. Monroe *et al*. using whole exome sequencing (WES) discovered compound heterozygous mutations of *NEK1* (c.464G>C, r.397_464del and c.1226G>A, p.Trp409X) in patients with oral-facial-digital syndrome (Mohr syndrome)[379]. *NEK1* has also been identified as the second gene to cause axial spondylometaphyseal dysplasia (axial SMD): the first identified gene was *CFAP410* (*C21orf2*)[380].

NEK1 was first implicated in ALS through an exome sequencing study by Cirulli *et al*. in 2015, which found *NEK1* heterozygous-LOF variants in 0.835% of cases and 0.091% of controls but only reached experiment wide significance (17,248 genes tested) by combining the discovery and replication analyses[62]. Brenner *et al*. in 2016 also utilised exome sequencing in 265 European FALS cases and 827 controls and found a statistically significant enrichment of *NEK1*

heterozygous-LOF mutations in the FALS patients: with an allele frequency of 0.57% in cases and 0.06% in controls. Rare missense mutations of NEK1 (MAF<0.01) were also found in cases (allele frequency of 3.2%) and controls (1.87%) but did not reach statistical significance. Similarly, even rarer missense mutations (MAF<0.001) were also identified, with an allele frequency of 1.13% in cases and 0.6% in controls but again this was not statistically significant[64].

Through Project MinE (and funding raised from the ice bucket challenge[381]), a study in 2016 by Kenna *et al*. identified *NEK1* variants which confer risk for ALS through rare variant burden analysis. Utilising exome sequencing, this study aimed to assess combined rare variant frequency of the gene within an ALS case and control cohort and demonstrated a statistically significant overrepresentation of heterozygous-LOF variants of *NEK1* in both FALS and SALS patients. In total 120 non-synonymous variants of *NEK1* were found; specifically, LOF variants in 1.2% of FALS, 1% of SALS and 0.17% of controls and missense mutations in 1.8% of FALS, 1.3% of SALS and 1.2% of controls. The same group also identified NEK1 p.Arg261His missense variant as a candidate risk variant for ALS, identifying this variant in 1.7% of FALS, 1.6% of SALS and 0.69% of controls. Overall, *NEK1* risk variants were identified in approximately 3% of European and European-American ALS patients[63]. A study in a Belgian ALS cohort also detected the p.Arg261His missense variant in cases (MAF=0.90%) and controls (MAF=0.33%) and two ALS specific LOF mutations were found. Interestingly, a third LOF mutation (p.Ser1036*) was found in two siblings with ALS but also in one unrelated control sample[65]; this variant was also found in one control sample in a previous study assessing exome variation in European FALS patients[64]. Thanks to the global

collaboration through Project MinE and the generation of large sample sizes, it has been possible to detect novel rare mutations in genes such as *NEK1* which further indicate the contribution of rare variants in the genetic architecture of ALS.

There has been a focus on SNPs in the aetiology of complex disease, with the majority of GWAS SNPs found in non-coding regions of the genome[382]. However, common SNPs (MAF>1%) are estimated to account for a fraction (~8%) of the heritability of ALS[11]. Following the emergence of next generation sequencing there has been a shift towards whole genome sequencing (WGS), with more attention now falling on larger forms of genetic variation such as indels and structural variants exhibited by repetitive elements. Repeat DNA polymorphisms, which have recently been discovered to be contributing to the aetiology of ALS, are often found in non-coding regions of the genome[41,42,86]. Non-coding DNA can have profound effects on the regulation of gene expression: regulating transcription, mRNA splicing and modifying enhancer, silencer and insulator sequences[137,138,383-385].

Mobile DNA constitutes a major source of non-coding genetic variation, accounting for approximately half of the human genome[322,386]. Of particular importance are the non-LTR retrotransposons, including LINE-1, *Alu* and SVA elements, known to still be active and mobile in the human genome[235,387,388]. While these elements can copy and paste across the genome it is also important to consider the regulatory properties they possess when they are not mobilising. Of particular interest are the youngest subclass of the non-LTR retrotransposons: Sine-VNTR-*Alu*s (SVAs). Our lab has previously investigated the genetic variation of

SVAs and the potential effect these polymorphisms have on transcription and gene regulation, specifically in loci associated with PD and ALS[224,389].

These hominid specific elements can exhibit genetic polymorphism in two distinct ways: sequence variation within the element itself and presence or absence of the entire element in the general population[223,224,390,391]. An exemplar of the importance of the polymorphism in SVAs as mentioned in Chapter 1, is the SVA retrotransposon insertion polymorphism within the *TAF1* gene causes X-linked Dystonia Parkinsonism (XDP)[296]; a neurodegenerative disease affecting individuals with maternal heritage from the Philippines[297]. This element is also polymorphic within the SVA itself and this genetic variation drives XDP progression, as the repeat unit size of 5' hexamer (CT element) is inversely correlated with age of onset for the disease[297].

Since the discovery of *NEK1* variants being implicated in ALS, there was a focus only on the coding regions of this gene, with past analyses concentrating on exome sequencing[62-65,366]. We decided to assess non-coding variation present at this locus; to see if we could uncover potential transcriptional regulatory domains and regions of variation important in ALS aetiology. Upon analysis of the *NEK1* locus we identified a human specific SVA element within the *NEK1* gene which we hypothesised could act as a potential modulator of gene expression at this locus and postulated the presence of variation within the SVA could support differential gene expression. We decided to address the genetic variation of this SVA in ALS cases and controls (MNDA UK and Project MinE cohorts) to decipher if there was any association with disease. The aim of this chapter was to focus on characterising

this non-coding region in *NEK1*, using a number of strategies, to determine if it could contain risk variants for ALS and also to highlight SVA retrotransposons as an important source of genetic variation within the human genome which may contribute to the susceptibility of complex diseases.

4.2 Hypothesis and aims

Hypothesis:

The SVA-D within *NEK1* is polymorphic and sequence variants of this element could be potential risk factors for ALS.

Aims:

Genotype the full length SVA-D in an MNDA cohort to identify potential risk factors for ALS by PCR.

Genotype the composite regions of the SVA by PCR: CT element, VNTR and Poly A tail to identify specific polymorphic regions.

Expand the PCR genotyping analysis bioinformatically via Isaac Variant Caller using WGS data from Project MinE.

Validate the Isaac Variant Caller data using PCR generated genotyping data.

4.3 Results

4.3.1 Bioinformatic analysis of the *NEK1* locus

The *NEK1* locus was assessed using UCSC Genome Browser to identify potential regulatory elements. This analysis of the *NEK1* gene allowed identification of an SVA retrotransposon within intron 11 of the protein coding isoform termed transcript 1 of *NEK1* in UCSC (hg19, chr4:170,490,244-170,492,723) (Figure 4.1). This element was found using the Repeatmasker data track (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org); a program which screens the reference genome for interspersed repeats such as retrotransposons and outputs a detailed annotation over their location, which can then be displayed on UCSC. Initial inspection showed that this SVA was of the subclass D and appeared to be missing both the canonical CT rich repeat (known as a CT element) and the poly adenylation signal (known as a poly A tail) (Figure 1.3). However, closer examination showed that Repeatmasker had incorrectly annotated these regions as separate simple repeats adjacent to the SVA. Therefore, when inspecting the primary DNA sequence of the SVA flanking sequence was included to account for this annotation error and thus encompass both the CT element and Poly A tail. This extended region we defined as the full length *NEK1* SVA-D.

The SVA-D was anti-sense with respect to the orientation of the NEK1 gene and was approximately 1.8kb in length. The vertebrate multiz alignment & conservation of 100 vertebrate species, from Phylogenetic Analysis with Space/Time models (PHAST program)[311], showed that this genomic region was not

conserved in chimps, gorillas, gibbons, rhesus macaques, rats and mice, indicating that this SVA-D was human specific (Figure 4.1A). To infer if this non-coding region had any regulatory function, data from The Encyclopedia of DNA Elements (ENCODE) was assessed: a project portal made available on UCSC which aims to define any genomic regions with function[312]. According to ENCODE data the SVA-D identified was directly adjacent to DNaseI hypersensitive clusters (regions of open chromatin) and histone marks, specifically layered H3K4Me1 marks, which are often found downstream of transcriptional start sites and associated with regulatory domains such as enhancers, and H3K27Ac marks, which are associated with active promoters and enhancers[312] (Figure 4.1B). There are gaps in the ENCODE data directly over the SVA element (Figure 4.1B), as repetitive DNA is difficult to map due to amplification noise and can lead to inaccurate data interpretation, therefore these problematic regions are often discarded and ignored in downstream analyses[392].

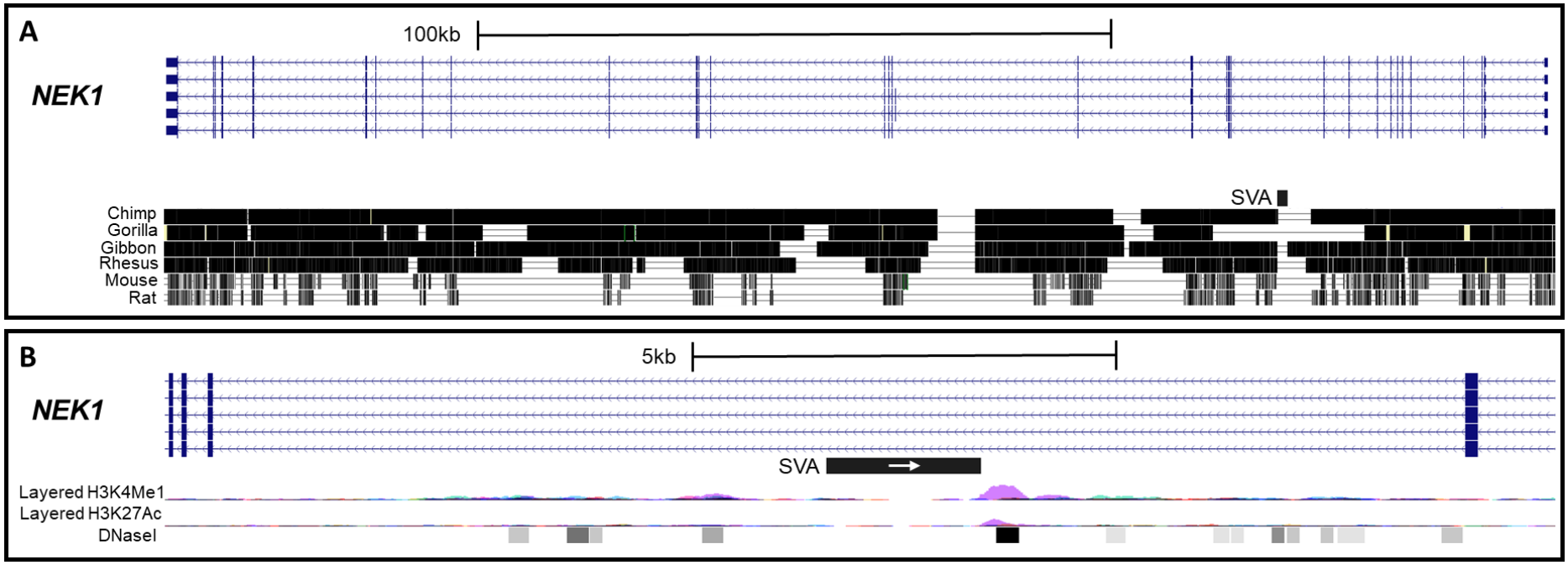


Figure 4.1. The *NEK1* gene contains a human specific SVA retrotransposon.

A: *NEK1* loci present on chromosome 4 (UCSC, hg19: chr4:170,490,244-170,492,723). There are five RefSeq curated transcripts (protein coding isoforms) of *NEK1*. An SVA element was identified within an intron of *NEK1*. This SVA is of the D subclass, is present only in humans and is anti-sense to the orientation of *NEK1* (indicated by white arrow). **B**: Zoomed visualisation of the SVA present within *NEK1*. ENCODE data from UCSC shows the levels of enrichment of histone marks within this locus, specifically signals for mono-methylation H3K4Me1 and acetylation H3K27Ac, both associated with regulatory elements. DNaseI hypersensitive clusters show sections of open and accessible DNA.

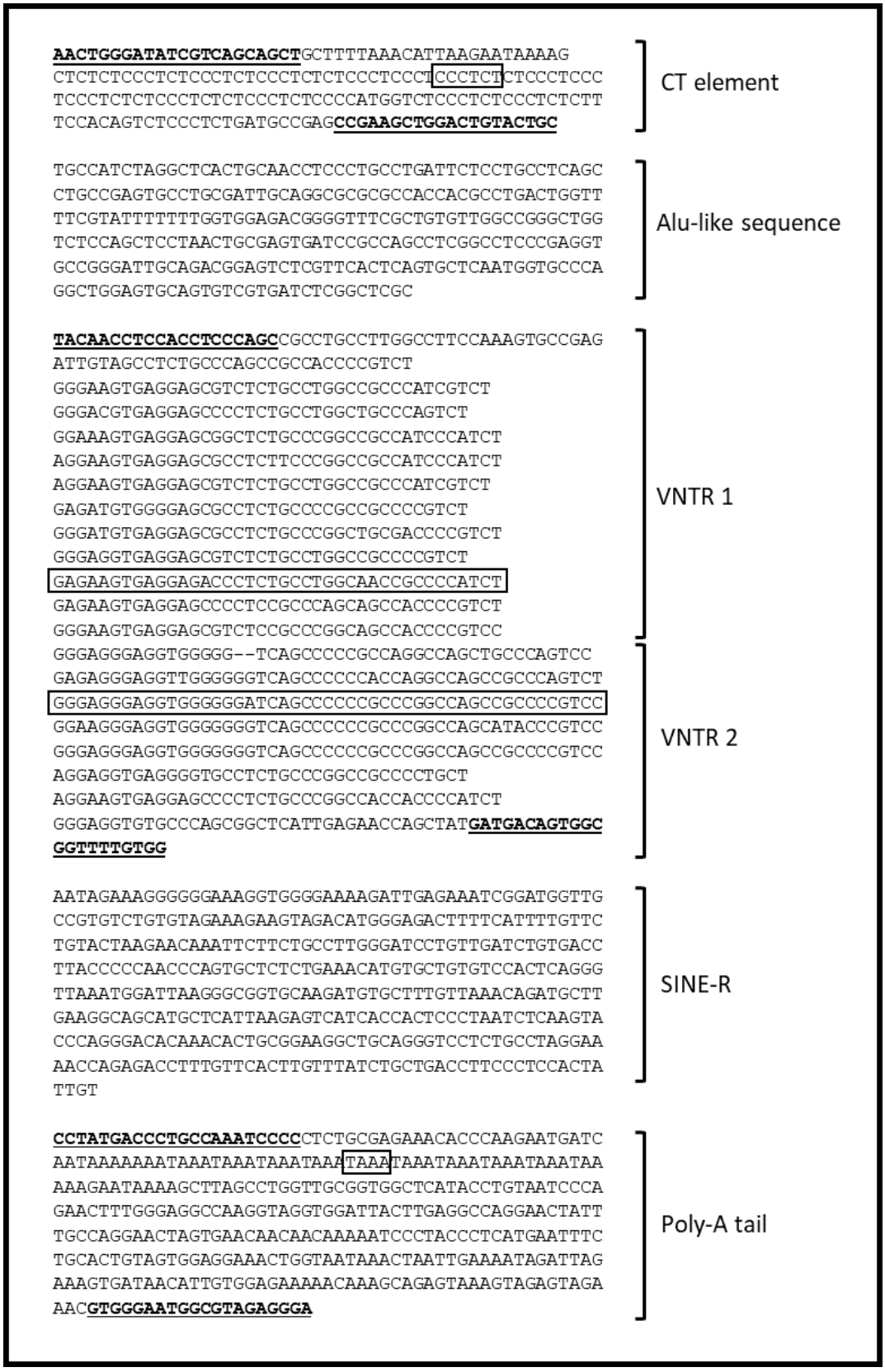


Figure 4.2. Annotation of the *NEK1* SVA-D composite regions.

DNA sequence of the *NEK1* SVA-D taken from UCSC (hg19) and split into its respective compartments in the 5' to 3' orientation (top to bottom). The SVA-D

began with a canonical CT hexamer repeat (CT element), followed by an Alu-like region, two VNTRs, the SINE-R region (which defines the SVA subclass) and finally a 3' Poly-A tail. Primers that were subsequently designed for PCR are shown in bold and underlined. All repeats are boxed and were aligned by eye and are therefore arbitrary.

The SVA within *NEK1* was found to be built of distinct composite domains: 5' CT-rich repeat (known as a CT element), an Alu-like sequence, two VNTR domains, a SINE-R region and a 3' poly adenylation signal (poly A tail). It was hypothesised that due to the repetitive nature of the sequences it was possible that the 5' CT element, central VNTRs and 3' poly A tail could each exhibit sequence polymorphisms within the general population, a phenomenon which is consistent with previous work[223,224]. The CT element in the reference genome was 189 bp in length and contained the canonical "CCCTCT" hexamer repeat which has been previously described[220,225]; constituting 19 hexamer repeats in total. This SVA also contained two central VNTRs, the first of which contained 11 imperfect repeats of approximately 40 bp and the second which was built of 5 repeats of approximately 49 bp. The 3' poly A tail of this SVA contained 10 repeats of TAAA and was approximately 60 bp long. The aim of this study was to assess the potential for these regions to exhibit genetic variation.

4.3.2 Characterising genetic variation of the *NEK1* SVA-D

Previous studies have shown that SVA elements are polymorphic in the human population[224,297,389], therefore we decided to genotype this non-coding element found in *NEK1* (a recently discovered ALS risk gene) in a MNDA cohort of

ALS patients and controls. Initial genotyping of the full length SVA by PCR showed no variation on agarose gels (Figure 4.3B1). However, as the full length SVA is a 1.8kb amplicon it was difficult to determine small genetic changes, for example hexamer repeats of the CT element, therefore primers were designed for the regions known to be potentially polymorphic in SVAs[220,223,224] (the CT element, VNTR and Poly A tail) (Figure 4.3A) and these domains were then separately genotyped by PCR in a Motor Neurone Disease Association (MNDA) cohort of ALS cases and controls. In total, four alleles of the CT element, three alleles of the VNTR and six alleles of the poly A tail were identified (Figure 4.3B).

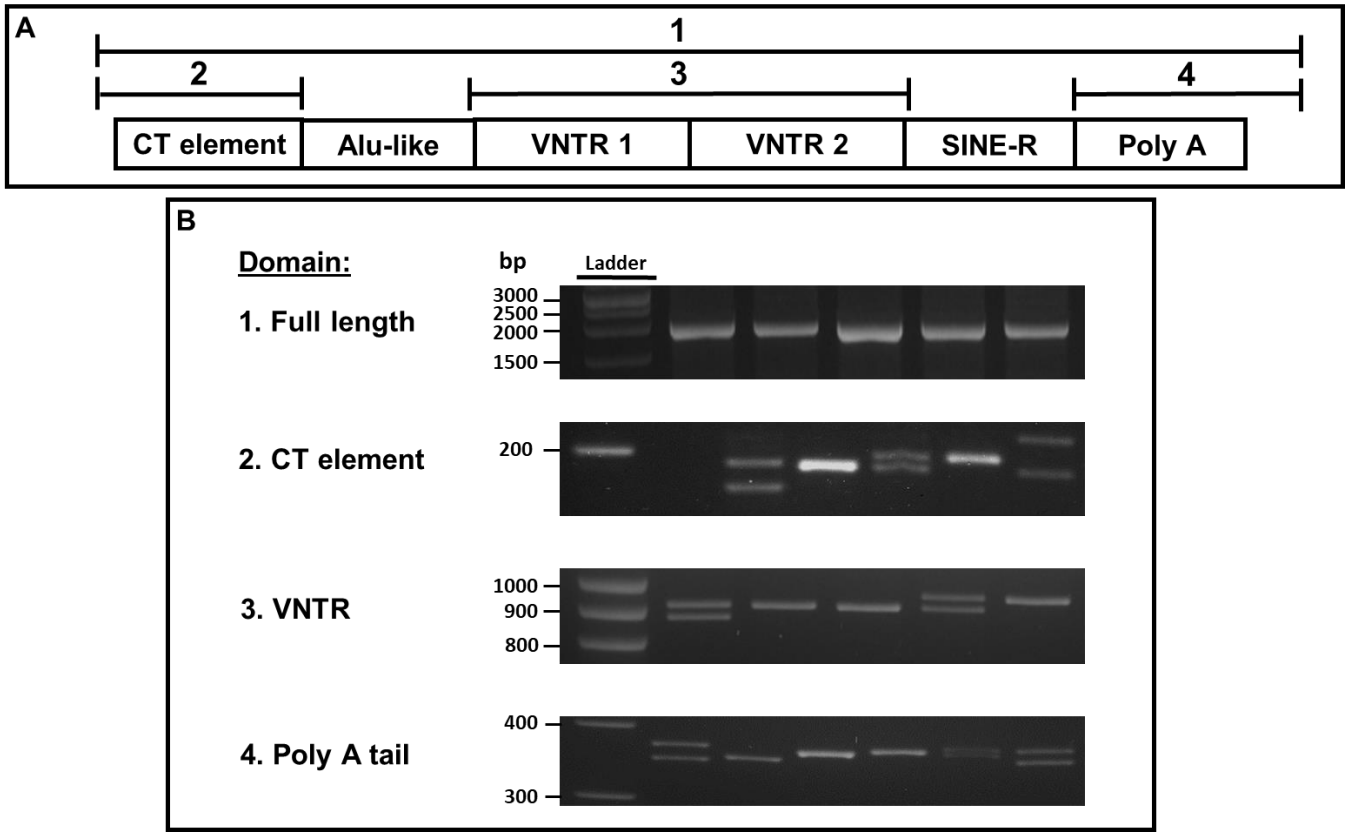


Figure 4.3 Structure and PCR design of the *NEK1* SVA-D.

A: Structurally, the SVA contained a 5' $(CCCTCT)_n$ hexamer repeat, an Alu-like region, two variable number tandem repeats (VNTRs), a SINE-R region and a 3' Poly A tail. Location of PCR amplicons are indicated by numbered lines: 1 = 1877 bp, 2 = 189 bp, 3 = 888 bp, 4 = 366 bp (SVA

composite regions are not to scale). **B**: PCR amplification and gel electrophoresis of the SVA-D within the *NEK1* locus in an ALS and matched control cohort**. 1B**: Full length *NEK1* SVA-D (1877 bp). Samples were run at 100V on 1% agarose for 1 hour. There was no clear sign of polymorphism in the full length SVA. **2B**: CT element of the SVA (189 bp). Samples were run at 100V on 3% agarose for 4.5 hours. The CT element of the *NEK1* SVA-D was found to be polymorphic and four alleles were identified. **3B**: VNTRs of the NEK1 SVA-D (888 bp). Samples run at 100V on 1.2% agarose for 2 hours. The VNTR was also found to be polymorphic and three alleles were identified. **4B**: Poly A tail of the SVA-D (366 bp). Samples run at 100V on 3% agarose for 4 hours. The Poly A tail of the *NEK1* SVA-D was found to be polymorphic, with six alleles identified. The five samples shown for each region of the SVA were not from the same individuals but were all from the same MNDA cohort.

Table 4.1. Allele and genotype frequencies of the *NEK1* SVA CT element in an ALS cohort and matched controls.

A: The four identified alleles of the *NEK1* SVA-D CT element in an ALS cohort (n = 976) and matched controls (n = 992). Alleles 1 and 4 of the CT were found only in the ALS cohort. There was no significant difference in allele frequency between the ALS cohort and matched controls (Fisher's exact test). **B:** The five identified genotypes of the *NEK1* SVA-D CT element in an ALS cohort (n = 488) and matched controls (n = 496). There was no significant difference in genotype frequency between the ALS cohort and matched controls (Fisher's exact test).

A

Cohort / Allele	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1	1	0.10	0	0.00	1	0.10	0.50
2	802	82.17	833	83.97	1635	-1.80	0.31
3	170	17.42	159	16.03	329	1.39	0.43
4	3	0.31	0	0.00	3	0.31	0.12
Total	976	100.00	992	100.00	1968	0.00	N/A

B

Cohort / Genotype	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1,2	1	0.20	0	0.00	1	0.20	0.50
2,2	327	67.01	347	69.96	674	-2.95	0.34
2,3	144	29.51	139	28.02	283	1.48	0.62
2,4	3	0.61	0	0.00	3	0.61	0.12
3,3	13	2.66	10	2.02	23	0.65	0.53
Total	488	100.00	496	100.00	984	0.00	N/A

Genomic DNA from the blood of ALS patients (n = 488) and matched controls (n = 496) of an MNDA cohort was genotyped and categorised into the alleles observed for the CT element of the *NEK1* SVA-D and the frequency of the four alleles was calculated. Table 4.1 shows allele 1-4 listed from smallest to largest. Allele 2 was the most frequent in this population, accounting for 82.17% of ALS patients and 83.97% of matched controls, but no significant difference in allele frequency between ALS cases and controls was observed (Fisher's exact test, p-value = 0.31). The second most frequent variant, allele 3, was observed in 17.42% of ALS patients and 16.03% of controls and again there was no significant difference in allele frequency between ALS cases and controls (Fisher's exact test, p-value = 0.43). Alleles 1 and 4 were rare (MAF<1%) and only observed in the ALS population, accounting for 0.10% and 0.31% respectively (Table 4.1). Genotype 2,2 was the most common genotype observed, accounting for 67.01% in ALS cases and 69.96% of controls. Overall, no significant difference in genotype frequency was observed across case and control (Table 4.1) and no relationship with ALS was observed. Due to alleles 1 and 4 only being present in ALS patients we decided to expand this analysis into a larger sample size, through utilising WGS data from Project MinE. This analysis will be outlined in detail later on in this chapter.

Table 4.2. Genotype and allele frequencies of the *NEK1* SVA VNTR in an ALS cohort and matched controls.

A: The four identified genotypes for the *NEK1* SVA-D VNTR in an ALS cohort (n = 48) and matched controls (n = 45). **B**: The three identified alleles for the NEK1 SVA-D VNTR in an ALS cohort (n = 96) and matched controls (n=90). There was no significant difference in genotype or allele frequency between the ALS cohort and matched controls (Fisher's exact test).

A

Cohort / Allele	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1	2	2.08	2	2.22	4	-0.14	1.00
2	94	97.92	84	93.33	178	4.58	0.16
3	0	0.00	4	4.44	4	-4.44	0.05
Total	96	100.00	90	100.00	186	0.00	N/A

B

Cohort / Genotype	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1,2	2	4.17	2	4.44	4	-0.28	1.00
2,2	46	95.83	40	88.89	86	6.94	0.26
2,3	0	0.00	2	4.44	2	-4.44	0.23
3,3	0	0.00	1	2.22	1	-2.22	0.48
Total	48	100.00	45	100.00	93	0.00	N/A

The *NEK1* SVA-D VNTR was also genotyped within the same MNDA cohort of ALS patients (n = 48) and controls (n = 45) and genotype and allele frequency were calculated (Table 4.2). Allele 2 was the most frequent allele in this population, accounting for 97.9% of ALS patients and 93.3% of matched controls, but there was no significant difference in allele frequency (Fisher's exact test, p-value = 0.16). Allele 1 was present in 2.1% of ALS patients and 2.2% of controls, no significant difference in allele frequency between ALS cases and controls was observed (Fisher's exact test, p-value = 1.00). Allele 3 was only observed in the control population, accounting for 4.44% of the population. All VNTR variants found were common (MAF>1%) and no ALS specific variants were found. However, allele 3 was only present in controls. There was no significant difference in allele distribution across the two populations (Fisher's exact test) and no association with ALS was observed.

Table 4.3. Genotype and allele frequencies of the *NEK1* SVA Poly A tail in an ALS cohort and matched controls.

A: The fifteen identified genotypes of the NEK1 SVA D Poly A tail in an ALS cohort (n = 42) and matched controls (n = 43). **B**: The six identified alleles of the NEK1 SVA D Poly A tail in an ALS cohort (n = 84) and matched controls (n = 86). There was no significant difference in genotype or allele frequency between the ALS cohort and matched controls (Fisher's exact test).

A

Cohort / Allele	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1	1	1.19	1	1.16	2	0.03	1.00
2	13	15.48	17	19.77	30	-4.29	0.55
3	7	8.33	6	6.98	13	1.36	0.78
4	20	23.81	22	25.58	42	-1.77	0.86
5	38	45.24	37	43.02	75	2.21	0.88
6	5	5.95	3	3.49	8	2.46	0.49
Total	84	100.00	86	100.00	170	0.00	N/A

B

Cohort / Genotype	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1,4	0	0.00	1	2.33	1	-2.33	1.00
1,5	1	2.38	0	0.00	1	2.38	0.49
2,2	0	0.00	2	4.65	2	-4.65	0.49
2,4	6	14.29	3	6.98	9	7.31	0.31
2,5	6	14.29	8	18.60	14	-4.32	0.77
2,6	1	2.38	2	4.65	3	-2.27	1.00
3,3	1	2.38	2	4.65	3	-2.27	1.00
3,4	1	2.38	0	0.00	1	2.38	0.49
3,5	4	9.52	2	4.65	6	4.87	0.43
4,4	4	9.52	7	16.28	11	-6.76	0.52
4,5	4	9.52	4	9.30	8	0.22	1.00
4,6	1	2.38	0	0.00	1	2.38	0.49
5,5	11	26.19	11	25.58	22	0.61	1.00
5,6	1	2.38	1	2.33	2	0.06	1.00
6,6	1	2.38	0	0.00	1	2.38	0.49
Total	42	100.00	43	100.00	85	0.00	N/A

The Poly A tail of the *NEK1* SVA-D was genotyped in the same MNDA cohort of ALS patients (n = 42) and controls (n = 43) and genotype and allele frequency were calculated (Table 4.3). Allele 5 was the most common allele, present in 45.24% of ALS patients and 43.02% of controls, and no significant difference in allele frequency was observed (Fisher's exact test, p-value = 0.88). Allele 4 accounted for 23.81% of ALS patients and 25.58% of controls, but again no significant difference was found (Fisher's exact test, p-value = 0.86). Furthermore, allele 2 was present in 15.48% of cases and 19.77% of controls (no significant difference, Fisher's exact test, p-value = 0.55), while allele 3 was only identified in 8.33% of cases and 6.98% of controls (no significant difference; Fisher's exact test, p-value = 0.78). Allele 6 only accounted for 5.95% of cases and 3.49% of controls (Fisher's Exact test, p-value = 0.49), and allele 1 was the rarest variant, being identified only once in each cohort (Fisher's Exact test, p-value = 1.00). Overall, no significant difference in frequency of any genotype was observed between cases and controls (Fisher's exact test) (Table 4.3) and again no association with ALS was observed.

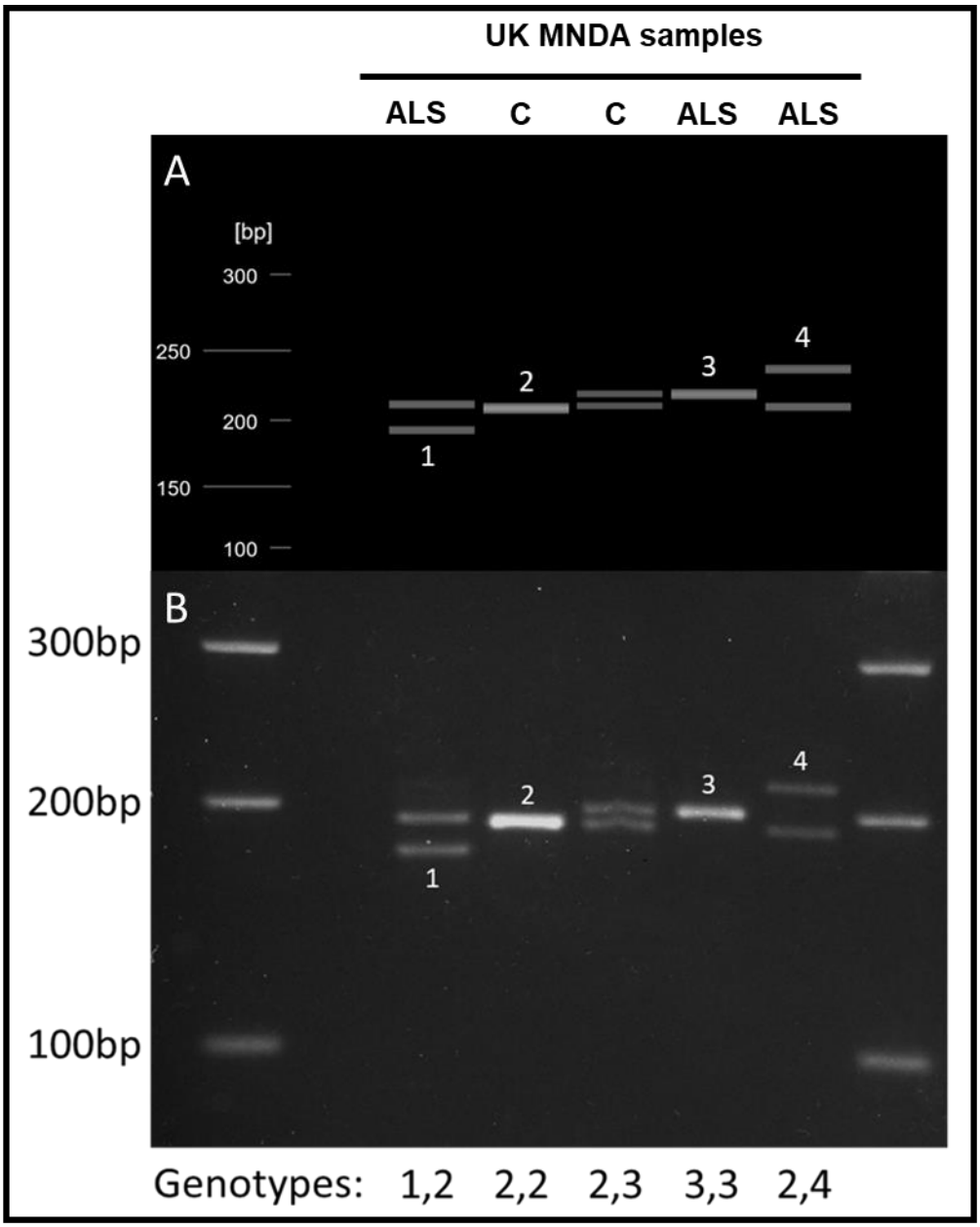


Figure 4.4. *NEK1* SVA-D CT element genotyping

PCR amplification, gel electrophoresis and calculation of allele and genotype frequency of the *NEK1* SVA-D CT element in an MNDA cohort (ALS and matched control) (n = 982). **A**: gel capillary electrophoresis performed using the QIAxcel advanced system and electronic gel image generated using the QIAxcel ScreenGel software showing all four alleles of the CT element. **B**: agarose gel electrophoresis of the CT element; samples run on 3% agarose at 100V for 4.5 hours.

As shown in Figure 4.4 the CT element of the *NEK1* SVA-D was found to be polymorphic and four alleles were identified and the estimated amplicon sizes were 171 bp, 189 bp, 207 bp and 213 bp (equating to 16, 19, 20 and 23 hexamer repeats respectively) (Figure 4.4).

4.3.3 The *NEK1* SVA-D CT element genotype is the same across brain and blood of the same ALS patient

As previously discussed, one of the major risk variants of ALS is an intronic repeat expansion of GGGGCC within *C9orf72*. This intronic VNTR has been shown to be variable between neuronal and non-neuronal tissues of the same ALS patient[393]. To address if a similar phenomenon was possible with the *NEK1* SVA CT element, PCR was performed over this domain using motor cortex and blood DNA from the same ALS patient (n = 7) (please refer to Chapter 3, Section 3.3. for the same analysis in the *CFAP410* VNTR; Figure 3.9). However, in this small sample size, the genotype of the CT element was the same in both the motor cortex and blood from the same ALS patient (Figure 4.5).

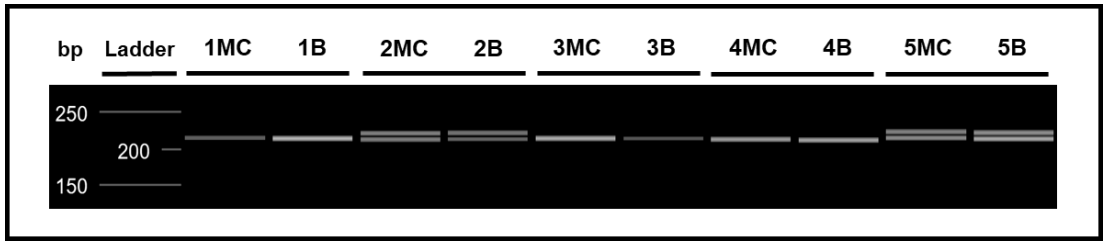


Figure 4.5 *NEK1* SVA-D CT element genotyping in matched brain and blood.

PCR amplification and gel capillary electrophoresis of the *NEK1* SVA-D CT element within matched motor cortex (MC) and blood (B) for 5 ALS patients. There was no variation seen in the genotype of the CT element between motor cortex and blood

for each patient. Gel capillary electrophoresis performed using the QIAxcel advanced system and electronic gel image generated using the QIAxcel ScreenGel software.

4.3.4 Sequencing the *NEK1* SVA-D CT element variants

To characterise the different alleles of the *NEK1* SVA-D CT element the full length SVA was cloned into pCR®-Blunt vector (ThermoFisher) and the CT element was sequenced from each clone (Figure 4.7). The full length SVA was cloned rather than the CT element alone as the CT variants were difficult to resolve on an agarose gel (predicted to be repeats of 6 bp) and thus had to be run on 3% agarose gels for 4 hours to in order to visually resolve the polymorphisms (Figure 4.4). This was impractical for the application of cloning, as it was stated in the Wizard® SV Gel and PCR Clean-Up System (Promega) protocol that excising and purifying DNA from agarose gels above 1% would lead to reduced yield of DNA. We therefore decided to amplify the full length SVA and run this on a 1% agarose gel to ensure a high yield of DNA from the clean-up protocol (Chapter 2 Section 2.2.3.3 Ligation of DNA fragments into pCR®-Blunt intermediate vector). Following successful ligation of the SVA into pCR®-Blunt vector (ThermoFisher) (Chapter 2 Section 2.2.3.3 Ligation of DNA fragments into pCR®-Blunt intermediate vector) and transformation into chemically competent DH5α *E.coli* (Chapter 2 Section 2.2.3.6) the plasmid DNA was extracted and purified (Chapter 2 Section 2.2.4.1) and sequenced using dye-terminator Sanger sequencing by Source Bioscience (Chapter 2 Section 2.2.6) (Figure 4.6).

4.3.5 Troubleshooting the sequencing of allele 4 of the *NEK1* SVA-D CT element

Initially two sequencing reactions for allele 4 of the CT element failed, with two separate samples stalling at position 205 (Figure 4.6A and B). To address this issue a sequencing primer on the anti-sense strand was designed, attempting to sequencing from the reverse direction (Figure 4.6C). This strategy yielded high quality sequencing read, avoiding secondary structure generated on the sense strand.

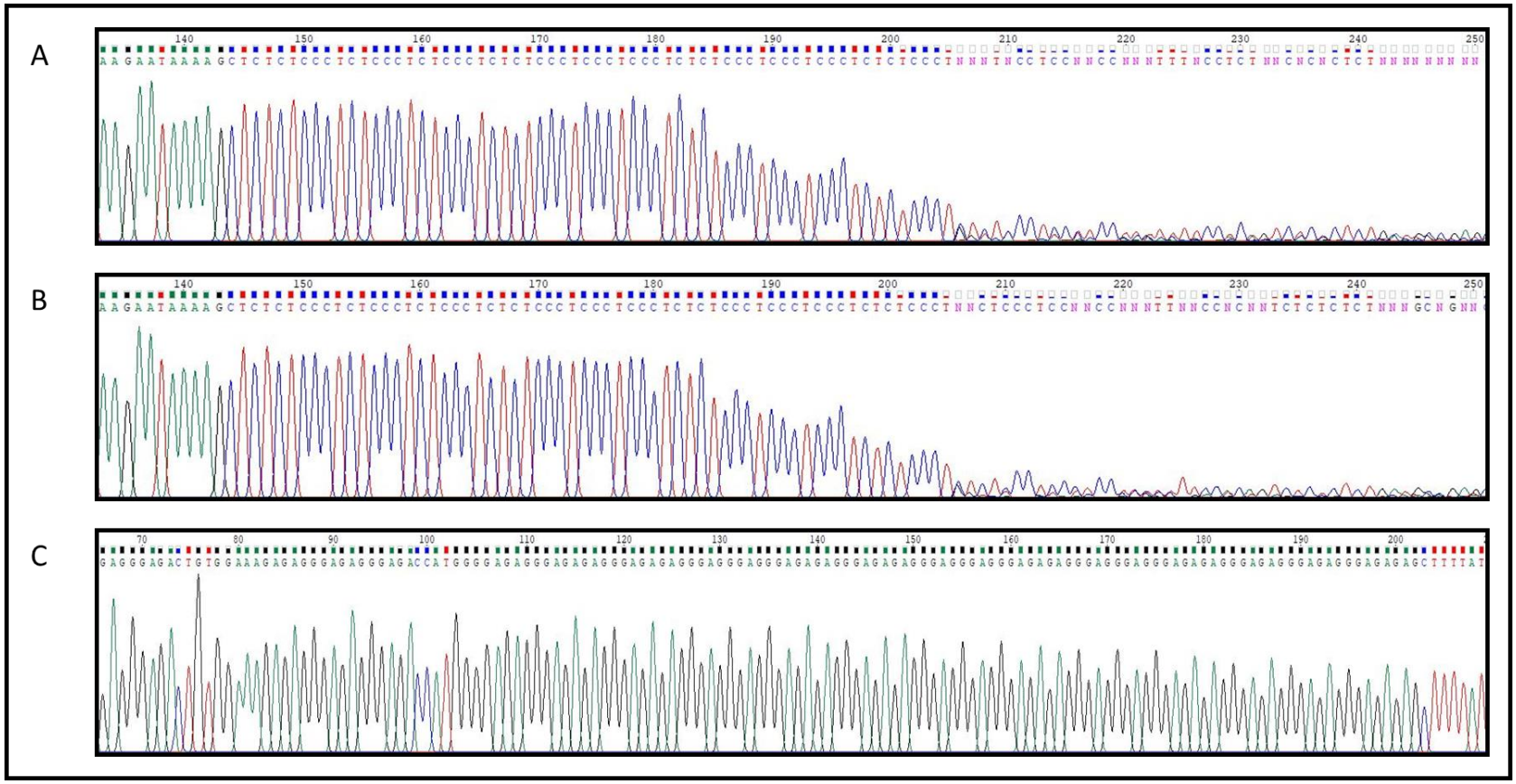


Figure 4.6 Allele 4 of the *NEK1* SVA-D CT element stalls Sanger sequencing reaction on the sense strand.

Dye-terminator Sanger sequencing of allele 4 of the NEK1 SVA-D CT element. **A**: Failed sequencing reaction for one of the MNDA ALS patients with allele 4 of the *NEK1* SVA-D CT element. **B**: Failed sequencing reaction from another MNDA ALS patient with allele 4 of the *NEK1* SVA-D CT element. Primer used for both sequencing reactions was the same forward primer used for PCR amplification of the CT element and binds to the

sense strand. **C**: Successful sequencing reaction for MNDA ALS patient with allele 4. A new primer was designed to bind to the anti-sense strand of the CT element. All DNA sequences were visualised using Chromas (http://technelysium.com.au/wp/chromas/)

4.3.6 The *NEK1* SVA-D CT element consists of two octamer repeats

After successfully sequencing all four CT element variants, the sequences were aligned using the multiple sequence alignment programme, MUSCLE[394,395], and visualised in Jalview (https://www.jalview.org/)[396] (Figure 4.8B). It was found that the CT element was not solely built of the canonical "CCCTCT" hexamer repeat but larger repeat polymorphisms were also present: octamer repeats of either "CCCTCTCT" or "CCCTCCCT". Although both hexamer and octamer repeats were present and overlapping, not all observed expansions and contractions were divisible by 6 and therefore the polymorphisms were confirmed to be 8 bp repeat units. Compared to allele 2 (the reference variant) allele 1 equated to a 16 bp deletion (2 repeats), while allele 3 corresponded to an 8 bp expansion (1 repeat) and allele 4 was a 24 bp expansion (3 repeats). It was confirmed that alleles 1-4 consisted of 12 (98 bp), 14 (114 bp), 15 (122 bp) and 17 (138 bp) octamer repeats respectively (Figure 4.7).

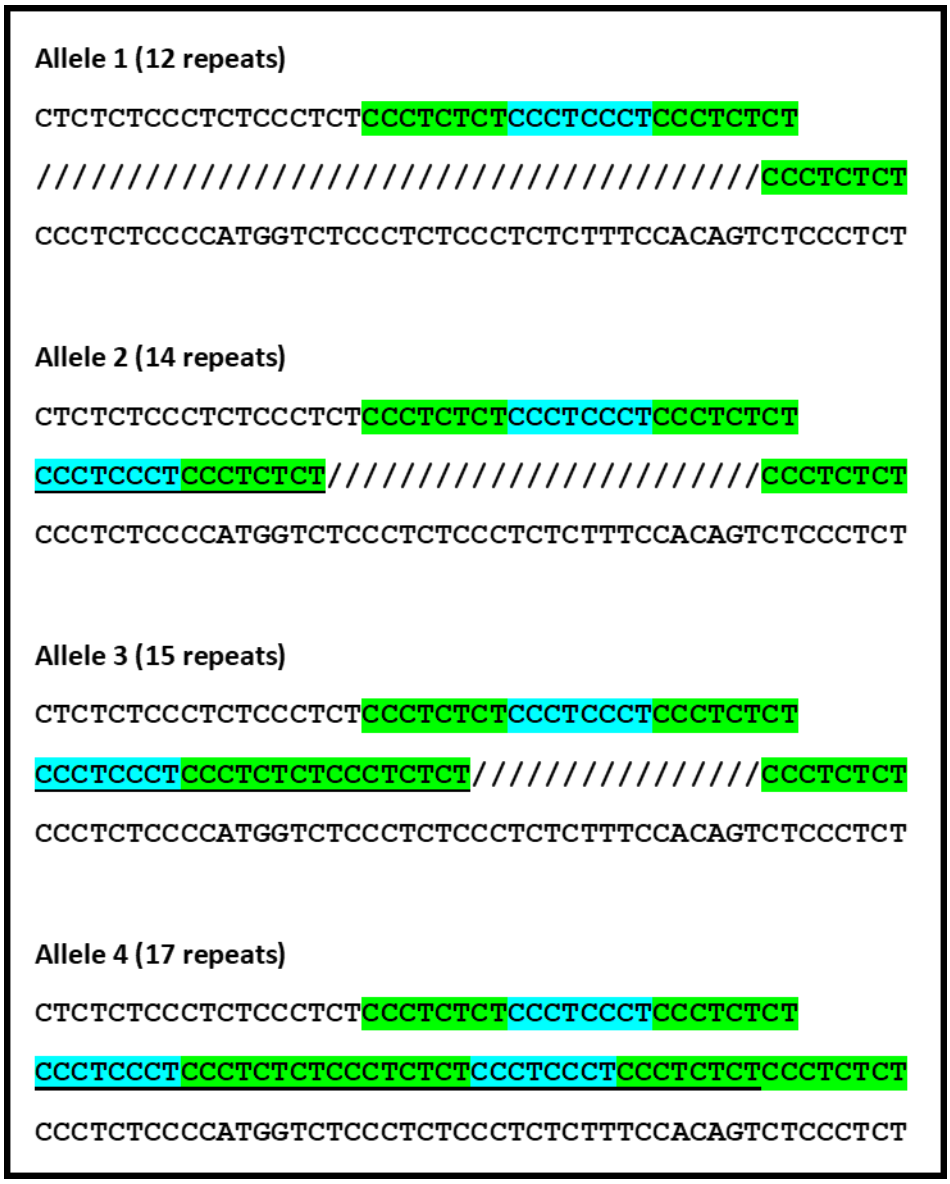


Figure 4.7. *NEK1* SVA-D CT element sequences.

The primary DNA sequence of the four alleles of the *NEK1* SVA-D CT element. All four alleles were confirmed by Sanger sequencing. Underlined regions indicate sequence expansions, which were octamer repeats (highlighted in green and blue and marked as forward slashes when not present).

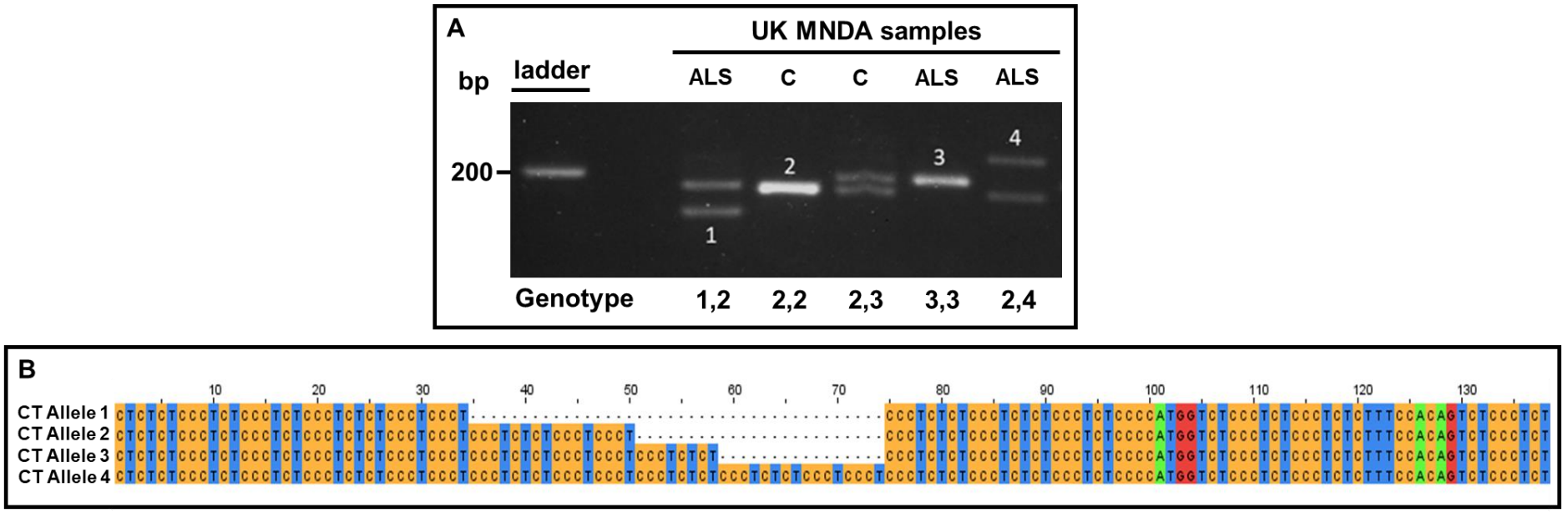


Figure 4.8. *NEK1* SVA-D CT element variant genotypes and aligned genomic sequences.

A: PCR amplification, agarose gel electrophoresis and genotyping of the NEK1 SVA-D CT element in a UK MNDA cohort. Samples run on 3% agarose gel at 100V for 4.5 hours. The CT element of the NEK1 SVA-D was found to be polymorphic and four alleles were identified (1-4). **B**: The primary DNA sequence of the four alleles of the *NEK1* SVA-D CT element. Alleles 1 and 4 were only found in ALS patients. All four alleles were validated through Sanger sequencing, aligned using MUSCLE[394,395] and visualised in Jalview[396]. Alignment gaps indicate sequence expansions of 8 bp repeats.

4.3.7 Expanding the *NEK1* SVA-D CT element analysis into the Project MinE UK dataset

Following the discovery of ALS specific variants in the MNDA cohort, the genotyping was expanded into a larger sample size of ALS cases and controls within Project MinE. This was made possible through collaboration with Dr Johnathan Cooper-Knock (University of Sheffield, SiTraN) and Professor Ammar Al-Chalabi and Dr Alfredo Iacoangeli (King's College London). Dr Cooper-Knock runs Working Group 6 (WG6) of Project MinE; a team of researchers who focus on the role of non-coding DNA in ALS. I was invited to run a retrotransposons subgroup within WG6, specifically to expand the *NEK1* SVA-D CT element analysis through utilising Project MinE WGS data.

To expand upon this study the *NEK1* SVA-D CT element was genotyped bioinformatically using WGS data available from Project MinE. Using the variant calling tool Isaac Variant Caller (IVC) from Illumina[315,316], Dr Iacoangeli at King's College London (KCL) was able to generate genome variant call format (VCF) files from the WGS data of ALS patients and controls within the UK dataset of Project MinE. Isaac Variant Caller (IVC) is part of the Isaac whole genome sequencing (WGS) workflow by Illumina and is specifically designed to call and genotype SNPs and indels in WGS data[315] (please refer to Chapter 2 Section 2.2.12.4 Isaac Variant Caller data analysis and manipulation for a detailed outline of this process). The aim was to specifically investigate if IVC could call the *NEK1* SVA-D CT element polymorphisms previously identified by PCR in the MNDA cohort (Figure 4.9).

There was some crossover of samples in the MNDA DNA bank and the UK dataset of Project MinE, so the PCR data generated from the MNDA cohort was used as a direct validation of the genotyping results from the IVC data of Project MinE. This validation analysis was performed on 206 samples for which both PCR data and IVC data. It was found that 199 of the 206 IVC results (96.6%) perfectly agreed (had an exact genotype match) with the previously generated PCR data (Table 4.4). Furthermore, only 1 sample was called completely incorrectly (0.49%) and 6 samples had one allele called correctly when compared to the PCR data (2.91%). Furthermore, due to a 96.6% genotype match when compared to the previous MNDA cohort PCR data we were confident with the results of IVC.

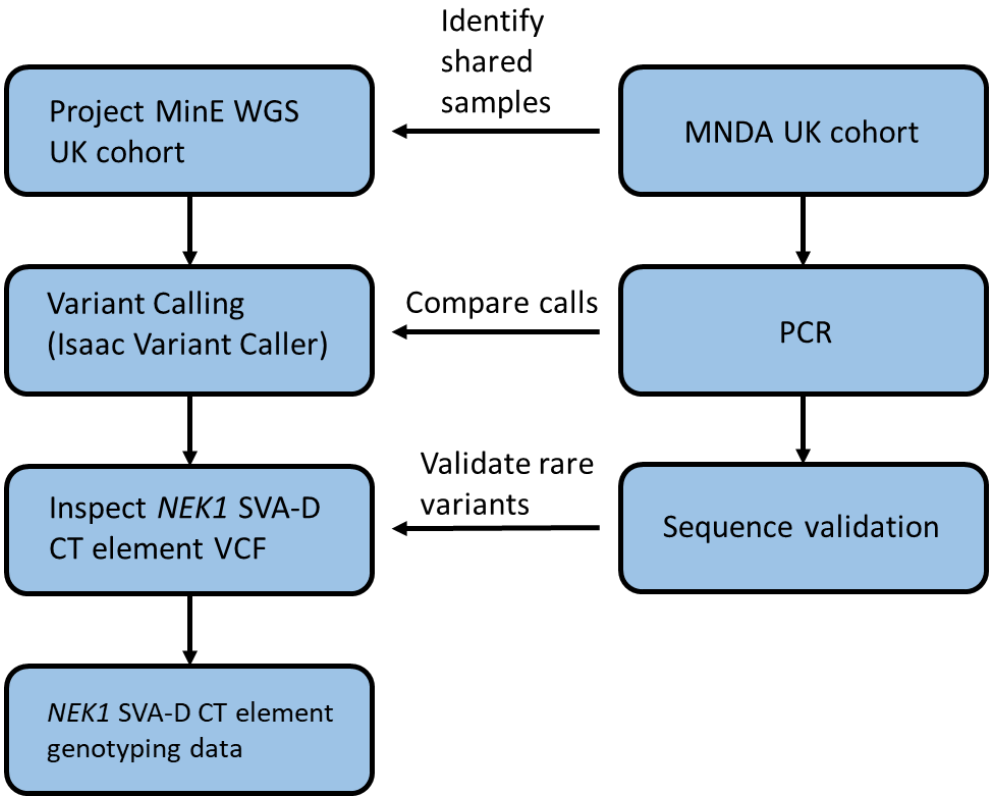


Figure 4.9 *NEK1* SVA-D CT element genotyping pipeline.

Experimental workflow of the genotyping analysis of the *NEK1* SVA-D CT element in two ALS patient and control populations. Shared samples were identified in these two

cohorts and used to compare IVC data to previous PCR data, thus validating IVC as a tool to genotype this genomic region.

Table 4.4 *NEK1* SVA-D CT element genotyping in MNDA UK and Project MinE UK shared dataset.

	Count	%
Both alleles correct	199	96.60
One allele correct	6	2.91
Both alleles incorrect	1	0.49
Total	206	100

Table 4.5 Isaac Variant Caller analysis of the *NEK1* SVA-D CT element in the Project MinE UK dataset.

	Allele 1 or 4 carriers	Other samples	Marginal Row Totals
Cases	13	1284	1297
Controls	0	500	500
Marginal Column Totals	13	1784	1797 (Grand Total)

Table 4.5 shows the initial screen of the UK dataset of Project MinE and includes the shared MNDA cohort ALS patient samples. After expanding this analysis to UK samples of Project MinE 10 more ALS patients with the rare variants were identified (allele 1 or 4); specifically 3 more patients with allele 1 (genotype 1,2) and 6 more patients with allele 4 (genotype 2,4). The ratio of the rare variants identified was similar when expanding from the MNDA cohort (1 patient with allele 1 and 3 patients with allele 4) to Project MinE (3 patients with allele 1 and 6 patients with allele 4), with allele 4 being more frequent in both cohorts. There was a statistically

significant difference in rare variant genotype frequency between ALS cases and controls (Fisher's Exact test; p-value = 0.0252).

4.3.8 ALS cases containing rare *NEK1* SVA-D CT element variants do not have rare *NEK1* coding variants which confer risk for ALS

NEK1 was identified as an ALS risk gene through rare variant burden analysis, leading to the discovery of an overrepresentation of LOF coding mutations in ALS patients compared to controls[63]. As ALS specific non-coding variants in *NEK1* were identified in this we wanted to determine if these polymorphisms were just tagging the rare coding mutations previously found[63]. To confirm this, WGS data for the UK dataset of Project MinE was utilised and the *NEK1* region was extracted with all coding mutations called by IVC in all ALS patients (Figure 4.10). These *NEK1* VCFs were then inspected for the known rare LOF variants which confer risk for ALS. *NEK1* coding mutations in the 13 ALS cases with the rare SVA CT element variants (allele 1 and 4) were inspected, to confirm if any of these patients had the rare LOF variants and thus to check if the rare SVA CT element variants were just inherited (in linkage) with these coding mutations (Supplementary Table 2). This is consistent with the rare variants of the *NEK1* SVA CT element potentially conferring risk for ALS.

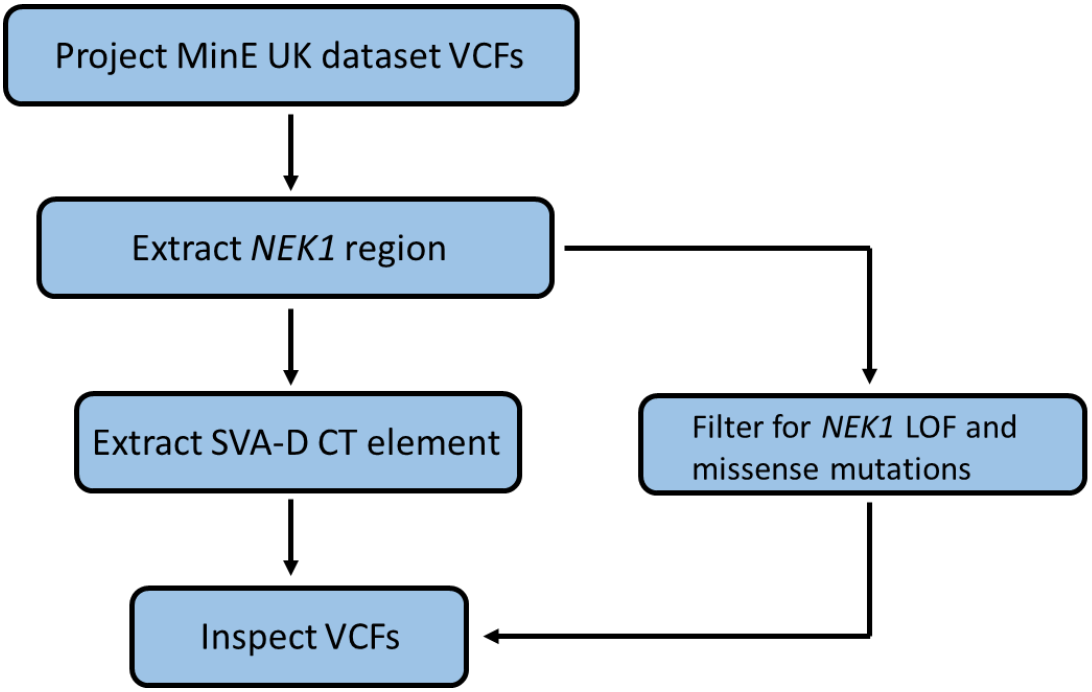


Figure 4.10 Outline of *NEK1* coding mutation analysis using Project MinE UK dataset.

Experimental pipeline using the Rosalind high performance computing (HPC) cluster at Kings College London to inspect *NEK1* coding mutations found in ALS patients of the UK dataset of Project MinE. All annotated coding variants within the *NEK1* gene were extracted using bcftools[319], and then the refined VCFs were inspected for the presence of any of the known missense or LOF ALS risk variants previously discovered.

Table 4.6. Coding mutations found within *NEK1* for the ALS cases containing the rare SVA-D CT element variant.

Illumina ID	MNDA ID	Genotype	Status	Variants in *NEK1*	rs number	Frequency
LP6008118-DNA_C01	BP6044	2,4	case	Missense	rs34540355	Common (MAF >1%)
LP6008118-DNA_F03	LNH0037	2,4	case	None	-	-
LP6008238-DNA_E02	BLI0099	2,4	case	Missense	rs34540355	Common (MAF >1%)
LP6008124-DNA_E07	SP3453	2,4	case	Splice region variant	rs33962953	N/A
LP6008124-DNA_H07	LP0383	2,4	case	Missense, Missense	rs34099167 and rs34540355	Common (MAF >1%)
LP6008235-DNA_A05	BP6387	2,4	case	Missense, Splice region variant	rs33933790, rs33962953	Common (MAF >1%), N/A
LP6008237-DNA_B12	LPO0067	2,4	case	Missense	rs34540355	Common (MAF >1%)
LP6008238-DNA_D06	BOX0037	2,4	case	Splice region variant	rs33962953	N/A
LP6008239-DNA_D07	BLI0147	2,4	case	Missense	rs34099167	Common (MAF >1%)
LP6008240-DNA_G10	SP3503	2,4	case	Missense, Missense	rs33933790, rs34540355	Common (MAF >1%)
LP6008121-DNA_H09	SP3399	1,2	case	None	-	-
LP6008237-DNA_E06	BLI0015	1,2	case	None	-	-
LP6008199-DNA_B03	LP0739	1,2	case	Frameshift variant	rs483352907	N/A

Table 4.6 shows the 13 ALS cases with the rare SVA-D CT element variants and all known coding mutations of *NEK1* found in these patients. Any cases which didn't have any coding mutations called by IVC were labelled as "None" and each coding variant was annotated based on type of mutation (missense, frameshift or splice region variant) and their respective rs numbers are shown. Each of the coding variants were also evaluated using dbSNP[321] to check their frequency. Variant frequency data was not available for rs33962953 and rs483352907, but the remaining variants were common (MAF>1%). The UK dataset samples containing rare *NEK1* coding ALS risk variants (Supplementary Table) were then identified and cross-referenced with the 13 ALS cases with rare SVA-D CT element variants. Overall, no rare *NEK1* missense or LOF variants which confer risk for ALS were found in any of the 13 cases which contain the rare SVA-D CT polymorphisms therefore confirming that the SVA CT element mutations identified in this study were not tagging rare coding variants previously identified.

4.3.9 *NEK1* SVA-D CT element genotyping analysis within other populations of Project MinE

The successful CT element genotyping analysis in the UK dataset of Project MinE was expanded by investigating other cohorts, including Spanish, Belgian, Irish, American, Dutch and Turkish populations within Project MinE (Table 4.7).

Expanding the *NEK1* SVA-D CT element genotyping into other cohorts of Project MinE (ALS cases, n = 1955 and controls, n = 986) identified more samples containing rare variants, specifically genotypes 1,2 and 2,4. Allele 4 was present in three other datasets: 4 Spanish ALS cases (1.6% of total cases in this cohort), 3 Irish ALS cases (1.1%) and 3 Dutch ALS cases (0.5%). However, allele 4 was also found in 3 Dutch controls (0.75%). Allele 1 was found only in the Spanish and American cohorts, being present in 1 Spanish control (0.4%) and 3 USA ALS cases (0.9%). With the addition of these results the Fisher's exact test did not reach statistical significance

Table 4.7. Expanding variant analysis of the *NEK1* SVA-D CT element into more cohorts of Project MinE.

Country	Cases/control	2,4	1,2
Spain	248/114	4/0	0/1
Belgium	374/180	0/3	0/0
Ireland	269/136	3/0	0/0
USA	320/81	0/0	3/0
Netherlands	596/400	3/2	0/0
Turkey	148/75	0/0	0/0
UK	1297/650	8/0	3/0
Total	3252/1636	18/5	6/1

(p-value = 0.34) and thus there was no significant difference in genotype frequency of the rare variants between case and control.

4.3.10 Validating the presence of rare *NEK1* SVA-D CT element variants in Dutch controls of Project MinE

Bioinformatically genotyping the *NEK1* SVA-D CT element in the Dutch cohort of Project MinE identified 3 controls with allele 4. As the UK dataset was validated using PCR, the same strategy was adopted for the Dutch population carrying this allele to confirm the IVC results. Unfortunately this could not be done for the Belgian cohort (which also contained controls with allele 4) as this DNA was no longer available. DNA of the 3 Dutch controls containing the 2,4 genotype of the *NEK1* SVA-D CT element was obtained and screened via PCR to determine if the IVC data was correct for these samples. UK ALS samples were used as positive controls for the 2,4 and 1,2 genotypes of the CT element and run alongside the Dutch control samples for comparison. For completion, a sample with the 2,3 genotype was also run alongside the Dutch controls to control for allele 3 being present in the Dutch samples (Figure 4.11).

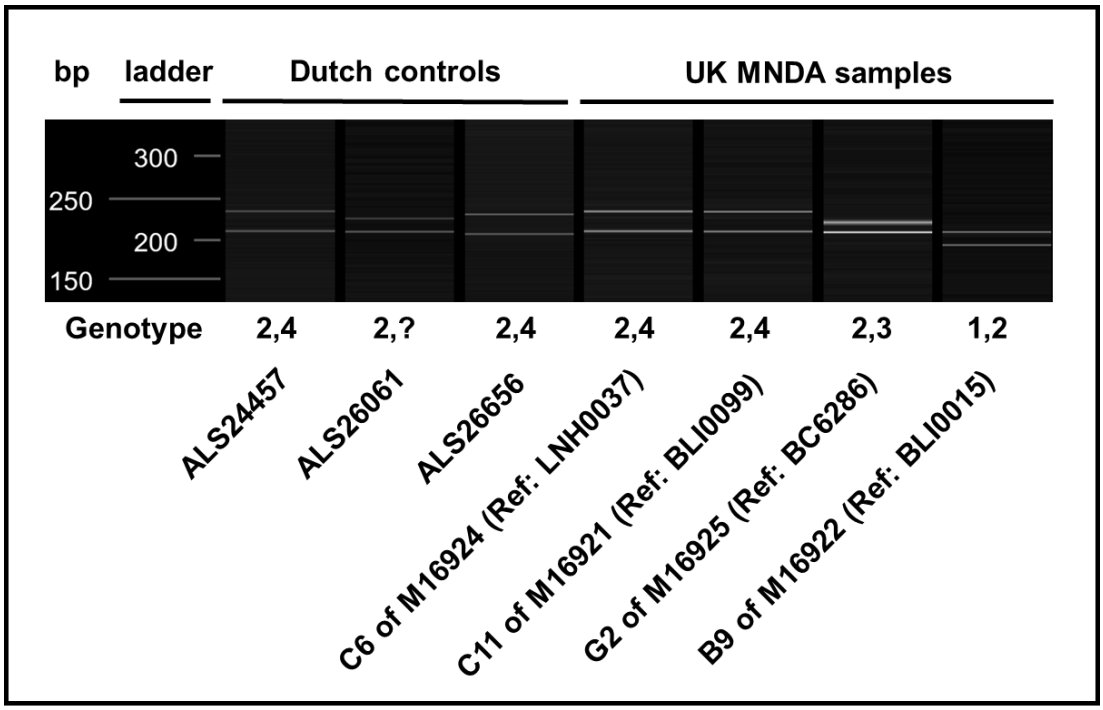


Figure 4.11 Validating Dutch Project MinE control samples.

PCR amplification, gel capillary electrophoresis and genotyping of Project MinE Dutch control samples. These samples were run alongside UK samples containing the 2,4 *NEK1* SVA-D CT element genotype to aid identification of alleles. Gel capillary electrophoresis performed using the QIAxcel advanced system and electronic gel image generated using the QIAxcel ScreenGel software.

Results from the PCR genotyping showed that 2 of the 3 Dutch controls were correctly called by IVC as the 2,4 genotype. However, sample ALS26061 was not genotype 2,4 and was thus called incorrectly by IVC. This sample did contain allele 2, but the other allele was not found in the MNDA cohort or UK dataset of Project MinE and was thus a novel variant (annotated as genotype 2,? In Figure 4.11). Further analysis was undertaken to determine the length and sequence of this polymorphism and results compared to the previously sequenced alleles in Figure 4.8. Although two of the Dutch controls appeared to contain allele 4 this was also verified by sequencing, to determine if they contained SNPs that differed from the ALS cases containing allele 4.

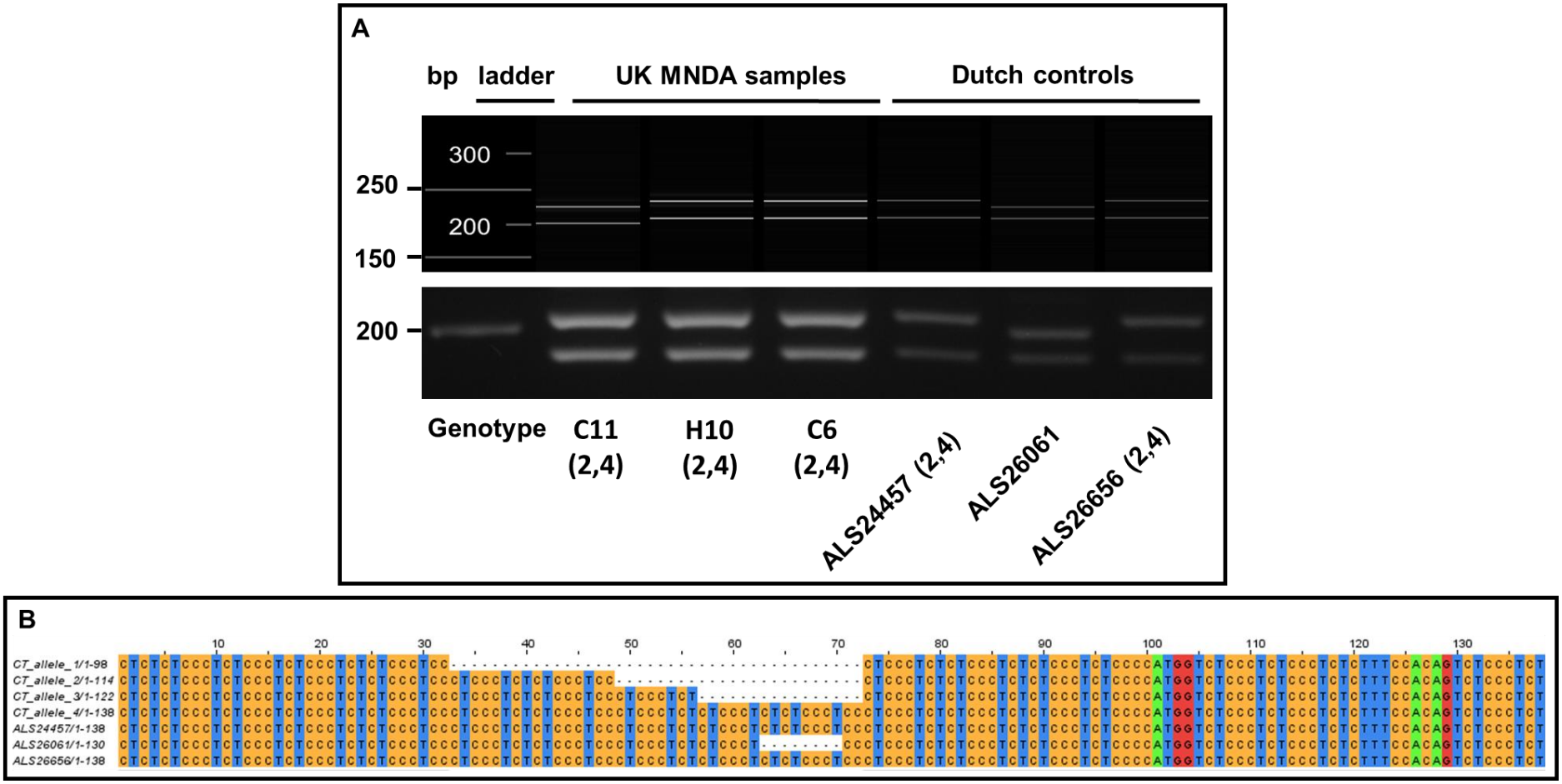


Figure 4.12 Genotyping and sequencing verification *NEK1* SVA-D CT element in Dutch control samples from Project MinE.

A: PCR amplification, agarose gel electrophoresis and gel capillary electrophoresis of the *NEK1* SVA-D CT element of MNDA UK samples and Dutch

control samples from Project Mine. Samples run on 3% agarose at 100V for 4.5 hours and gel capillary electrophoresis performed using the QIAxcel advanced system and electronic gel image generated using the QIAxcel ScreenGel software. **B**: Primary sequence of the *NEK1* SVA-D CT element. All five alleles have been validated through Sanger sequencing, aligned using MUSCLE[394,395] and visualised in Jalview[396]. The novel (hereafter allele 5) found in the Dutch control sample is indicated by the 8 bp gap in sample ALS26061 (130 bp, 16 repeats).

After successfully cloning and sequencing the CT element of the Dutch control samples they were aligned against the other known variants. The novel variant (130 bp, 16 repeats) found in ALS26061 contained one octamer "CTCTCCCT" repeat more than allele 3 and one less than allele 4. This novel variant will now be referred to as allele 5 for the remainder of this chapter. It was also found that the sequence of both Dutch controls containing allele 4 matched exactly with the ALS cases containing allele 4; there was no SNP variation between allele 4 of ALS cases and controls (Figure 4.12).

4.3.11 Visually inspecting the *NEK1* SVA-D CT element IVC calls

After the discovery of allele 5 in ALS26061, it was determined if this variant was actually present in the UK dataset of Project MinE also and if any other variants had been miscalled by IVC. To do this genome VCFs of the *NEK1* SVA-D CT element region were extracted for the 10 UK ALS patients previously determined to have allele 4 (Table 4.6). From this, the VCFs were visually inspected for the CT element expansion called by IVC and then these expansion variants were compared against the sequencing data generated in Figure 4.8 and Figure 4.12 to determine which allele the expansions matched. This analysis demonstrated that 8 of the 10 previously genotyped 2,4 samples were correct and matched the expansion size of allele 4. However, 2 of these cases actually contained the novel allele 5 (Table 4.8) originally discovered in Dutch control ALS26061 (Figure 4.12). Visual inspection and sequence comparison were also performed for the 3 ALS patients with the 1,2 genotypes to confirm that these genotypes were also called correctly by IVC. In all 3 cases the deletion size exactly matched the size of allele 1.

Table 4.8 13 UK ALS samples with rare *NEK1* SVA-D CT element variants.

Illumina ID	MNDA ID	Genotype	Status	VCF and Sanger sequencing agree
LP6008118-DNA_C01	BP6044	2,4	case	Yes
LP6008118-DNA_F03	LNH0037	2,4	case	Yes
LP6008238-DNA_E02	BLI0099	2,4	case	Yes
LP6008124-DNA_E07	SP3453	2,4	case	Yes
LP6008124-DNA_H07	LP0383	2,4	case	Yes
LP6008235-DNA_A05	BP6387	2,4	case	Yes
LP6008237-DNA_B12	LPO0067	2,4	case	Yes
LP6008238-DNA_D06	BOX0037	2,4	case	No
LP6008239-DNA_D07	BLI0147	2,4	case	No
LP6008240-DNA_G10	SP3503	2,4	case	Yes
LP6008121-DNA_H09	SP3399	1,2	case	Yes
LP6008237-DNA_E06	BLI0015	1,2	case	Yes
LP6008199-DNA_B03	LP0739	1,2	case	Yes

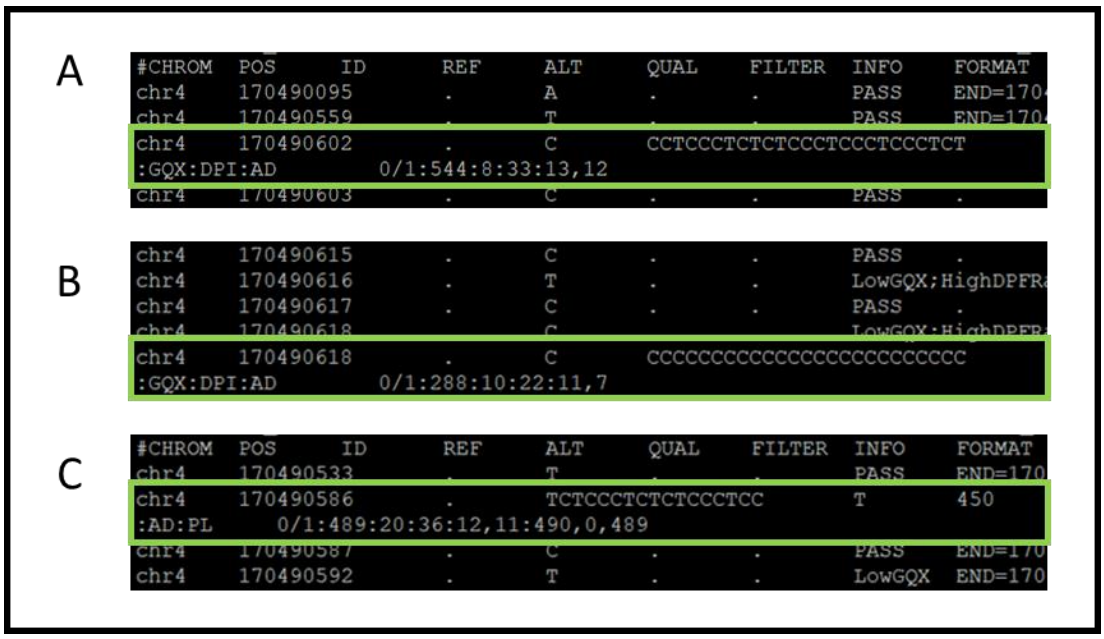


Figure 4.13. Example outputs from Isaac Variant Caller.

Project MinE UK cohort genotyping results from Isaac Variant Caller (VCF file output). Chromosome (#CHROM), position (POS), reference allele (REF) and alternate allele (ALT) are all shown. **A**: ALS patient containing the 2,4 genotype of the *NEK1* SVA-D CT element. In this call the reference allele (REF) is variant 2 and the alternate (ALT) allele is an expansion of 24 bp and therefore equates to variant 4. In this format, 0 = REF and 1 = ALT, meaning that 0/1 in this call corresponds to the 2,4 genotype. **B**: ALS patient containing the 2,4 genotype of the *NEK1* SVA-D CT element, with the ALT allele correctly calling an expansion of 24 bp, but poor sequence fidelity meant it was

called as an expansion of only cytosines (C). **C**: ALS patient containing the 1,2 genotype of the *NEK1* SVA-D CT element. In this call, the REF call is variant 1 and the ALT call is an expansion of 16 bp and thus equates to variant 2; 0,1 therefore corresponds to genotype 1,2 in this call.

4.3.12 Amended IVC results for the NEK1 SVA-D CT element in Project MinE

After the discovery of allele 5 within the UK dataset of Project MinE, the genotyping results for both the UK dataset (Table 4.9) and for the combination of cohorts (Table 4.10) were amended. Furthermore, an additional 150 control samples were included that had been analysed by PCR in the MNDA cohort (which were not part of Project MinE). For the UK dataset alone the rare CT element variants are only found in ALS patients, specifically 11 out of 1297 (0.85%), and the Fishers Exact test statistic was significant (p-value = 0.02). However this was not the case for the combined results of all cohorts (p-value = 0.25). Overall, the rare CT element variants (alleles 1 and 4) were found in 24 of 3252 ALS patients (0.73%) and in 7 of 1636 controls (0.43%).

Table 4.9 Amended results for Isaac Variant Caller analysis of the *NEK1* SVA-D CT element in the Project MinE UK dataset.

	Allele 1 or 4 carriers	Other samples	Marginal Row Totals
Cases	11	1286	1297
Controls	0	650	650
Marginal Column Totals	11	1934	1947

Table 4.10 Amended variant analysis of the *NEK1* SVA-D CT element in several cohorts of Project MinE.

Country	Cases/control	2,4	1,2
Spain	248/114	4/0	0/1
Belgium	374/180	0/3	0/0
Ireland	269/136	3/0	0/0
US	320/81	0/0	3/0
Netherlands	596/400	3/2	0/0
Turkey	148/75	0/0	0/0
UK	1297/650	8/0	3/0
Total	3252/1636	18/5	6/1

4.4 Discussion

In this study, a human specific SVA retrotransposon within *NEK1* was identified and ENCODE data suggested this could be a potential regulatory element at this locus. The *NEK1* SVA-D was variable in the population, with distinct polymorphisms in three specific regions; CT element, VNTR and Poly A tail (Figure 4.3). Genotyping identified four CT-element variants, three VNTR variants and six Poly-A tail variants, highlighting that this particular retrotransposon is highly polymorphic in the general population. Genotype screening of the CT element in an MNDA cohort of ALS cases and controls identified novel variants only found in ALS patients (Figure 4.4). This analysis was then expanded into the larger WGS dataset of Project MinE (n=4888), identifying both cases and controls with the two rare variants (Table 4.10). From cloning and sequencing this domain it was found that this element is built of 8 bp repeats; "CCCTCTCT" and "CCCTCCCT" (Figure 4.7). The CT element genotype was the same in both motor cortex and blood of the same ALS patient (n=5), confirming no expansion of this domain in the motor cortex. Ultimately, this study presents a validated high-throughput method of screening SVA CT-rich domain variation using IVC calls from Project MinE WGS data, while directly validating this analysis using PCR and Sanger sequencing.

The results presented here show the presence of the canonical "CCCTCT" hexamer repeat which has been previously identified[219,220,225,397], however this region in *NEK1* is expanding and contracting in larger repeats of "CCCTCTCT" and "CCCTCCCT. Due to the repetitive nature of this region, larger or smaller polymorphisms could be possible because of strand slippage during DNA replication,

caused by misalignment and mispairing of the alternative complementary repeats on newly synthesised DNA strands[398]. This particular CT-rich repeat was built of two perfect repeats all occurring in tandem, meaning this mispairing of DNA strands could occur at multiple locations and thus give rise to several expansions or deletions at this region (Figure 4.7). Previously it has been shown that tandem repeat genotypes are not stable across tissues, causing somatic mosaicism in the brain and thus adding another layer of complexity to characterising such variation. This phenomenon has been observed in ALS cases with the *C9orf72* repeat expansion, with expansions being much larger in neuronal tissues compared to non-neuronal tissues[393]. Westenberger *et al*. have found that CT element repeat size of the *TAF1* SVA is variable across brain and blood of the same person, with Southern Blot analysis identifying somatic expansions in the brains of two XDP patients[298]. The *NEK1* SVA-D CT element was genotyped in matched motor cortex and blood of 5 ALS patients. It was found that the genotype of the element was stable across the two tissues from the same patient: the CT element genotype was the same in both motor cortex and blood from the same ALS patient (n = 5), providing no evidence that this domain undergoes expansion in the motor cortex in disease, albeit this was only a small number of individuals tested (Figure 4.5). As this was only performed in a small sample size a future study would include expansion of this analysis, to screen more ALS matched motor cortex and blood samples for the *NEK1* SVA CT element, to more definitively determine if there are any somatic expansions of this CT element in the brains of ALS patients.

In line with the hypothesis that SVA variants could be potential risk factors for ALS, novel rare CT element variants were discovered (MAF<1%) that were present

only in ALS patients (UK MNDA cohort, n=984). One patient with allele 1 (allele frequency = 0.10%) and three patients with allele 4 (allele frequency = 0.31%) were identified (Table 4.1). These results indicated that these two variants were ALS specific, however this initial study was only performed in 984 samples, a relatively small sample size. Savage *et al*. previously evaluated an SVA retrotransposon in the promoter region of *FUS* (Fused in sarcoma) by genotyping the SVA in 241 SALS and 228 controls. Two alleles were identified in this cohort, with the polymorphism being associated with the VNTR region of the element but no significant difference in genotype frequency across SALS and controls was found[224]. The results from the PCR genotyping experiment presented in this chapter agree with this previous research as we also found no significant difference in genotype frequency of the SVA element between ALS and control samples (Table 4.1, Table 4.2 and Table 4.3). However, the initial study in the MNDA cohort did reveal CT element variants specific to ALS cases (Table 4.1), a novel finding when compared to the previous study[224].

The *NEK1* SVA CT element genotyping was expanded into Project MinE by screening this region using Isaac Variant Caller (IVC). IVC is a variant calling pipeline designed by Illumina as part of their Isaac WGS package and has been previously used to identify SNPs and small indels[316]. The results presented here demonstrated that this pipeline can call indels of up to 24 bp in length (panels A and B of Figure 4.13). Furthermore, IVC results showed a 96.6% agreement with the optimised PCR assay, indicating a high degree of confidence in the calls generated (Table 4.4). However, one pitfall of assessing indels using this variant caller is the poor fidelity of the variant sequences; while IVC did call the CT element expansion and deletion sizes correctly, the sequence did not perfectly match our Sanger sequencing data (panel B of Figure

4.13). Therefore, IVC can be used to screen for expansions and deletions of a certain size, but the specific sequence of these variants cannot be accurately called.

After the successful use of IVC to expand the CT element analysis into the UK dataset it was expanded into other datasets of Project MinE. Unfortunately, this replication study led to the discovery of both allele 1 and 4 in control samples, albeit in a specific geographical area (Table 4.10) and a further variant was also identified (Figure 4.12). This result was not consistent with our study in the UK dataset of Project MinE, however it was concluded that both rare variants were not concomitant with the presence of disease. This result is perhaps not surprising, as all previously discovered missense and LOF *NEK1* ALS risk variants were also found in controls[62-65,366,399]. It is important to note that two previous studies identified the same rare LOF variant (p.Ser1036*) in both FALS cases and unrelated controls[64,65]. It is important to remember that ALS is a complex disease, involving numerous factors such as the potential of multiple mutations (oligogenic inheritance), pleiotropy, penetrance and environmental factors which can contribute to disease manifestation[115]. These factors have been shown to fit within a multi-step model of ALS, consistent with approximately (on average) six sequential pathological steps (ie possessing a genetic mutation will still require the other steps to cause disease)[114-116]; it has also been found that but the number of steps does change dependent on the mutated gene, therefore differing between patient subgroups[116]. This model has helped to explain the fact not all individuals with a genetic mutation (even highly penetrant ones) always develop ALS, plus the adult onset nature of the disease (even in familial cases)[114].

In the full set of available samples of Project MinE (at the time of writing) there was no significant difference in frequency of rare CT element variant between ALS cases and controls (Fisher's exact test, p-value = 0.25). Overall due to this result we cannot definitively say the variants identified confer risk for ALS, however they are still more prevalent in cases and with a further increase in sample size it could be possible to see a statistically significant enrichment of these variants. Further studies should expand this analysis into even more samples of Project MinE (when they become available) and also other populations such as Asian populations, as rare coding *NEK1* variants have also been identified in Chinese and Japanese ALS patients[366,399,400]. One study, using exome sequencing data of a Chinese cohort, specifically found rare coding mutations in 1.8% of cases and 0.4% of controls[366]. A separate study in a Chinese population also identified six rare heterozygous-LOF and three rare missense mutations in *NEK1*, accounting for 2.7% of ALS cases[400]. Seven LOF *NEK1* variants have also been identified in Japanese SALS patients (1.57%) and found to be associated with increase ALS risk[399]. It would be interesting to see if the rare CT element variants were also found in Chinese and Japanese ALS patients.

The rare variants identified in this study were always found to be heterozygous (1,2 and 2,4 genotype) (Table 4.10). This result agrees with previous studies which identified an enrichment of heterozygous-LOF mutations[62-65,366]. Due to this finding, a haploinsufficiency mechanism has been proposed, meaning that one copy of the mutation would be sufficient to drive disease phenotype[64]. It could be possible that a dominant genetic mechanism applies to the noncoding heterozygous CT element variants found in this project, but these mutations would need to be functionally tested to prove this, such as using reporter gene assays. Functionally

these non-coding mutations could induce dysregulation of *NEK1* gene expression, potentially altering affinity for transcription factors, or inducing secondary structure formation which could stall transcription or alter splicing efficiency at this region, an effect previously exhibited by intronic repeat variation[150,153,180]. A study in 2018 by Nguyen *et al*. identified other ALS gene mutations in 7 of the 13 identified *NEK1* variant carriers (54%). This observation was found to be significantly higher than expected, suggesting that the disease could arise from the accumulation of several mutations in multiple genes (a multiple-hit hypothesis/oligogenic inheritance)[65]. It is important to take this hypothesis into consideration and in future assess if the patients with the rare CT element variants have additional ALS gene mutations and confirm if they are therefore genetically distinct to the matched controls. In the 13 ALS patients which harboured the rare CT element variants no known *NEK1* coding mutations, which confer ALS risk, were identified (Table 4.8); indicating that the SVA mutations are not tagging known ALS risk variants.

This project presents a first attempt at using IVC to genotype an SVA retrotransposon (at the time of writing) and thus offers a novel and high-throughput method to screen such elements. While we have shown that genotyping SVAs is possible using IVC, it is important to note that the SVA analysed had been previously characterised and sequence validated and therefore we could confirm the IVC calls by comparing against PCR data and Sanger sequencing (Figure 4.4 and Figure 4.8). Uncharacterised *de novo* SVA variation generated from this method should be interpreted with caution and be subject to PCR validation. Ultimately the IVC pipeline can be utilised for high-throughput screening of SVA elements using WGS data, but

the regions investigated should be characterised prior to this, so that one knows what variants they are looking for.

The data presented in this chapter shows that IVC can be used in conjunction with PCR data and Sanger sequencing to aid genotyping large cohorts but should not be used in isolation to identify *de novo* variation, as indicated by the two ALS patients whom were mis-genotyped and actually contained allele 5 rather than allele 4 (Table 4.8). With this result in mind IVC should not be used to globally genotype SVA retrotransposons within the human genome without first having prior knowledge of the size all the composite regions of each SVA element. Presently, the definitive way to characterise these elements is through PCR and directed sequencing, which is labour intensive, time consuming and expensive as there are approximately 3000 SVAs in the reference genome. Due to the highly repetitive nature, high GC content and length of SVA retrotransposons, it can be difficult to map the tandem repeat variation accurately to unique locations using current short read WGS data[182]. In addition, flanking sequence is required to map SVA elements back to a specific location of the genome and as current WGS generates short (~150 bp) reads[401], meaning that only the terminal domains (5' CT element and 3' poly A tail) of SVAs can be accurately genotyped using programs such as IVC, as central domains (such as the VNTR) cannot be accurately mapped with short sequencing reads and are therefore not well defined. It is important to note that this work only focused on one domain (CT element) and that other sections of the *NEK1* SVA elements are variable and may harbour potential risk variants of ALS. Further analysis on other domain variants in the *NEK1* SVA that could highlight other indels which may, in conjunction with other mutations, modify transcriptional regulation at the *NEK1* locus. As read mappability

improves and WGS becomes more readily available it would be possible to more rigorously address polymorphism in SVAs. Software tools such as ExpansionHunter have now been used to accurately call *C9orf72* repeat expansions in ALS patients and provides one possible method to genotype tandem repeats genome-wide; however this appears to only be accurate for perfect repeats and therefore imperfect tandem variation (or emergence of several different tandem repeats in close proximity as shown in Figure 4.7) could prove more challenging to genotype[182]. Long read sequencing technology such as the MinION instrument (Oxford Nanopore) may help pave a way to more accurately genotype structural variation on a global scale, as a previously well validated human cell line (GM12878) and human genome (NA12878) have both now been successfully whole-genome sequenced and assembled using this technology[402,403]. Long read sequencing would facilitate genotyping of all composite regions of specific SVA elements, allowing for a more complete assessment of the potential genetic variation harboured by these retrotransposons.

The studies performed in this chapter highlight the need to further assess and characterise retrotransposon genetic variation in ALS and suggests such elements could be an untapped source of genetic risk variants. This work demonstrates it is possible to evaluate such regions by using PCR and variant caller pipelines together. Future studies should investigate the functional consequences of the identified SVA variants, to understand their role in gene regulation.

Chapter 5: Assessing function of an SVA retrotransposon as a potential regulatory element in the *NEK1* locus

5.1 Introduction

As the genetic polymorphisms of the *NEK1* SVA-D have been addressed in Chapter 4, Chapter 5 will focus on the functional capacity of the SVA and examine the potential of this element to serve as a transcriptional regulatory domain. There are two proposed mechanisms of action for SVA: acting as source regulatory domains in the germline without the need for retrotransposition and as inducers of insertional mutagenesis via their mobilisation[232]. SVAs can function as modulators of gene expression and transcriptional regulatory domains either within promoters or intronic regions of genes, which has been studied *in vitro* and *in vivo*[223,224]. They influence epigenetic and transcriptional parameters at the locus in which they are found without the need for retrotransposition. These elements exhibit two forms of polymorphism: variation within the components that make up these elements and as present or absent with respect to the reference human genome[231].

Savage *et al*. have shown that SVA elements are transcriptional regulators both *in vitro* and *in vivo*. Their initial discovery of a variable number tandem repeat (VNTR) in the promoter region of *FUS* led them to determine that it was in fact part of an SVA element: with the central region being made up of a tandem repeat (TR) and the discovered VNTR. The SVA was genotyped using PCR and two VNTR alleles (long and short) were identified. To test functionality, pGL3P (Promega) reporter gene constructs containing the SVA and central isolated TR/VNTR region alone were generated (long and short variants of each). In SKNAS cells when compared to empty pGL3P vector, a significant difference in reporter gene expression was observed. Both the SVA constructs led to a significant decrease in reporter gene activity, with a

significant difference between the long and short SVA variants also being observed. The TR/VNTR constructs elicited the opposite response, leading to a significant increase in luciferase expression when compared to empty pGL3P vector. Savage *et al*. also generated *in vivo* reporter gene constructs: the *FUS* proximal promoter and both the long SVA and the isolated long TR/VNTR region were cloned upstream of a GFP reporter gene. These constructs were then injected into the neural tube of chick embryos and then transfected into cells by electroporation. Overall it was found the proximal promoter of *FUS* did not drive GFP expression in the neural tube of this model, but GFP expression was generated by both the SVA and VNTR constructs thus highlighting that these domains can function as transcriptional activators *in vivo*[224].

Since the discovery of the anti-sense SVA insertion within *TAF1* being the cause of XDP there has been a focus on this particular SVA element, with several groups studying the functional implications of this insertion and how it might lead to dysregulation of the *TAF1* locus. Makino *et al*. identified the disease specific SVA insertion within intron 32 of *TAF1* back in 2007 and hypothesised that this insertion caused XDP by altering expression of *TAF1* due to the retrotransposon mediating methylation changes. They measured mRNA expression of 12 isoforms of TAF1, one of which was neuron specific (TA14-391). It was found that TA14-391 expression was significantly decreased in the caudate, accumbens and cortex of XDP patients (n=3) when compared to controls (n=3)[296]. Bragg *et al*. have assessed the genetic variation of the *TAF1* SVA in 140 XDP patients and identified polymorphic variation in the CT-element region, with hexamer repeat number ranging from 35 to 52 across the patients. Furthermore, they found an inverse correlation between repeat length and age of onset of disease: as repeat length increased, age of onset decreased. Bragg *et*

al. then went on to test this SVA in a luciferase reporter gene assay in SH-SY5Y cells, assessing if CT element length variation had any functional effect on Firefly luciferase activity. To perform this they cloned the SVA into the pGL3B reporter gene vector, generating constructs containing the SVA bearing 35, 41 and 52 repeats and a construct lacking the CT element (all constructs were generated in both the forward and reverse orientation with respect to the firefly luciferase reporter gene). Interestingly, they found that all forward orientation SVA constructs containing the CT element led to a significant repression of basal luciferase expression: the reverse orientation constructs elicited the opposite effect, inducing an increase in luciferase activity but this effect was only statistically significant in the constructs bearing 41 and 52 hexamer repeats.

SVA elements are hypothesised and predicted to form DNA secondary structures including G4 quadruplexes[297], which are known to impact the dynamics of transcription[404-407]. Bragg *et al*. predicted the G4 potential of each position of the *TAF1* SVA using QGRS Mapper. Using this in silico prediction tool they found that the antisense strand of hexamer repeat (AGAGGG) generated the largest G4 prediction score (G-score) out of all positions of the SVA and the G-score increased as the hexamer repeat number expanded. They hypothesise that G4 quadruplex formation could stall RNA polymerase II, thus repressing transcription and altering expression of *TAF1*[297].

To date there have been two studies which have removed the aforementioned SVA using the CRISPR Cas9 system to determine if this retrotransposon is driving the reduction in *TAF1* expression that was previously seen

by Makino *et al*.[408,409]. Rakovic *et al*. excised the TAF1 SVA element using CRISPR in previously generated induced pluripotent stem cells (iPSCs) from 3 XDP patients and controls and then differentiated these lines into both cortical neurons and spiny projection neurons (SPNs). From this, they measured expression of total *TAF1* and the neuron specific TA14-391 isoform (*nTAF1*) and compared across XDP cases and controls. Overall, the removal of the SVA led to a significant increase in total *TAF1* expression in XDP patient derived iPSCs when compared to control iPSCs. However, this result was not replicated in the cortical neurons or SPNs as there was no significant difference in total *TAF1* or *nTAF1* expression in SVA knock out cells compared to unedited control cells[409].

Aneichyk *et al*. recently performed transcriptional profiling of *TAF1* in XDP patient-derived cell lines and uncovered *de novo TAF1* transcripts. One of these was an isoform (termed TAF1-32i) containing exon 32 spliced to a cryptic exon found within intron 32 which then terminated approximately 716 bp upstream of the SVA insertion. They tested the effect that excision of the SVA had on this observed retention of intron 32, utilising the CRISPR Cas9 system to knockout the SVA from XDP patient derived iPSCs, which were then differentiated into neural stem cells (NSCs), induced cortical neurons (iNs) and also NSC-derived cortical and GABAergic neurons. They then used quantitative reverse-transcription PCR (RT-PCR) to measure relative expression of intron 32 of *TAF1* while also generating RNA-seq counts of intron 32. Removal of the *TAF1* SVA resulted in a significant decrease in expression of intron 32 in fibroblasts, iPSCs and NSCs: a decrease was observed in iNs, NSC-derived neurons and GABAergic neurons but this did not reach statistical significance. Furthermore, normalisation of total *TAF1* expression in response to the SVA excision

was also observed. These results helped uncover a potential mechanism behind the SVA-mediated reduction of *TAF1* expression in XDP, through retention of intron 32[408].

The studies in this chapter examine if the SVA found within *NEK1* exhibits functional properties, specifically addressing its capacity to regulate transcription and gene expression. Similarly to the studies mentioned above, we tested this SVA *in vitro* using reporter gene assays and by excising the SVA element using CRISPR and measuring gene expression in response to this modification.

5.2 Hypothesis and aims

Hypothesis:

The SVA-D within *NEK1* is a functional regulatory domain.

The SVA-D regulates *NEK1* on the transcriptional level, driving differential expression in a tissue specific manner.

Aims:

Clone the *NEK1* SVA-D into several reporter gene constructs.

Test the functionality of the SVA-D *in vitro* using a luciferase reporter gene assay in multiple cell lines.

Generate single and double knock outs of the *NEK1* SVA-D in the human cell line HEK293.

Characterise expression of *NEK1* transcripts on UCSC via RT-PCR.

Measure *NEK1* and *CLCN3* gene expression of the modified CRISPR lines via qPCR.

5.3 Results

The aim of this study was to address the potential for the SVA-D in *NEK1* to be a functional regulatory element by utilising two *in vitro* strategies; generating reporter gene constructs of the SVA and testing these in a number of cell lines and also knocking out the SVA-D in HEK293 cells using the CRISPR Cas9 system.

5.3.1 The *NEK1* SVA-D shows functional properties in the pGL3-P vector in several cell lines.

To test the functionality of the *NEK1* SVA-D as a transcriptional regulator it was cloned into the reporter gene construct pGL3-P (Promega) and then tested in a reporter gene assay. Initially the full length SVA-D was amplified using PCR (Table 2.2), the DNA was then purified and ligated into the intermediate pCR®-Blunt vector (ThermoFisher). The SVA insert from this vector was then excised and ligated it into the multiple cloning site (upstream of the SV40 promoter) of pGL3-P (Promega) (Figure 2.2). This SVA element was cloned twice, in both the sense (forward) and anti-sense (reverse) orientation with respect to the SV40 promoter of pGL3-P (Figure 5.2A), to test if orientation of the SVA had any effect on transcriptional regulation in this model. The presence of the SVA in both constructs was confirmed using restriction digest (Figure 5.1) and Sanger sequencing (Appendix 3). Please refer to Chapter 2 section 2.2.3.7 Restriction enzyme digests for a detailed overview of this process. A vector map of this construct was generated using SnapGene software (from Insightful Science; available at snapgene.com) (Supplementary Figure 7).

All cells were plated in 24-well format, 100,000 cells per well and were transfected with Turbofect™ transfection reagent (Table 2.6) (please refer to Chapter

2 Section 2.2.8 Transfection of plasmid DNA into cultured cells). 48 hours post transfection, the Dual-Luciferase® Reporter Assay (Promega) was used to measure the luminescent signal generated by the pGL3-P-SVA-D constructs and directly compared to the signal generated by the pGL3-P vector alone (empty vector) (Figure 5.2). In this assay the pRL-TK vector (Promega), expressing Renilla (*Renilla reinormis*) luciferase, served as an internal control to allow normalisation and account for variation in cell number, cell death and transfection efficiency (please refer to Chapter 3 for the same assay procedure for *CFAP410* VNTR containing constructs).

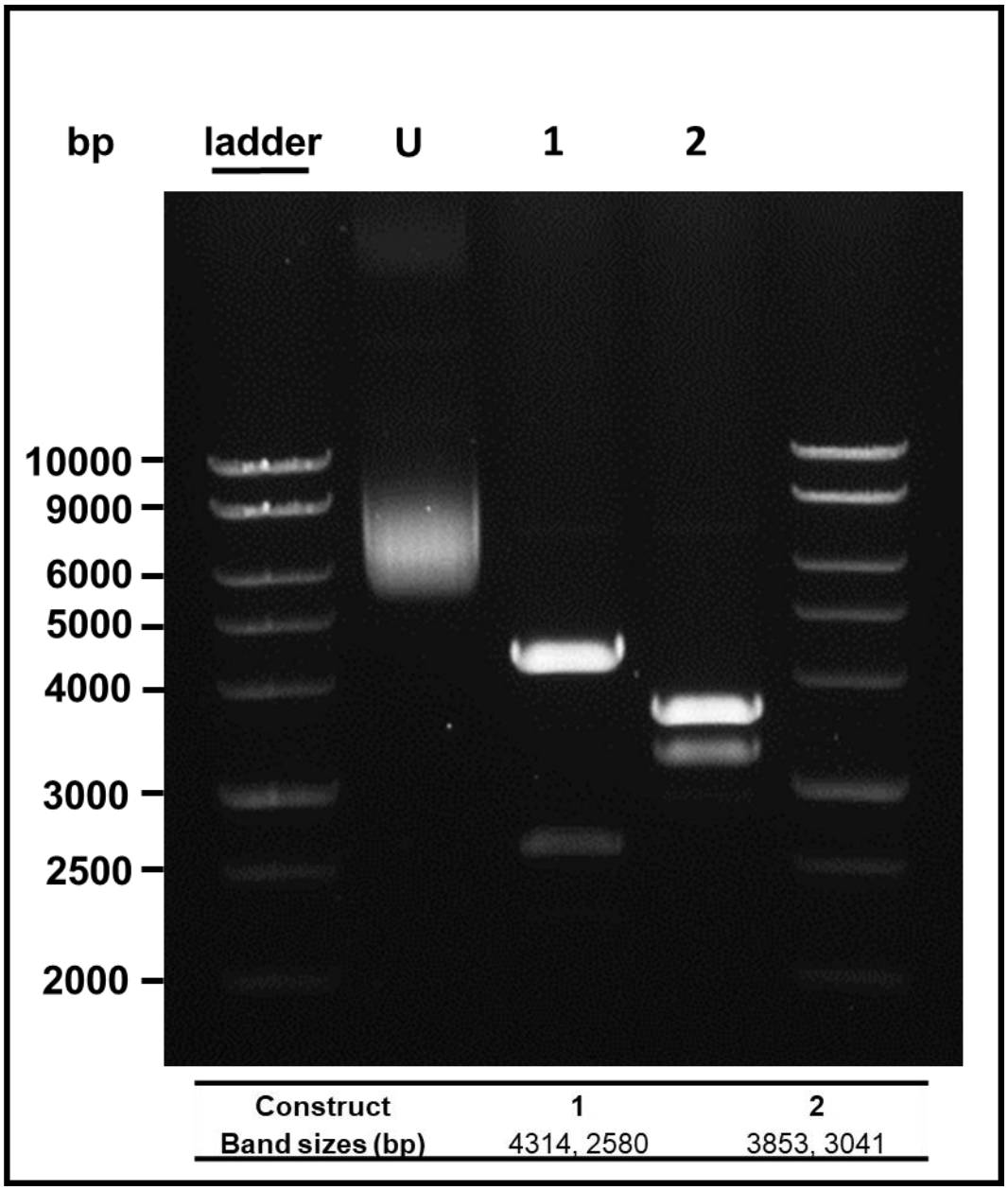


Figure 5.1. Validation of *NEK1* SVA-D pGL3-P constructs.

Restriction enzyme digest of *NEK1* SVA-D containing pGL3-P constructs. Constructs were cut with *BamHI* and run on a 1% agarose gel for 2 hours at 120V. The SVA-D was cloned in both the sense (forward) and anti-sense (reverse) orientation with respect to the SV40 promoter of the pGL3-P vector. Expected band sizes shown in table below gel image. U = uncut, 1 = Sense, 2 = Anti-sense.

Luciferase activity was measured 48 hours post transfection (as previously described in Chapter 2 Section 2.2.9 Luciferase reporter gene assays). When compared to the empty pGL3-P vector, the construct containing the SVA-D in the sense orientation showed a 3.8 fold decrease in luciferase expression (0.26±0.03, Mann-Whitney U test, p-value = 3.66E-05). Similarly, the construct containing the SVA-D in the anti-sense orientation showed a 4.1 fold decrease in luciferase expression (0.24±0.02, Mann-Whitney U test, p-value = 3.66E-05). There was no significant difference in reporter gene expression between SVA-D in the sense orientation and the SVA-D in the anti-sense orientation (Mann-Whitney U test, p-value = 0.13). Overall, in HEK293 cells, both SVA-D constructs showed a significant decrease in fold activity compared to the empty pGL3-P vector (Figure 5.2B).

The SVA-D reporter gene constructs were also tested in SH-SY5Y and SKNAS cell lines, to see if there were cell specific expression patterns. However, a similar effect to HEK293 cells was observed when tested in SH-SY5Y, with both orientations of the SVA-D causing a decrease in fold activity. The SVA-D sense orientation construct showed a 3.4 fold decrease in luciferase expression (0.29±0.03, Mann-Whitney U test, p-value = 3.66E-05), while the anti-sense orientation construct only showed a 2.7 fold decrease in expression (0.37±0.05, Mann-Whitney U test, p-value = 4.78E-04). There was no significant difference in luciferase activity between SVA-D in the sense orientation and the SVA-D in the anti-sense orientation (Mann-Whitney U test, p-value = 0.24). Overall, in SH-SY5Y cells, there was a significant decrease in fold activity when compared to the empty pGL3-P vector (Figure 5.2C).

In contrast, when tested in SKNAS, the SVA-D only showed a decrease in fold activity in one orientation. No significant difference in luciferase expression was seen between the SVA-D sense orientation and the empty pGL3-P vector (0.91±0.12, Mann-Whitney U test, p-value = 0.12). However, again a significant decrease in luciferase activity was seen: a 1.4 fold decrease in fold activity when tested in the anti-sense orientation (0.71±0.03, Mann-Whitney U test, p-value = 1.89E-05). No significant difference in luciferase expression was observed between the SVA-D in the sense orientation and anti-sense orientation (Mann-Whitney U test, p-value = 0.87). Overall, in SKNAS cells, only the SVA-D anti-sense construct showed a significant decrease in fold activity when compared to the empty pGL3P vector (Figure 5.2D).

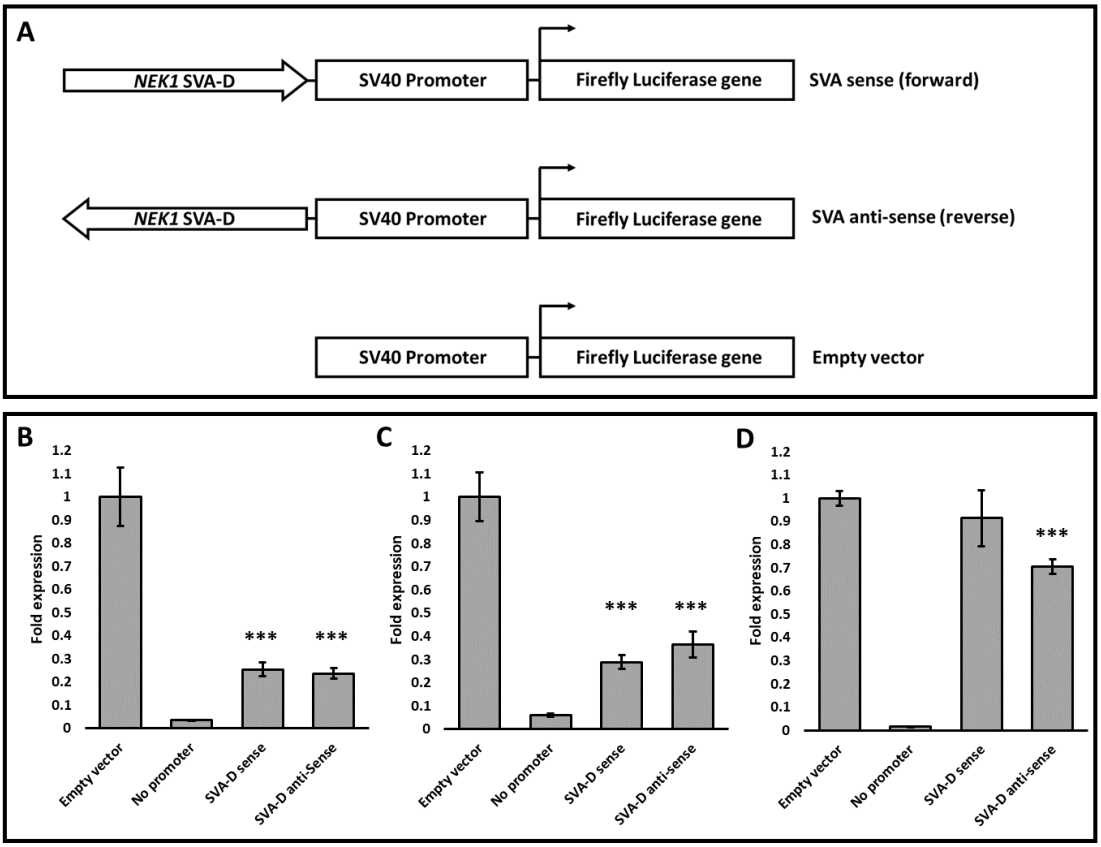


Figure 5.2. The *NEK1* SVA-D shows repressive effects in multiple cell lines.

A: Schematic of the *NEK1* SVA-D containing luciferase reporter constructs. The SVA-D was cloned in both the sense (forward) and anti-sense (reverse) orientation with

respect to the SV40 minimal promoter of the construct. The SVA-D is anti-sense with respect to the promoter of the *NEK1* gene and therefore the anti-sense orientation is the endogenous orientation of the SVA in the human genome. **B**: The fold activity of the *NEK1* SVA-D in the sense (forward) and anti-sense (reverse) orientation within the pGL3-P vector normalised to the internal control Renilla Luciferase in HEK293 (biological replicate n = 3, technical replicate per assay n = 4). **C**: The fold activity of the *NEK1* SVA-D in the sense (forward) and anti-sense (reverse) orientation within the pGL3-P vector normalised to the internal control Renilla Luciferase in SH-SY5Y (biological replicate n=3, technical replicate per assay n = 4). **D**: The fold activity of the *NEK1* SVA-D in the sense (forward) and anti-sense (reverse) orientation within the pGL3-P vector normalised to the internal control Renilla Luciferase in SKNAS (biological replicate n = 4, technical replicate per assay n = 4). The vector labelled no promoter is pGL3-B and was included as a negative control. Mann-Whitney U test was used to compare SVA containing constructs against each other and to empty vector (pGL3-P). ***P<0.001.

5.3.2 The *NEK1* SVA-D shows functional properties and an orientation bias in the pSHM06 vector in several cell lines

To determine if the *NEK1* SVA could regulate or alter efficiency of splicing it was cloned into the pSHM06 vector[304]. The pSHM06 vector was originally generated from a Renilla luciferase cDNA reporter gene construct, with the addition of exons 6 and 7 from the human *triose phosphate isomerase* (*TPI*) gene, which were inserted at both the 5' and 3' of the Renilla luciferase gene; producing a fusion protein with identical TPI peptides at each terminus of Renilla. Nott *et al*., generated three versions of this construct: an intronless version, one with intron 6 cloned between the set of exons 6 and 7 at the 5' flank of Renilla (5' intron) and one with intron 6 cloned within the set of exons 6 and 7 at the 3' end of Renilla luciferase (3' intron). These constructs were used to test the effect that the addition, and location, of intron

6 had on Renilla luciferase expression. When tested in HeLa cells and compared against the intronless pSHM06 vector, both intron constructs enhanced Renilla luciferase expression: the 5' intron version of pSHM06 induced a ~13 fold increase and the 3' intron construct elicited a ~2 fold increase in Renilla expression. Both intron constructs also led to an increase in total TPI/Renilla mRNA, concluding that addition of an intron in this vector had an enhancive effect on both TPI/Renilla expression and mRNA levels[304].

The 5' intron version of pSHM06 vector was gifted to us by Gerald Schumann (Paul Ehrlich Institute, Germany) and utilised to determine if presence of the *NEK1* SVA in intron 6 would have any effect on luciferase expression, compared to the empty pSHM06 vector (5' intron version) (Figure 2.3 and Figure 5.4A). The presence of the SVA in both constructs was confirmed using restriction digest (Figure 5.3) and Sanger sequencing (Appendix 3). A vector map of this construct was also generated using Snapgene (Figure 2.3 and Supplementary Figure 8). The Dual-Luciferase® Reporter Assay (Promega) was used to measure the luminescent signal generated by the luciferase reaction of the SVA-D containing constructs and directly compared to the signal generated by the pSHM06 vector containing the CMV promoter alone (empty vector) (Figure 2.3). The pMLuc-2 vector[145], expressing Firefly (*Photinus pyralis*) luciferase, was used as an internal control to normalise all samples, accounting for variation in cell number, cell death and transfection efficiency.

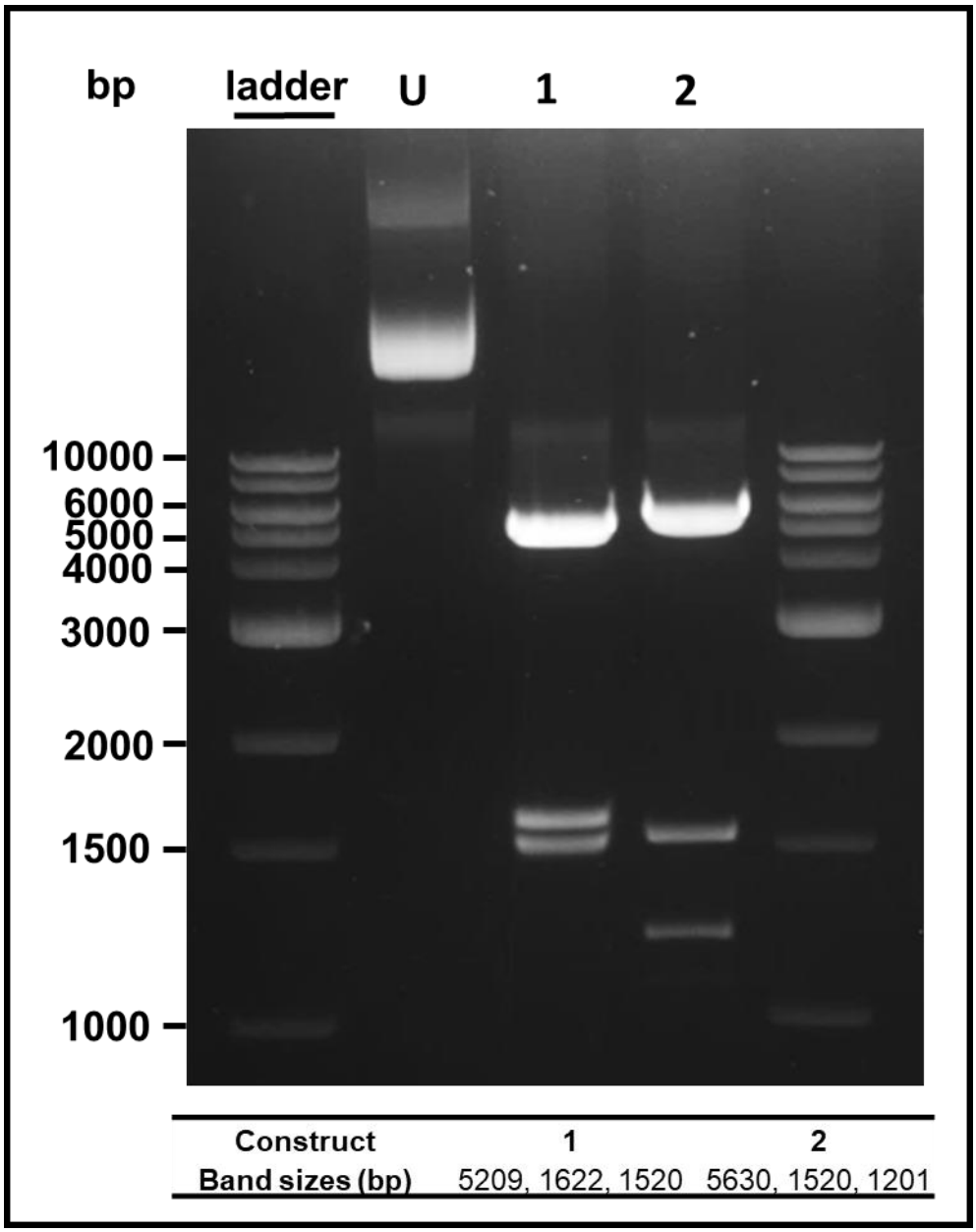


Figure 5.3. Validation of *NEK1* SVA-D pSHM06 constructs.

Restriction enzyme digest of *NEK1* SVA-D containing pSHM06 constructs. Constructs were cut with *BamHI* and run on a 1% agarose gel for 2 hours at 120V. The SVA-D was cloned in both the sense (forward) and anti-sense (reverse) orientation with respect to the CMV promoter of the pSHM06 vector. Expected band sizes shown in a table below gel image. U = uncut, 1 = Sense, 2 = Anti-sense.

When transfected into HEK293 cells and compared to the transfected empty pSHM06 vector, the construct containing the SVA-D in the sense orientation showed a 10-fold decrease in luciferase expression (0.10±0.005, Mann-Whitney U test, p-value = 3.66E-05). While we also saw the same trend in the construct containing the SVA-D in the anti-sense orientation, we only saw a 1.6 fold decrease in activity (0.64±0.05, Mann-Whitney U test, p-value = 1.11E-03). There was also a significant difference in reporter gene activity between the two SVA constructs (Mann-Whitney U test, p-value = 3.66E-05). Overall we saw a statistically significant decrease in fold expression of luciferase in both SVA-D containing constructs in HEK293 cells (Figure 5.4B).

To again examine cell specific expression profiles, we performed the luciferase assay in both SH-SY5Y and SKNAS cells. When tested in SH-SY5Y, we again observed an orientation specific repression in luciferase activity, with a 4.8 fold decrease in the construct containing the SVA in the sense orientation (0.25±0.02, Mann-Whitney U test, p-value = 3.66E-05). But no significant change in reporter gene expression was observed in the construct containing the SVA in the anti-sense orientation (0.98±0.07, Mann-Whitney U test, p-value = 0.98). A significant difference in luciferase expression between the sense and anti-sense SVA constructs was also observed (Mann-Whitney U test, p-value = 3.66E-05) (Figure 5.4C).

Similarly, this orientation specific trend was observed in SKNAS cells, with a statistically significant 5.3 fold decrease in luciferase activity in the sense SVA construct when compared to the empty pSHM06 vector (0.19±0.01, Mann-Whitney U test, p-value = 9.39E-04), but the 1.3 fold decrease observed when testing the anti-

sense SVA was not statistically significant (0.78±0.09, Mann-Whitney U test, p-value = 0.08). There was a significant difference in luciferase expression between the sense and anti-sense SVA constructs (Mann-Whitney U test, p-value = 5.38E-03) (Figure 5.4D)

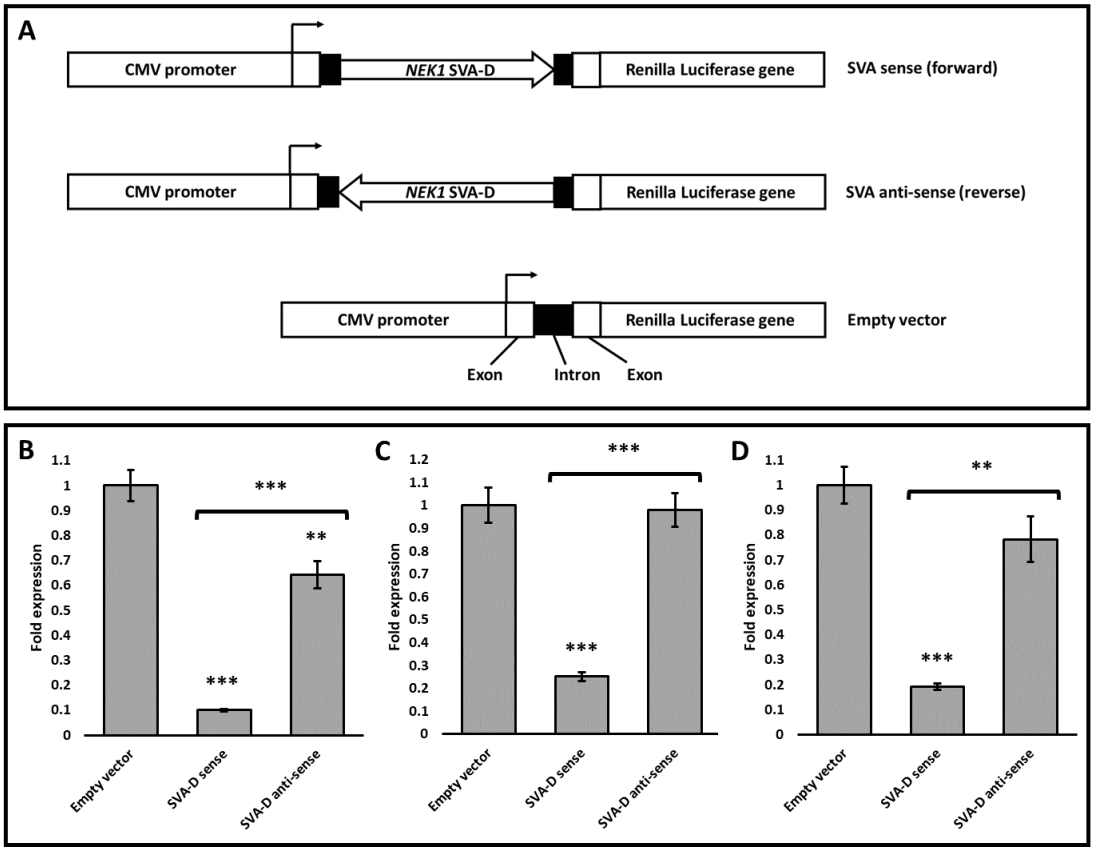


Figure 5.4. The *NEK1* SVA-D shows functional properties in pSHM06 vector in several cell lines.

A: Schematic of the *NEK1* SVA-D containing luciferase reporter constructs. The SVA-D was cloned in both the sense (forward) and anti-sense (reverse) orientation with respect to the CMV promoter of the construct and is present within intron 6 (shown in black). The SVA-D is anti-sense with respect to the promoter of the *NEK1* gene and therefore the anti-sense orientation is the endogenous orientation of the SVA in the human genome. **B**: The fold activity of the *NEK1* SVA-D in the sense (forward) and anti-sense (reverse) orientation within the pSHM06 vector normalised to the internal

control Firefly Luciferase in HEK293 (biological replicate n=3, technical replicate per assay n=4). **C**: The fold activity of the *NEK1* SVA-D in the sense (forward) and anti-sense (reverse) orientation within the pSHM06 vector normalised to the internal control Firefly Luciferase in SH-SY5Y (biological replicate n=3, technical replicate per assay n=4). **D**: The fold activity of the *NEK1* SVA-D in the sense (forward) and anti-sense (reverse) orientation within the pSHM06 vector normalised to the internal control Firefly Luciferase in SKNAS (biological replicate n=2, technical replicate per assay n=4). Mann-Whitney U test was used to compare SVA containing constructs to each other and to the empty vector (pSHM06). **P<0.01***P<0.001.

5.3.3 *NEK1* SVA-D CRIPSR knockout design and optimisation

While it had been shown that the *NEK1* SVA-D was functional within two distinct reporter gene constructs (Figure 5.2 and Figure 5.4) we wanted to also determine if the SVA was having an actual regulatory effect at the *NEK1* locus. To test the *NEK1* SVA in this context it was knocked out using the clustered regularly interspaced palindromic repeats (CRISPR) Cas9 system and *NEK1* gene expression was measured in response to this modification. *CLCN3* gene expression was also assessed: this is found to determine if the SVA could be regulating neighbouring genes.

The CRISPR experiments were performed in the HEK293 cell line, as it is well documented that this cell line has a high transfection efficiency (personal communication with colleague Ben Middlehurst)[410,411]. The ATCC website does not describe any rearrangements, markers or changes in ploidy for chromosome 4 (https://www.lgcstandards-atcc.org/products/all/crl-1573.aspx?geo_country=gb#characteristics), supported by karyotyping performed

by Binz *et al*. showing that chromosome 4 was diploid in HEK293[412]. Before attempting to knock out the SVA, *NEK1* expression in HEK293 was confirmed using RT-PCR (Figure 5.6). Primers were designed within exons that are common to all five full length isoforms of *NEK1*, therefore measuring total expression of this gene. The same strategy was also adopted for *CLCN3*.

CRISPR is an RNA-guided adaptive immune response within bacteria and archaea, used to destroy invading phage viruses and mobile elements[413]. These repetitive regions, originally obtained from previously invading exogenic elements, encode CRISPR RNA (crRNA) and trans-activating RNA (tracrRNA) which program and direct activity of Cas nuclease proteins to target and destroy invading protospacer elements and thus provide immunity[413-415]. By artificially fusing the crRNA and tracrRNA into a single guide RNA (gRNA), this system has now been repurposed and is routinely used to conduct targeted genetic modifications[416]. To knock out the *NEK1* SVA the RNA-guided CRISPR Cas9 nuclease system was utilised to stimulate double strand breaks (DSBs) either side of the SVA and excise the element (Figure 5.5). This project utilised the Type II CRISPR-Cas system, which was derived from *Streptococcus pyogenes*, using a Cas9 nuclease (spCas9) expressed by the pSpCas9(BB)-2A-GFP vector (Figure 2.4)[305]. The gRNA in the CRISPR-Cas9 system utilises a 20 bp spacer sequence which binds the target site, thus orchestrating a targeted modification by Cas9. To generate successful editing, the 20 bp target site must come immediately before a protospacer adjacent motif (PAM), specifically a 5'-NGG (N being any base) for spCas9.

Guide RNA (gRNA) sequences (20 bp oligos) were designed using the gRNA design tool (http://crispr.mit.edu/) (please refer to Chapter 2 Section 2.2.10.1 Guide RNA design). These guides were complementary to the chosen 20 bp target sequences (flanking regions surrounding the SVA): guides 1 and 2 were 395 bp and 750 bp upstream, respectively: guides 3 and 4 were 99 bp and 139 bp downstream, respectively (Figure 5.5). All four guides and two non-target control guides were then cloned into the pSpCas9(BB)-2A-GFP vector (Figure 2.4). This vector will then transcribe the necessary guide RNA sequence and scaffold which can then direct the Cas9 nuclease to the target site and induce a double strand break (DSB). This experiment used a dual-target approach, by co-transfecting two separate pSpCas9(BB)-2A-GFP vectors each containing a gRNA sequence. These two vectors each expressed a gRNA which directed the Cas9 nuclease to a target sequence, generating DSBs at the 5' and 3' ends of the SVA, excising the element. Each DSB was then repaired by non-homologous end joining (NHEJ).

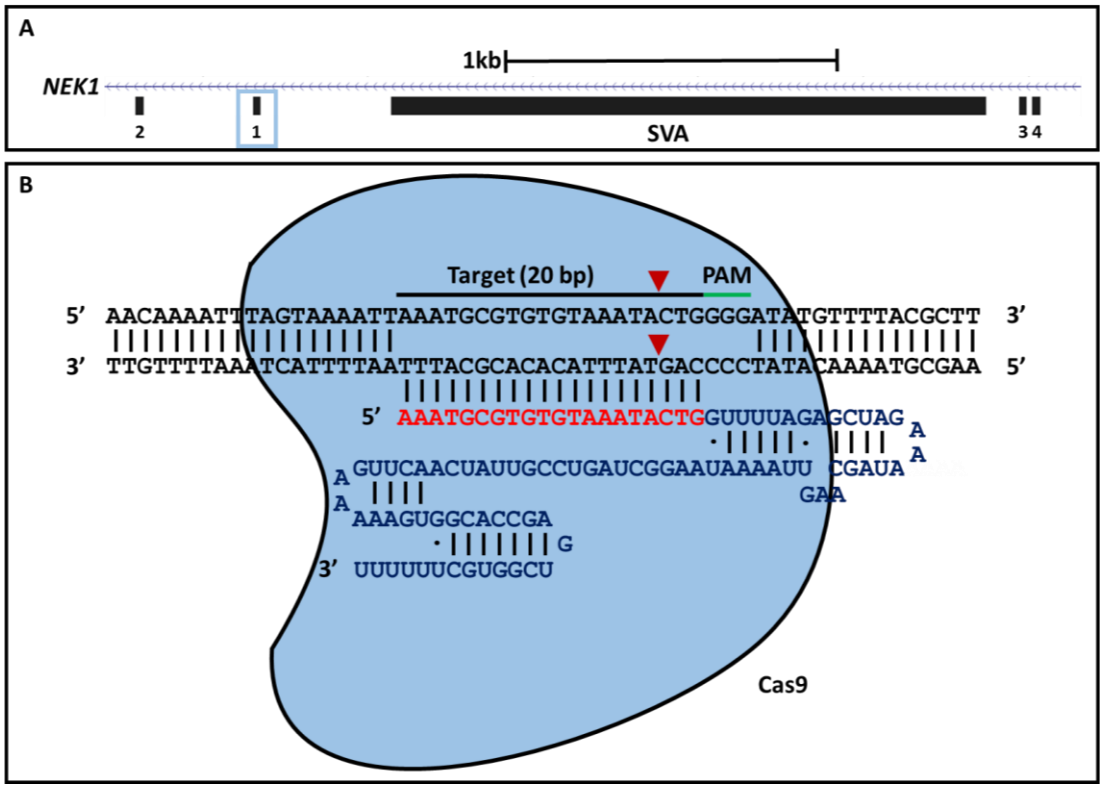


Figure 5.5. RNA-guided CRISPR Cas9 nuclease schematic.

A: The *NEK1* locus overlaid with locations of the designed guide RNAs (gRNA) used to excise the SVA element. **B**: Schematic of the CRISPR Cas 9 machinery and gRNA expressed by the pSpCas9(BB)-2A-GFP plasmid. The DNA region of interest is shown in black, with the 20 bp target sequence shown adjacent to a PAM site (green line). The designed 20 bp oligo (red) is complementary to the target site of interest and fused to an RNA scaffold (dark blue), constituting the guide RNA (gRNA). The gRNA then pairs with the complementary target, directing the Cas9 protein (blue) to this region which then stimulates a DSB approximately 3 bp upstream of the PAM site (red arrow). (Adapted from Ran *et al*., 2013)[305].

5.3.4 *NEK1* is expressed in HEK293 cells

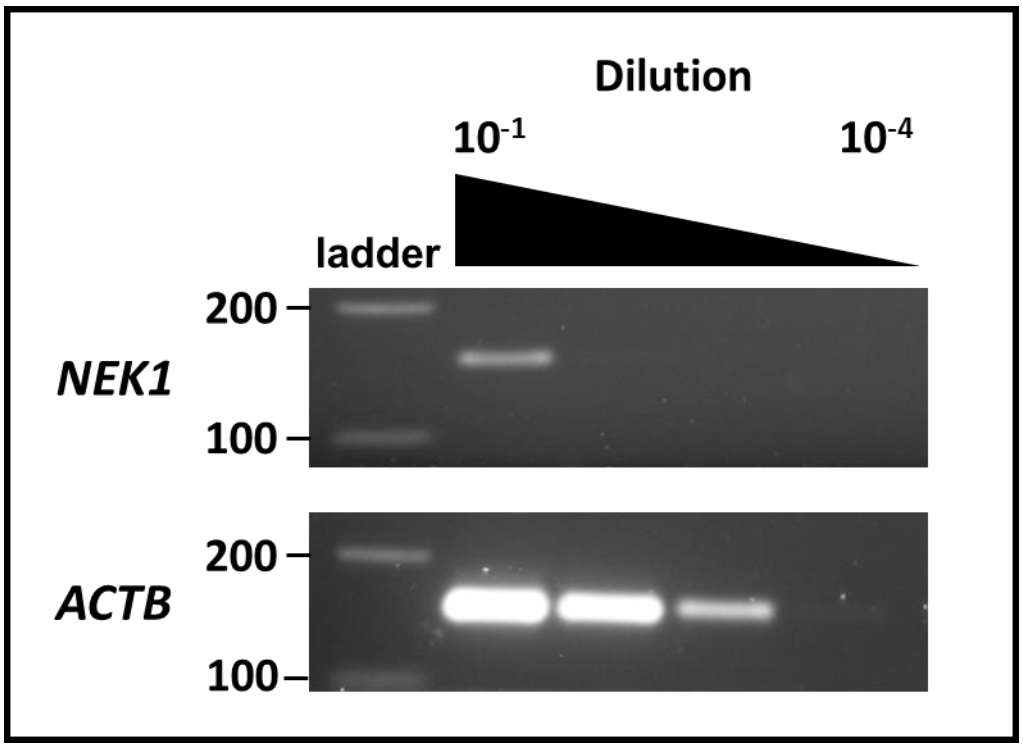


Figure 5.6 *NEK1* is expressed in HEK293 cells.

Reverse transcription PCR using cDNA generated from HEK293 cells for total *NEK1* expression and *ACTB*. 1:10-1:10000 cDNA dilutions were used for both targets. Primers were designed in exons common to all five full length isoforms of *NEK1* (amplicon size of 157 bp) and therefore constitutes total expression in this cell line.

5.3.5 *NEK1* SVA CRISPR KO experimental outline

A total of four guides were designed to target the flanks of the SVA-D: two upstream and two downstream. These guides were cloned into the pSpCas9(BB)-2A-GFP vector using Golden Gate cloning technique (Please refer to Chapter 2 section

2.2.10.2 Golden Gate cloning for a detailed overview of this technique). Once cloned, the vector insertion site was sequenced to determine presence of the guide (Appendix 3). Each of the four guides were co-transfected as pairs in HEK293 cells (a total of four combinations) and 48 hours post transfection genomic DNA (gDNA) was extracted and purified. PCR was used to determine presence or absence of the SVA in the transfected population: a forward primer was designed 1207 bp upstream of the SVA and a reverse primer was designed 434 bp downstream of the SVA. When the region was unmodified the amplicon/region equated to 3432 bp (the larger band in Figure 5.7B) and the presence of smaller amplicons equates to the region being modified (SVA being excised by the different gRNA combinations). Overall, all four guide combinations did successfully excise the SVA. Densitometry was performed on all four guide combinations using ImageJ[417]: a rectangle was drawn over each lane and the integrative density was calculated for each lane (constituting the total integrative density for that lane/cell population). The integrative density for each modified band was then calculated and the ratio of modified:whole lane signal was calculated. Each ratio was then then normalised against the 1,3 guide lane to determine which lane had the highest signal of modified DNA and then modified signal % was then determined (Supplementary Table 3). The most efficient combination (determined through densitometry, Supplementary Table 3) was guides 1 and 3 (1089 bp band in Figure 5.7), therefore this pair were therefore taken forward to generate the SVA KO lines. Guides 1 and 3 were also chosen as they were the only guide combination which knocked out the SVA alone; all other guide combinations knocked out neighbouring LINE and *Alu* elements at this locus and thus were not a

true and accurate representation of the regulatory effect of the SVA alone at this region.

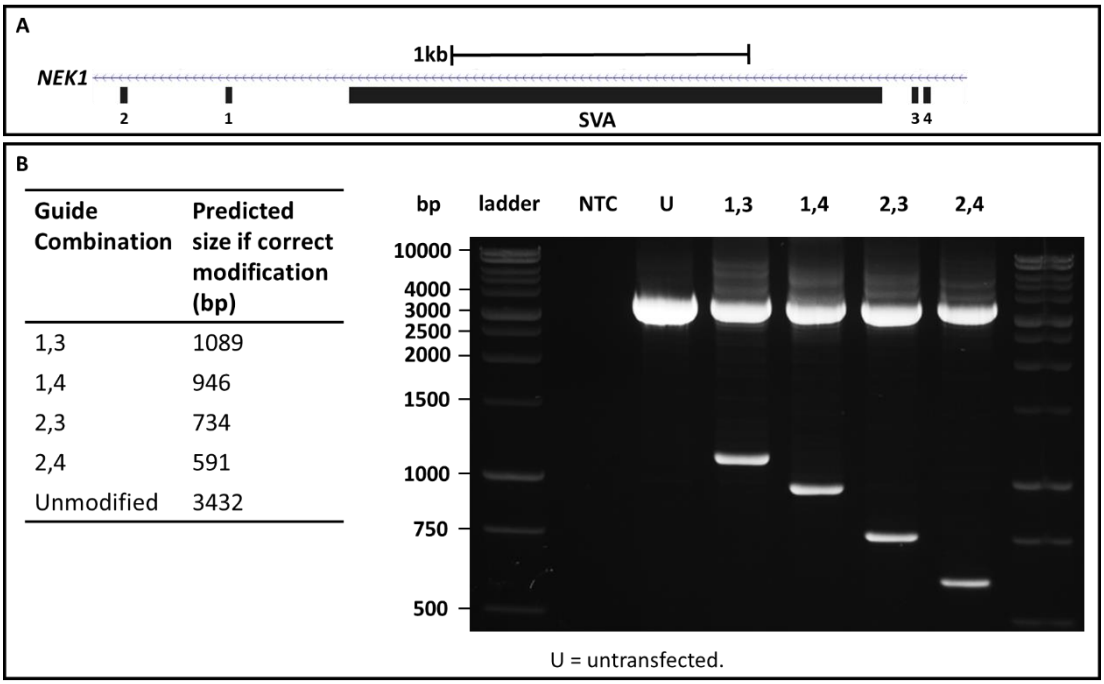


Figure 5.7. CRISPR guide RNA modification verification.

A: Visual representation of the NEK1 SVA-D and the four gRNAs designed for the SVA KO experiment. B: PCR amplification and gel agarose electrophoresis of the wider intronic region containing the SVA-D. Each lane corresponds to the combination of guide RNAs used to knock out the SVA and their respective amplicon sizes; unmodified and modified. Guides 1 and 3 were taken forward. NTC = no template control, U = untransfected.

Once the CRISPR mediated SVA KO had been validated, wildtype HEK293 cells were seeded at 100,000 cells per well (24-well format). After 24-hours, guides 1 and 3 (ΔSVA) were co-transfected using Turbofect™ Transfection Reagent (ThermoFisher) (Table 2.9). Two non-target guide RNAs were also cloned into pSpCas9(BB)-2A-GFP vector, were co-transfected and used as a negative control within this experiment. Please refer to Chapter 2 Section 2.2.10.4 Single cell seeding and clonal expansion for exact transfection details. To generate clonal populations of modified cells, the

wildtype and transfected cells were then seeded at a low-density; single cells were left to divide into small colonies and transferred into separate wells, expanded over time, and then genotyped to confirm if SVA excision had occurred. 48 hours post transfection, each condition (untransfected/wildtype, non-target guide and ΔSVA) was plated on sterile and TC treated 10 cm petri dishes (both 1000 and 2000 cells per condition). Media changes were performed every 2 days and all cells were monitored for 14 days (or until visible colonies of cells could be seen). Following this, colonies were picked with a sterile pipette tip and placed into separate wells of a 96-well plate. Overall, 220 colonies of cells co-transfected with guides 1 and 3 were picked and 204 survived (survival rate = 93%). Once the well of the 96-well plate was 90% confluent, half of the cells were taken forward to a fresh 96-well plate and the other half were used to genotype the cell line. DNA was extracted from each cell line (n = 204) using the DirectPCR® Lysis Reagent (Viagen Biotech). This crude lysis system allowed us to screen multiple samples in a time effective manner, with 1 µl of cell crude lysis being used in a preliminary screen using the *NEK1* SVA presence/absence PCR (Figure 5.9). Once SVA KO cells were validated through PCR they were expanded to 24 well and then 6-well plates. These cells were then split into two 6-well plates; one for freezing and long terms storage in liquid nitrogen and one for downstream applications (RT-PCR and qPCR) (Figure 5.8). Untransfected (wild type) and non-target guide cell lines were also picked and clonally expanded (n=3 per condition).

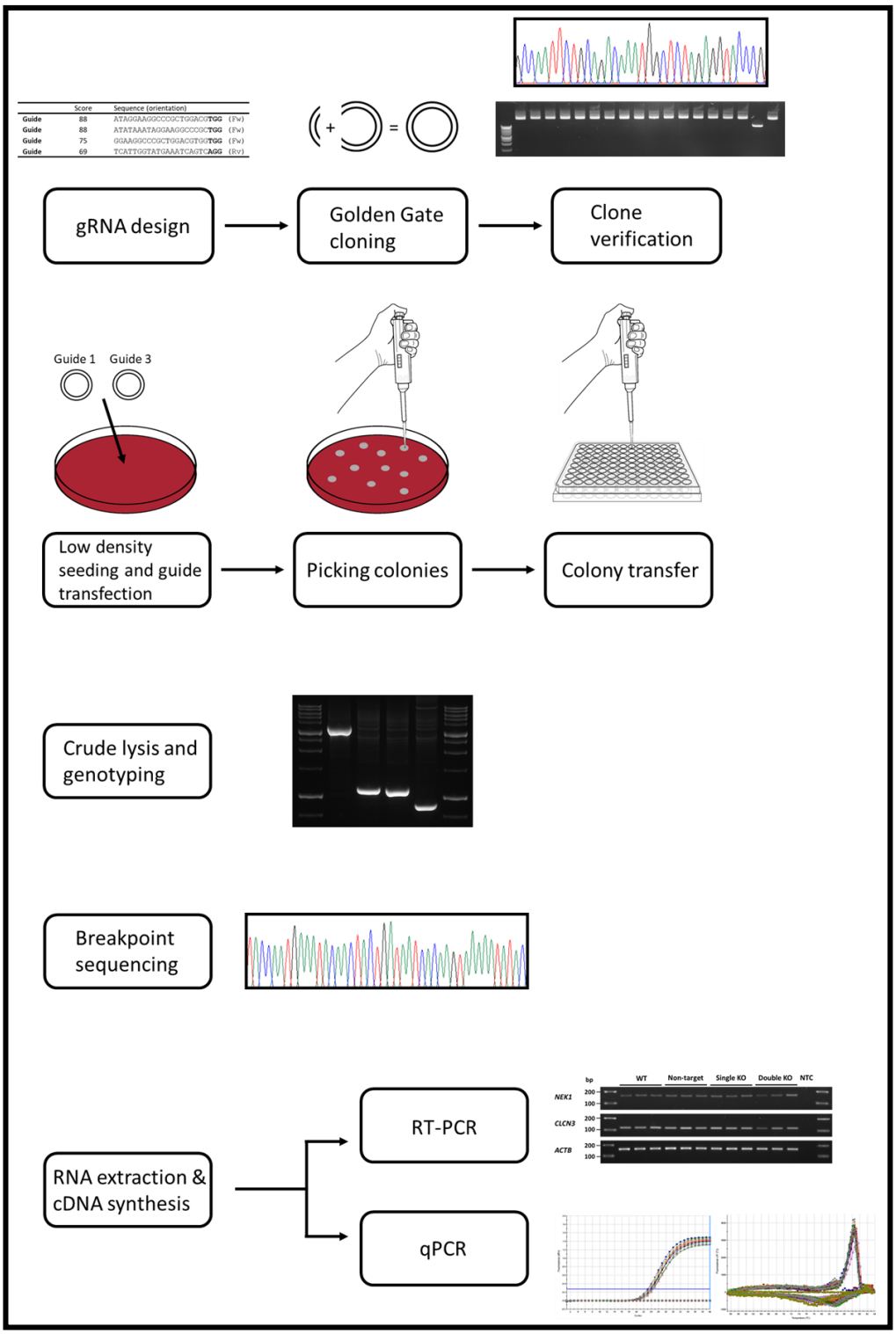


Figure 5.8. *NEK1* SVA-D CRISPR project pipeline.

A pipeline of the steps involved in the process of generating SVA KO cell lines. The first step of gRNA design, cloning and clone verification was finished in 1-2 weeks. Low density seeding, colony picking and colony transfer was accomplished in approximately 3 weeks (but is dependent on colony number). Clonal expansion and genotyping took approximately 3 weeks (also dependent on colony number). RNA

extraction, cDNA synthesis was finished in a couple of days, but optimisation of the downstream applications took 2-3 weeks.

After screening the 204 clonal colonies, 20 lines were identified which were heterozygous for the SVA KO (9.8% efficiency) and 3 lines were identified which were homozygous for the SVA KO (1.5% efficiency). To validate and confirm the accuracy of the preliminary screen and to ensure high quality and purity of genomic DNA for future experiments, DNA was next extracted, isolated and purified using the GenElute™ Mammalian Genomic DNA Miniprep kit (Sigma). This validation step identified one cell line to be unmodified which had previously been genotyped (from crude lysis) and found to be heterozygous for the SVA KO (Supplementary Figure 9).

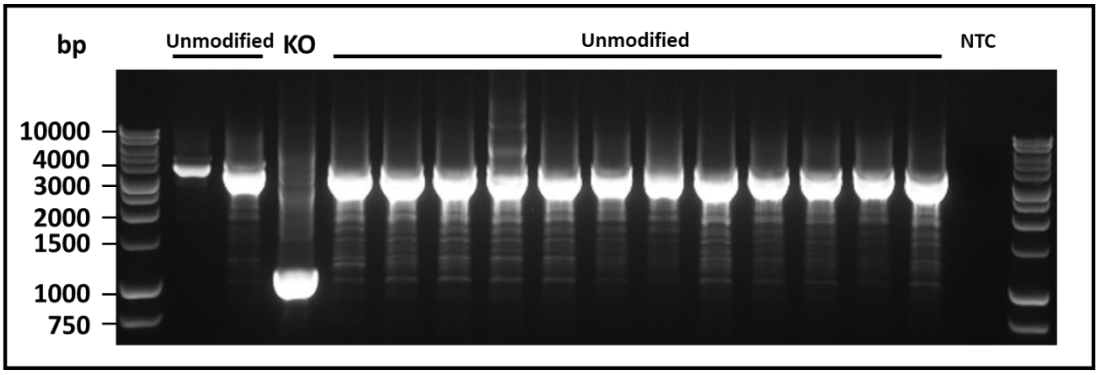


Figure 5.9. The *NEK1* SVA was successfully removed from HEK293 cell line.

PCR amplification and agarose gel electrophoresis of isolated gDNA from crude lysate of CRISPR colonies. The unmodified amplicon (SVA present) is 3432 bp and edited (SVA absent) is 1089 bp. The absence of the unmodified band in lane 4 indicates that all alleles of the SVA have been removed in this cell line. Samples were run on a 1% agarose gel at 110V for 1 hour. KO = knockout, NTC = non template control.

5.3.6 Validation of *NEK1* SVA CRISPR KO lines

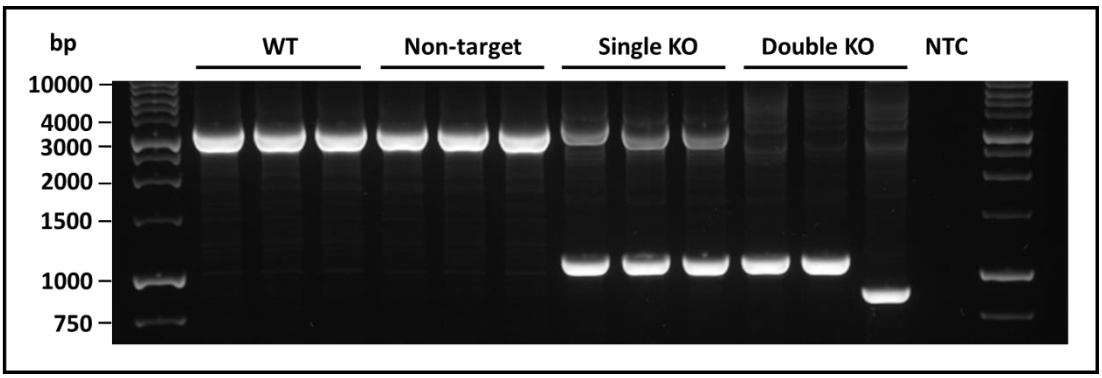


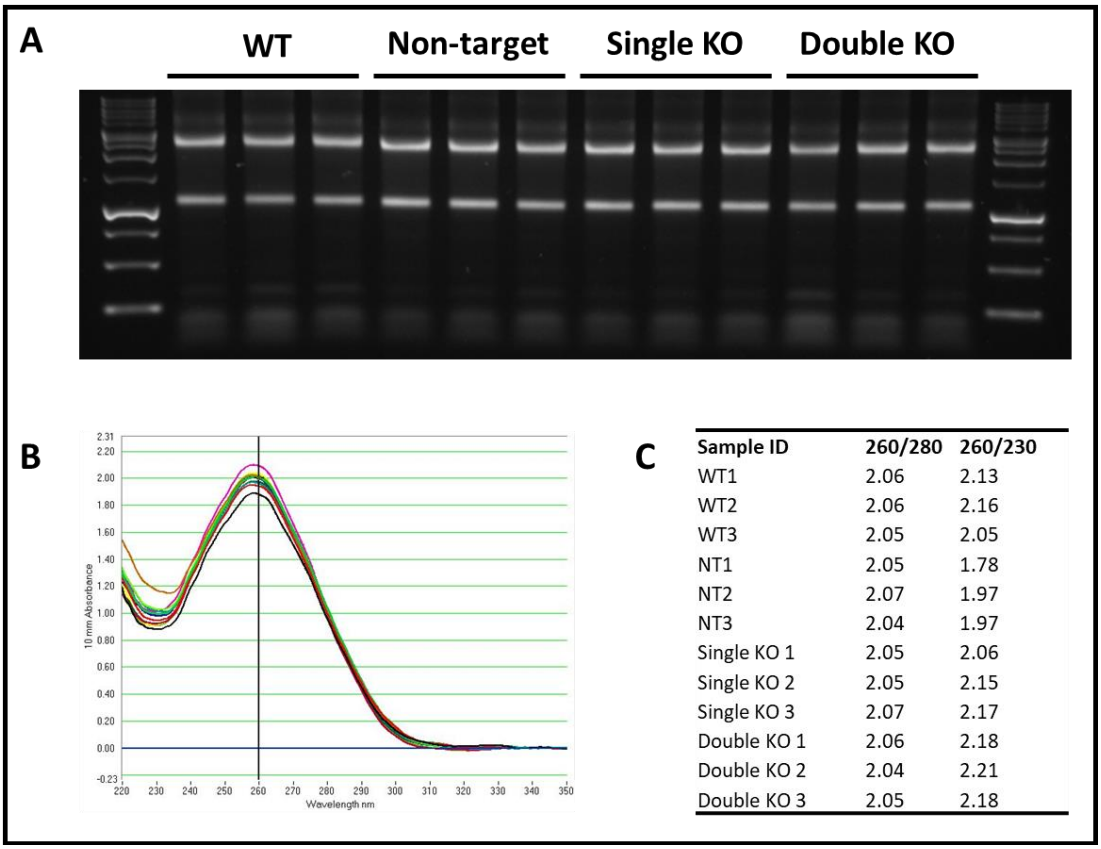
Figure 5.10. *NEK1* SVA KO cell line validation.

PCR amplification and agarose gel electrophoresis of purified gDNA from untreated (wildtype; WT), non-target, single SVA KO and double SVA KO HEK293 cell lines (per condition, n=3). The unmodified amplicon (SVA present) is 3432 bp and edited (SVA absent) is 1089 bp, with one double KO. Samples were run on a 1% agarose gel at 110V for 1.5 hours. KO = knockout, NTC = no template control.

Interestingly, one of the homozygous SVA deletion cell lines had a larger deletion than predicted (Figure 5.10, lane 14). Due to the NHEJ repair mechanism being error prone the targeted breakpoints produced in all six modified cell lines were sequenced, determining if any indels were generated during the repair process (Appendix 3).

5.3.7 RT-PCR analysis of SVA KO lines

RNA from wildtype, non-target, single KO and double KO (per condition, n = 3) was isolated from mammalian cell lines using the Monarch Total RNA Miniprep kit (NEB). To determine the RNA quality and purity was adequate for downstream processes, two quality control checks were performed. Firstly, RNA purity and quality were assessed using a NanoDrop™ spectrophotometer. All RNA samples had 260/280 ratios between 2.00-2.10 and 260/230 ratios between 1.78 and 2.21 indicating that all samples were high quality and contained no contaminants (please refer to Chapter 2 Section 2.2.5 Nucleic acid quality control for more details). Secondly, all RNA samples were normalised and run on an agarose gel to determine if any degradation of ribosomal subunits had occurred. Two clear bands were observed, confirming that both the 28S and 18S ribosomal RNA subunits were intact and there was no RNA degradation in any of the samples (Figure 5.11). Once the RNA had undergone quality control it was normalised through dilution with nuclease free water then converted to complementary DNA (cDNA) using the GoScript™ Reverse Transcription System (Promega), which was then used for reverse transcription PCR (RT-PCR). Please refer to Chapter 2 Section 2.2.11.1 RNA extraction and quality control and Section 2.2.11.2 cDNA synthesis for more details on these protocols.



Sample ID	260/280	260/230
WT1	2.06	2.13
WT2	2.06	2.16
WT3	2.05	2.05
NT1	2.05	1.78
NT2	2.07	1.97
NT3	2.04	1.97
Single KO 1	2.05	2.06
Single KO 2	2.05	2.15
Single KO 3	2.07	2.17
Double KO 1	2.06	2.18
Double KO 2	2.04	2.21
Double KO 3	2.05	2.18

Figure 5.11. *NEK1* SVA KO cell line RNA integrity and purity assessment.

RNA quality control assessment of the 12 HEK293 cell lines used throughout the CRISPR project (wildtype, non-target control, single KO and double KO). **A**: Gel agarose electrophoresis to determine if RNA degradation had occurred. All samples were run on 1% agarose at 120V for 50 minutes. **B**: Spectral output from a NanoDrop™ spectrophotometer. **C**: NanoDrop™ 260/280 and 260/230 ratios, used to indicate RNA quality and purity via detection of contaminants (protein and phenol) in each sample prep. WT = wildtype, KO = knockout.

To determine if the SVA knock out resulted in any gene expression changes, *NEK1* and *CLCN3* expression was assessed using RT-PCR and compared to the reference genes *ACTB* and *GAPDH*. Initially, *ACTB*, *GAPDH* and *UBC* were all tested by RT-PCR of CRISPR SVA KO cell lines, to test if these reference genes were stable and thus suitable as qPCR reference genes. Both *GAPDH* and *UBC* have previously been tested and proven to be stable reference genes in HEK293[418]. The primers obtained for *UBC* from Zhang *et al*.[419] were not specific and therefore were discarded from this

experiment (Supplementary Figure 10). However, both *ACTB* and *GAPDH* were found to be stable across all CRISPR modified cells and controls (wildtype and non-target) under basal conditions. To assess total expression, primers were designed within common exons of all main isoforms of both target genes (*NEK1* and *CLCN3*) and the reference genes *ACTB* and *GAPDH*. No change in *ACTB* and *GAPDH* expression across the 12 different cell lines was observed, confirming that this reference gene is stable under basal conditions and thus is suitable as a reference gene for qPCR. Moreover, no clear changes in *NEK1* or *CLCN3* expression were seen in any of the single KO lines (Figure 5.12, lanes 8-10). Minor changes in *NEK1* and *CLCN3* expression were only observed within the homozygous KO (double KO) lines (Figure 5.12, lanes 11-13). Primers were also designed which amplified all 5 full length isoforms of *NEK1* but yielded separately sized amplicons for each isoform. Only isoforms 1, 2 and 5 were found to be expressed in HEK293 and no clear difference in expression was observed between edited and unedited cells (Supplementary Figure 12).

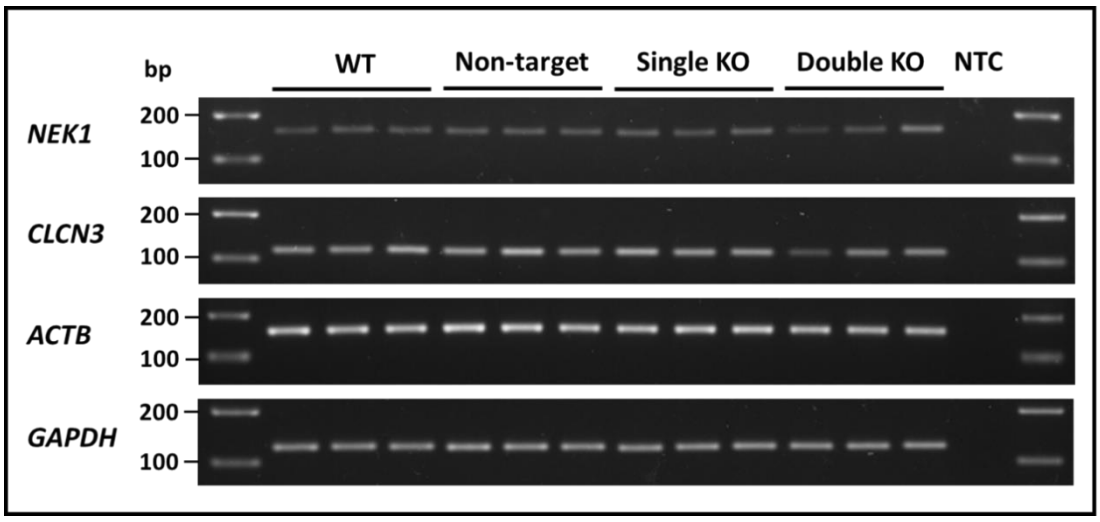


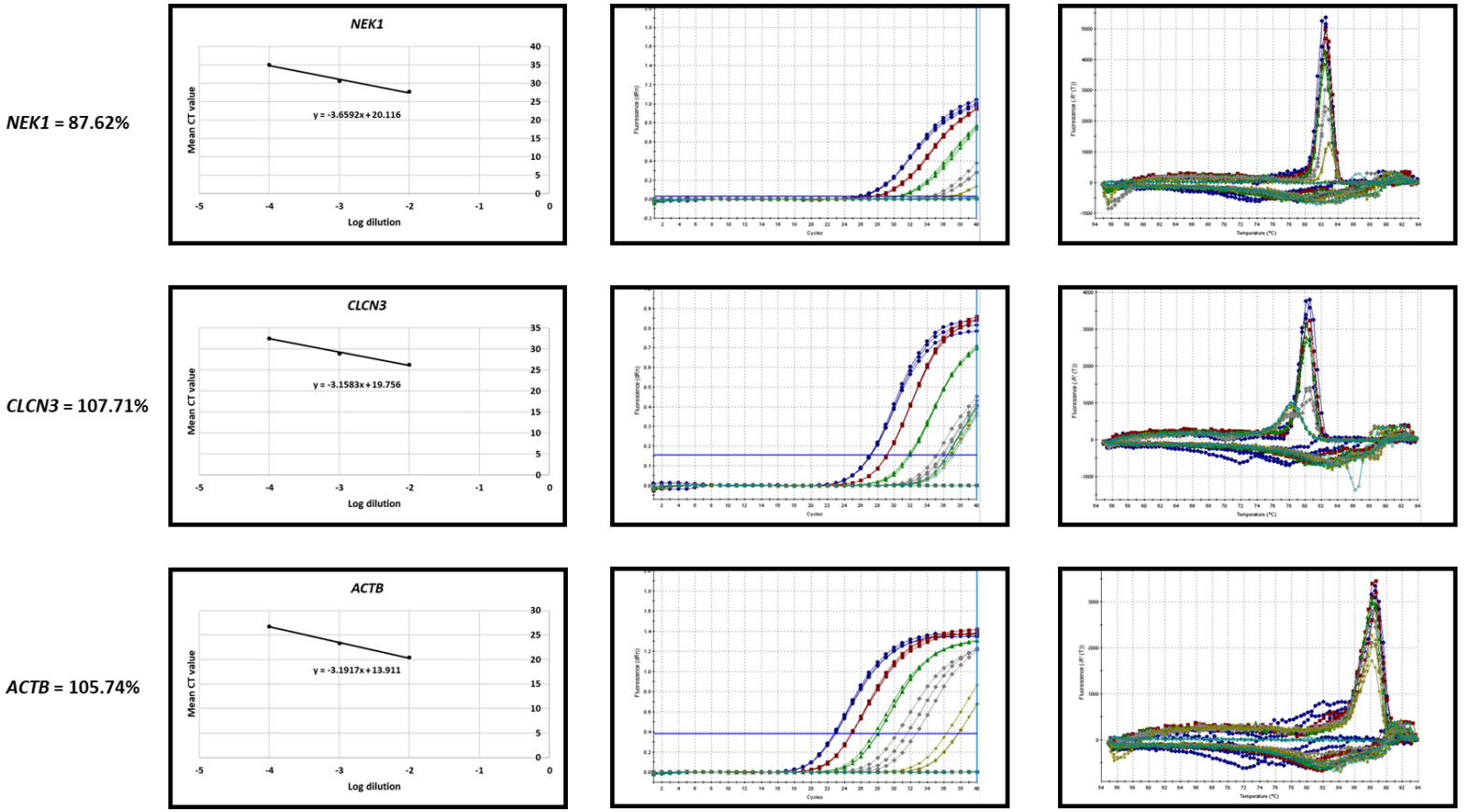
Figure 5.12. *NEK1* and *CLCN3* RT-PCR analysis in SVA KO cell lines.

Reverse transcription PCR analysis of *NEK1*, *CLCN3*, *ACTB* and *GAPDH* mRNA from untreated (wildtype; WT), non-target, single SVA KO and double SVA KO HEK293 cell

lines (per condition, n = 3). Samples run on 2% agarose gel at 110V for 1 hour. NTC = no template control.

5.3.8 qPCR primer efficiency and specificity assessment

For validation of primers for qPCR, amplification efficiencies were calculated for *NEK1*, *CLCN3* and *ACTB* (please refer to Chapter 2 Section 2.2.11.4.2 Testing primer efficiencies for details). Primer efficiencies for both *CLCN3* and *ACTB* exceeded 100% (107.71% and 105.74%, respectively), but fell within the accepted range of 90-110% efficiency for the ΔΔCT method of qPCR. The *NEK1* primer efficiency was slightly below 90% (87.62%) but was taken forward due to time constraints. The efficiency of primers for *GAPDH* previously used in RT-PCR (Figure 5.12) were also tested but fell far below the 90% cut off (63.78%) and thus this reference gene was not taken forward in qPCR. Melt curves were also generated to confirm that only single amplicons were being generated in each PCR reaction (Figure 5.13); this was also confirmed though gel agarose electrophoresis of RT-PCR products (Figure 5.12). A small secondary peak was observed in the CLCN3 melt curve but this result was interpreted as not all regions of the CLCN3 amplicon melted immediately and was a multi-state process and therefore was not due to multiple amplicons being generated in the reaction, which was supported by the observation of a single band for CLCN3 on an agarose gel (Figure 5.12). Relative gene expression was then assessed using the delta delta Ct (ΔΔCt) method as previously described in Chapter 2 Section 2.2.11.4.3 Relative quantification of gene expression.



***NEK1* = 87.62%**

***CLCN3* = 107.71%**

***ACTB* = 105.74%**

Figure 5.13. Quality control for qPCR.

Standard curves for *NEK1*, *CLCN3* and *ACTB* used to calculate efficiency of primers used in qPCR. Amplification plots for *NEK1*, *CLCN3* and *ACTB* qPCR reactions. Melt curves generated after qPCR for *NEK1*, *CLCN3* and *ACTB* amplicons.

5.3.9 *NEK1* and *CLCN3* gene expression analysis of SVA KO lines

When compared to the wildtype (WT, unmodified), no significant difference in *NEK1* expression was observed in the non-target (NT) cell lines (Mann-Whitney U Test, p-value = 0.66); the same result was also observed for CLCN3 expression (Mann-Whitney U Test, p-value = 1.00), demonstrating that the non-target guides utilised in this process had no significant effect on *NEK1* or *CLCN3* expression. Given that these lines went through an identical transfection process and establishment of clonal cell lines, they served as the best control for this experiment and therefore all results were normalised to these cell lines. The heterozygous SVA deletion (n = 3) led to a 1.23 fold increase in relative *NEK1* gene expression when compared to the non-target cell lines (n = 3), but this observation was not statistically significant (Mann-Whitney U test, p-value = 0.19): there was no significant difference in *NEK1* expression between the non-target and the heterozygous SVA KO cell lines. Excision of the SVA from all copies of chromosome 4 (homozygous SVA KO) (n = 3) led to 1.69 fold increase in relative *NEK1* expression when compared to non-target control cell lines (n = 3), but this observation also did not reach statistical significance (Mann-Whitney U test, p-value = 0.19). Heterozygous excision of the SVA (n = 3) elicited a 1.24 fold increase in relative *CLCN3* expression when compared to non-target control cell lines (n = 3), but this result did not reach statistical significance (Mann-Whitney U test, p-value = 0.38). Homozygous deletion of the SVA (n = 3) only induced a 1.09 fold increase in relative *CLCN3* expression when compared to non-target control cell lines (n = 3), therefore was not statistically significant (Mann-Whitney U test, p-value = 1.00). Overall, excision of the SVA had no significant effect on either *NEK1* or *CLCN3* gene expression.

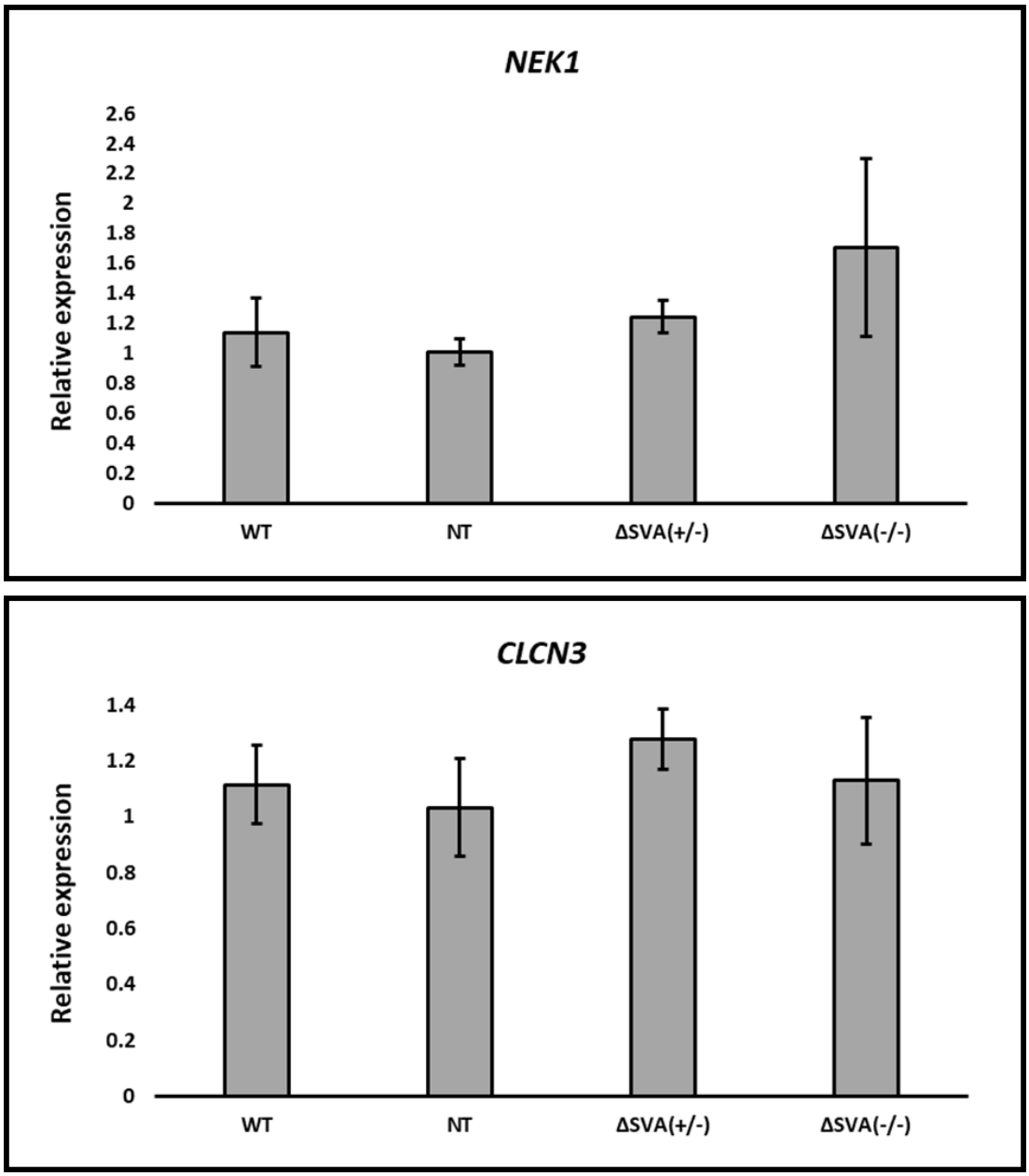


Figure 5.14. Relative expression of *NEK1* and *CLCN3* in response to SVA-D knock out.

A: Interval plot of gene expression of *NEK1* relative to *ACTB* reference gene. **B**: Interval plot of gene expression of *CLCN3* relative to *ACTB* reference gene. Gene expression analysis via qPCR was performed in control (wild type; WT), non-target guide (NT), heterozygous ΔSVA and homozygous ΔSVA HEK293 cell lines (each group n=3). Relative expression was calculated using the ΔΔCt method and was normalised to the non-target (NT) group. There was no significant difference between relative *NEK1* expression of the non-target (NT) lines and the heterozygous ΔSVA lines (Mann-Whitney U test). There was also no significant difference between the relative *NEK1*

expression of the non-target (NT) lines and the homozygous ΔSVA lines (Mann-Whitney U test). The same result was also found for relative *CLCN3* expression in both SVA KOs compared to non-target (NT) lines (Mann-Whitney U Test). No outliers were present in this data.

5.3.10 The larger homozygous deletion also excised an Alu element

The results from the qPCR data showed that one of the homozygous SVA KO lines elicited a large increase in *NEK1* gene expression (Figure 5.15, panel B). This particular cell line had undergone a larger deletion than predicted (an extra 254 bp were cleaved). Based on the position of the guide RNA sequences, the expected deletion size was 2309 bp: the observed larger deletion corresponded to 2563 bp. Furthermore, this enlarged excision overlapped with an AluSq2 element (hg19, chr4:170489810-170490120); according to the RepeatMasker track on UCSC (hg19) this element is 311 bp in length and found 450 bp upstream of the SVA. A total of 200 bp of this element was cleaved, causing a partial deletion of this Alu on all copies of chromosome 4 in this cell line (Figure 5.15) (Supplementary Figure 11). Inspection of this chromatogram also led to the discovery of a 16 bp insertion of GGCAACAACAAAAATC, which we hypothesise was adding during the NHEJ repair process (Appendix 3).

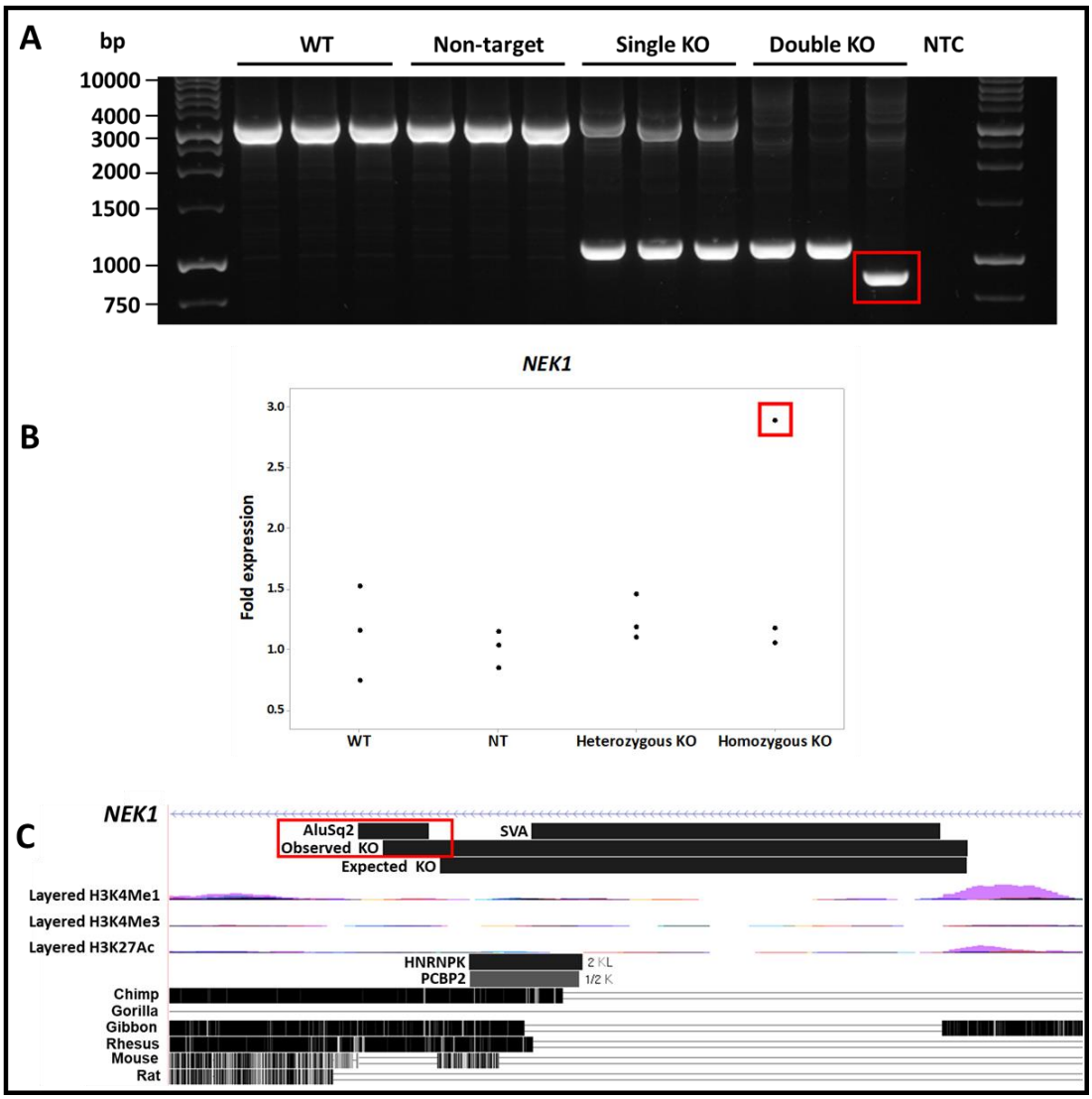


Figure 5.15. The larger homozygous KO cleaved an Alu element.

A: PCR amplification and gel agarose electrophoresis of extracted and purified DNA from CRISPR cell lines: wild-type (WT), non-target guides, single SVA KO (heterozygous KO) and double SVA KO (homozygous KO). **B**: individual value plot showing relative *NEK1* gene expression in each set of cell lines (n = 3 for each). **C**: Visual representation of the modified region from UCSC (hg19), showing expected and observed modified regions, overlaid with SVA and AluSq2 position, ENCODE data (histone marks and transcription factor binding) and conservation in primates. Red boxes indicate the larger modification observed in one of the homozygous SVA KO lines, corresponding with the observed increase in *NEK1* expression and excision of part of an *Alu*Sq2 element.

From overlaying the *Alu*Sq2 with ENCODE and conservation data, no clear histone marks or transcription factor binding was observed but this element was conserved in several species of primate. This element was also genotyped using Isaac Variant Caller data (as previously described in Chapter 2 Section 2.2.12.4 Isaac Variant Caller data analysis and manipulation), but no indels were observed within this *Alu*. As the third homozygous SVA KO was not genetically identical to other KO lines we decided to exclude this cell line and repeat the gene expression analysis (Figure 5.16). Once this third cell line was excluded, the heterozygous SVA KO lines (n = 2) elicited a 1.36 fold increase in *NEK1* expression compared to non-target controls (n = 2), but this was not statistically significant (Mann Whitney U test, p-value = 0.25). The homozygous SVA KO lines (n = 2) induced a 1.18 fold increase in relative *NEK1* gene expression when compared to non-target controls (n = 2) (Mann Whitney U test, p-value = 0.25). In this same analysis, the heterozygous SVA KO cell lines led to a 1.48 fold increase in relative *CLCN3* gene expression, but the result was not statistically significant (Mann Whitney U test, p-value = 0.25); homozygous SVA KO only led to a 1.07 fold increase and was therefore was not statistically significant (Mann Whitney U test, p-value = 1.00).

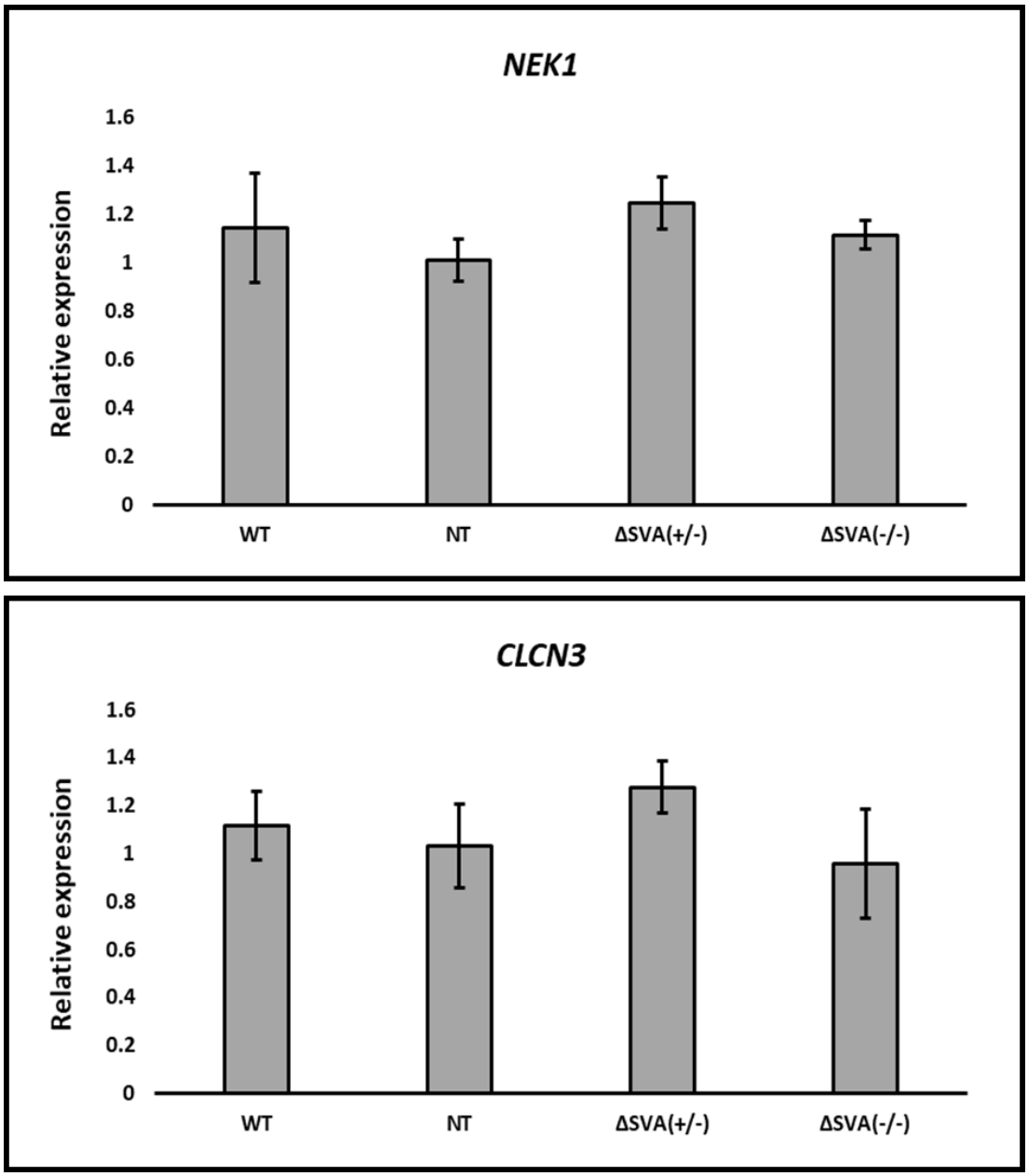


Figure 5.16. Amended relative expression of *NEK1* and *CLCN3* in response to SVA-D knock out.

A: Interval plot of gene expression of *NEK1* relative to *ACTB* reference gene, with the third homozygous deletion cell line (ΔSVA(-/-)) removed. **B**: Interval plot of gene expression of *CLCN3* relative to *ACTB* reference gene, with the third homozygous deletion cell line (ΔSVA(-/-)) removed. Gene expression analysis via qPCR was performed in control (wild type; WT) (n = 2), non-target guide (NT) (n = 2), heterozygous ΔSVA (n = 2) and homozygous ΔSVA (n = 2) HEK293 cell line. Relative expression was calculated using the ΔΔCt method and was normalised to the non-target (NT) group. There was no significant difference between relative *NEK1* expression of the non-target (NT) lines and the heterozygous ΔSVA lines (Mann-

Whitney U Test). There was also no significant difference between the relative *NEK1* expression of the non-target (NT) lines and the homozygous ΔSVA lines (Mann-Whitney U Test). The same result was also found for relative *CLCN3* expression in both SVA KOs compared to non-target (NT) lines (Mann-Whitney U Test). No outliers were present in this data.

5.4 Discussion

The studies in this chapter aimed to address the potential for an SVA retrotransposon within the *NEK1* locus to act as a transcriptional regulatory domain. Two main *in vitro* strategies were adopted: luciferase reporter gene assays and SVA knock out using the CRISPR Cas9 system. The *NEK1* SVA (reference genome sequence) was successfully cloned into two separate reporter gene constructs, pGL3-P and pSHM06, in both the forward and reverse orientation with respect to the vector promoters. The *NEK1* SVA when tested within the pGL3-P vector elicited significant repression in reporter gene expression in HEK293 and SH-SY5Y cell lines, but this response was not observed in SKNAS. When cloned into pSHM06, the SVA also exhibited repressive effects on reporter gene expression, but in an orientation specific manner: repression was only observed when the SVA was in the sense (forward) orientation. Utilising the CRISPR Cas9 system the *NEK1* SVA was successfully knocked out in HEK293 cells. In this chapter an optimised CRISPR pipeline was presented which can be adopted to knock out SVA elements, with a heterozygous KO efficiency of 10% and a homozygous KO efficiency of 1%. Excision of the *NEK1* SVA did not induce any statistically significant changes to either *NEK1* or *CLCN3* gene expression. One homozygous KO line demonstrated a mild increase in *NEK1* gene expression but had also partially cleaved an *Alu*Sq2 element upstream of the SVA (Figure 5.15). Once this cell line was removed, overall *NEK1* gene expression dropped, and no clear difference compared to the non-target control was observed (Figure 5.16).

The reporter gene data presented here further supports that SVA elements are functional regulatory domains, inducing repression of reporter gene activity in an orientation specific manner (Figure 5.2 and Figure 5.4). When tested in HEK293 and SH-SY5Y, the *NEK1* SVA in pGLP-3 facilitated a significant decrease in luciferase gene expression, in both the forward and reverse orientation. In SKNAS, the SVA in the sense orientation did not alter reporter gene activity, but a significant decrease in expression was observed in the reverse orientation. Savage *et al*. have previously shown that SVAs can induce cell specific and orientation specific regulatory effects *in vitro*. By utilising luciferase assays they found that the *PARK7* SVA cloned into pGL3-P exhibited orientation specific functional activity in both SKNAS and MCF-7, with distinct regulatory effects in each line. When compared to empty pGL3-P, the SVA in the sense orientation induced no effect on reporter gene expression in SKNAS, but a significant increase in reporter gene activity was observed in MCF-7. In the reverse orientation, the SVA induced a significant decrease in reporter gene expression in both cell lines[223].

Savage *et al*. also tested an SVA found upstream of *FUS* and found that when it was cloned into pGL3-P it caused a significant decrease in reporter gene activity, with repeat length of the central VNTR modulating expression and the longest variant exhibiting the strongest repression[224]. Bragg *et al*. have also tested SVA function *in vitro* by cloning the *TAF1* SVA into pGL3-B. Although we cannot directly compare to our data due to them using a different vector, Bragg *et al*. did observe orientation specific expression profiles and showed that CT-element repeat number also modulated reporter gene activity. When tested in SH-SY5Y and compared to empty pGL3-B vector the *TAF1* SVA in the forward orientation caused a significant repression

of luciferase activity: this effect was strongest in the longest variant of the CT element and complete removal of the CT-element resulted in a moderate but statistically significant rescue in repression. Alternatively, in the reverse orientation the *TAF1* SVA facilitated an increase in reporter gene expression, but this effect was only significant in the 52 and 41 repeat CT-element SVAs, not the 35 repeat or CT-element deleted variant. In U2OS cells they observed a similar repression in the reverse orientation, but deletion of the CT-element resulted in an increase in reporter gene activity: an increase in luciferase expression was observed in all constructs in the reverse orientation[297]. It has been previously shown that intronic retrotransposons can cause transcriptional interference and formation of aberrant transcripts through several mechanisms, including blocking RNA polymerase II progression, inducing retention of introns, causing forced exonisation and cryptic polyadenylation and also facilitating premature transcript termination[408,420-422]. We hypothesise that the *NEK1* SVA could also induce transcriptional interference, with repetitive DNA within the CT-element and VNTR regions causing formation of alternative DNA secondary structures, such as G4 quadruplexes, which may impede progression of the RNA polymerase II (RNAP II) complex[297], thus inhibiting transcription and reducing pre-mRNA levels. Westenberger *et al*. have shown that repeat size of the CT-element within the *TAF1* SVA is associated with *TAF1* expression, with increased repeat size being significantly correlated with decreased *TAF1* expression[298]. Future work would include cloning multiple alleles of the *NEK1* SVA to test if repeat unit size affected reporter gene expression: In particular, it would be of interest to clone and test alleles 1 and 4 of the CT element which were found in ALS patients and thus confirm if these rare

variants posed a functional consequence and therefore could highlight a potential mechanism of *NEK1* dysfunction.

The intronic SVA of *TAF1* has now been shown to induce a reduction in *TAF1* expression, now postulated to be due to retention of intron 32[408]. As the SVA within *NEK1* is intronic, the pSHM06 vector was implemented to test if this SVA would have any effect on luciferase expression when present within an intron. When cloned in the sense orientation, the *NEK1* SVA facilitated a significant decrease in reporter gene expression when compared to empty pSHM06. A trend towards a decrease in reporter gene activity was observed when the SVA was present in the reverse orientation, but this decrease only reached statistical significance when tested in HEK293. Overall, the *NEK1* SVA exhibited a significant repressive effect on luciferase gene expression in the sense orientation, but the mechanism of action is unknown and should be investigated. As previously mentioned, Nott *et al*. discovered that addition of intron 6 from the *TPI* gene into pSHM06 vector led to an increase in luciferase gene expression and accumulation of TPI/Renilla mRNA. They hypothesise that this observation could be due to intron-containing transcripts having a higher processing efficiency, bypassing surveillance and subsequent degradation from nuclear exosomes[304]. There is growing evidence that some introns can regulate gene expression and increase accumulation of mRNA; a phenomenon referred to as intron mediated enhancement[423-427]. IME has been shown to boost rate of transcription, such as by inducing increased recruitment of RNA polymerase II[428], and involves numerous factors: intron orientation, presence of stimulating sequence elements and proximity to the transcription initiation site[426-430], thus indicating that it is a complex process and cannot be initiated by all introns[423,426-428]. Alternatively, intron

retention within mRNA can occur, leading to translation being abolished, increased mRNA instability and reduction in gene expression[408]. When compared to the empty pSHM06 vector, which contains intron 6 of *TPI* at the 5' end of Renilla luciferase (Figure 2.3), a significant repression in luciferase expression was observed in the constructs containing the NEK1 SVA in the sense orientation (cloned into intron 6) (Figure 5.4); this result could indicate that the enhancive effect on mRNA accumulation and luciferase expression previously observed by Nott et al., was abolished once the NEK1 SVA was introduced. It could be possible that the SVA is disrupting an IME-stimulating sequence within intron 6 of TPI, as the presence of particular sequence motifs has previously shown to be an important factor in initiating the mechanism of IME[426,429,430]. Furthermore, the significant repressive effect only occurred in the sense orientation (), supporting the notion that orientation of the intronic sequence is an important factor in driving IME[427]. As previously mentioned, the SVA could be facilitating transcriptional interference, with the potential for the SVA to form DNA secondary structure which could inhibit progression of RNAP II[297,422], or possibly recruiting transcription factors which could repress transcription[220,223]. Alternatively, the SVA could perhaps be having a post-transcriptional effect, altering efficiency of processes such pre-mRNA splicing, processing and export of pre-mRNA and cytoplasmic mRNA stability, all of which are previously postulated mechanisms of action for retrotransposons[188,281,431]; alteration of these processes could therefore reduce levels of mature mRNA available for translation. One key limitation in all reporter gene assays performed in this PhD is the lack of an additional negative control: an insert that is the same length as the VNTRs and SVAs being tested but is a different sequence composition. Future experiments

should include a construct containing an insert of identical size but of different sequence composition, to confirm that the inclusion of any stretch of DNA does not drive a significant change in luciferase expression in this model.

In this chapter we were also able to successfully remove the SVA from *NEK1* using the CRISPR Cas9 system and measure total *NEK1* and *CLCN3* gene expression in response to this modification (Figure 5.14 and Figure 5.16). As a lab we are the first to date (at the time of writing) to successfully remove reference genome SVAs and measure gene expression of the respective loci to test if such elements serve as cis regulatory domains *in vitro*. We were able to produce multiple heterozygous SVA knockouts (KOs) and three homozygous KOs. In the heterozygous KOs we observed a trend towards an increase (1.23 fold) in *NEK1* expression but this result did not reach significance (n = 3). Similarly, we observed a 1.69 fold increase in *NEK1* relative expression in response to removal of all copies of the SVA, but again this did not reach significance (n = 3). These results fit in with the response observed in our reporter gene data, showing that the SVA induced a decrease in reporter gene expression (Figure 5.2 and Figure 5.4), further supporting a role for SVAs as transcriptional repressors. Furthermore, we hypothesise that modest gene expression alterations observed in this study are not surprising, due to agreement with previous work by Rakovic *et al*. where the *TAF1* SVA knockout only resulted in modest rescue in *TAF1* expression (1.48 fold increase) in iPSCs. Although the result was statistically significant in iPSCs, only a mild trend towards an increase was found in spiny projection neurons and no clear trend was observed in cortical neurons[409]. From this result it is possible to infer that SVAs function in a cell specific manner, which has been previously supported by a study from Trizzino *et al*. in which they discovered

expression of SVA associated genes was significantly increased in adipose tissue, but significantly decreased in the liver[267].

One of the homozygous *NEK1* SVA KO lines generated here consisted of a larger than predicted deletion, resulting in partial cleavage of a neighbouring *Alu*Sq2 element (Figure 5.15). Plotting of individual values showed that this cell line drove the increase in *NEK1* gene expression observed via qPCR, suggesting this *Alu* might be functional at this locus. Previous studies have shown that *Alu* elements possess regulatory function, by modulating splicing efficiency[281], inducing alternative splicing, but also acting at the DNA level as TF binding sites and a source of DNA methylation[269,431]. However, the lack of observable histone marks and transcription factor binding over this *Alu* does not infer that this element is a functional regulatory region at the *NEK1* locus: further work would need to be conducted to conclude whether this element is of functional relevance, such as testing this *Alu* in luciferase reporter gene assays and excision of this element alone using CRISPR. Removal of the third homozygous SVA KO line from the qPCR experiment caused the increase in relative *NEK1* expression to drop to 1.10 fold, suggesting that the increase in *NEK1* expression observed in this particular cell line was only due to partial excision of the neighbouring *Alu* element and that the SVA alone had no effect on *NEK1* expression. Overall, removal of at least one copy of SVA (but not all copies) led to minor increases in *NEK1* gene expression, but these effects did not reach statistical significance: removal of all copies of the SVA did not facilitate any effect on *NEK1* expression (Figure 5.14 and Figure 5.16). Heterozygous KO of the SVA also induced a minor trend towards an increase in *CLCN3* expression, specifically a 1.24 fold increase: homozygous SVA KO did not induce any change in *CLCN3* expression. Overall, a small

increase in *CLCN3* expression was observed in the heterozygous SVA KO cell lines when compared to non-target controls, but did not reach statistical significance (Figure 5.14 and Figure 5.16). Although mild increases in both *NEK1* and *CLCN3* were observed in the heterozygous SVA KO lines (n=3) (1.23 and 1.24 fold respectively) when compared to non-target controls (n=3), this trend was not observed when compared against wildtype (unmodified) cells. Ultimately, it must be concluded that the removal of the SVA alone had no effect on *NEK1* or *CLCN3* expression in HEK293 cells.

The CRISPR project presented here presents a novel approach to study the functional capacity of SVA elements. Future work could include increasing the sample size of KO lines in this study: a total of 22 heterozygous KO lines were generated so the remaining 17 lines could be analysed and included in a future qPCR project. Previous work has shown that repetitive DNA can modulate gene expression in a stimulus inducible manner[146]. Future work may also include introducing a stimulus to the SVA KO cell lines, to determine if removal of the SVA could alter *NEK1* gene expression in response to the challenge. *NEK1* has previously been shown to be vital in response to genotoxic agents such as cisplatin and hydrogen peroxide[369,432]; Pelegrini *et al*. have shown that when compared to wild type cells *NEK1* deficient HEK293T cells exhibit reduced DNA repair and cell viability in response to cisplatin[432]. A further study by Melo-Hanchuk *et al*. has also demonstrated that *NEK1* deficient HEK293T cells display greatly impaired DNA repair capacity in response to cisplatin[369]. It would be of interest to see if knocking out the SVA alters how *NEK1* responds to a stimulus such as cisplatin, helping to better understand the potential regulatory mechanisms that this intronic SVA could influence at this region. Furthermore, we

only measured total *NEK1* gene expression, therefore qPCR primers for all 5 isoforms of *NEK1* should be designed to test if the SVA displays isoform specific regulation. Rakovic *et al*. measured TAF1 protein levels in response to SVA removal, but no significant difference in normalised levels of TAF1 protein between edited and unedited cells was observed[409]: measuring NEK1 protein levels in response to the SVA excision was not performed here and therefore would be on interest in the future. It has been previously shown that SVAs exhibit tissues-specific regulatory properties, acting as enhancers and repressors in distinct tissues[267]. As previously mentioned, Rakovic *et al*. observed cell-specific expression changes in response to the removal of the *TAF1* SVA: only a significant difference in *TAF1* gene expression was observed in iPSCs[409]. With this result in mind, it would be beneficial to generate *NEK1* SVA KOs in other cell lines, investigating whether excision of the SVA facilitates cell-specific gene expression profiles.

Ultimately, we have shown that the *NEK1* SVA is functional in two reporter gene assays by inducing a significance decrease in luciferase expression in both models, suggesting this element can act as a repressor *in vitro*. Future work should include testing multiple alleles of the SVA to confirm that repeat size of this element can modulate gene expression. Excision of this SVA using the CRISPR Cas9 system did not lead to any significance alterations in *NEK1* or *CLCN3* gene expression. With this result in mind it cannot be conclusively claimed that this element is a cis-regulatory element within the *NEK1/CLCN3* locus, but further work should be carried out which should include measuring *NEK1* isoform expression via qPCR and measuring protein levels via western blot.

Chapter 6: Thesis summary

Conclusions

The aim of this PhD thesis was to address non-coding repetitive DNA elements in the form of both VNTRs (static) and transposable (mobile) elements in recently discovered ALS risk loci: to characterise the genetic variation exhibited by these repetitive regions and to determine if it was possible to delineate ALS specific variants. This work has also addressed the functional capacity of these domains *in vitro*, to determine the potential role that both VNTRs and SVAs can play on transcription and gene regulation. Three main studies were performed: firstly focussing on two VNTRs at the *REST* and *CFAP410* loci, characterising the tandem repeat polymorphisms and testing the functional capacity of those domains. Secondly, identification of a reference SVA element within an intron of *NEK1*, specifically characterising genetic variation of this element in ALS patients and controls. Thirdly, investigation of the potential function of the *NEK1* SVA *in vitro* to confirm if it could act as a cis-regulatory domain.

To characterise genetic polymorphisms of VNTRs with high accuracy and reproducibility, optimised protocols for both gel agarose and gel capillary electrophoresis were set up and utilised. The VNTR found at the *REST* locus was a microsatellite, consisting of 3 bp tandem repeats, and thus agarose gel electrophoresis was not sufficient to accurately resolve each variant (Figure 3.2). To address this issue, the QIAxcel advanced system was implemented which utilises gel capillary electrophoresis, allowing one to resolve this region to 1 bp resolution and also facilitated high throughput screening of samples (Figure 3.3). Genotyping on the MNDA cohort led to the discovery of a 6-repeat variant in an ALS patient (Table 3.1),

a variant which had been previously been identified in an Alzheimer's patient[352]. To confirm this observation was correct the 6-repeat VNTR was sequenced and aligned to the 7, 9 and 12-repeat VNTRs (Figure 3.1). Following this, the functional capacity of the VNTR was tested by cloning the domain within pGL3-B and testing it in a reporter gene assay, concluding that the *REST* VNTR could drive reporter gene expression in an allele dependent manner in SH-SY5Y cells, highlighting the potential for the VNTR to act as a promoter at this locus. Ultimately, we present an optimised and high-throughput protocol for genotyping the *REST* VNTR with high precision, while also displaying that the VNTR can drive reporter gene expression and repeat length can modulate expression profiles *in vitro*.

In Chapter 3 we also discovered an intronic VNTR within the *CFAP410* gene, a locus which has recently become associated with ALS risk[11]. This VNTR was a human specific minisatellite, constituting repeats of 22 and 35 bp, with a total of 7 alleles identified in this study. Due to the length of this VNTR it could not be resolved on the QIAxcel advanced system with high resolution: the largest variant (allele 7) was not detected by this system as it fell outside of the detection range for the alignment markers. To resolve this, we assessed the *CFAP410* VNTR using agarose gels, allowing us to distinguish each variant but at the cost of high resolution, making accurate sizing of each variant difficult (Figure 3.8). To address this problem, subcloning and subsequent sequencing of each VNTR was attempted. Unfortunately, due to the difficulties of cloning highly repetitive and GC-rich DNA only variants 2, 4, 5 and 7 were successfully cloned (Figure 3.10). Two alleles, variants 2 and 7, were found to be unique to ALS patients in this study, however the assessed cohort was small: determining if these variants are associated with ALS would require a larger cohort

and due to the complexity and size of this VNTR the genotyping would need to be performed by gel electrophoresis.

In an attempt to expand the cohort size other methods of genotyping the *CFAP410* VNTR were investigated, to take advantage of the large volume of WGS data available. Using Isaac Variant Caller (IVC), only the two common variants (alleles 4 and 5) were identified: when compared to gel electrophoresis data only a 46.8% agreement between methodologies was observed (n = 235), indicating the drawback and inaccuracy of IVC when trying to characterise larger and multiple imperfect tandem repeats. This is probably due to IVC only being qualified to accurately genotype small indels and SNPs, furthermore this variant caller has a cut-off of 50 bp and therefore variants 1 and 7 would not be called. As variant 7 missed the threshold for IVC we decided to check this region with Manta Structural Variant Caller, a tool designed to discover and score structural variants using paired and split read sequencing data[433], but this indel was not present in the VCF file of the assessed ALS patient (n = 1). In conclusion, due to the complexity (built of two different repeat units which are GC-rich) and size of the *CFAP410* VNTR, currently available bioinformatic pipelines are not suitable or optimal for genotyping this region and therefore PCR followed by gel electrophoresis must be performed. This process is time consuming but once optimised is efficient and consistent and can be validated through cloning and sequencing of the variants, as shown in this chapter.

The functionality of the *CFAP410* VNTR was also investigated in Chapter 3, as we were able to successfully clone this VNTR into both pGL3-P and pGL3-B (Figure 3.12). When the commonest variants were tested in the forward (endogenous)

orientation within pGL3-P we observed variant specific profiles: compared to the empty vector, allele 5 led to a 1.71 fold increase in luciferase activity whereas allele 4 induced a 2.64 fold increase. An orientation specific effect was also observed: in the reverse orientation, allele 5 drove a 1.2 fold decrease in reporter gene expression, but no significant difference in luciferase activity was observed when compared to the empty vector (Figure 3.13). When tested within pGL3-B allele 5 in the endogenous orientation only drove a 7.38 fold increase in reporter gene activity: when present in the reverse orientation a 71.48 fold increase in activity was observed, the latter of which was 2.4 fold higher than the activity of pGL3-P (containing an SV40 promoter) (Figure 3.14). From assessment on UCSC were we able to intersect the VNTR with transcript data from Ensembl, which led to the identification of a shorter isoform starting downstream of the VNTR (Figure 3.11). Using GTEx portal we were able to assess expression of each isoform of *CFAP410* in multiple tissues and found that the shorter isoform is ubiquitously expressed across all tissues. Based on our VNTR-pGL3-B luciferase data we hypothesise that the *CFAP410* VNTR could drive transcription and function as a promoter for this shorter transcript, but future work would need to confirm this.

A human specific SVA retrotransposon was discovered in the *NEK1* locus, which was investigated using both PCR and genetic variant caller software to define polymorphisms which were hypothesised could be associated with ALS risk. Initial analysis of this SVA concluded that it was a highly polymorphic region, containing multiple variants of the 5' CT-element, central VNTR and 3' Poly A tail (Figure 4.3). Contemporaneously with this work Bragg *et al*. have previously investigated an SVA insertion found within *TAF1* which is associated with XDP, discovering that CT-

element repeat size was inversely correlated with age of onset of disease[297]. Genotyping analysis within an MNDA cohort led to the discovery two ALS specific variants of the CT-element (allele 1 and 4), the first ALS specific variants found within an SVA retrotransposon to date. This investigation was then extended into the UK dataset of Project MinE, analysing whole genome sequencing data (WGS) using Isaac Variant Caller (IVC). This high-throughput approach was initially applied to cases which had already been genotyped to determine if it was possible to accurately detect the previously identified rare polymorphisms found within the CT-element of the SVA. We were able to validate the IVC methodology to distinguish between the common variants and the rare ALS-specific variants of the *NEK1* SVA CT-element: this was the first time this variant caller pipeline has been used to call retrotransposon polymorphisms to date, with the successful identification of variants which were found to be significantly associated with ALS in the UK cohort of Project MinE.

Several studies have now focussed on the functional capacity of SVA elements, concluding that they can act as transcriptional regulatory domains both *in vitro* and *in vivo*[223,224,297,298,408,409]. In Chapter 5 we investigated the potential functionality of the SVA found within *NEK1*, constituting two key *in vitro* approaches: reporter gene assays and genetic modification using CRISPR. We were able to successfully clone the *NEK1* SVA into two reporter gene constructs, pGL3P and pSHM06. When tested in pGL3P the *NEK1* SVA induced a significant repression on luciferase activity in several cell lines (Figure 5.2). Similarly, in pSHM06 the SVA facilitated a significant reduction in reporter gene expression in several cell lines, but only in the sense (forward) orientation (Figure 5.4). It was concluded that the SVA could act as a repressor *in vitro*, which could be due to the repetitive nature of the

element facilitating formation of alternative DNA structures which may impede transcription: this mechanism has been postulated for the *TAF1* SVA, with the potential for the CT-element to form G-quadruplexes (G4) which could stall RNA polymerase II progression and thus repress transcription[297]. SVAs also harbour binding sites for transcription factors (such as SP1 and CTCF)[219,225,276], which could induce repression of transcription; thus there is the potential for polymorphic repeat size within the SVA to affect binding affinities of TFs. The *NEK1* SVA may also induce intron retention in a similar manner to the *TAF1* SVA[408], or alter pre-mRNA splicing efficiency and thus modulate mRNA accumulation, but further experimentation would be required to confirm this.

We were able to successfully remove the SVA element within *NEK1* in a cell line and measure gene expression via qPCR in response to this modification. This work (at the time of writing) is amongst the first reporting removal of reference SVAs from the human genome to test if they can act as cis regulatory elements and present an optimised pipeline for this process, with a heterozygous KO efficiency of 10% and a homozygous KO efficiency of 1%. Excision did not lead to any significant changes in *NEK1* or *CLCN3* gene expression, however modest trends towards an increase in expression were observed. Based on these results we cannot definitively state that the SVA is acting as a cis regulatory element in this model, but other parameters including specific isoform expression and protein levels were not measured in response to removal of the SVA and must be investigated in the future. Previous work by Zabolotneva *et al*. has shown that CpG methylation found within SVAF1s is important in the regulation of transcriptional activation[218], so epigenetic parameters

such as DNA methylation of the *NEK1* SVA and the surrounding locus need to be taken into consideration and should be assessed in further experiments.

This work led to the discovery of multiple rare VNTR and SVA variants which were only present in ALS patients, highlighting that rare non-coding variation in repetitive and mobile regions could be a source of missing heritability in ALS. The projects presented here have also generated a number of resources, including optimised PCR protocols for GC-rich VNTRs and SVAs elements, a protocol for high-throughput genotyping of microsatellite VNTRs, an optimised IVC pipeline to assess SVA CT-element variation, VNTR and SVA reporter gene constructs, SVA KO cell lines and optimised qPCR protocols to measure relative expression of *NEK1* and *CLCN3*. With these resources now available a number of studies should be implemented to compliment and reinforce the data generated in this PhD project.

Ongoing work and future projects

REST VNTR

With the identification of a rare variant (6-copy number repeat) in ALS, the aim was to expand the genotyping analysis of the *REST* VNTR. Due to short repeat size and simplicity of the repeats it has been possible to accurately genotype the *REST* VNTR using Isaac Variant Caller, allowing us to expand this analysis into the UK dataset of Project MinE.

Another method of interest is haplotyping this region using SNP data. It has been possible to genotype the *REST* VNTR within the North American Brain Expression Cohort (NABEC): containing genome-wide genotyping and RNA-seq data of human cerebral frontal cortex from neurologically healthy individuals (available

here: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001300.v1.p1). With genotyping data available, it would be possible to generate tagging SNPs for the *REST* VNTR: SNPs which are inherited (in linkage) with a particular allele of the VNTR[138]. Generating tagging SNPs for each variant of the VNTR (identifying a distinct SNP for each variant) will remove the requirement of directly genotyping the VNTR and therefore save time and resource. It would then also be possible to correlate the SNP genotyping data with the available RNAseq data, investigating whether certain VNTR variants are associated with distinct *REST* gene expression profiles.

CFAP410 VNTR

With the identification of two rare variants in ALS patients, the aim would be to expand the *CFAP410* VNTR genotyping into a larger sample size. Unfortunately, due to the complexity and size of these tandem repeats, genotyping with IVC was found to be inaccurate and is therefore not recommended for this region. Alternatively, we would attempt to generate tagging SNPs for each VNTR variant and then correlate the genotyping with RNA-Seq data to test for any association between VNTR polymorphisms and expression profiles at this locus. Further functional work should also be carried out, including excision of the VNTR using the CRISPR Cas9 system and measurement of expression of the short isoform (ENST00000462742) downstream of the VNTR to conclude whether this region can function as a promoter. Variants 2 and 7, which were ALS specific in this study, should also be tested in a reporter gene assay to confirm if they can either drive transcription (in pGL3-B) or regulate transcription (in pGL3-P).

NEK1 SVA

Collaboration with Kings College London has facilitated the set up and optimisation of a high-throughput pipeline to genotype the rare variants of the *NEK1* SVA CT element in WGS data using IVC; the next step is to genotype more samples in Project MinE once they become available. In particular, it would be of interest to increase the sample size of the UK cohort to see if the association between the rare variants and ALS can be replicated. To compliment this work, variants 1 and 4 of the CT element must be tested functionally, including generating reporter gene constructs containing these variants and testing them within a luciferase assay. Future work could also include generating lymphoblastoid cell lines from ALS patients with the rare CT element variants, then removing the SVA using CRISPR and measuring *NEK1* expression in response to the modification. Alternatively, one could generate patient derived iPSCs then differentiate these into motor neurons; knocking out the SVA in both the iPSCs and motor neurons would be of interest, to determine if the SVA is active in certain cell types as previously shown in other studies[408,409]. It is also important to mention that several other retrotransposons have been identified in the *NEK1* locus, specifically 5 retrotransposon insertion polymorphisms (RIPs) (Supplementary Figure 4). We have begun to validate these RIPs (Supplementary Figure 6) and this work will be continued in the future; it would be of importance to determine if the ALS patients harbouring the rare *NEK1* CT element variants also contained any of these RIPs, to determine if they are genetically distinct to the healthy individuals who also contained the rare CT element variants.

The ultimate goal of this work was to identify novel ALS risk variants in non-coding regions and to better understand the potential function that mobile and

repetitive DNA may possess at ALS risk loci. We hope that the work presented here will raise the profile of both VNTRs and non-LTR retrotransposons in ALS, helping to better understand the role they could play in gene regulation and also the potential for them to be a missing source of heritability in ALS.

Bibliography

1 Hardiman, O. *et al.* Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers* **3**, 1-19, doi:doi:10.1038/nrdp.2017.71 (2017).

2 *Amyotrophic Lateral Sclerosis (ALS) Fact Sheet | National Institute of Neurological Disorders and Stroke*, <https://www.ncbi.nlm.nih.gov/pubmed/> (2013).

3 LI, G., GA, R., J, R. & NR, C. Clinical Spectrum of Amyotrophic Lateral Sclerosis (ALS). *Cold Spring Harbor perspectives in medicine* **7**, doi:10.1101/cshperspect.a024117 (2017).

4 Ajroud-Driss, S. & Siddique, T. Sporadic and hereditary amyotrophic lateral sclerosis (ALS). *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1852**, 679-684, doi:https://doi.org/10.1016/j.bbadis.2014.08.010 (2015).

5 Nguyen, H. P., Van Broeckhoven, C. & van der Zee, J. ALS Genes in the Genomic Era and their Implications for FTD. *Trends Genet* **34**, 404-423, doi:10.1016/j.tig.2018.03.001 (2018).

6 Ranganathan, R. *et al.* Multifaceted Genes in Amyotrophic Lateral Sclerosis-Frontotemporal Dementia. *Frontiers in Neuroscience* **14**, 21, doi:10.3389/fnins.2020.00684 (2020).

7 Longinetti, E. & Fang, F. Epidemiology of amyotrophic lateral sclerosis: an update of recent literature. *Curr Opin Neurol* **32**, 771-776, doi:10.1097/wco.0000000000000730 (2019).

8 Taylor, J. P., Brown Jr, R. H. & Cleveland, D. W. Decoding ALS: from genes to mechanism. *Nature* **539**, 197-206, doi:10.1038/nature20413 (2016).

9 Pochet, R. Genetics and ALS: Cause for Optimism. *Cerebrum* **2017** (2017).

10 Mejzini, R. *et al.* ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Frontiers in Neuroscience* **13**, doi:10.3389/fnins.2019.01310 (2019).

11 van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* **48**, 1043-1048, doi:10.1038/ng.3622 http://www.nature.com/ng/journal/v48/n9/abs/ng.3622.html#supplementary-information (2016).

12 Rosen, D. R. *et al.* Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59-62, doi:doi:10.1038/362059a0 (1993).

13 Pansarasa, O. *et al.* in *Int J Mol Sci* Vol. 19 (2018).

14 Bunton-Stasyshyn, R. K., Saccon, R. A., Fratta, P. & Fisher, E. M. SOD1 Function and Its Implications for Amyotrophic Lateral Sclerosis Pathology: New and Renascent Themes. *Neuroscientist* **21**, 519-529, doi:10.1177/1073858414561795 (2015).

15 Zou, Z.-Y. *et al.* Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. doi:10.1136/jnnp-2016-315018 (2017).

16 Neumann, M. *et al.* Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. doi:10.1126/science.1134108 (2006).

17 Arai, T. *et al.* TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochem Biophys Res Commun* **351**, 602-611, doi:10.1016/j.bbrc.2006.10.093 (2006).

18 Prasad, A., Bharathi, V., Sivalingam, V., Girdhar, A. & Patel, B. K. Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis. *Front Mol Neurosci* **12**, doi:10.3389/fnmol.2019.00025 (2019).

19 Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123-133, doi:10.1016/j.cell.2015.09.015 (2015).
20 Kabashi, E. *et al.* TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nature Genetics* **40**, 572-574, doi:doi:10.1038/ng.132 (2008).
21 Sprovieri, T. *et al.* A novel S379A TARDBP mutation associated to late-onset sporadic ALS. *Neurological Sciences* **40**, 2111-2118, doi:doi:10.1007/s10072-019-03943-y (2019).
22 Tan, A. Y., Riley, T. R., Coady, T., Bussemaker, H. J. & Manley, J. L. TLS/FUS (translocated in liposarcoma/fused in sarcoma) regulates target gene transcription via single-stranded DNA response elements. *Proc Natl Acad Sci U S A* **109**, 6030-6035, doi:10.1073/pnas.1203028109 (2012).
23 Picchiarelli, G. *et al.* FUS-mediated regulation of acetylcholine receptor transcription at neuromuscular junctions is compromised in amyotrophic lateral sclerosis. *Nat Neurosci* **22**, 1793-1805, doi:10.1038/s41593-019-0498-9 (2019).
24 Wang, W. Y. *et al.* Interaction of FUS and HDAC1 Regulates DNA Damage Response and Repair in Neurons. *Nat Neurosci* **16**, 1383-1391, doi:10.1038/nn.3514 (2013).
25 Mastrocola, A. S., Kim, S. H., Trinh, A. T., Rodenkirch, L. A. & Tibbetts, R. S. The RNA-binding Protein Fused in Sarcoma (FUS) Functions Downstream of Poly(ADP-ribose) Polymerase (PARP) in Response to DNA Damage*. *J Biol Chem* **288**, 24731-24741, doi:10.1074/jbc.M113.497974 (2013).
26 Ratti, A. & Buratti, E. Physiological functions and pathobiology of TDP-43 and FUS/TLS proteins. *J Neurochem* **138 Suppl 1**, 95-111, doi:10.1111/jnc.13625 (2016).
27 Lagier-Tourenne, C. *et al.* Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci* **15**, 1488-1497, doi:10.1038/nn.3230 (2012).
28 Ito, D., Taguchi, R., Deguchi, M., Ogasawara, H. & Inoue, E. Extensive splicing changes in an ALS/FTD transgenic mouse model overexpressing cytoplasmic fused in sarcoma. *Sci Rep* **10**, doi:10.1038/s41598-020-61676-x (2020).
29 Kamelgarn, M. *et al.* ALS mutations of FUS suppress protein translation and disrupt the regulation of nonsense-mediated decay. doi:10.1073/pnas.1810413115 (2018).
30 J, L.-E. *et al.* ALS/FTD-Linked Mutation in FUS Suppresses Intra-axonal Protein Synthesis and Drives Disease Without Nuclear Loss-of-Function of FUS. *Neuron* **100**, doi:10.1016/j.neuron.2018.09.044 (2018).
31 Udagawa, T. *et al.* FUS regulates AMPA receptor function and FTLD/ALS-associated behaviour via GluA1 mRNA stabilization. *Nature Communications* **6**, 1-13, doi:doi:10.1038/ncomms8098 (2015).
32 Nolan, M., Talbot, K. & Ansorge, O. Pathogenesis of FUS-associated ALS and FTD: insights from rodent models. *Acta Neuropathologica Communications* **4**, 1-13, doi:doi:10.1186/s40478-016-0358-8 (2016).
33 Kwiatkowski, T. J., Jr. *et al.* Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **323**, 1205-1208, doi:10.1126/science.1166066 (2009).
34 Vance, C. *et al.* Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. *Science* **323**, 1208-1211, doi:10.1126/science.1165942 (2009).
35 Belzil, V. V. *et al.* Mutations in FUS cause FALS and SALS in French and French Canadian populations. *Neurology* **73**, 1176-1179, doi:10.1212/WNL.0b013e3181bbfeef (2009).

36 Rademakers, R. *et al.* FUS GENE MUTATIONS IN FAMILIAL AND SPORADIC AMYOTROPHIC LATERAL SCLEROSIS. *Muscle Nerve* **42**, 170-176, doi:10.1002/mus.21665 (2010).
37 Deng, H., Gao, K. & Jankovic, J. The role of FUS gene variants in neurodegenerative diseases. *Nat Rev Neurol* **10**, 337-348, doi:10.1038/nrneurol.2014.78 (2014).
38 Morita, M. *et al.* A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology* **66**, 839-844, doi:10.1212/01.wnl.0000200048.53766.b4 (2006).
39 Vance, C. *et al.* Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. *Brain* **129**, 868-876, doi:10.1093/brain/awl030 (2006).
40 Shatunov, A. *et al.* Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *The Lancet Neurology* **9**, 986-994, doi:10.1016/S1474-4422(10)70197-6 (2010).
41 DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron* **72**, 245-256, doi:10.1016/j.neuron.2011.09.011 (2011).
42 Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257-268, doi:10.1016/j.neuron.2011.09.010 (2011).
43 Shi, Y. *et al.* Haploinsufficiency leads to neurodegeneration in C9ORF72 ALS/FTD human induced motor neurons. *Nature Medicine* **24**, 313, doi:10.1038/nm.4490 https://www.nature.com/articles/nm.4490#supplementary-information (2018).
44 Farg, M. A. *et al.* C9ORF72, implicated in amytrophic lateral sclerosis and frontotemporal dementia, regulates endosomal trafficking. *Hum Mol Genet* **23**, 3579-3595, doi:10.1093/hmg/ddu068 (2014).
45 Amick, J., Roczniak-Ferguson, A., Ferguson, S. M. & Brill, J. C9orf72 binds SMCR8, localizes to lysosomes, and regulates mTORC1 signaling. *https://doi.org/10.1091/mbc.e16-01-0003*, doi:10.1091/mbc.e16-01-0003 (2016).
46 Nassif, M., Woehlbier, U. & Manque, P. A. The Enigmatic Role of C9ORF72 in Autophagy. *Front Neurosci* **11**, doi:10.3389/fnins.2017.00442 (2017).
47 Babić Leko, M. *et al.* Molecular Mechanisms of Neurodegeneration Related to C9orf72 Hexanucleotide Repeat Expansion. *Behav Neurol* **2019**, doi:10.1155/2019/2909168 (2019).
48 Iacoangeli, A. *et al.* C9orf72 intermediate expansions of 24–30 repeats are associated with ALS. *Acta Neuropathologica Communications* **7**, 1-7, doi:doi:10.1186/s40478-019-0724-4 (2019).
49 Veldink, J. H. ALS genetic epidemiology 'How simplex is the genetic epidemiology of ALS?'. doi:10.1136/jnnp-2016-315469 (2017).
50 Wroe, R., Butler, A. W.-L., Andersen, P. M., Powell, J. F. & Al-Chalabi, A. ALSOD: The Amyotrophic Lateral Sclerosis Online Database. *http://dx.doi.org/10.1080/17482960802146106*, doi:10.1080/17482960802146106 (2009).
51 Theunissen, F. *et al.* Structural Variants May Be a Source of Missing Heritability in sALS. *Front Neurosci* **14**, doi:10.3389/fnins.2020.00047 (2020).
52 A, S. & A, A.-C. The genetic architecture of ALS. *Neurobiology of disease* **147**, doi:10.1016/j.nbd.2020.105156 (2021).
53 M, G. & RH, B. Genetics of Amyotrophic Lateral Sclerosis. *Cold Spring Harbor perspectives in medicine* **8**, doi:10.1101/cshperspect.a024125 (2018).

54 S, B. *et al.* A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *Brain : a journal of neurology* **137**, doi:10.1093/brain/awu138 (2014).
55 Gitcho, M. A. *et al.* TDP-43 A315T mutation in familial motor neuron disease. *Ann Neurol* **63**, 535-538, doi:10.1002/ana.21344 (2008).
56 JO, J. *et al.* Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis. *Nature neuroscience* **17**, doi:10.1038/nn.3688 (2014).
57 MJ, G. *et al.* ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nature genetics* **38**, doi:10.1038/ng1742 (2006).
58 N, T. *et al.* Mutational analysis reveals the FUS homolog TAF15 as a candidate gene for familial amyotrophic lateral sclerosis. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **156B**, doi:10.1002/ajmg.b.31158 (2011).
59 IR, M. *et al.* TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. *Neuron* **95**, doi:10.1016/j.neuron.2017.07.025 (2017).
60 HJ, K. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature* **495**, doi:10.1038/nature11922 (2013).
61 J, C. *et al.* Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis. *Human molecular genetics* **21**, doi:10.1093/hmg/dds116 (2012).
62 Cirulli, E. T. *et al.* Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**, 1436-1441, doi:10.1126/science.aaa3650 (2015).
63 Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet* **48**, 1037-1042, doi:10.1038/ng.3626 (2016).
64 Brenner, D. *et al.* NEK1 mutations in familial amyotrophic lateral sclerosis. *Brain* **139**, e28-e28, doi:10.1093/brain/aww033 (2016).
65 Nguyen, H. P. *et al.* NEK1 genetic variability in a Belgian cohort of ALS and ALS-FTD patients. *Neurobiol Aging* **61**, 255.e251-255.e257, doi:10.1016/j.neurobiolaging.2017.08.021 (2018).
66 A, O. *et al.* SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain : a journal of neurology* **133**, doi:10.1093/brain/awp325 (2010).
67 CL, S. *et al.* Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Human molecular genetics* **18**, doi:10.1093/hmg/ddn375 (2009).
68 P, C. *et al.* Abnormal SMN1 gene copy number is a susceptibility factor for amyotrophic lateral sclerosis. *Annals of neurology* **51**, doi:10.1002/ana.10104 (2002).
69 YZ, C. *et al.* DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *American journal of human genetics* **74**, doi:10.1086/421054 (2004).
70 JO, J. *et al.* Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* **68**, doi:10.1016/j.neuron.2010.11.036 (2010).
71 AL, N. *et al.* A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *American journal of human genetics* **75**, doi:10.1086/425287 (2004).
72 H, M. *et al.* Mutations of optineurin in amyotrophic lateral sclerosis. *Nature* **465**, doi:10.1038/nature08971 (2010).
73 KL, W. *et al.* CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. *Nature communications* **7**, doi:10.1038/ncomms11253 (2016).
74 F, F. *et al.* SQSTM1 mutations in familial and sporadic amyotrophic lateral sclerosis. *Archives of neurology* **68**, doi:10.1001/archneurol.2011.250 (2011).

75 MJ, G. *et al.* A novel candidate region for ALS on chromosome 14q11.2. *Neurology* **63**, doi:10.1212/01.wnl.0000144344.39103.f6 (2004).
76 A, N. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, doi:10.1016/j.neuron.2018.02.027 (2018).
77 C, M. *et al.* Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. *Neurology* **63**, doi:10.1212/01.wnl.0000134608.83927.b1 (2004).
78 BN, S. *et al.* Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron* **84**, doi:10.1016/j.neuron.2014.09.027 (2014).
79 CH, W. *et al.* Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature* **488**, doi:10.1038/nature11280 (2012).
80 BN, S. *et al.* Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. *Science translational medicine* **9**, doi:10.1126/scitranslmed.aad9157 (2017).
81 CL, L. *et al.* A pathogenic peripherin gene mutation in a patient with amyotrophic lateral sclerosis. *Brain pathology (Zurich, Switzerland)* **14**, doi:10.1111/j.1750-3639.2004.tb00066.x (2004).
82 DA, F., GA, R., A, K. & JP, J. Polymorphism in the multi-phosphorylation domain of the human neurofilament heavy-subunit-encoding gene. *Gene* **132**, doi:10.1016/0378-1119(93)90211-k (1993).
83 CY, C. *et al.* Deleterious variants of FIG4, a phosphoinositide phosphatase, in patients with ALS. *American journal of human genetics* **84**, doi:10.1016/j.ajhg.2008.12.010 (2009).
84 JE, L. *et al.* Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proceedings of the National Academy of Sciences of the United States of America* **106**, doi:10.1073/pnas.0812937106 (2009).
85 S, H. *et al.* A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nature genetics* **29**, doi:10.1038/ng1001-166 (2001).
86 Lattante, S. *et al.* ATXN1 intermediate-length polyglutamine expansions are associated with amyotrophic lateral sclerosis. *Neurobiology of Aging* **64**, 157.e151-157.e155, doi:10.1016/j.neurobiolaging.2017.11.011 (2018).
87 AC, E. *et al.* Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature* **466**, doi:10.1038/nature09320 (2010).
88 A, A.-S., F, A.-M. & S, B. A mutation in sigma-1 receptor causes juvenile amyotrophic lateral sclerosis. *Annals of neurology* **70**, doi:10.1002/ana.22534 (2011).
89 MA, v. E. *et al.* Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nature genetics* **41**, doi:10.1038/ng.442 (2009).
90 JC, S. *et al.* Progranulin mutations and amyotrophic lateral sclerosis or amyotrophic lateral sclerosis-frontotemporal dementia phenotypes. *Journal of neurology, neurosurgery, and psychiatry* **78**, doi:10.1136/jnnp.2006.109553 (2007).
91 N, P. *et al.* ALS phenotypes with mutations in CHMP2B (charged multivesicular body protein 2B). *Neurology* **67**, doi:10.1212/01.wnl.0000231510.89311.8b (2006).
92 MA, v. E. *et al.* Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nature genetics* **40**, doi:10.1038/ng.2007.52 (2008).
93 J, M. *et al.* Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. *Proceedings of the National Academy of Sciences of the United States of America* **107**, doi:10.1073/pnas.0914128107 (2010).
94 D, L. *et al.* VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nature genetics* **34**, doi:10.1038/ng1211 (2003).

95 MA, v. E. *et al.* ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *The Lancet. Neurology* **6**, doi:10.1016/S1474-4422(07)70222-3 (2007).
96 CA, D. *et al.* Detection of a novel frameshift mutation and regions with homozygosis within ARHGEF28 gene in familial amyotrophic lateral sclerosis. *Amyotrophic lateral sclerosis & frontotemporal degeneration* **14**, doi:10.3109/21678421.2012.758288 (2013).
97 A, S. *et al.* Paraoxonase gene polymorphisms and sporadic ALS. *Neurology* **67**, doi:10.1212/01.wnl.0000219565.32247.11 (2006).
98 XS, W. *et al.* Increased incidence of the Hfe mutation in amyotrophic lateral sclerosis and related cellular consequences. *Journal of the neurological sciences* **227**, doi:10.1016/j.jns.2004.08.003 (2004).
99 SMK, F. *et al.* Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nature neuroscience* **22**, doi:10.1038/s41593-019-0530-0 (2019).
100 J, C.-K. *et al.* Mutations in the Glycosyltransferase Domain of GLT8D1 Are Associated with Familial Amyotrophic Lateral Sclerosis. *Cell reports* **26**, doi:10.1016/j.celrep.2019.02.006 (2019).
101 Blauw, H. M. *et al.* NIPA1 polyalanine repeat expansions are associated with amyotrophic lateral sclerosis. *Hum Mol Genet* **21**, 2497-2502, doi:10.1093/hmg/dds064 (2012).
102 Polubriaginof, F. C. G. *et al.* Disease heritability inferred from familial relationships reported in medical records. *Cell* **173**, 1692-1704 e1611, doi:10.1016/j.cell.2018.04.032 (2018).
103 McLaughlin, R. L. *et al.* Heritability of Amyotrophic Lateral Sclerosis: Insights From Disparate Numbers. *JAMA Neurology* **72**, 857-858, doi:10.1001/jamaneurol.2014.4049 (2015).
104 Ryan, M. *et al.* Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurology* **76**, 1367-1374, doi:10.1001/jamaneurol.2019.2044 (2019).
105 Al-Chalabi, A. *et al.* An estimate of amyotrophic lateral sclerosis heritability using twin data. *J Neurol Neurosurg Psychiatry* **81**, 1324-1326, doi:10.1136/jnnp.2010.207464 (2010).
106 Reference, G. H. *What is heritability?*, <https://www.ncbi.nlm.nih.gov/pubmed/> (2020).
107 Yu, B. & Pamphlett, R. Environmental insults: critical triggers for amyotrophic lateral sclerosis. *Translational Neurodegeneration* **6**, 1-10, doi:doi:10.1186/s40035-017-0087-3 (2017).
108 Niccoli, T. *et al.* Ageing as a risk factor for ALS/FTD. *Human Molecular Genetics* **26**, doi:10.1093/hmg/ddx247 (2017).
109 Chiò, A. *et al.* ALS phenotype is influenced by age, sex, and genetics. doi:10.1212/WNL.0000000000008869 (2020).
110 FC, G., BB, T., NR, W. & E, A. Cardiovascular disease, psychiatric diagnosis and sex differences in the multistep hypothesis of amyotrophic lateral sclerosis. *European journal of neurology*, doi:10.1111/ene.14554 (2020).
111 AM, M., A, B., R, B., A, Y. & EO, T. Pesticide exposure as a risk factor for amyotrophic lateral sclerosis: a meta-analysis of epidemiological studies: pesticide exposure as a risk factor for ALS. *Environmental research* **117**, doi:10.1016/j.envres.2012.06.007 (2012).
112 MD, W., J, G., NR, C., J, L. & D, K. A meta-analysis of observational studies of the association between chronic occupational exposure to lead and amyotrophic lateral sclerosis. *Journal of occupational and environmental medicine* **56**, doi:10.1097/JOM.0000000000000323 (2014).

113 Al-Chalabi, A. & Visscher, P. M. Common genetic variants and the heritability of ALS. *Nature Reviews Neurology* **10**, 549-550, doi:doi:10.1038/nrneurol.2014.166 (2014).
114 A, A.-C. *et al.* Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *The Lancet. Neurology* **13**, doi:10.1016/S1474-4422(14)70219-4 (2014).
115 Al-Chalabi, A., Berg, L. H. v. d. & Veldink, J. Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. *Nature Reviews Neurology* **13**, 96-104, doi:doi:10.1038/nrneurol.2016.182 (2016).
116 A, C. *et al.* The multistep hypothesis of ALS revisited: The role of genetic mutations. *Neurology* **91**, doi:10.1212/WNL.0000000000005996 (2018).
117 Felbecker, A. *et al.* Four familial ALS pedigrees discordant for two SOD1 mutations: are all SOD1 mutations pathogenic? *J Neurol Neurosurg Psychiatry* **81**, 572-577, doi:10.1136/jnnp.2009.192310 (2010).
118 van Blitterswijk, M. *et al.* Evidence for an oligogenic basis of amyotrophic lateral sclerosis. *Hum Mol Genet* **21**, 3776-3784, doi:10.1093/hmg/dds199 (2012).
119 van Blitterswijk, M. *et al.* VAPB and C9orf72 mutations in 1 familial amyotrophic lateral sclerosis patient. *Neurobiol Aging* **33**, 2950.e2951-2954, doi:10.1016/j.neurobiolaging.2012.07.004 (2012).
120 Cady, J. *et al.* Amyotrophic lateral sclerosis onset is influenced by the burden of rare variants in known amyotrophic lateral sclerosis genes. *Ann Neurol* **77**, 100-113, doi:10.1002/ana.24306 (2015).
121 Bury, J. J. *et al.* Oligogenic inheritance of optineurin (OPTN) and C9ORF72 mutations in ALS highlights localisation of OPTN in the TDP-43-negative inclusions of C9ORF72-ALS. *Neuropathology* **36**, 125-134, doi:10.1111/neup.12240 (2016).
122 O, G. *et al.* Rare homozygosity in amyotrophic lateral sclerosis suggests the contribution of recessive variants to disease genetics. *Journal of the neurological sciences* **402**, doi:10.1016/j.jns.2019.05.006 (2019).
123 Beck, J. *et al.* in *Am J Hum Genet* Vol. 92 345-353 (2013).
124 Brown, R. H., Al-Chalabi, A. & Longo, D. L. Amyotrophic Lateral Sclerosis. *http://dx.doi.org.liverpool.idm.oclc.org/10.1056/NEJMra1603471*, doi:NJ201707133770211 (2017).
125 Sproviero, W. *et al.* ATXN2 trinucleotide repeat length correlates with risk of ALS. *Neurobiol Aging* **51**, 178 e171-179, doi:10.1016/j.neurobiolaging.2016.11.010 (2017).
126 Fournier, C. *et al.* Interrupted CAG expansions in ATXN2 gene expand the genetic spectrum of frontotemporal dementias. *Acta Neuropathologica Communications* **6**, 1-4, doi:doi:10.1186/s40478-018-0547-8 (2018).
127 Tazelaar, G. H. P. *et al.* ATXN1 repeat expansions confer risk for amyotrophic lateral sclerosis and contribute to TDP-43 mislocalization. *Brain Communications*, doi:10.1093/braincomms/fcaa064 (2020).
128 Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur J Hum Genet* **26**, 1537-1546, doi:10.1038/s41431-018-0177-4 (2018).
129 *Working Groups - Project MinE*, <https://www.projectmine.com/research/working-groups/> (2020).
130 TJ, T. & SL, S. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics* **13**, doi:10.1038/nrg3117 (2011).
131 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 Genetics* **7**, e1002384, doi:10.1371/journal.pgen.1002384 (2011).
132 Hannan, A. J. in *EBioMedicine* Vol. 31 3-4 (2018).

133 Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted Genotyping of Variable Number Tandem Repeats with adVNTR. doi:10.1101/221754 (2018).
134 Castillo-Lizardo, M. *et al.* Replication slippage of the thermophilic DNA polymerases B and D from the Euryarchaeota Pyrococcus abyssi. *Frontiers in Microbiology* **5**, doi:10.3389/fmicb.2014.00403 (2014).
135 Breen, G., Collier, D., Craig, I. & Quinn, J. Variable number tandem repeats as agents of functional regulation in the genome. *Ieee Engineering in Medicine and Biology Magazine* **27**, 103-+, doi:10.1109/Emb.2008.915501 (2008).
136 Ohno, S. So much "junk" DNA in our genome. *Brookhaven Symp Biol* **23**, 366-370 (1972).
137 Hannan, A. J. TRPing up the genome: Tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease. *Discov Med* **10**, 314-321 (2010).
138 Haddley, K. *et al.* Molecular Genetics of Monoamine Transporters: Relevance to Brain Disorders. *Neurochemical Research* **33**, 652-667, doi:10.1007/s11064-007-9521-8 (2008).
139 Haddley, K., Bubb, V. J., Breen, G., Parades-Esquivel, U. M. & Quinn, J. P. in *Curr Top Behav Neurosci* Vol. 12 503-535 (2012).
140 Brookes, K. J. The VNTR in complex disorders: The forgotten polymorphisms? A functional way forward? *Genomics* **101**, 273-281, doi:https://doi.org/10.1016/j.ygeno.2013.03.003 (2013).
141 Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature Genetics* **30**, 315-320, doi:doi:10.1038/ng836 (2002).
142 Martin, P., Makepeace, K., Hill, S. A., Hood, D. W. & Moxon, E. R. in *Proc Natl Acad Sci U S A* Vol. 102 3800-3804 (2005).
143 TY, H. *et al.* Molecular pathogenesis of Gilbert's syndrome: decreased TATA-binding protein binding affinity of UGT1A1 gene promoter. *Pharmacogenetics and Genomics* **17**, 229-236, doi:10.1097/fpc.0b013e328012d0da (2007).
144 Zukic, B. *et al.* Functional analysis of the role of the TPMT gene promoter VNTR polymorphism in TPMT gene transcription. *http://dx.doi.org/10.2217/pgs.10.7*, doi:10.2217/pgs.10.7 (2010).
145 Ali, F. R. *et al.* Combinatorial interaction between two human serotonin transporter gene variable number tandem repeats and their regulation by CTCF. *J Neurochem* **112**, 296-306, doi:10.1111/j.1471-4159.2009.06453.x (2010).
146 Vasiliou, S. A. *et al.* The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro. *Addict Biol* **17**, 156-170, doi:10.1111/j.1369-1600.2010.00288.x (2012).
147 Warburton, A., Breen, G., Rujescu, D., Bubb, V. J. & Quinn, J. P. Characterization of a REST-Regulated Internal Promoter in the Schizophrenia Genome-Wide Associated Gene MIR137. *Schizophr Bull* **41**, 698-707, doi:10.1093/schbul/sbu117 (2015).
148 Warburton, A., Breen, G., Bubb, V. J. & Quinn, J. P. A GWAS SNP for Schizophrenia Is Linked to the Internal MIR137 Promoter and Supports Differential Allele-Specific Expression. *Schizophr Bull* **42**, 1003-1008, doi:10.1093/schbul/sbv144 (2016).
149 Manca, M. *et al.* The Regulation of Monoamine Oxidase A Gene Expression by Distinct Variable Number Tandem Repeats. *Journal of Molecular Neuroscience* **64**, 459-470, doi:10.1007/s12031-018-1044-z (2018).
150 Roeck, A. D. *et al.* An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathologica* **135**, 827-837, doi:doi:10.1007/s00401-018-1841-z (2018).
151 Holder, I. T. *et al.* Intrastrand triplex DNA repeats in bacteria: a source of genomic instability. *Nucleic Acids Res* **43**, 10126-10142, doi:10.1093/nar/gkv1017 (2015).

152 Kobayashi, T. in *DNA Replication, Recombination, and Repair: Molecular Mechanisms and Pathology* (eds Fumio Hanaoka & Kaoru Sugasawa) 235-247 (Springer Japan, 2016).
153 Hefferon, T. W., Groman, J. D., Yurk, C. E. & Cutting, G. R. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3504-3509, doi:10.1073/pnas.0400182101 (2004).
154 Brazda, V., Fojta, M. & Bowater, R. P. Structures and stability of simple DNA repeats from bacteria. *Biochem J* **477**, 325-339, doi:10.1042/bcj20190703 (2020).
155 Hall, A. C., Ostrowski, L. A., Pietrobon, V. & Mekhail, K. Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. https://doi.org/10.1080/19491034.2017.1292193, doi:10.1080/19491034.2017.1292193 (2017).
156 Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* **48**, 22-29, doi:10.1038/ng.3461 (2016).
157 Quilez, J. *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res* **44**, 3750-3762, doi:10.1093/nar/gkw219 (2016).
158 Verkerk, A. J. M. H. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905-914, doi:10.1016/0092-8674(91)90397-H (1991).
159 Kremer, E. *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. doi:10.1126/science.1675488 (1991).
160 JD, B. *et al.* Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68**, doi:10.1016/0092-8674(92)90154-5 (1992).
161 Harley, H. G. *et al.* Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature* **355**, 545-546, doi:doi:10.1038/355545a0 (1992).
162 Buxton, J. *et al.* Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. *Nature* **355**, 547-548, doi:doi:10.1038/355547a0 (1992).
163 Aslanidis, C. *et al.* Cloning of the essential myotonic dystrophy region and mapping of the putative defect. *Nature* **355**, 548-551, doi:doi:10.1038/355548a0 (1992).
164 Mahadevan, M. *et al.* Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255**, 1253-1255, doi:10.1126/science.1546325 (1992).
165 Fu, Y. H. *et al.* An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**, 1256-1258, doi:10.1126/science.1546326 (1992).
166 MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971-983, doi:10.1016/0092-8674(93)90585-E (1993).
167 Orr, H. T. *et al.* Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat Genet* **4**, 221-226, doi:10.1038/ng0793-221 (1993).
168 Ryan, C. P. & Department of Anthropology, N. U., Evanston, IL, USA. Tandem repeat disorders. *Evolution, Medicine, and Public Health* **2019**, 17-17, doi:10.1093/emph/eoz005 (2019).
169 Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics* **19**, 286-298, doi:doi:10.1038/nrg.2017.115 (2018).
170 Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansion in human-specific NOTCH2NLC associated with neuronal intranuclear inclusion disease. doi:10.1101/515635 (2019).

171 Lehesjoki, A.-E. & Kälviäinen, R. Unverricht-Lundborg Disease. doi:https://www.ncbi.nlm.nih.gov/books/NBK1142/ (2014).
172 Gendron, T. F. *et al.* in *Acta Neuropathol* Vol. 126 829-844 (2013).
173 Mizielinska, S. *et al.* in *Acta Neuropathol* Vol. 126 845-857 (2013).
174 Green, K. M. *et al.* RAN translation at C9orf72 -associated repeat expansions is selectively enhanced by the integrated stress response. *Nature Communications* **8**, 1-13, doi:doi:10.1038/s41467-017-02200-0 (2017).
175 Zu, T. *et al.* in *Proc Natl Acad Sci U S A* Vol. 108 260-265 (2011).
176 Chang, Y.-J., Jeng, U.-S., Chiang, Y.-L., Hwang, I.-S. & Chen, Y.-R. Glycine-Alanine Dipeptide Repeat from C9orf72 Hexanucleotide Expansions Forms Toxic Amyloids Possessing Cell-to-cell Transmission Property. doi:10.1074/jbc.M115.694273 (2016).
177 Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. doi:10.1074/jbc.REV119.007678 (2020).
178 Freibaum, B. D. & Taylor, J. P. The Role of Dipeptide Repeats in C9ORF72-Related ALS-FTD. *Front Mol Neurosci* **10**, doi:10.3389/fnmol.2017.00035 (2017).
179 Lam, D. *et al.* Genotype-dependent associations between serotonin transporter gene (SLC6A4) DNA methylation and late-life depression. *BMC Psychiatry* **18**, 1-10, doi:doi:10.1186/s12888-018-1850-4 (2018).
180 Pacheco, A., Berger, R., Freedman, R. & Law, A. J. A VNTR Regulates miR-137 Expression Through Novel Alternative Splicing and Contributes to Risk for Schizophrenia. *Scientific Reports* **9**, 1-12, doi:doi:10.1038/s41598-019-48141-0 (2019).
181 Tazelaar, G. H. P. *et al.* Association of NIPA1 repeat expansions with amyotrophic lateral sclerosis in a large international cohort. *Neurobiol Aging* **74**, 234.e239-234.e215, doi:10.1016/j.neurobiolaging.2018.09.012 (2019).
182 Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* **27**, 1895-1903, doi:10.1101/gr.225672.117 (2017).
183 Course, M. M. *et al.* Evolution of a Human-Specific Tandem Repeat Associated with ALS. *The American Journal of Human Genetics* **0**, doi:10.1016/j.ajhg.2020.07.004 (2020).
184 Jönsson, M. E. *et al.* Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. *Trends in Genetics* **0**, doi:10.1016/j.tig.2020.05.004 (2020).
185 Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biology* **19**, 1-12, doi:doi:10.1186/s13059-018-1577-z (2018).
186 McClintock, B. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A* **36**, 344-355 (1950).
187 Ravindran, S. Barbara McClintock and the discovery of jumping genes. doi:10.1073/pnas.1219372109 (2012).
188 Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, aac7247, doi:10.1126/science.aac7247 (2016).
189 Kazazian, H. H. & Moran, J. V. Mobile DNA in Health and Disease. *N Engl J Med* **377**, 361-370, doi:10.1056/NEJMra1510092 (2017).
190 Platt, R. N., Vandewege, M. W. & Ray, D. A. in *Chromosome Res* Vol. 26 25-43 (2018).
191 AL, S. *et al.* Retrotransposons in the Development and Progression of Amyotrophic Lateral Sclerosis. *Journal of neurology, neurosurgery, and psychiatry* **90**, doi:10.1136/jnnp-2018-319210 (2019).
192 MJ, D., O, M. S. & C, B. Transposons: A Blessing Curse. *Current opinion in plant biology* **42**, doi:10.1016/j.pbi.2018.01.003 (2018).

193 Ayarpadikannan, S. & Kim, H.-S. The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genomics & Informatics* **12**, 98-104, doi:10.5808/gi.2014.12.3.98 (2014).
194 B, M., L, R. & MM, S. Transposable Elements and Their Epigenetic Regulation in Mental Disorders: Current Evidence in the Field. *Frontiers in genetics* **10**, doi:10.3389/fgene.2019.00580 (2019).
195 Linker, S. B., Marchetto, M. C., Narvaiza, I., Denli, A. M. & Gage, F. H. Examining non-LTR retrotransposons in the context of the evolving primate brain. *BMC Biol* **15**, doi:10.1186/s12915-017-0409-z (2017).
196 GJ, F. & JL, G.-P. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. *Trends in genetics : TIG* **33**, doi:10.1016/j.tig.2017.07.004 (2017).
197 Goodier, J. L. Restricting retrotransposons: a review. *Mobile DNA* **7**, 1-30, doi:doi:10.1186/s13100-016-0070-z (2016).
198 N, G. & E, T. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Frontiers in immunology* **9**, doi:10.3389/fimmu.2018.02039 (2018).
199 Douville, R. N. & Nath, A. Human endogenous retroviruses and the nervous system. *Handb Clin Neurol* **123**, 465-485, doi:10.1016/b978-0-444-53488-0.00022-5 (2014).
200 Thompson, P. J., Macfarlan, T. S. & Lorincz, M. C. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. doi:10.1016/j.molcel.2016.03.029 (2016).
201 Dolei, A., Ibba, G., Piu, C. & Serra, C. Expression of HERV Genes as Possible Biomarker and Target in Neurodegenerative Diseases. *Int J Mol Sci* **20**, doi:10.3390/ijms20153706 (2019).
202 P, K. *et al.* Human Endogenous Retroviruses in Neurological Diseases. *Trends in molecular medicine* **24**, doi:10.1016/j.molmed.2018.02.007 (2018).
203 Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221-225, doi:10.1038/nature14308 (2015).
204 Brudek, T. *et al.* B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity. *Retrovirology* **6**, 1-13, doi:doi:10.1186/1742-4690-6-104 (2009).
205 Perron, H. *et al.* Molecular characteristics of Human Endogenous Retrovirus type-W in schizophrenia and bipolar disorder. *Translational Psychiatry* **2**, doi:doi:10.1038/tp.2012.125 (2012).
206 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, doi:10.1002/ana.22149 (2011).
207 Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* **7**, 307ra153, doi:10.1126/scitranslmed.aac8201 (2015).
208 Deininger, P. *et al.* A comprehensive approach to expression of L1 loci. *Nucleic Acids Research* **45**, doi:10.1093/nar/gkw1067 (2016).
209 Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. *Nature Reviews Genetics* **20**, 760-772, doi:doi:10.1038/s41576-019-0165-8 (2019).
210 Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. doi:10.1073/pnas.0831042100 (2003).
211 Beck, C. R. *et al.* LINE-1 Retrotransposition Activity in Human Genomes. *Cell* **141**, 1159-1170, doi:10.1016/j.cell.2010.05.021 (2010).
212 AM, D. *et al.* Primate-specific ORF0 Contributes to Retrotransposon-Mediated Diversity. *Cell* **163**, doi:10.1016/j.cell.2015.09.025 (2015).
213 A, B. & GG, S. in *Methods in molecular biology (Clifton, N.J.)* Vol. 1400 (Methods Mol Biol, 2016).

214 Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res* **40**, 1666-1683, doi:10.1093/nar/gkr863 (2012).
215 V, A., H, K., S, S. & O, W. Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. *Molecular cell* **60**, doi:10.1016/j.molcel.2015.10.003 (2015).
216 Kim, S., Cho, C. S., Han, K. & Lee, J. Structural Variation of Alu Element and Human Disease. *Genomics Inform* **14**, 70-77, doi:10.5808/gi.2016.14.3.70 (2016).
217 C, A., AM, R.-E. & PL, D. Alu elements: an intrinsic source of human genome instability. *Current opinion in virology* **3**, doi:10.1016/j.coviro.2013.09.002 (2013).
218 Zabolotneva, A. A. *et al.* Transcriptional regulation of human-specific SVAF1 retrotransposons by cis-regulatory MAST2 sequences. *Gene* **505**, 128-136, doi:https://doi.org/10.1016/j.gene.2012.05.016 (2012).
219 Hancks, D. C. & Kazazian, H. SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol* **20**, 234-245, doi:10.1016/j.semcancer.2010.04.001 (2010).
220 Gianfrancesco, O., Bubb, V. J. & Quinn, J. P. SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides* **64**, 3-7, doi:10.1016/j.npep.2016.09.006 (2017).
221 Hancks, D. C., Ewing, A. D., Chen, J. E., Tokunaga, K. & Kazazian, H. H. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res* **19**, 1983-1991, doi:10.1101/gr.093153.109 (2009).
222 Gianfrancesco, O. *et al.* The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome. *Int J Mol Sci* **20**, doi:10.3390/ijms20235977 (2019).
223 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 Evolutionary Biology* **13**, 101, doi:10.1186/1471-2148-13-101 (2013).
224 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, doi:10.1371/journal.pone.0090833 (2014).
225 Wang, H. *et al.* SVA Elements: A Hominid-specific Retroposon Family. *Journal of Molecular Biology* **354**, 994-1007, doi:https://doi.org/10.1016/j.jmb.2005.09.085 (2005).
226 Richardson, S. R. *et al.* The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, doi:10.1128/microbiolspec.MDNA3-0061-2014 (2015).
227 SL, M., Sandy.martin@ucdenver.edu & Department of Cell and Developmental Biology, U. o. C. S. o. M., Aurora, CO, USA. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. *RNA Biology* **7**, 706-711, doi:10.4161/rna.7.6.13766 (2010).
228 Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research* **22**, 2328-2338, doi:10.1101/gr.145235.112 (2012).
229 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75, doi:10.1038/nature15394 https://www.nature.com/articles/nature15394#supplementary-information (2015).
230 Bachiller, S., del-Pozo-Martín, Y. & Carrión, Á. M. L1 retrotransposition alters the hippocampal genomic landscape enabling memory formation. *Brain, Behavior, and Immunity* **64**, 65-70, doi:https://doi.org/10.1016/j.bbi.2016.12.018 (2017).
231 Quinn, J. P. & Bubb, V. J. SVA retrotransposons as modulators of gene expression. *Mobile Genetic Elements* **4**, e32102, doi:10.4161/mge.32102 (2014).
232 Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. & Gage, F. H. The Role of Transposable Elements in Health and Diseases of the Central Nervous System. *The*

Journal of Neuroscience **33**, 17577-17586, doi:10.1523/JNEUROSCI.3369-13.2013 (2013).

233 Singer, T., McConnell, M. J., Marchetto, M. C. N., Coufal, N. G. & Gage, F. H. LINE-1 Retrotransposons: Mediators of Somatic Variation in Neuronal Genomes? *Trends in neurosciences* **33**, 345-354, doi:10.1016/j.tins.2010.04.001 (2010).

234 Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534-537, doi:http://www.nature.com/nature/journal/v479/n7374/abs/nature10531.html#supplementary-information (2011).

235 Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228-239, doi:10.1016/j.cell.2015.03.026 (2015).

236 Kubo, S. *et al.* in *Proc Natl Acad Sci U S A* Vol. 103 8036-8041 (2006).

237 Macia, A. *et al.* Engineered LINE-1 retrotransposition in nondividing human neurons. *Genome Res* **27**, 335-348, doi:10.1101/gr.206805.116 (2017).

238 JV, M. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, doi:10.1016/s0092-8674(00)81998-4 (1996).

239 Sanchez-Luque, F. J., Richardson, S. R. & Faulkner, G. J. in *Transposons and Retrotransposons: Methods and Protocols* (ed Jose L. Garcia-Pérez) 47-77 (Springer New York, 2016).

240 Ewing, A. D. Transposable element detection from whole genome sequence data. *Mobile DNA* **6**, 24, doi:10.1186/s13100-015-0055-3 (2015).

241 Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable elements. *Nature Reviews Genetics* **19**, 688-704, doi:doi:10.1038/s41576-018-0050-x (2018).

242 Ewing, A. D. & Kazazian, H. H. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Research* **20**, 1262-1270, doi:10.1101/gr.106419.110 (2010).

243 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics* **10**, 691-703, doi:10.1038/nrg2640 (2009).

244 Feusier, J. *et al.* Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res* **29**, 1567-1577, doi:10.1101/gr.247965.118 (2019).

245 Evsikov, A. V. & Marín de Evsikova, C. Friend or Foe: Epigenetic Regulation of Retrotransposons in Mammalian Oogenesis and Early Development. *Yale J Biol Med* **89**, 487-497 (2016).

246 FJ, S.-L. *et al.* LINE-1 Evasion of Epigenetic Repression in Humans. *Molecular cell* **75**, doi:10.1016/j.molcel.2019.05.024 (2019).

247 Zamudio, N. *et al.* DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev* **29**, 1256-1270, doi:10.1101/gad.257840.114 (2015).

248 Lertkhachonsuk, R., Paiwattananupant, K., Tantbirojn, P., Rattanatanyong, P. & Mutirangura, A. LINE-1 Methylation Patterns as a Predictor of Postmolar Gestational Trophoblastic Neoplasia. *BioMed Research International* **2015**, doi:https://doi.org/10.1155/2015/421747 (2015).

249 Tahara, S. *et al.* Lower LINE-1 methylation is associated with promoter hypermethylation and distinct molecular features in gastric cancer. *https://doi.org/10.2217/epi-2019-0091*, doi:10.2217/epi-2019-0091 (2019).

250 S, B. *et al.* LINE-1 retrotransposon encoded ORF1p expression and promoter methylation in oral squamous cell carcinoma: a pilot study. *Cancer Genetics* **244**, 21-29, doi:10.1016/j.cancergen.2020.01.050 (2020).

251 Jacobs, F. M. J. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242-245,

doi:10.1038/nature13760
http://www.nature.com/nature/journal/v516/n7530/abs/nature13760.html#supplementary-information (2014).
252 Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. doi:10.1242/dev.132605 (2017).
253 Koito, A. *et al.* Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. *Frontiers in Microbiology* **4**, doi:10.3389/fmicb.2013.00028 (2013).
254 Lindič, N. *et al.* Differential inhibition of LINE1 and LINE2 retrotransposition by vertebrate AID/APOBEC proteins. *Retrovirology* **10**, 1-16, doi:doi:10.1186/1742-4690-10-156 (2013).
255 Tóth, K. F., Pezic, D., Stuwe, E. & Webster, A. The piRNA Pathway Guards the Germline Genome Against Transposable Elements. *Adv Exp Med Biol* **886**, 51-77, doi:10.1007/978-94-017-7417-8_4 (2016).
256 SJ, R. & J, L. Transposons and the PIWI pathway: genome defense in gametes and embryos. *Reproduction (Cambridge, England)* **156**, doi:10.1530/REP-18-0218 (2018).
257 Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics* **20**, 89-108, doi:doi:10.1038/s41576-018-0073-3 (2018).
258 McLaughlin, R. N. & Malik, H. S. Genetic conflicts: the usual suspects and beyond. *J Exp Biol* **220**, 6-17, doi:10.1242/jeb.148148 (2017).
259 Van Valen, L. (Theory, 1973).
260 Papkou, A. *et al.* The genomic basis of Red Queen dynamics during rapid reciprocal host–pathogen coevolution. doi:10.1073/pnas.1810402116 (2019).
261 Carroll, L. *Through the looking glass: And what Alice found there*. (Rand, McNally, 1917).
262 T, S., A, Z., G, C. & P, L. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature reviews. Genetics* **18**, doi:10.1038/nrg.2017.7 (2017).
263 Feschotte, C. The contribution of transposable elements to the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405, doi:10.1038/nrg2337 (2008).
264 Jangam, D., Feschotte, C. & Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet* **33**, 817-831, doi:10.1016/j.tig.2017.07.011 (2017).
265 Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554, doi:doi:10.1038/nature21683 (2017).
266 J, P. *et al.* Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell stem cell* **24**, doi:10.1016/j.stem.2019.03.012 (2019).
267 Trizzino, M., Kapusta, A. & Brown, C. D. Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics* **19**, doi:10.1186/s12864-018-4850-3 (2018).
268 Kellner, M. & Makałowski, W. Transposable elements significantly contributed to the core promoters in the human genome. *Science China Life Sciences* **62**, 489-497, doi:doi:10.1007/s11427-018-9449-0 (2019).
269 Polak, P. & Domany, E. in *BMC Genomics* Vol. 7 133 (2006).
270 Jiang, J.-C. & Upton, K. R. Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines. *Mobile DNA* **10**, 1-14, doi:doi:10.1186/s13100-019-0158-3 (2019).

271 Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences* **104**, 18613-18618, doi:10.1073/pnas.0703637104 (2007).
272 Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**, 1752-1762, doi:10.1101/gr.080663.108 (2008).
273 Schmid, C. D. & Bucher, P. in *PLoS One* Vol. 5 (2010).
274 Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* **42**, 631-634, doi:doi:10.1038/ng.600 (2010).
275 Schmidt, D. *et al.* in *Cell* Vol. 148 335-348 (2012).
276 Pugacheva, E. M. *et al.* The cancer-associated CTCFL/BORIS protein targets multiple classes of genomic repeats, with a distinct binding and functional preference for humanoid-specific SVA transposable elements. *Epigenetics & Chromatin* **9**, 35, doi:10.1186/s13072-016-0084-2 (2016).
277 V, S. & J, W. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **375**, doi:10.1098/rstb.2019.0347 (2020).
278 Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. doi:10.1101/gr.168872.113 (2014).
279 Garcia-Perez, J. L., Widmann, T. J. & Adams, I. R. The impact of transposable elements on mammalian development. *Development* **143**, 4101-4114, doi:10.1242/dev.132639 (2016).
280 Rodríguez-Martín, C. *et al.* Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. *Journal of Human Genetics* **61**, 463-466, doi:doi:10.1038/jhg.2015.173 (2016).
281 Payer, L. M. *et al.* in *Nucleic Acids Res* Vol. 47 421-431 (2019).
282 A, A. *et al.* A global reference for human genetic variation. *Nature* **526**, doi:10.1038/nature15393 (2015).
283 T, L. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, doi:10.1038/nature12531 (2013).
284 L, W., L, R., L, M.-R. & IK, J. Human Population-Specific Gene Expression and Transcriptional Network Modification With Polymorphic Transposable Elements. *Nucleic acids research* **45**, doi:10.1093/nar/gkw1286 (2017).
285 Spirito, G., Mangoni, D., Sanges, R. & Gustincich, S. Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data. *BMC Bioinformatics* **20**, 1-13, doi:doi:10.1186/s12859-019-3113-x (2019).
286 C, G., NA, Z. & C, F. Contribution of unfixed transposable element insertions to human regulatory variation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **375**, doi:10.1098/rstb.2019.0331 (2020).
287 Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. doi:10.1126/sciadv.aaw1668 (2019).
288 Kentepozidou, E. *et al.* Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biology* **21**, 1-19, doi:doi:10.1186/s13059-019-1894-x (2020).
289 Krumm, A. & Duan, Z. Understanding the 3D genome: emerging impacts on human disease. *Semin Cell Dev Biol* **90**, 62-77, doi:10.1016/j.semcdb.2018.07.004 (2019).
290 Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* **20**, 535-550, doi:doi:10.1038/s41580-019-0132-4 (2019).

291 EM, P. *et al.* CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proceedings of the National Academy of Sciences of the United States of America* **117**, doi:10.1073/pnas.1911708117 (2020).
292 Diehl, A. G., Ouyang, N. & Boyle, A. P. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nature Communications* **11**, 1-18, doi:doi:10.1038/s41467-020-15520-5 (2020).
293 R, F. *et al.* TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation. *Molecular cell* **77**, doi:10.1016/j.molcel.2019.10.020 (2020).
294 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, doi:10.1038/332164a0 (1988).
295 Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mobile DNA* **7**, 9, doi:10.1186/s13100-016-0065-9 (2016).
296 Makino, S. *et al.* in *Am J Hum Genet* Vol. 80 393-406 (2007).
297 Bragg, D. C. *et al.* Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proceedings of the National Academy of Sciences* **114**, E11020 (2017).
298 Westenberger, A. *et al.* A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism. *Ann Neurol* **85**, 812-822, doi:10.1002/ana.25488 (2019).
299 Stacey, S. N. *et al.* in *Hum Mol Genet* Vol. 25 1008-1018 (2016).
300 W, L., Y, J., L, P., M, H. & J, D. Transposable elements in TDP-43-mediated neurodegenerative disorders. *PloS one* **7**, doi:10.1371/journal.pone.0044099 (2012).
301 Krug, L. *et al.* Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. *PLOS Genetics* **13**, e1006635, doi:10.1371/journal.pgen.1006635 (2017).
302 Prudencio, M. *et al.* Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. *Human Molecular Genetics* **26**, 3421-3431, doi:10.1093/hmg/ddx233 (2017).
303 Tam, O. H. *et al.* Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. doi:10.1101/574509 (2019).
304 NOTT, A., MEISLIN, S. H. & MOORE, M. J. A quantitative analysis of intron effects on mammalian gene expression. doi:10.1261/rna.5250403 (2003).
305 Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281-2308, doi:doi:10.1038/nprot.2013.143 (2013).
306 Takagi, M. *et al.* Characterization of DNA polymerase from Pyrococcus sp. strain KOD1 and its application to PCR. *Appl Environ Microbiol* **63**, 4504-4510 (1997).
307 Mizuguchi, H., Nakatsuji, M., Fujiwara, S., Takagi, M. & Imanaka, T. Characterization and application to hot start PCR of neutralizing monoclonal antibodies against KOD DNA polymerase. *J Biochem* **126**, 762-768, doi:10.1093/oxfordjournals.jbchem.a022514 (1999).
308 S, P., S, S., B, J. & T, B. Role of membrane potential on artificial transformation of E. coli with plasmid DNA. *Journal of biotechnology* **127**, doi:10.1016/j.jbiotec.2006.06.008 (2006).
309 M, R., M, S., F, N., S, A. & H, M. Impact of heat shock step on bacterial transformation efficiency. *Molecular biology research communications* **5** (2016).
310 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343-345, doi:doi:10.1038/nmeth.1318 (2009).

311 Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* **12**, 41-51, doi:10.1093/bib/bbq072 (2010).
312 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:doi:10.1038/nature11247 (2012).
313 Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, doi:10.1038/ng.2653 (2013).
314 Ovcharenko, I., Nobrega, M. A., Loots, G. G. & Stubbs, L. in *Nucleic Acids Res* Vol. 32 W280-286 (2004).
315 Illumina. *Isaac Whole Genome Sequencing v2 User Guide*. (2014).
316 Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041-2043, doi:10.1093/bioinformatics/btt314 (2013).
317 P, D. *et al.* The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**, doi:10.1093/bioinformatics/btr330 (2011).
318 H, L. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, doi:10.1093/bioinformatics/btp352 (2009).
319 H, L. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27**, doi:10.1093/bioinformatics/btr509 (2011).
320 Li, H. & Program in Medical Population Genetics, T. B. I. o. H. a. M., Cambridge, MA 02142, USA. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718-719, doi:10.1093/bioinformatics/btq671 (2011).
321 Sherry, S. T. *et al.* in *Nucleic Acids Res* Vol. 29 308-311 (2001).
322 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
323 Li, R. D. *et al.* The SMYD3 VNTR 3/3 polymorphism confers an increased risk and poor prognosis of hepatocellular carcinoma in a Chinese population. *Pathol Res Pract* **214**, 625-630, doi:10.1016/j.prp.2018.04.005 (2018).
324 Rockowitz, S. & Zheng, D. Significant expansion of the REST/NRSF cistrome in human versus mouse embryonic stem cells: potential implications for neural development. *Nucleic Acids Res* **43**, 5730-5743, doi:10.1093/nar/gkv514 (2015).
325 Zhao, Y. *et al.* Brain REST/NRSF Is Not Only a Silent Repressor but Also an Active Protector. *Molecular Neurobiology* **54**, 541-550, doi:doi:10.1007/s12035-015-9658-4 (2016).
326 Hwang, J. Y. & Zukin, R. S. REST, a master transcriptional regulator in neurodegenerative disease. *Curr Opin Neurobiol* **48**, 193-200, doi:10.1016/j.conb.2017.12.008 (2018).
327 Carminati, E. *et al.* Mild Inactivation of RE-1 Silencing Transcription Factor (REST) Reduces Susceptibility to Kainic Acid-Induced Seizures. *Front Cell Neurosci* **13**, doi:10.3389/fncel.2019.00580 (2019).
328 Song, Z., Zhao, D., Zhao, H. & Yang, L. NRSF: an Angel or a Devil in Neurogenesis and Neurological Diseases. *Journal of Molecular Neuroscience* **56**, 131-144, doi:doi:10.1007/s12031-014-0474-5 (2014).
329 Rockowitz, S. *et al.* in *PLoS Comput Biol* Vol. 10 (2014).
330 Mozzi, A. *et al.* REST , a master regulator of neurogenesis, evolved under strong positive selection in humans and in non human primates. *Scientific Reports* **7**, 1-9, doi:doi:10.1038/s41598-017-10245-w (2017).
331 Baldelli, P. & Meldolesi, J. The Transcription Repressor REST in Adult Neurons: Physiology, Pathology, and Diseases1,2,3. *eNeuro* **2**, doi:10.1523/eneuro.0010-15.2015 (2015).

332 Rodenas-Ruano, A., Chávez, A. E., Cossio, M. J., Castillo, P. E. & Zukin, R. S. REST-dependent epigenetic remodeling promotes the in vivo developmental switch in NMDA receptors. *Nat Neurosci* **15**, 1382-1390, doi:10.1038/nn.3214 (2012).
333 Song, Z. *et al.* in *Cell Death Dis* Vol. 9 (2018).
334 L, L. *et al.* REST overexpression in mice causes deficits in spontaneous locomotion. *Scientific reports* **8**, doi:10.1038/s41598-018-29441-3 (2018).
335 Hu, X.-L. *et al.* Conditional Deletion of NRSF in Forebrain Neurons Accelerates Epileptogenesis in the Kindling Model. *Cerebral Cortex* **21**, 2158-2165, doi:10.1093/cercor/bhq284 (2011).
336 Liu, M. *et al.* Neuronal conditional knockout of NRSF decreases vulnerability to seizures induced by pentylenetetrazol in mice. *Acta Biochimica et Biophysica Sinica* **44**, 476-482, doi:10.1093/abbs/gms023 (2012).
337 M, Y. *et al.* NRSF/REST neuronal deficient mice are more vulnerable to the neurotoxin MPTP. *Neurobiology of aging* **34**, doi:10.1016/j.neurobiolaging.2012.06.002 (2013).
338 Lu, T. *et al.* REST and Stress Resistance in Aging and Alzheimer's Disease. *Nature* **507**, 448-454, doi:10.1038/nature13163 (2014).
339 Khan, A. O., Eisenberger, T., Nagel-Wolfrum, K., Wolfrum, U. & Bolz, H. J. C21orf2 is mutated in recessive early-onset retinal dystrophy with macular staphyloma and encodes a protein that localises to the photoreceptor primary cilium. doi:10.1136/bjophthalmol-2015-307277 (2015).
340 Suga, A. *et al.* Identification of Novel Mutations in the LRR-Cap Domain of C21orf2 in Japanese Patients With Retinitis Pigmentosa and Cone-Rod Dystrophy. *Invest Ophthalmol Vis Sci* **57**, 4255-4263, doi:10.1167/iovs.16-19450 (2016).
341 Scott, H. S. *et al.* Characterization of a novel gene, C21orf2, on human chromosome 21q22.3 and its exclusion as the APECED gene by mutation analysis. *Genomics* **47**, 64-70, doi:10.1006/geno.1997.5066 (1998).
342 Lai, C. K. *et al.* in *Mol Biol Cell* Vol. 22 1104-1119 (2011).
343 GA, B., NF, B., J, L. & K, M. Type III adenylyl cyclase localizes to primary cilia throughout the adult mouse brain. *The Journal of comparative neurology* **505**, doi:10.1002/cne.21510 (2007).
344 Ma, X., Peterson, R. & Turnbull, J. Adenylyl Cyclase type 3, a marker of primary cilia, is reduced in primary cell culture and in lumbar spinal cord in situ in G93A SOD1 mice. *BMC Neuroscience* **12**, 1-11, doi:doi:10.1186/1471-2202-12-71 (2011).
345 Ma, X., Turnbull, P., Peterson, R. & Turnbull, J. Trophic and proliferative effects of Shh on motor neurons in embryonic spinal cord culture from wildtype and G93A SOD1 mice. *BMC Neuroscience* **14**, 1-9, doi:doi:10.1186/1471-2202-14-119 (2013).
346 Fang, X. *et al.* The NEK1 interactor, C21ORF2, is required for efficient DNA damage repair. *Acta Biochim Biophys Sin (Shanghai)* **47**, 834-841, doi:10.1093/abbs/gmv076 (2015).
347 Abu-Safieh, L. *et al.* Autozygome-guided exome sequencing in retinal dystrophy patients reveals pathogenetic mutations and novel candidate disease genes. *Genome Res* **23**, 236-247, doi:10.1101/gr.144105.112 (2013).
348 Wheway, G. *et al.* An siRNA-based functional genomics screen for the identification of regulators of ciliogenesis and ciliopathy genes. *Nat Cell Biol* **17**, 1074-1087, doi:10.1038/ncb3201 (2015).
349 Wang, Z. *et al.* in *PLoS One* Vol. 11 (2016).
350 Dao, T. P., Majumdar, A. & Barrick, D. Capping motifs stabilize the leucine-rich repeat protein PP32 and rigidify adjacent repeats. *Protein Sci* **23**, 801-811, doi:10.1002/pro.2462 (2014).

351 Y, W. *et al.* An Amyotrophic Lateral Sclerosis-Associated Mutant of C21ORF2 Is Stabilized by NEK1-Mediated Hyperphosphorylation and the Inability to Bind FBXO3. *iScience* **23**, doi:10.1016/j.isci.2020.101491 (2020).
352 Manca, M. *Role of Variable Number Tandem Repeats (VNTRs) on gene expression in the CNS* PhD thesis, University of Liverpool, (2016).
353 McLaughlin, R. L. *et al.* Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nat Commun* **8**, 14774, doi:10.1038/ncomms14774 (2017).
354 Diekstra, F. P. *et al.* C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: a genome-wide meta-analysis. *Ann Neurol* **76**, 120-133, doi:10.1002/ana.24198 (2014).
355 Karch, C. M. *et al.* in *JAMA Neurol* Vol. 75 860-875 (2018).
356 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
357 Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Research* **30**, 38-41, doi:10.1093/nar/30.1.38 (2002).
358 Mitsuhashi, S., Frith, M. C. & Matsumoto, N. Genome-wide survey of tandem repeats by nanopore sequencing shows that disease-associated repeats are more polymorphic in the general population. doi:10.1101/2019.12.19.883389 (2019).
359 Boivin, M., Willemsen, R., Hukema, R. K. & Sellier, C. Potential pathogenic mechanisms underlying Fragile X Tremor Ataxia Syndrome: RAN translation and/or RNA gain-of-function? *Eur J Med Genet* **61**, 674-679, doi:10.1016/j.ejmg.2017.11.001 (2018).
360 Glineburg, M. R., Todd, P. K., Charlet-Berguerand, N. & Sellier, C. Repeat-associated non-AUG (RAN) translation and other molecular mechanisms in Fragile X Tremor Ataxia Syndrome. *Brain Res* **1693**, 43-54, doi:10.1016/j.brainres.2018.02.006 (2018).
361 Tian, Y. *et al.* in *Am J Hum Genet* Vol. 105 166-176 (2019).
362 Orafidiya, F. A. & McEwan, I. J. Trinucleotide repeats and protein folding and disease: the perspective from studies with the androgen receptor. *http://dx.doi.org/10.4155/fso.15.47*, doi:10.4155/fso.15.47 (2015).
363 Seixas, A. I. *et al.* in *Am J Hum Genet* Vol. 101 87-103 (2017).
364 Paulson, H. Repeat expansion diseases. *Handb Clin Neurol* **147**, 105-123, doi:10.1016/b978-0-444-63233-3.00009-9 (2018).
365 Paredes, U. M., Quinn, J. P. & D'Souza, U. M. Allele-specific transcriptional activity of the variable number of tandem repeats in 5' region of the DRD4 gene is stimulus specific in human neuronal cells. *Genes Brain Behav* **12**, 282-287, doi:10.1111/j.1601-183X.2012.00857.x (2013).
366 Gratten, J. *et al.* Whole-exome sequencing in amyotrophic lateral sclerosis suggests NEK1 is a risk gene in Chinese. *Genome Med* **9**, 97, doi:10.1186/s13073-017-0487-0 (2017).
367 Meirelles, G. V. *et al.* "Stop Ne(c)king around": How interactomics contributes to functionally characterize Nek family kinases. *World J Biol Chem* **5**, 141-160, doi:10.4331/wjbc.v5.i2.141 (2014).
368 Fry, A. M., Bayliss, R. & Roig, J. Mitotic Regulation by NEK Kinase Networks. *Front Cell Dev Biol* **5**, doi:10.3389/fcell.2017.00102 (2017).
369 Melo-Hanchuk, T. D. *et al.* NEK1 kinase domain structure and its dynamic protein interactome after exposure to Cisplatin. *Sci Rep* **7**, doi:10.1038/s41598-017-05325-w (2017).
370 Chen, Y., Craigen, W. J. & Riley, D. J. Nek1 regulates cell death and mitochondrial membrane permeability through phosphorylation of VDAC1. *Cell Cycle* **8**, 257-267 (2009).

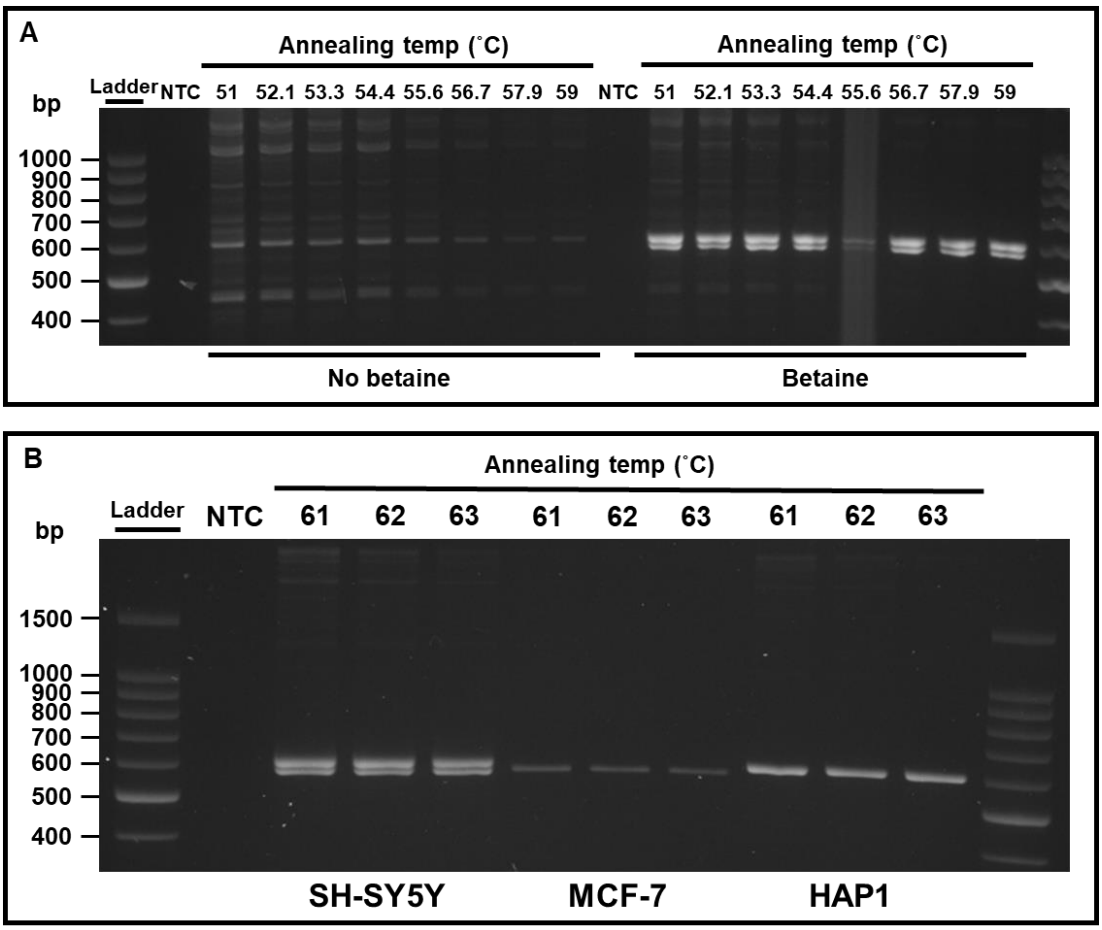
371 Chen, Y., Chen, P.-L., Chen, C.-F., Jiang, X. & Riley, D. J. Never-in-mitosis related kinase 1 functions in DNA damage response and checkpoint control. *Cell cycle (Georgetown, Tex.)* **7**, 3194-3201 (2008).
372 Chen, Y. *et al.* Mutation of NIMA-related kinase 1 (NEK1) leads to chromosome instability. *Molecular Cancer* **10**, 1-13, doi:doi:10.1186/1476-4598-10-5 (2011).
373 Spies, J. *et al.* Nek1 Regulates Rad54 to Orchestrate Homologous Recombination and Replication Fork Stability. *Mol Cell* **62**, 903-917, doi:10.1016/j.molcel.2016.04.032 (2016).
374 Tavares, E. M., Wright, W. D., Heyer, W. D., Le Cam, E. & Dupaigne, P. In vitro role of Rad54 in Rad51-ssDNA filament-dependent homology search and synaptic complexes formation. *Nat Commun* **10**, doi:10.1038/s41467-019-12082-z (2019).
375 Chen, Y., Gaczynska, M., Osmulski, P., Polci, R. & Riley, D. J. Phosphorylation by Nek1 Regulates Opening and Closing of Voltage Dependent Anion Channel 1. *Biochemical and biophysical research communications* **394**, 798-803, doi:10.1016/j.bbrc.2010.03.077 (2010).
376 Camara, A. K. S., Zhou, Y., Wen, P. C., Tajkhorshid, E. & Kwok, W. M. Mitochondrial VDAC1: A Key Gatekeeper as Potential Therapeutic Target. *Front Physiol* **8**, 460, doi:10.3389/fphys.2017.00460 (2017).
377 Upadhya, P., Birkenmeier, E. H., Birkenmeier, C. S. & Barker, J. E. Mutations in a NIMA-related kinase gene, Nek1, cause pleiotropic effects including a progressive polycystic kidney disease in mice. *Proc Natl Acad Sci U S A* **97**, 217-221 (2000).
378 Thiel, C. *et al.* NEK1 Mutations Cause Short-Rib Polydactyly Syndrome Type Majewski. *Am J Hum Genet* **88**, 106-114, doi:10.1016/j.ajhg.2010.12.004 (2011).
379 Monroe, G. R. *et al.* Compound heterozygous NEK1 variants in two siblings with oral-facial-digital syndrome type II (Mohr syndrome). *European Journal of Human Genetics* **24**, 1752-1760, doi:doi:10.1038/ejhg.2016.103 (2016).
380 Wang, Z. *et al.* Axial spondylometaphyseal dysplasia is also caused by NEK1 mutations. *Journal of Human Genetics* **62**, 503-506, doi:doi:10.1038/jhg.2016.157 (2017).
381 Sohn, E. Fundraising: The Ice Bucket Challenge delivers. *Nature* **550**, doi:doi:10.1038/550S113a (2017).
382 Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748-1759, doi:10.1101/gr.136127.111 (2012).
383 Siavrienė, E. *et al.* A novel CHD7 variant disrupting acceptor splice site in a patient with mild features of CHARGE syndrome: a case report. *BMC Medical Genetics* **20**, 1-7, doi:doi:10.1186/s12881-019-0859-y (2019).
384 Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature* **533**, 95-99, doi:10.1038/nature17939 (2016).
385 Szafranski, P. *et al.* Association of rare non-coding SNVs in the lung-specific FOXF1 enhancer with a mitigation of the lethal ACDMPV phenotype. *Hum Genet* **138**, 1301-1311, doi:10.1007/s00439-019-02073-x (2019).
386 Tang, W., Mun, S., Joshi, A., Han, K. & Liang, P. Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Research* **25**, 521-533, doi:10.1093/dnares/dsy022 (2018).
387 Guichard, E. *et al.* Impact of non-LTR retrotransposons in the differentiation and evolution of anatomically modern humans. *Mobile DNA* **9**, 1-19, doi:doi:10.1186/s13100-018-0133-4 (2018).
388 Muñoz-Lopez, M. *et al.* LINE-1 retrotransposition impacts the genome of human pre-implantation embryos and extraembryonic tissues. *bioRxiv*, 522623, doi:10.1101/522623 (2019).

389 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 Evolutionary Biology* **13**, 1-12, doi:doi:10.1186/1471-2148-13-101 (2013).
390 Clayton, E. A. *et al.* Patterns of Transposable Element Expression and Insertion in Cancer. *Front Mol Biosci* **3**, 76, doi:10.3389/fmolb.2016.00076 (2016).
391 Ha, H., Loh, J. W. & Xing, J. Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA). *Mobile DNA* **7**, 15, doi:10.1186/s13100-016-0072-x (2016).
392 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* **9**, 1-5, doi:doi:10.1038/s41598-019-45839-z (2019).
393 Nordin, A. *et al.* Extensive size variability of the GGGGCC expansion in C9orf72 in both neuronal and non-neuronal tissues in 18 patients with ALS or FTD. *Human Molecular Genetics* **24**, 3133-3142, doi:10.1093/hmg/ddv064 (2015).
394 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
395 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1-19, doi:doi:10.1186/1471-2105-5-113 (2004).
396 Waterhouse, A. M. *et al.* Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).
397 Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian Jr, H. H. in *Am J Hum Genet* Vol. 73 1444-1451 (2003).
398 Lovett, S. T. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Molecular Microbiology* **52**, 1243-1253, doi:10.1111/j.1365-2958.2004.04076.x (2004).
399 Naruse, H. *et al.* Loss-of-function variants in NEK1 are associated with an increased risk of sporadic ALS in the Japanese population. *Journal of Human Genetics*, 1-5, doi:doi:10.1038/s10038-020-00830-9 (2020).
400 Shu, S. *et al.* Mutation screening of NEK1 in Chinese ALS patients. *Neurobiol Aging* **71**, 267.e261-267.e264, doi:10.1016/j.neurobiolaging.2018.06.022 (2018).
401 Slotkin, R. K. The case for not masking away repetitive DNA. *Mobile DNA* **9**, 1-4, doi:doi:10.1186/s13100-018-0120-9 (2018).
402 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338-345, doi:doi:10.1038/nbt.4060 (2018).
403 Bowden, R. *et al.* Sequencing of human genomes with nanopore technology. *Nature Communications* **10**, 1-9, doi:doi:10.1038/s41467-019-09637-5 (2019).
404 Cogoi, S., Department of Biomedical Science and Technology, S. o. M. P. l. K., 33100 Udine, Italy, Xodo, L. E. & Department of Biomedical Science and Technology, S. o. M. P. l. K., 33100 Udine, Italy. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Research* **34**, 2536-2549, doi:10.1093/nar/gkl286 (2006).
405 L, Z., W, T., J, Z., M, X. & G, Y. Investigation of G-quadruplex formation in the FGFR2 promoter region and its transcriptional regulation by liensinine. *Biochimica et biophysica acta. General subjects* **1861**, doi:10.1016/j.bbagen.2017.01.028 (2017).
406 N, K. The Interplay between G-quadruplex and Transcription. *Current medicinal chemistry* **26**, doi:10.2174/0929867325666171229132619 (2019).
407 J, S., S, A. & S, B. The Structure and Function of DNA G-Quadruplexes. *Trends in chemistry* **2**, doi:10.1016/j.trechm.2019.07.002 (2020).

408 Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-909.e821, doi:10.1016/j.cell.2018.02.011 (2018).
409 Rakovic, A. *et al.* Genome editing in induced pluripotent stem cells rescues TAF1 levels in X-linked dystonia-parkinsonism. *Mov Disord* **33**, 1108-1118, doi:10.1002/mds.27441 (2018).
410 Avci-Adali, M. *et al.* Optimized conditions for successful transfection of human endothelial cells with in vitro synthesized and modified mRNA for induction of protein expression. *Journal of Biological Engineering* **8**, 1-11, doi:doi:10.1186/1754-1611-8-8 (2014).
411 Ooi, A., Wong, A., Esau, L., Lemtiri-Chlieh, F. & Gehring, C. A Guide to Transient Expression of Membrane Proteins in HEK-293 Cells for Functional Characterization. *Frontiers in Physiology* **7**, doi:10.3389/fphys.2016.00300 (2016).
412 RL, B. *et al.* Identification of novel breakpoints for locus- and region-specific translocations in 293 cells by molecular cytogenetics before and after irradiation. *Scientific reports* **9**, doi:10.1038/s41598-019-47002-0 (2019).
413 Mir, A., Edraki, A., Lee, J. & Sontheimer, E. J. Type II-C CRISPR-Cas9 Biology, Mechanism and Application. *ACS Chem Biol* **13**, 357-365, doi:10.1021/acschembio.7b00855 (2018).
414 Barrangou, R. The roles of CRISPR-Cas systems in adaptive immunity and beyond. *Curr Opin Immunol* **32**, 36-41, doi:10.1016/j.coi.2014.12.008 (2015).
415 Ceasar, S. A., Rajan, V., Prykhozhij, S. V., Berman, J. N. & Ignacimuthu, S. Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. *Biochim Biophys Acta* **1863**, 2333-2344, doi:10.1016/j.bbamcr.2016.06.009 (2016).
416 Karvelis, T. *et al.* Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biology* **16**, 1-13, doi:doi:10.1186/s13059-015-0818-7 (2015).
417 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**, 671-675, doi:doi:10.1038/nmeth.2089 (2012).
418 F, A. Normalization of Gene Expression by Quantitative RT-PCR in Human Cell Line: comparison of 12 Endogenous Reference Genes. *Ethiopian journal of health sciences* **28**, doi:10.4314/ejhs.v28i6.9 (2018).
419 Zhang, C. *et al.* Selection of reference genes for gene expression studies in human bladder cancer using SYBR-Green quantitative polymerase chain reaction. *Oncol Lett* **14**, 6001-6011, doi:10.3892/ol.2017.7002 (2017).
420 N, S., B, M., A, H.-W. & G, A. Characteristics of transposable element exonization within human and mouse. *PloS one* **5**, doi:10.1371/journal.pone.0010907 (2010).
421 K, K., J, B., A, H., P, N. & M, S. Intronic L1 retrotransposons and nested genes cause transcriptional interference by inducing intron retention, exonization and cryptic polyadenylation. *PloS one* **6**, doi:10.1371/journal.pone.0026099 (2011).
422 K, K. & M, S. Intronic retroelements: Not just "speed bumps" for RNA polymerase II. *Mobile genetic elements* **2**, doi:10.4161/mge.20774 (2012).
423 JE, G. & AB, R. The enduring mystery of intron-mediated enhancement. *Plant science : an international journal of experimental plant biology* **237**, doi:10.1016/j.plantsci.2015.04.017 (2015).
424 O, S. How introns enhance gene expression. *The international journal of biochemistry & cell biology* **91**, doi:10.1016/j.biocel.2017.06.016 (2017).
425 AB, R. Introns as Gene Regulators: A Brick on the Accelerator. *Frontiers in genetics* **9**, doi:10.3389/fgene.2018.00672 (2019).
426 Gallegos, J. E. & Rose, A. B. An intron-derived motif strongly increases gene expression from transcribed sequences through a splicing independent mechanism

in Arabidopsis thaliana. *Scientific Reports* **9**, 1-9, doi:doi:10.1038/s41598-019-50389-5 (2019).
427 M, L. Intron-Mediated Enhancement: A Tool for Heterologous Gene Expression in Plants? *Frontiers in plant science* **7**, doi:10.3389/fpls.2016.01977 (2017).
428 M, L. *et al.* The 5'UTR Intron of Arabidopsis GGT1 Aminotransferase Enhances Promoter Activity by Recruiting RNA Polymerase II. *Plant physiology* **172**, doi:10.1104/pp.16.00881 (2016).
429 AB, R., T, E., G, P. & I, K. Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *The Plant cell* **20**, doi:10.1105/tpc.107.057190 (2008).
430 Parra, G. *et al.* Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. *Nucleic Acids Research* **39**, 5328-5337, doi:10.1093/nar/gkr043 (2011).
431 LL, C. & L, Y. ALUternative Regulation for Gene Expression. *Trends in cell biology* **27**, doi:10.1016/j.tcb.2017.01.002 (2017).
432 Pelegrini, A. L. *et al.* Nek1 silencing slows down DNA repair and blocks DNA damage-induced cell cycle arrest. *Mutagenesis* **25**, 447-454, doi:10.1093/mutage/geq026 (2010).
433 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222, doi:10.1093/bioinformatics/btv710 (2016).
434 Stewart, C. *et al.* A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genetics* **7**, e1002236, doi:10.1371/journal.pgen.1002236 (2011).
435 Helman, E. *et al.* Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Research* **24**, 1053-1063, doi:10.1101/gr.163659.113 (2014).
436 Kuhn, A. *et al.* Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 8131-8136, doi:10.1073/pnas.1401532111 (2014).
437 Iskow, R. C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253-1261, doi:10.1016/j.cell.2010.05.020 (2010).
438 Shukla, R. *et al.* Endogenous Retrotransposition Activates Oncogenic Pathways in Hepatocellular Carcinoma. *Cell* **153**, 101-111, doi:10.1016/j.cell.2013.02.032 (2013).

Supplementary Material



Supplementary Figure 1. *CFAP410* VNTR PCR optimisation

Optimisation process for the PCR of the *CFAP410* VNTR. **A**: Gel agarose electrophoresis of samples that underwent a gradient PCR (using 8 different annealing temperatures): used to determine an optimal annealing temperature for PCR primers. Addition of betaine determined that these samples were heterozygous, by linearising any secondary structure generated by the VNTR which may have restricted amplification of both alleles. **B**: Veriflex PCR (using three different annealing temperatures), trying higher annealing temperatures to attempt to reduce non-specific binding of primers: 63 °C generated the least amount of non-specific amplification.

Supplementary Table 1. Allele and genotype frequencies of the *UNC13A* VNTR in an ALS cohort and matched controls.

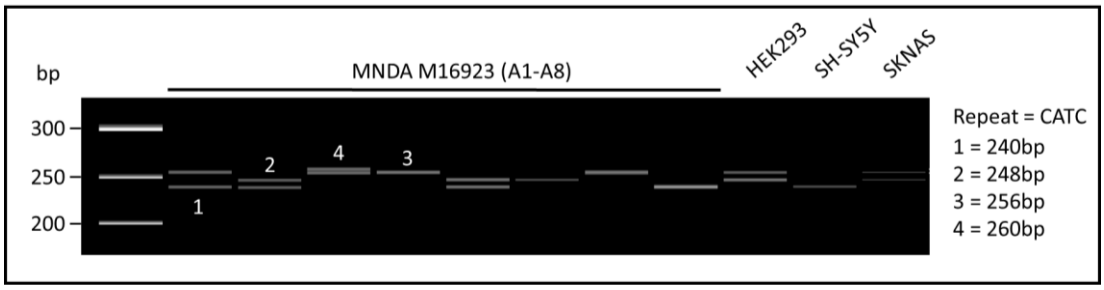
A: The four observed alleles of the UNC13A VNTR in and ALS cohort (n = 196) and matched controls (n = 178). Allele 4 was only found in the control cohort. There is no significant difference in allele frequency between the ALS cohort and matched controls (Fisher's exact test). **B**: The seven observed genotypes of the UNC13A VNTR in and ALS cohort (n = 98) and matched controls (n = 89). There is no significant difference in genotype frequency between the ALS cohort and matched controls (Fisher's exact test).

A

Cohort / Allele	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1	93	47.45	84	47.19	177	0.26	1.00
2	34	17.35	29	16.29	63	1.05	0.89
3	69	35.20	64	35.96	133	-0.75	0.91
4	0	0.00	1	0.56	1	-0.56	0.48
Total	196	100.00	178	100.00	374	0.00	N/A

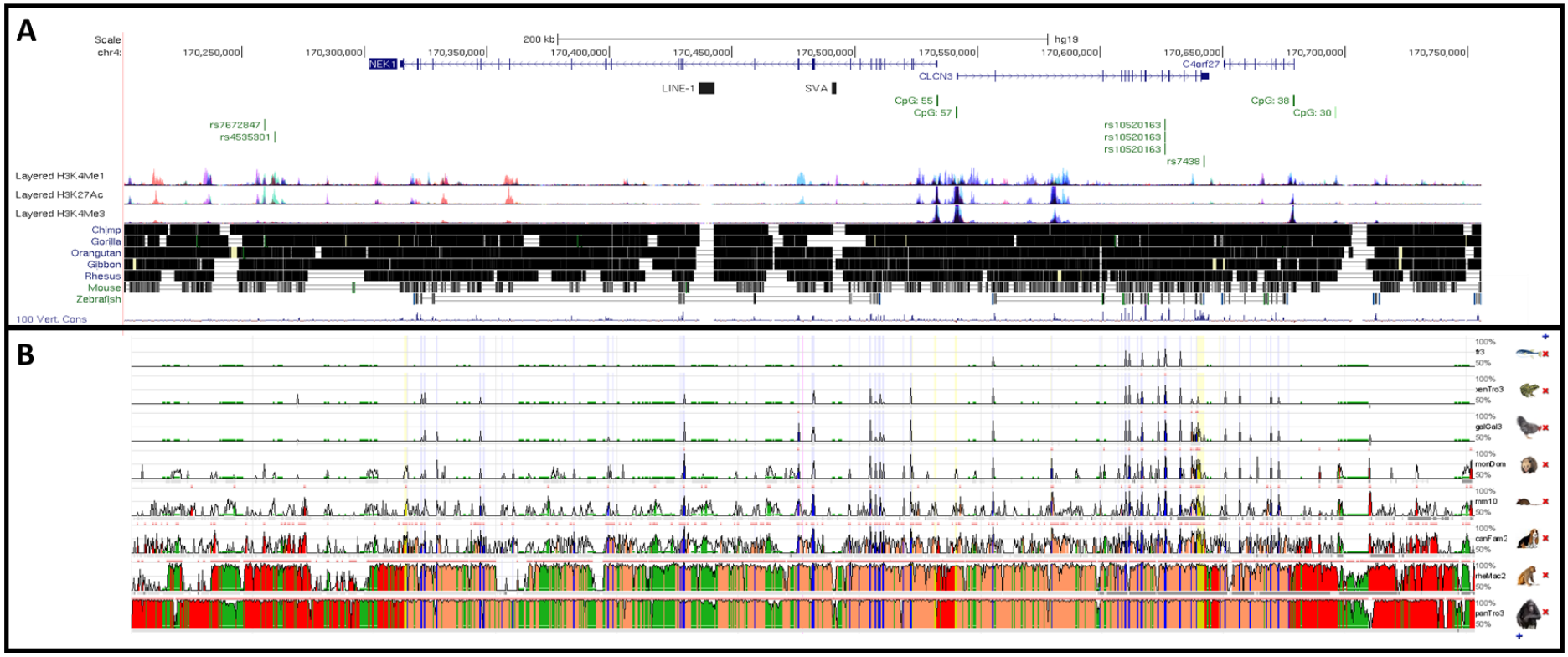
B

Cohort / Genotype	ALS cohort		Control cohort		Total Cases	% Difference (ALS - Control)	p-value (Fisher's exact test)
	Count	%	Count	%			
1,1	24	24.49	21	23.60	45	0.89	1.00
1,2	13	13.27	16	17.98	29	-4.71	0.42
1,3	32	32.65	26	29.21	58	3.44	0.64
2,2	5	5.10	1	1.12	6	3.98	0.21
2,3	11	11.22	11	12.36	22	-1.14	0.82
3,3	13	13.27	13	14.61	26	-1.34	0.83
3,4	0	0.00	1	1.12	1	-1.12	0.48
Total	98	100.00	89	100.00	187	0.00	N/A



Supplementary Figure 2. *UNC13A* VNTR genotyping in MNDA cohort.

PCR amplification and gel capillary electrophoresis of the *UNC13A* VNTR performed on MNDA samples using the QIAxcel advanced system and electronic gel image generated using the QIAxcel ScreenGel software. A total of four variants of the *UNC13A* VNTR were identified in this cohort (n = 187).



Supplementary Figure 3. *NEK1/CLCN3* locus overlaid with evolutionary conserved regions (ECRs) and human specific elements.

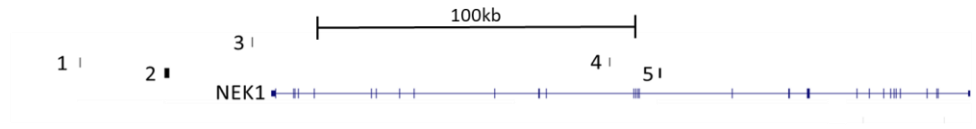
Visualisation of the *NEK1/CLCN3* locus (**A**) (chr4:170,201,958-170,755,566;UCSC/hg19) overlaid with species conservation from ECR browser (**B**). There are several transcripts according to genome browser UCSC hg19 (not shown). ENCODE data from UCSC shows the levels of enrichment of histone marks within this locus, specifically signals for mono-methylation H3K4Me1 and acetylation H3K27Ac, often associated with regulatory

regions[312]. There are two human specific elements within the *NEK1* gene, in the form of an SVA and a LINE-1. The LINE-1 is part of known active subclass of these elements (L1HS) and is full length; the SVA is part of the D subclass and is anti-sense to the orientation of the *NEK1* gene.

Supplementary Table 2. Project MinE UK dataset ALS samples with known coding *NEK1* mutations which confer risk for ALS.

Sample
LP6008119-DNA_B10
LP6008123-DNA_H02
LP6008124-DNA_D07
LP6008124-DNA_H08
LP6008125-DNA_G04
LP6008125-DNA_G07
LP6008197-DNA_G07
LP6008198-DNA_C09
LP6008200-DNA_B04
LP6008234-DNA_D04
LP6008236-DNA_A06
LP6008236-DNA_C02
LP6008237-DNA_E08
LP6008239-DNA_G12
LP6008464-DNA_C06

A

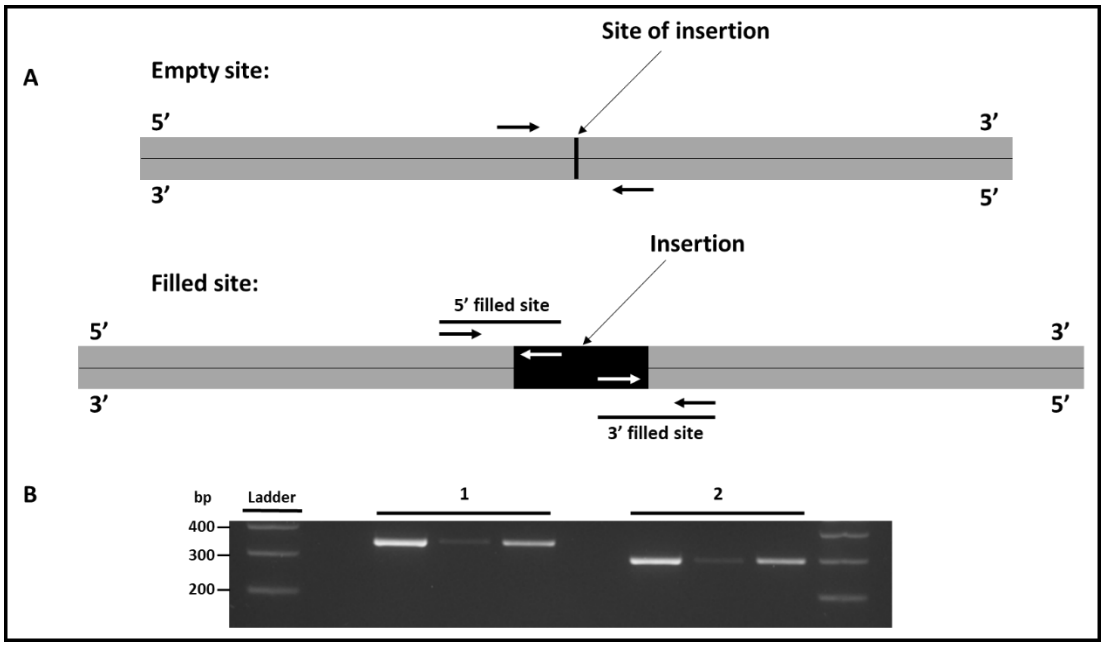


B

RIP	Type	Location	Coordinates (hg19)	Studies	Sample size	Presence count
1	LINE-1	60kb downstream of NEK1	chr4:170254290-170254337	1. CS2011 (Stewart *et al.*, 2011) 2. EH2014 (Helman *et al.*, 2014)	1. 185 2. 200	1. N/A 2. 2
2	LINE-1	32kb downstream of NEK1	chr4:170280811-170282453	1. AE2010 (Ewing and Kazazian Jr, 2010) 2. AK2014 (Kuhn *et al.*, 2014) 3. RI2010 (Iscow *et al.*, 2010) 4. RS2013 (Shukla *et al.*, 2013) 5. SS2012 (Solyom *et al.*, 2012)	1. 25 2. 20 3. 76 4. 19 5. 16	1. N/A 2. 20 3. N/A 4. 2 5. N/A
3	SVA	7kb downstream of NEK1	chr4:170308291-170308390	1. Sudmant *et al.*, 2015	1. 105	1
4	Alu	Within intron 23 of NEK1	chr4:170420484-170420485	1. EH2014 (Helman *et al.*, 2014)	1. 200	1
5	LINE-1	Within intron 19 of NEK1	chr4:170436196-170436659	1. AK2014 (Kuhn *et al.*, 2014)	1. 20	1

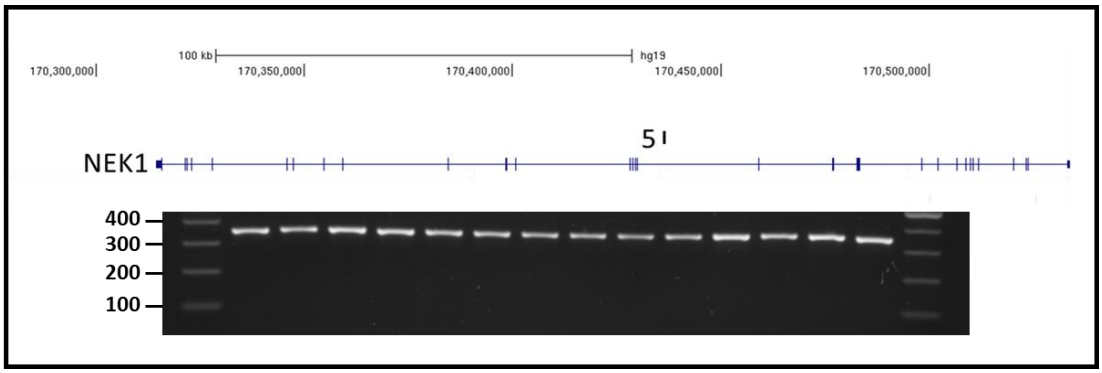
Supplementary Figure 4. Retrotransposon insertion polymorphisms within the *NEK1* locus.

The *NEK1* gene is present on chromosome 4 (chr4: 170,198,712-170,558,496 hg19). **A**: Retrotransposon insertion polymorphisms (RIPs) are numbered and shown above isoform 1 of the *NEK1* gene. **B**: Table listing details of these known RIPs in the *NEK1* locus; type of insertion, location with respect to *NEK1*, the studies they have been found in, the sample size within each study and the respective frequency of each insertion. Coordinates (hg19) for RIPs 1, 2, 4 and 5 were taken from a BED file from (Ewing, 2015[240]) and RIP 3 coordinates (hg19) were taken from a BED file generated in TeBreak. These studies use a wide variety of software tools for the detection of TEs in whole genome sequencing data, which have been reviewed by Ewing, 2015[240]. Studies from the table: Stewart *et al.*, 2011[434]; Helman *et al.*, 2014[435]; Ewing and Kazazian Jr, 2010[242]; Kuhn *et al.*, 2014[436]; Iscow *et al.*, 2010[437]; Shukla *et al.*, 2013[438]; Solyom *et al.*, 2012[228]; Sudmant *et al.*, 2015[229].



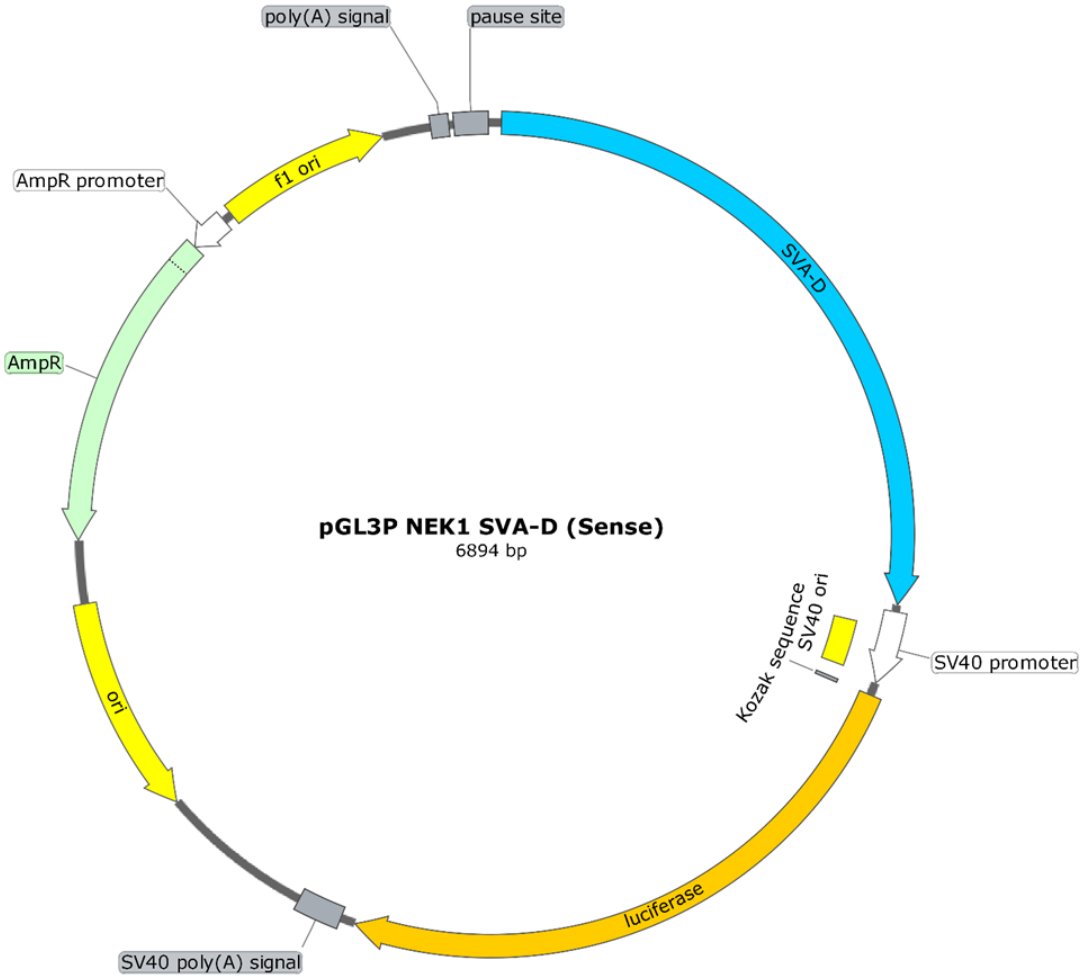
Supplementary Figure 5. PCR validation of RIPs.

A: schematic of RIP insertion validation via PCR. **B**: PCR amplification and gel electrophoresis of the filled site fragments for RIP 5 (LINE-1 element) of the *NEK1* locus. **B1**: the 5' filled site of the LINE-1 RIP. **B2**: 3' filled site of the LINE-1 RIP. Samples run on 2% agarose.



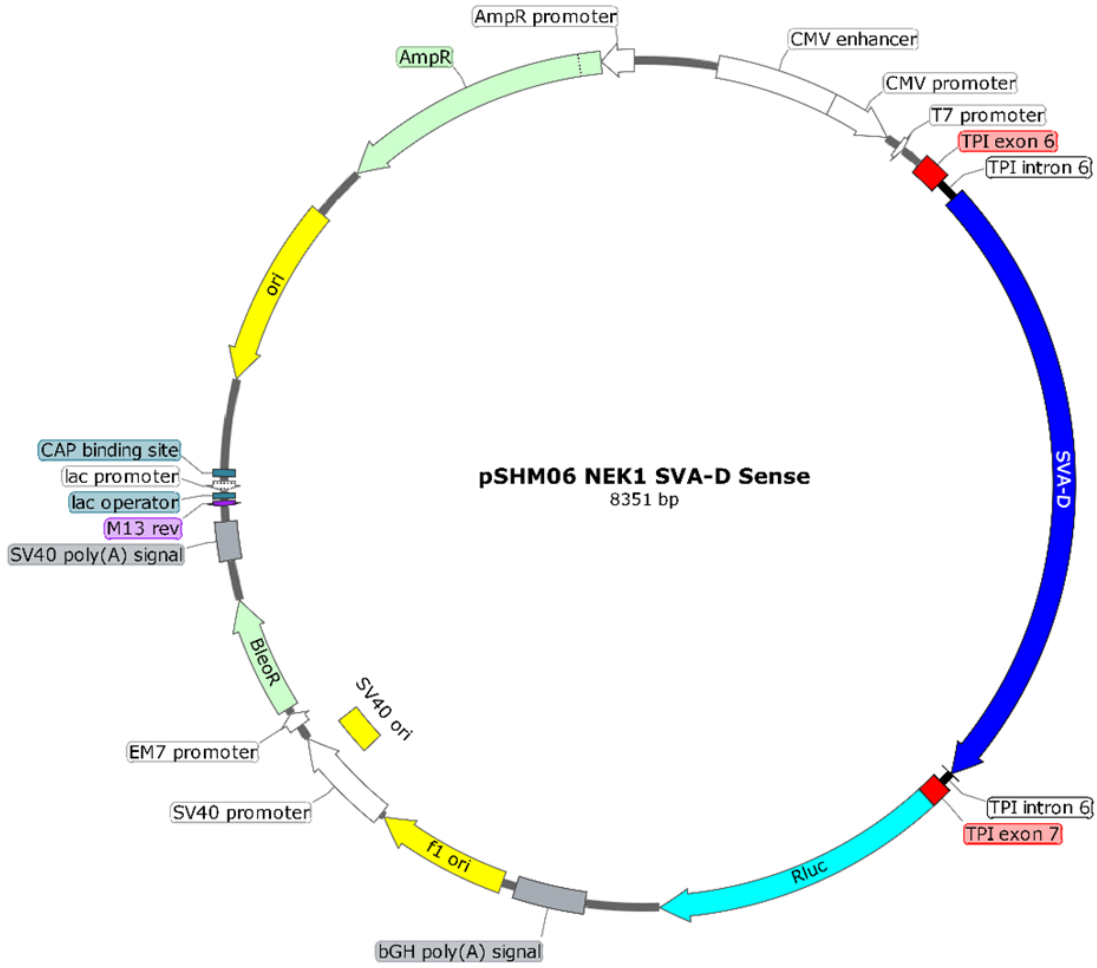
Supplementary Figure 6. PCR validation of RIP5 within *NEK1.*

PCR amplification and gel electrophoresis of the 5' filled site fragment for RIP 5 (LINE-1 element) of the *NEK1* locus in an ALS and matched control cohort (n = 96). Samples run on 1.5% agarose at 100V for 1 hour. All samples contain RIP 5.



Supplementary Figure 7. pGL3P/*NEK1* SVA-D vector map.

pGL3-Promoter (pGL3-P) contains a minimal SV40 promotor and a firefly luciferase reporter gene. The *NEK1* SVA-D shown above (blue) is present in the forward (sense) orientation with respect to the SV40 promotor: a second construct was designed with the SVA-D in the reverse (anti-sense orientation) with respect to the SV40 promotor (not shown).



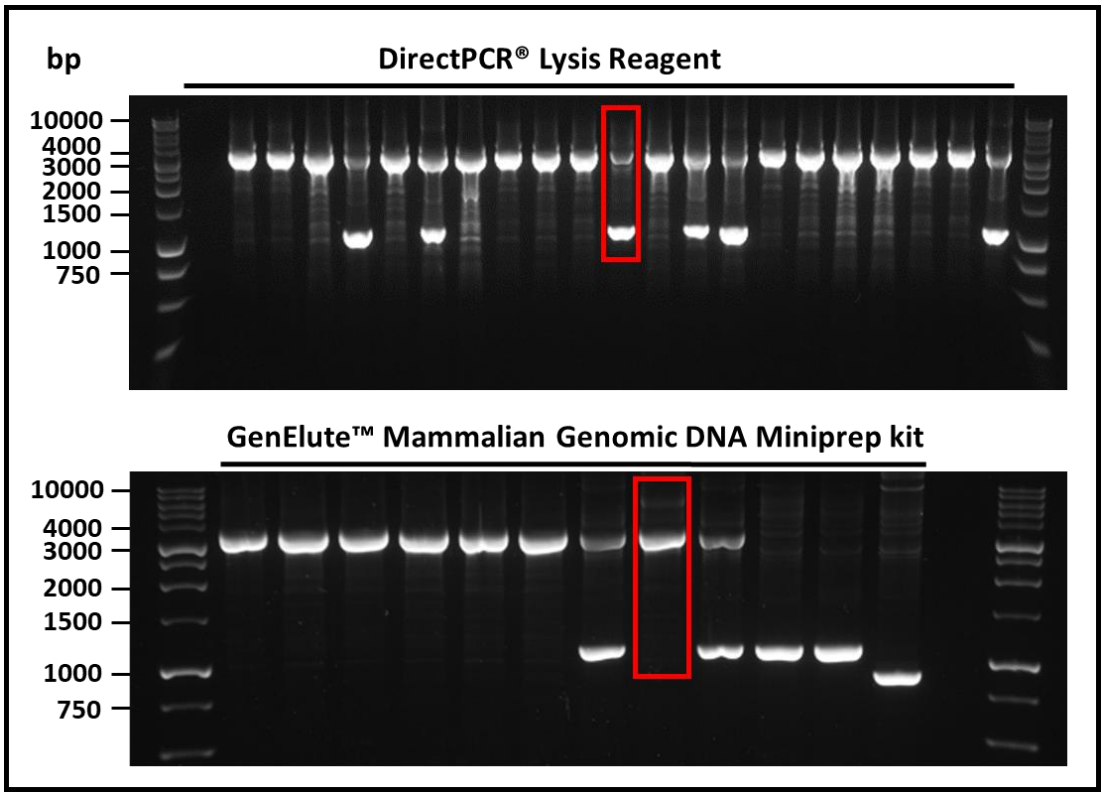
Supplementary Figure 8. pSHM06/*NEK1* SVA-D vector map.

pSHM06 contains a high expression CMV promotor and a renilla luciferase reporter gene, flanked by exons 6 and 7 of the *triosphosphate isomerase* (*TPI*) gene (red boxes). Intron 6 of *TPI* (shown as a black line) indicates the site of integration for the SVA-D (shown in dark blue). The *NEK1* SVA-D shown above is present in the forward (sense) orientation with respect to the CMV promotor: a second construct was designed with the SVA-D in the reverse (anti-sense orientation) with respect to the CMV promotor (not shown).

Supplementary Table 3. Densitometry for CRISPR guide modification bands.

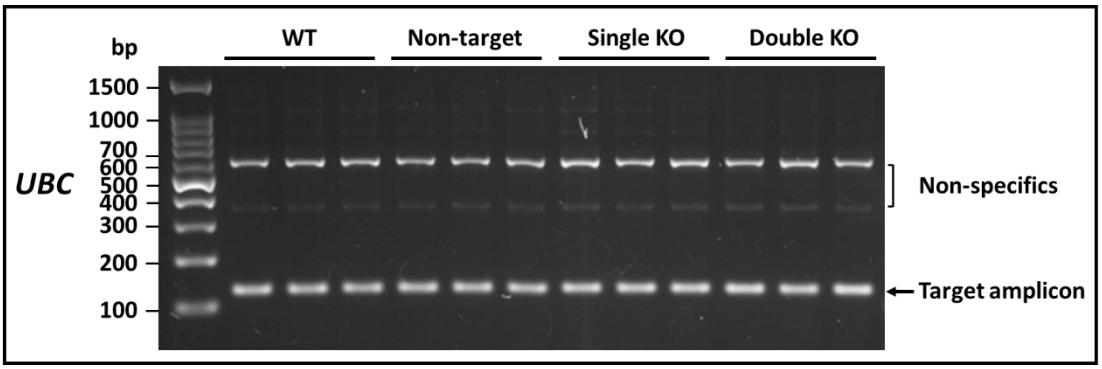
Quantification of agarose gel bands using ImageJ[417]. Integrative density was calculated for whole lane and the modified band for each of the four CRISPR guide combinations. Ratio of modified:whole signal was calculated and normalised to the ratio signal generated for guides 1,3. Modified signal was also calculated as a % for each guide combination.

Guide combination		Integrative density	Modified:lane ratio	Normalised ratio	% modified
1,3	whole lane	104.74	0.15	1.00	15.16
	modified band	15.88			
1,4	whole lane	104.24	0.14	0.94	14.26
	modified band	14.87			
2,3	whole lane	100.03	0.13	0.86	13.01
	modified band	13.01			
2,4	whole lane	86.08	0.12	0.78	11.81
	modified band	10.17			



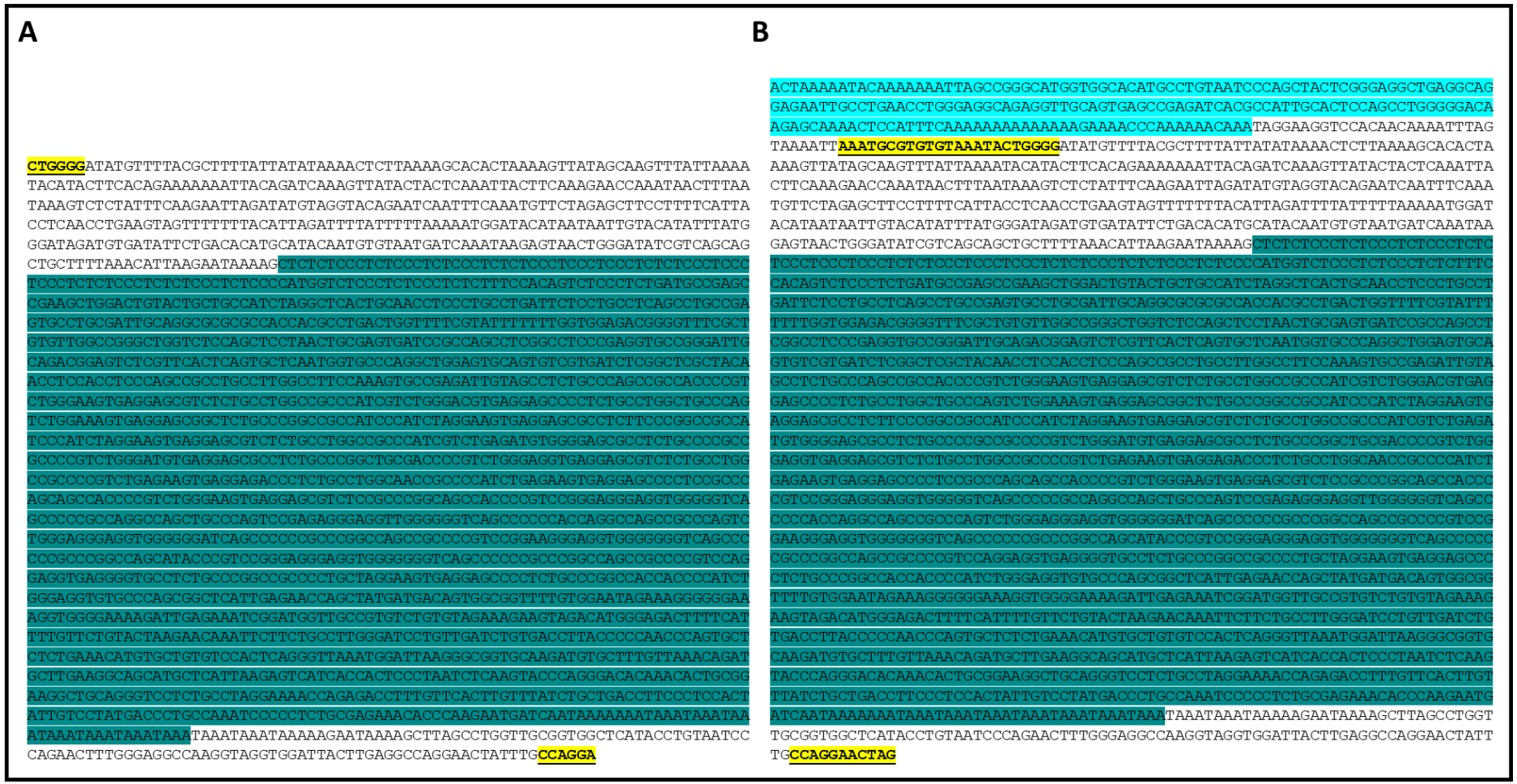
Supplementary Figure 9. *NEK1* SVA CRISPR KO genotyping disagreement between reagents.

Gel agarose electrophoresis of the *NEK1* SVA-D CRIPSR KO region: unmodified amplicon=3432 bp, modified amplicon=1089 bp. Top panel shows several cell lines where the template was generated using the DirectPCR Lysis Reagent (Viagen). The red box indicates a cell line which was genotyped as being heterozygous for the SVA KO modification (presence of both a modified and unmodified allele). Bottom panel is the same amplified region, but the template was purified gDNA extracted using the GenElute™ Mammalian Genomic DNA Miniprep kit (Sigma). The red box indicates the same cell line from the top panel, but using the purified gDNA template was genotyped as being unmodified on all alleles. The other cell lines were not shared across gels.



Supplementary Figure 10. Discarded RT-PCR primers.

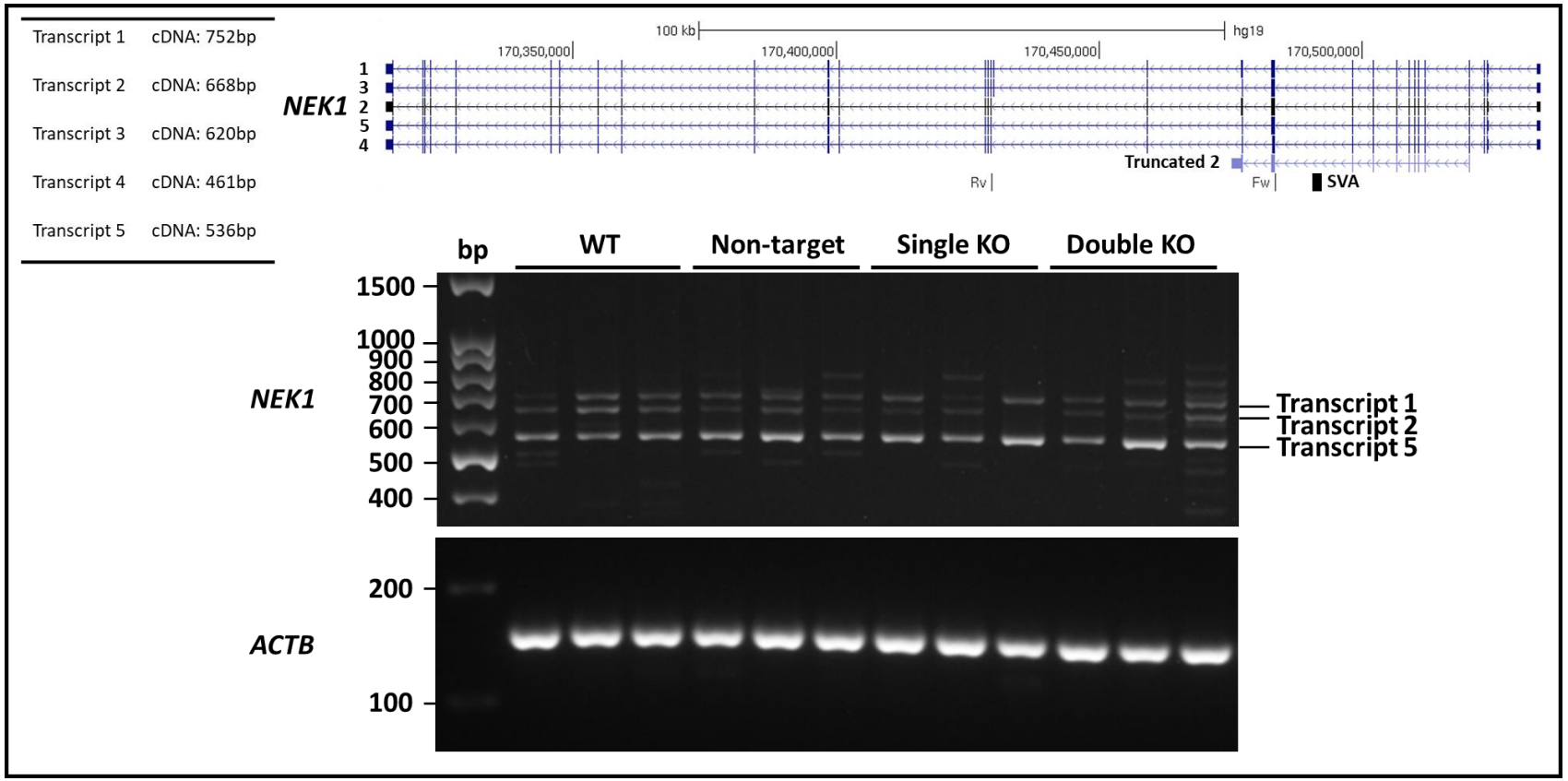
Reverse transcription PCR analysis of *UBC* mRNA from untreated (wildtype; WT), non-target, single SVA KO and double SVA KO HEK293 cell lines (per condition, n = 3). Samples run on 2% agarose gel at 110V for 1 hour. Target amplicon and two non-specifics are labelled, indicating that these primers were not optimal and were therefore discarded from this experiment.



Supplementary Figure 11. *NEK1* SVA KO sequence.

Sequence of the *NEK1* region excised using the CRISPR Cas 9 system. **A**: Expected size of the designed SVA excision (2309 bp), double strand

breaks occurred exactly 3 bp upstream of PAM sites. **B**: Larger modification (2563 bp) observed in one of the three homozygous SVA KO cell lines, also cleaving 200 bp of a 311 bp AluSq2 element (highlighted in light blue). Guide RNA sequences highlighted in yellow, SVA element highlighted in teal.



Supplementary Figure 12. *NEK1* transcript expression.

Reverse transcription PCR analysis of *NEK1* transcript and *ACTB* expression, mRNA from untreated (wildtype; WT), non-target, single SVA KO and double SVA KO HEK293 cell lines (per condition, n=3). Location of forward and reverse primer and SVA are shown. Samples run on 2% agarose gel at 110V for 1 hour. Only transcript 1 (752 bp), 2 (668 bp) and 5 (536 bp) were expressed in HEK293. WT = wildtype, KO = knockout.

Appendices

All documents are available upon request, please contact Professor John Quinn (jquinn@liverpool.ac.uk)

Appendix 1

Human DNA

1. MNDA UK DNA cohort information

Appendix 2

Isaac Variant Caller

1. BASH scripts used for analysis of Isaac Variant Caller data (txt files)

Appendix 3

Sequencing data

1. Sequencing verification (chromatograms) for *CFAP410* and *REST* VNTR luciferase constructs (pGL3B and pGL3P).
2. Sequencing verification (chromatograms) for *NEK1* SVA constructs (pGL3P and pSHM06).
3. Sequencing of CRISPR breakpoints (chromatograms), to confirm *NEK1* SVA KO in HEK293 (single and double KOs).