CELLULAR MINIGENE MODELS FOR THE STUDY OF ALLELIC EXPRESSION OF THE TAU GENE AND ITS ROLE IN PROGRESSIVE SUPRANUCLEAR PALSY

Victoria Anne Kay

Thesis submitted in the fulfilment of the degree of
Doctor of Philosophy

Reta Lila Weston Institute of Neurological Studies
Institute of Neurology
University College London

March 2013

Declaration

I, **Victoria Kay**, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Progressive supranuclear palsy (PSP) belongs to a group of neurodegenerative disorders that are characterised by hallmark pathology consisting of intra-neuronal aggregates of the microtubule-associated protein, tau. In PSP, these aggregates are almost exclusively composed of one of the two major tau protein isoform groups normally expressed at similar levels in the healthy brain, indicating a role for altered isoform regulation in PSP aetiology.

Although no causal mutations have been identified, common variation within the gene encoding tau, *MAPT*, has been highly associated with PSP risk. The A-allele of the rs242557 single nucleotide polymorphism has been repeatedly shown to significantly increase the risk of developing PSP. Its location within a distal region of the *MAPT* promoter region is significant, with independent studies – including this one – demonstrating that the rs242557-A allele alters the function of a transcription regulatory domain. As transcription and alternative splicing processes have been shown to be co-regulated in some genes, it was hypothesised that the rs242557-A allele could directly affect *MAPT* alternative splicing through its differential effect on transcription.

This project describes an investigation into the molecular mechanism linking the *MAPT* association with the tau isoform dysregulation characteristic of PSP. The design, construction and *in vitro* investigation of minigenes representing common *MAPT* variants will be presented in detail and will demonstrate that promoter identity plays an important role in regulating the alternative splicing of *MAPT* transcripts. The specific role of the rs242557 polymorphism in *MAPT* transcription and splicing are investigated and the two alleles of the polymorphism are shown to differentially influence these two molecular processes, providing a plausible mechanism linking the two phenomena known to be associated with PSP – a common genetic variant within the *MAPT* promoter region and detrimental changes to tau isoform production.

Acknowledgements

There are many people who – knowingly or otherwise – made an important contribution to the completion of this thesis. I'd like to thank everyone at the Reta Lila Weston Institute and Queen's Square Brain Bank for their support, advice and company over the past three and a half years.

In particular, I am hugely indebted to Rohan de Silva and Roberto Simone, who taught me everything I know and were a valuable source of knowledge, experience and encouragement. Special mentions should also be given to Geshanti Hondhamuni, Adam Mamais, Rina Bandopadhyay and Badma Segarane, for making my PhD experience a very enjoyable one.

I'd like to thank my wonderful family for their constant love and support, especially my two nephews, William and Jacob, for always making me smile. I would also like to thank Tessa Radwan, Lucy Dalton-Griffin and Sharon M^{c}Lean, for their unflinching support and encouragement.

The writing of this thesis was fuelled by endless cups of tea, often made by my friend and flatmate, Jenny Berlin.

Table of Contents

Table of Figures

List of Tables

1 Introduction

1.1 Overview

It is becoming increasingly apparent that the molecular processes responsible for normal gene expression do not occur independently of each other and are, in fact, co-regulated. A growing number of studies have described mechanistic links between transcription and splicing and have shown that an alteration in the regulation of transcription can simultaneously affect the alternative splicing of its mRNA [1-9]. It therefore follows that genetic variation that modifies the transcription rate of a gene can simultaneously affect the inclusion rate of its alternatively spliced downstream exons [10], leading to alterations in protein isoform production that are the hallmarks of a wide variety of diseases.

The tauopathies are a heterogeneous group of neurodegenerative disorders that are characterised by intra-neuronal aggregates of the microtubule-associated protein, tau. Pathological examination of the aggregates has shown that the tightly controlled balance of the two major tau isoform groups (3R- and 4R-tau) is altered in the tauopathy brain. This provided the first indication that disturbances in 3R- and 4R-tau homeostasis could be associated with disease pathogenesis. Genetic studies have further confirmed the link between tau dysfunction and the tauopathies, in the first instance by identifying a number of highly penetrant dominant mutations in the gene encoding tau (*MAPT*) that cause familial frontotemporal dementia with parkinsonism linked to chromosome 17 (FTLD-17). It is the discovery, however, that common variation in *MAPT* – in the absence of pathogenic mutations – can influence an individual's risk of developing a tauopathy that has opened up new avenues in the search for a molecular link between the genetic and pathological findings.

The strongest association of common *MAPT* variation is with the 4R tauopathy, progressive supranuclear palsy (PSP). The key polymorphism – denoted rs242557 – that drives the association of *MAPT* with PSP lies within a highly conserved region located approximately 47kb downstream to the *MAPT* core promoter (exon

0), yet upstream to the first coding exon. Studies have shown that this region has the potential to exert influence on *MAPT* transcription and that the two alleles of this highly associated polymorphism differentially alter the extent of this influence [11, 12]. When added to the neuropathological changes in tau isoform expression observed in the tauopathy brain, these findings indicate that *MAPT* is a likely candidate for the exhibition of co-regulation of transcription and alternative splicing.

This project describes an investigation into the molecular mechanism linking the *MAPT* association with the tau isoform dysregulation characteristic of PSP. The design, construction and *in vitro* investigation of minigenes representing common *MAPT* variants will be presented in detail and will demonstrate that promoter identity plays an important role in the regulation of the alternative splicing of *MAPT* exons, exerting influence over the delicate balance of 3R- and 4R-tau expression. The specific role of the rs242557 polymorphism in *MAPT* transcription and splicing will also be investigated and the two alleles of the polymorphism shown to differentially influence these two molecular processes, providing a plausible mechanism linking the two phenomena known to be associated with PSP – a common genetic variant in the *MAPT* promoter and detrimental changes to tau isoform production.

Many of the findings described here are currently being written up for publication.

1.2 The regulation of mammalian gene expression

1.2.1 Genetic elements in the regulation of expression

Mammalian gene expression is a multi-layered process involving a series of highly regulated and inter-related steps. The correct functioning of these processes is dependent upon precise signals situated at specific locations throughout the gene. Most mammalian protein-coding genes can be split into three major sections: the 5’ intronic region lying upstream to the first coding exon, the coding region and the 3’ intronic region lying downstream to the final STOP codon. Each

section contains signalling motifs relating to different, though inter-related stages of gene expression.

The 5’ intronic region contains the core promoter, often around exon 0, which forms the principal site at which transcription is initiated. Numerous *cis*-acting regulatory domains located within this region regulate transcription rate by enhancing or repressing core promoter activity.

The protein-coding exons are relatively short (on average around 150 nucleotides) and are separated by long stretches of non-coding sequence – called introns – that are removed from mRNA transcripts in a process called splicing. Splicing brings the exons into alignment for translation into protein and regulates the differential inclusion of certain exons in a developmental and/or tissue-specific manner. Thus, a single pre-mRNA transcript can be spliced in numerous different ways to produce a heterogeneous population of mature mRNAs. Current estimates suggest that over 90% of mammalian genes are alternatively spliced [13].

The 3’ intronic region (3’UTR) has important roles in the further processing, transport and stability of the mRNA transcript, containing signal sequences for polyadenylation, cellular localisation and degradation [14]. Well-characterised localisation signals are particularly important in neuronal cells due to their unique morphology, with axons extending out over particularly long distances from the cell body and nucleus. The correct sub-cellular localisation of mRNA transcripts is vital for maintaining the polarity of neurons, which in turn is vital for neuronal function.

1.2.2 The machineries involved in gene expression

1.2.2.1 Overview

Gene expression begins in the nucleus with transcription, where an RNA intermediate is synthesised from the genomic DNA template. During transcription, the nascent transcript undergoes a series of processing steps that result in the addition of a 5’ cap, the removal of introns and the cleavage and

polyadenylation of the 3' end. The mature mRNA transcript is then released from the transcription machinery and transported into the cytoplasm where it is translated into protein. This whole process is precisely regulated and undergoes surveillance, with incorrectly processed or mutant transcripts identified and subject to degradation or nonsense-mediated decay (NMD). Although each reaction is catalysed by different machineries, there are physical and functional interactions between them (figure 1.1).

Figure 1.1 The complex and inter-related steps involved in mammalian gene expression. ***Taken unchanged from Maniatis and Read (2002) [2].***

1.2.2.2 Transcription

Messenger RNA (mRNA) is a vital component of the gene expression process, forming an intermediate between the DNA template and its expressed protein product. Precursor mRNA (pre-mRNA) is synthesised from the DNA template by transcription, one of the most highly regulated cellular processes. Transcription is catalysed by a DNA-dependent RNA polymerase, of which there are three distinct types in eukaryotic cells, each with a specific function [15]:

- RNA polymerase I (Pol I) transcribes ribosomal RNA precursors that eventually form the primary site of protein synthesis in the cell
- RNA polymerase II (Pol II) is the major RNA polymerase that transcribes mRNA from protein-coding genes
- RNA polymerase III (Pol III) primarily transcribes transfer RNA, a necessary component of the machinery that translates the mature mRNA into protein

Of particular importance in the expression of protein-coding genes is the carboxy-terminal domain (CTD) of Pol II. The CTD is not only vital for transcription, but supports the recruitment and regulation of the independent machineries responsible for the capping, splicing and polyadenylation of pre-mRNA transcripts [16, 17]. The mammalian CTD consists of 52 tandem repeats of a heptapeptide motif: YSPTSPS. Dynamic phosphorylation and dephosphorylation of specific residues within the motif is vital to CTD function, providing signals to the processing machineries regarding the progress of the transcription complex and regulating the recruitment of specific processing factors to the nascent transcript [18, 19]. Thus, the exact pattern of CTD post-translational modification is referred to as 'the CTD code' and changes as Pol II moves through the different stages of transcription (figure 1.2) [20-22].

Figure 1.2 The 'CTD code of phosphorylation.
The five stages of transcription (A-E) are initiated by the 'CTD code' of phosphorylation and each stage is associated with different processing factors. A detailed description is given below. ***Taken unchanged from Montes et al (2012) [4].***

Transcription can be split into five separate stages (A-E), each involving specific modifications to the CTD of Pol II which leads to its interaction with the various processing machineries [4]. A summary of the different phosphorylation states of Pol II throughout transcription is given in figure 1.2, which is described below:

A. Hypophosphorylated Pol II (IIa) associates with transcription factors (denoted TFIID, B, E, F and H) to from the pre-initiation complex (PIC) at the gene promoter [23, 24].

B. Transcription is initiated following the phosphorylation of the serine residues at positions 5 (represented in red; figure 1.2) and 7 (green) of the CTD heptapeptide motif. This is catalysed by the kinase CDK7, which forms part of the transcription factor TFIIH. When the nascent transcript is approximately 20-40 nucleotides in length, the TFII initiation factors dissociate from the CTD, disbanding the PIC [19]. Serine-5 phosphorylation also signals the recruitment of the 5' capping machinery (pink oval) [22, 25, 26], with the addition of the cap preventing immediate degradation of the nascent transcript. A net reduction in serine-5 phosphorylation releases the capping machinery. This is followed by the CDK9- and CDK12/13-catalysed phosphorylation of serine-2 (blue), which converts Pol II into its hyperphosphorylated form (IIo) and signals the switch into the elongation phase of transcription [19, 22].

C. During elongation, hyperphosphorylated Pol II associates with specific elongation factors such as P-TEFb and TAT-SF1 [2, 27]. These factors recruit the splicing machinery (orange oval) and initiate intron removal, a more detailed description of which is given in sections 1.2.2.2 and 1.2.2.3.

D. As the transcription machinery moves towards the 3' end of the gene, the serine-5 residues are dephosphorylated by Ser5 phosphatase and this initiates the recruitment of polyadenylation factors (blue square). One such factor, CstF, plays an important role in the final stages of transcription, facilitating 3' cleavage and polyadenylation, transcription termination and transcript release. Throughout elongation, the activity of CstF is inhibited by its association with elongation factor PC4. As the transcription machinery nears the 3' end of the gene, PC4 releases CstF, allowing it to become functional [2, 28].

E. Final dephosphorylation of serine-2 residues by Ser2 phosphatase results in dissociation of the transcription machinery for re-initiation or recycling. Mutant pre-mRNAs, such as those incorrectly spliced, fail to release normally from the transcription machinery and instead accumulate at the site of transcription where they are targeted for degradation [29].

1.2.2.3 Splicing

Splicing, the process by which introns are removed from the pre-mRNA transcript, requires the formation of a spliceosome – a large complex comprising five core small ribonucleoprotein particles (snRNPs) denoted U1, U2, U4, U5 and U6 along with up to 200 other proteins [30]. The spliceosome components are recruited in a highly regulated, unidirectional order [31, 32] to specific recognition sequences at the 5' and 3' ends of the intron (the 5' and 3' splice sites or SS). The 5' SS defines the boundary between the upstream exon and the start of the intron and is usually signalled by a 'GU' dinucleotide motif. The 3' SS defines the boundary between the intron and the downstream exon and most commonly comprises an 'AG' dinucleotide motif. Two additional 3' intronic motifs are required for intron excision: the branch point sequence (BPS) consisting of a single 'A' nucleotide, and a polypyrimidine (Py) tract [4, 33, 34].

Spliceosome assembly requires the formation of a series of intermediate complexes in a four-step assembly process that results in the excision of the intron and the alignment of the upstream and downstream exons (figure 1.3A) [4]. The first complex is denoted complex 'E' (the commitment complex) and is formed by the binding of the U1 snRNP to the 5' GU dinucleotide signal and the co-operative binding of the SF1 and U2AF splicing factors to the BPS, the Py tract and the 3' AG motif [33, 34]. In the presence of ATP, the U2 snRNP binds to the BPS, forming the pre-spliceosomal complex 'A' [35]. A tri-snRNP comprising U5-U6-U4 binds to both the U1 and U2 proteins bound at the 5'SS and 3'SS respectively, forming a loop that brings the two exons into close proximity within complex 'B'. Subsequent RNA-RNA and RNA-protein rearrangements results in the release of U4 and U1 and the formation of complex 'C' (the catalytic

complex). Two *trans*-esterification reactions result in the excision of the intron and the ligation of the upstream and downstream exons [31].

Figure 1.3 The assembly of the mammalian spliceosome on pre-mRNA.
A: Spliceosome assembly requires the formation of four complexes denoted E, A, B and C. Each complex comprises interactions between specific splicing factors including the major snRNPs; B: Intronic and exonic splicing enhancers (ISE/ESE) and silencers (ISS/ESS) regulate splice site competition in alternative splicing. *Taken unchanged from Montes et al (2012) [4]*.

1.2.2.4 Alternative splicing

In an added layer of complexity, most mRNAs can potentially be spliced in a number of ways to produce different mature RNA messages coding for different isoforms of the same protein. This occurs via a process called alternative splicing, where specific exons are differentially removed from a subset of pre-mRNA transcripts. This is possible due to variation in the strength of splicing signals at intron-exon boundaries producing competition for spliceosome assembly. The 5' GU motif, the 3' AG motif, the BPS and the Py tract are all poorly conserved [36, 37] and the 5' and 3' splicing signals are open to modulation by numerous *cis*-acting silencer and enhancer sequences (figure 1.3B). These regulatory sequences can be either intronically or exonically located and are bound by specific RNA-binding proteins that interact with the spliceosome to facilitate or inhibit exon recognition, thus increasing or decreasing the splicing signal respectively [38]. A strong splice site will out-compete a weaker splice site for spliceosome assembly, leading to the weaker site being skipped and the intervening exon excised along with the introns. The pattern of exon inclusion/exclusion is defined by the 'splicing code', which is still not completely characterised and can vary among different tissues and at different stages of development [39-41].

1.2.3 Co-regulation of transcription and alternative splicing

1.2.3.1 Overview

It is now widely understood that transcription and splicing are not independent processes and are, in fact, physically and functionally coupled. In 1988 two groups used electron microscopy to show that intron removal occurred in nascent transcripts that were still tethered to their DNA templates [42, 43], a finding supported by further experiments comparing the splicing patterns of chromatin-tethered nascent RNA with that of RNA released into the nucleoplasm post-transcription [1]. Furthermore, fluorescent *in-situ* hybridisation of RNA molecules (RNA-FISH) using probes to distinguish between processed and unprocessed species demonstrated that intron removal occurs at or very close to the transcriptionally active template [44, 45].

Co-transcriptional coupling plays a vital role in the regulation of splicing and alternative splicing patterns, either by recruiting regulatory splicing factors to the nascent pre-mRNA (‘physical’ coupling) or by modulating splice site competition through the alteration of transcription rate (‘kinetic’ coupling). These two models are distinct, but not mutually exclusive.

1.2.3.2 Physical coupling

Physical coupling describes a mechanism by which the transcription and splicing machineries physically interact with each other and with components of the chromatin template. RNA polymerase II (Pol II) plays a major role in this coupling. As described earlier, the CTD of Pol II is vital in overseeing the production of mature mRNA transcripts, driving each stage of maturation from transcription and processing to transcript release and transport (section 1.2.2.2). The ‘CTD code’ of dynamic post-translational modifications sends important signals to the spliceosome regarding the progress of the transcription complex, allowing for the tight regulation of splicing factor recruitment and release throughout transcription [20-22].

A number of studies have confirmed that the CTD is required for pre-mRNA splicing, most notably by demonstrating that truncation of the CTD causes a significant reduction in splicing efficiency *in vivo* [17, 46, 47]. It has also been shown that Pol II-dependent initiation of transcription directly leads to the recruitment of splicing factors to the transcription site, but only when CTD integrity is maintained [48-50]. Further *in vitro* evidence has revealed that purified hypophosphorylated Pol II inhibits splicing during the initiation stage of transcription, whereas hyperphosphorylated Pol II activates splicing during the elongation phase [51].

The CTD alone is not sufficient to fully regulate splicing and its role is believed to be dependent on a number of adaptor or ‘coupling’ factors that are thought to interact with transcription and splicing components to regulate the physical coupling of the two machineries [32, 52, 53]. This, however, does appear to be

achieved through interaction of the coupling factors with the hyperphosphorylated CTD [7]. One class of coupling factors are the family of serine/arginine-rich proteins (SR proteins), of which splicing regulation is one of their primary functions. SR proteins have been shown to physically interact with the CTD of Pol II [54, 55] and *in vitro* studies have revealed that these proteins can partially enhance the co-transcriptional splicing of pre-mRNA transcripts, a role thought to involve the recruitment of the early spliceosome to the nascent transcripts [52, 56]. In addition, cell depletion of the essential splicing factor SC35 inhibited the recruitment of transcription elongation factor P-TEFb (section 1.2.2.2 C) to Pol II, reducing the level of CTD phosphorylation and impairing transcription elongation [57].

1.2.3.3 Kinetic coupling

Following the initiation of transcription, Pol II pauses at a site approximately 30-50 nucleotides downstream to the transcription start site. It is believed that this promoter-proximal pause – initially identified in the transcription of heatshock genes in *Drosophila* [58] – acts as a checkpoint to ensure only Pol II transcription complexes that have assembled correctly are allowed to enter the elongation phase of transcription [58-61]. This modulation of elongation rate has important implications for splicing events that are co-transcriptionally regulated. Following synthesis of a splice site, there is a certain period of time in which the spliceosome can functionally assemble on the site before it is subject to competition from a downstream splice site. Thus, a fast rate of elongation shortens the so-called 'window of opportunity' for spliceosome assembly and increases the likelihood of two splice sites being presented to the splicing machinery at the same time. In this scenario a weak site loses out to a stronger splice site, thus linking elongation rate to alternative splicing (figure 1.4) [3, 7, 62].

The first evidence of kinetic coupling was reported in 1988, with the discovery that the rate of transcription could influence alternative splicing by altering the secondary structure of the mRNA transcript [63]. It had previously been shown that exons lying within stem loop structures were more likely to be skipped [63-

66], due to the inability of the splicing machinery to access the splice sites. There is, however, a short period of time between transcription and RNA folding in which spliceosome assembly can take place. Thus, the slower the transcription rate (and the larger the stem loop), the more time the spliceosome has to assemble before sequestration of the 5' splice site into the loop [63].

Figure 1.4 The effect of transcription elongation rate on splice site recognition. Alternative exons are often preceded by a weak 3' splice signal (SS) which is subject to competition from a strong downstream 3'SS when elongation rate is high. Constitutive exons are preceded by a strong 3'SS and therefore out-compete their downstream counterparts regardless of elongation rate. *Taken unchanged from Kornblihtt et al (2004) [7].*

The artificial introduction of pause sites downstream to weak alternative splice sites has been shown to significantly increase exon recognition and inclusion [67]. The best evidence supporting the kinetic model, however, was gained from the study of the fibronectin extra domain I (EDI). The 3' splice site proximal to the 5' end of the EDI exon is degenerate [7] and therefore requires a low rate of Pol II elongation for preferential recognition and inclusion in the mature RNA transcript. When mutant Pol II enzymes were used to drive cell transcription *in vitro*, mutants conferring low rates of transcription elicited greater inclusion of the EDI exon than those demonstrating higher processivity [68]. Furthermore, the *C4*

Pol II mutant [69] was later shown to be associated with changes in alternative splicing of the *ultrabithorax* gene in *Drosophila*, highlighting a potential physiological and developmental function for kinetic coupling [68].

More recent evidence has shown that both the 5' and 3' ends of genes contain transcriptional pause sites, with Pol II pausing correlated with the recruitment of processing factors [70]. Pol II has also been shown to accumulate within the body of genes and this is greatest within the terminal exon, approximately 250 nucleotides upstream to the poly(A) site [71].

There is also a role for coupling factors within the kinetic model, acting as checkpoint regulators and influencing the length of Pol II pausing [72]. Indeed, putative coupling factor TCERG1 is thought to promote exon skipping in the *Bcl-x* gene by relieving Pol II pausing [5]. This may have important repercussions for disease, as insufficient time spent at pause sites increases the likelihood of the production of transcripts containing errors and/or being incorrectly processed.

1.2.3.4 Local regulation of co-transcriptional splicing by chromatin

The above sections have described global mechanisms for the co-regulation of transcription and alternative splicing. There are, however, tissue-, cell- and time-specific differences in the regulation of co-transcriptional splicing and these are believed to be influenced by local chromatin modifications. The major component of chromatin is the nucleosome, which describes a short stretch of DNA (approximately 147bp) wrapped around an octamer core of four histones (H3, H4, H2A and H2B) [73]. The precise positioning of the nucleosomes along the gene and reversible modifications to the core histones – including methylation, acetylation and phosphorylation [74] – have been shown to modulate transcription rate [75]. Compacted chromatin is a repressor of transcription, limiting access of the transcription machinery to the DNA template. Acetylation of the N-terminal tails of certain histones causes the chromatin to 'open' and is therefore regarded as a positive marker of transcription. This is because histone acetylation neutralises the charge of the basic histone proteins, leading to relaxation of DNA-protein

interactions and allowing access to the transcription machinery. Acetylated histone tails additionally promote transcription by acting as binding platforms for transcription factors [76].

A more intriguing finding, however, concerns the role of chromatin in co-transcriptional exon recognition and splicing. The average length of a mammalian exon is approximately 150 base pairs – strikingly similar to the length of DNA wrapped around each histone octamer [77]. Indeed, it has been shown that nucleosome positioning and histone modifications are closely correlated with the intron-exon structure of genes, with nucleosomes particularly concentrated around alternatively spliced exons (figure 1.5) [77, 78]. It is thought that nucleosomes act as 'speed bumps' and therefore an increase in accumulation causes a reduction in elongation rate; presumably followed by an increase in alternative exon recognition. Supporting evidence includes the finding that inhibition of the chromatin remodelling enzyme topoisomerase I by camptothecin results in Pol II pausing and increased splicing factor recruitment [79], and additional remodelling factors, SW1/SNF, have been shown to promote cluster exon inclusion in the CD44 gene [80].

Figure 1.5 Nucleosome positioning and intron-exon structure.
A: Transcription elongation is affected by changes in chromatin organisation brought about by chromatin remodelling factors (blue ovals), histone tail modifications (green star) and/or nucleosome position. A low elongation rate favours exon inclusion (yellow) whereas a high rate favours exclusion (red). B: Alternative splicing can also be influenced independently of transcription rate by the promotion of splicing factor recruitment through interactions between histone modifications (red star) and chromatin adaptors (orange/green shapes). ***Taken unchanged from Montes et al (2012) [4]*****.**

Histone modifying enzymes have also been shown to interact with components of the splicing machinery. The gene encoding the histone acetyltransferase (HAT) subunit of Gcn5 – a component of the transcription co-activating complex STAGA, which loosens chromatin and facilitates pre-initiation complex (PIC) formation – was found to share genetic interactions with the genes of two U2 snRNP-associated proteins Msl1 and Lea1. In fact it was shown that Gcn5 HAT activity was required for the co-transcriptional recruitment of U2 snRNP and its downstream interactors to the splicing branch point [76].

The 'histone code' refers to the specific pattern of post-translational histone modifications and provides important information required for the regulation of gene expression [81-83]. Such modifications play an active role in the coupling of transcription and alternative splicing and there is a growing body of evidence in support of this hypothesis. As described above, histone acetylation is required for the assembly of the spliceosome and recent studies have shown that deacetylase inhibitors, which prevent the reversal of histone acetylation, significantly alter the alternative splicing pattern of both reporter and endogenous genes [84, 85]. Acetylation, however, is not the only method of histone modification that influences splicing. Phosphorylation of histone H3 triggers the release of the SR protein coupling factor from chromatin during the cell cycle [86], whereas histone H3 tri-methylation has been shown to enhance the recruitment of splicing components [87]. In fact, enrichment of certain histone modifications at the intron-exon boundaries is thought to play an active role in exon recognition, with accumulation of histone 4 and histone 2B lysine methylation marking the 5' end and histone 3 tri-methylation marking the 3' end of exons [88, 89].

Thus, the histone code has an important role to play in the regulation of alternative splicing, with changes to histone methylation shown to influence the rate of alternative exon inclusion. In 2000, Carstens and colleagues studied histone modifications of the *FGFR2* gene and found that differences in the methylation pattern of histone 3 determined the differential inclusion of exons IIIb and IIIc. Tri-methylation of lysine residue 36 and mono-methylation of lysine 4

were enriched when exon IIIc was included, however, lysine 27 methylation and lysine 4 tri-methylation was correlated with the preferential inclusion of exon IIIb [90, 91].

Histone modifications, however, are unlikely to act alone and may function in conjunction with other factors to regulate splicing. Genome-wide association studies have shown that the transcriptional repressor and chromatin insulator CTCF binds downstream to alternative exons, in direct correlation with Pol II accumulation [92]. This suggests that insulation of chromatin – which prevents its conversion into an open structure – affects alternative splicing by modulating transcription elongation rate. An example of this is given by the *CD45* gene, where the specific histone methylation pattern that inhibits exon 5 inclusion is in complete opposition to the CTCF methylation pattern that promotes exon 5 inclusion [4]. As histone methylation patterns have been shown to fluctuate during development, this may provide a mechanism for the tissue-specific regulation of alternative splicing events through the differential recruitment of the CTCF chromatin modifier.

Thus, histone modifications have a role to play in both the physical and kinetic models of transcriptional coupling, with careful regulation of the conversion between closed and open chromatin conformations conferred by the histone modification code [93, 94] shown to affect both splicing factor recruitment and Pol II elongation rate.

1.2.3.5 The role of the promoter in alternative splicing regulation

The above sections describe the strong evidence supporting the coupling of the transcription machinery with the independent processing machineries, providing plausible cellular mechanisms for the co-transcriptional regulation of alternative splicing. There is another aspect to this regulation, however, which acts at the DNA level and is of particular relevance to the role of *MAPT* in PSP.

In 1997, Cramer and colleagues used the fibronectin gene to demonstrate that differences in promoter structure can influence alternative splicing patterns [10]. They created a number of constructs in which expression of the fibronectin gene (*FN*) was driven by different promoter elements *in vitro*. The *FN* gene contains the alternative exon EDI, the splicing of which changes during development and varies between cell types. The EDI exon contains a splicing enhancer, which increases the recognition of its sub-optimal upstream 3' splice site. When expression was driven by the α-1 globin promoter, the ratio of EDI+/EDI- was low, with exclusion favoured. When this promoter was replaced by the CMV promoter, EDI inclusion significantly increased (figure 1.6, left-hand lane). To ensure that differences in transcription start site did not influence alternative splicing, expression was driven by two variants of the *FN* proximal promoter, with one mutated to increase transcriptional activity. Both *FN* variants demonstrated significantly increased EDI inclusion compared to the α-1 globin promoter, however, the mutant promoter exhibited a 2.7-fold increase in EDI+/EDI- ratio compared to its wildtype counterpart (figure 1.6, right-hand lane) [10].

Figure 1.6 Fibronectin (*FN*) EDI exon inclusion with different promoters.
EDI+/EDI- ratio determined by reverse-transcription PCR (RT-PCR), Southern blot and Northern blot. α-1gb = α-1 globin promoter; CMV = cytomegalovirus promoter; wt = wildtype fibronectin promoter; mut = mutated fibronectin promoter. ***Taken unchanged from Cramer et al (1997) [10].***

This, for the first time, highlighted the importance of promoter structure in the regulation of alternative splicing. In fact, only five single nucleotide mutations in a 220bp promoter element were sufficient to significantly alter the EDI splicing pattern. Perhaps a more intriguing finding of this study was that up-regulation of transcription from each promoter type, leading to an increase in the overall abundance of mRNA transcripts, did not affect EDI splicing ratio. This suggests that the strength of the promoter is not relevant to the regulation of alternative

splicing in this gene; instead it is the nature of the promoter and thus the physical interactions with the processing machineries that regulates expression.

Further evidence supporting the role of promoter specificity in mRNA processing was gained following the creation of chimaeric constructs in which genes normally transcribed by Pol II were put under the control of promoters usually expressed by Pol III. *In vitro* analysis in mouse and kidney cells revealed that Pol III-synthesised mRNA was not subject to splicing, with transcripts still containing their introns despite the presence of consensus 5' and 3' splicing signals. Neither were transcripts polyadenylated, even though native 3' cleavage and poly(A) motifs were present [95].

Together, the evidence presented here for a co-transcriptional mechanism of mRNA splicing add layers of complexity to the study of gene expression in disease, as genetic variants found to differentially affect one process are likely to indirectly – or directly – affect numerous other processing pathways. This project aims to link together the transcription and splicing processes in the regulation of *MAPT* expression in an attempt to find a plausible mechanism that could form the pathway between a common *MAPT* variant and altered tau isoform expression in PSP.

1.3 The tauopathies

1.3.1 Overview

The tauopathies are a group of neurodegenerative diseases that are characterised neuropathologically by brain lesions comprising insoluble aggregates of tau protein. Alzheimer's disease (AD), progressive supranuclear palsy (PSP), corticobasal degeneration (CBD) and Pick's disease (PiD) are just a few examples of tauopathies, although significant clinical and pathological differences exist between them. The hallmark pathological feature linking these diseases is the abnormal intracellular accumulation of hyperphosphorylated tau and subsequent neuronal loss [96-98]. The specific factors that trigger tau aggregation remain

largely unknown. There is, however, now overwhelming evidence that common variation in the gene encoding tau (*MAPT*) can significantly influence disease risk, with a particularly strong effect observed for PSP [99-104].

1.3.2 Progressive supranuclear palsy (PSP)

PSP is a progressive neurodegenerative movement disorder, mainly sporadic, which commonly presents as atypical parkinsonism followed by dementia [105]. Diagnosed clinically as Richardson's syndrome (PSP-RS), classical symptoms include parkinsonism, supranuclear gaze palsy, postural instability with unexplained falls early in the disease course, and cognitive impairment [106, 107]. PSP is, however, a heterogeneous disorder and clinical variants of the PSP phenotype include PSP-parkinsonism (PSP-P), in which bradykinesia and dystonia are characteristic, PSP-pure akinesia with gait freezing (PSP-PAGF), and PSP-corticobasal syndrome (PSP-CBS) and PSP-non-fluent aphasia (PSP-NFA), where cortical degeneration is more pronounced [106-108]. Although rare (with a prevalence of 3.1-6.5 per 100,000 people [109]), PSP is the second most common cause of parkinsonism after Parkinson's disease [11, 110]. The average age at onset is 63 years with a disease duration of around six to seven years before eventual death [111]. At present disease-modifying options are limited and treatment instead focuses on the management of individual symptoms [107, 112].

PSP is classed as a primary tauopathy as tau is the only abnormal protein observed in the brain post mortem. This separates PSP from secondary tauopathies, such as Alzheimer's disease (AD) where tau pathology is accompanied by amyloid plaques [106]. Characteristic neuropathological features of PSP include neurofibrillary tangle (NFT) formation and neuronal loss in the basal ganglia, diencephalon and brainstem, with the substantia nigra, the globus pallidus and subthalamic nucleus most affected [105]. The NFTs in PSP are composed of straight filaments of hyperphosphorylated tau and are distinct from the paired helical filaments forming the NFTs of AD. Tau-positive inclusions in oligodendrocytes and tufted astrocytes are also typical of PSP [105].

PSP is sporadic in most cases, although approximately seven percent of patients have a positive family history of parkinsonism or dementia, consistent with an autosomal dominant pattern of inheritance [106]. The G303V mutation, located within exon 10 of the tau gene, has been associated with PSP in one large family and demonstrates autosomal dominant inheritance among the affected members. Several features atypical of PSP, however, including a much lower age of onset (average 40.3 years), has raised questions over the reliability of the PSP diagnosis in this study [106]. One member of a PSP family with autosomal dominant PSP was found to have a novel L284R mutation [113], and a recent study of *MAPT* in Asian PSP families identified four mutations in six individuals associated with sporadic early onset PSP, including one *de novo* mutation. The latter study reported that the *MAPT* mutations were only found in patients exhibiting abnormal eye movements – additional to supranuclear gaze palsy and not characteristic of sporadic PSP – and may suggest co-morbidity with an underlying, secondary disorder [114]. The above mutations occur extremely rarely in PSP and most sporadic cases do not have an identifiable genetic cause. There are, however, common genetic factors that can increase an individual's risk of developing PSP.

It is thought that a combination of environmental and hereditary factors modulate an individual's risk of developing PSP [108]. Repetitive brain trauma has been shown to cause the progressive tauopathy, dementia pugilistica [115], and a recent genome-wide association study (GWAS) has identified a number of genetic risk factors for PSP (see section 1.6.2.3) [104]. Demographic factors including gender, ethnicity, geographical location and occupation do not appear to influence PSP risk [104], though an association with low education levels has been reported [116]. The biggest risk factor for PSP, however, is located within the gene encoding the cellular protein, tau.

1.4 Microtubule-associated protein, tau

1.4.1 Function

Tau belongs to the family of microtubule-associated proteins (MAPs) that bind to tubulin and regulate its assembly into microtubules – the dynamic cytoskeletal tracks that are vital for cell transport, shape and polarity [117-119]. MAP expression is specific to cell type and individual MAPs can function either to stabilise or destabilise microtubules [120]. Activity is regulated by phosphorylation, with the addition of phosphate groups to specific protein residues resulting in its detachment from tubulin and subsequent microtubule de-stabilisation.

Tau is a developmentally regulated protein of around 50-65kD in size [68, 121]. It is most abundant in the neurons of the central nervous system (CNS) where it is enriched in axons [121-123]. It is also expressed to a much lesser extent in glia, astrocyctes and oligodendrocytes and in certain tissues outside the CNS [120, 124]. Its primary role is to promote and stabilise the polymerisation of tubulin into microtubules [125, 126], which is of particular importance in the long axonal extensions that are a unique feature of neurons.

The tau protein has been shown to function synergistically with the other major MAP family member, MAP1B. Double knock-out mice (*MAPT*-/-, *MAP1B*-/-) demonstrate severe defects in microtubule extension and neuronal migration, a phenotype that is much less severe in single knockout mice (*MAPT*+/+, *MAP1B*-/-) [127]. Indeed, tau knockout mice (*MAPT*-/-) do not exhibit major brain defects, suggesting the presence of a compensatory mechanism in which *MAPT* function is fulfilled by other MAPs [128].

1.4.2 Tau isoform expression

In the adult brain, the tau protein has six major isoforms, each characterised by the presence or absence of two N-terminal inserts and one of the four C-terminal microtubule binding repeat domains. The inclusion of the extra C-terminal

binding repeat domain produces 4R-tau isoforms, with 3R-tau the result of its specific exclusion. It has been shown that 4R-tau has a three-fold stronger binding affinity for microtubules than 3R-tau [124] and is also more fibrillogenic [129, 130], though all six tau isoforms are capable of forming pathological aggregates [131].

Tau isoform expression changes throughout development, with the foetal brain containing only the shortest 3R isoform as a result of the constitutive exclusion of the two N-terminal inserts and the C-terminal binding domain. In the healthy adult brain, all six tau isoforms are expressed, albeit in slightly different abundances. In 2003, Takuma and colleagues compared the tau isoform pattern in brain tissue from a 20-week human embryo and a 70-year old adult (figure 1.7) [132]. Using the pool-2 antibody against all six isoforms (figure 1.7A), foetal tau was shown to comprise only the shortest 0N3R isoform, as expected. The elderly brain, also as expected, contained all six isoforms, with the 1N3R and 1N4R isoforms most abundant (3rd and 4th bands from the top, right hand lane). The 2N isoforms (2N3R and 2N4R) were the least abundant and were barely visible on the Western blot (top two bands, right hand lane). The absence of 1N, 2N and 4R isoforms in foetal tau was confirmed using antibodies specific for these inserts (figure 1.7, panels B and C).

Figure 1.7 Tau isoform expression in the human brain
Western blots of tau protein extracted from human brain tissue of a 20-week embryo (E20w) and a 70-year old adult (70yr). A: Reactivity of global tau antibody, pool-2; B: Reactivity of three antibodies specifically targeting 0N (top), 1N (middle) and 2N (bottom) isoforms; C: Reactivity of an antibody specifically targeting 4R isoforms. ***Adapted from Takuma et al (2003) [132].***

This suggests that the tau isoforms have differing functions, with alterations in isoform production made to meet the changing tau-microtubule interactions

required throughout development [124]. A recent study of isoform expression in the adult brain has confirmed that ~50% of total tau comprises 1N isoforms, with ~40% representing 0N isoforms and just 10% containing both exons 2 and 3 (2N). This isoform pattern was consistent in most of the brain regions (figure 1.8; top panel), with the only significant change comprising a reduction in 0N3R in the cerebellum. The ratio of 4R- and 3R-tau expression was approximately equal in all brain regions, as expected [133]. This research, however, has been followed by a similar study, which reported a significant increase in 4R-tau expression in the occipital lobe and globus pallidus compared to four other regions (figure 1.8; bottom panel).

Figure 1.8 Expression of the tau isoforms in different adult brain regions.
Top: Expression of all six isoform (C-H) measured in cerebellum (CRBL); frontal cortex (FCTX); occipital cortex (OCTX); putamen (PUTM) and white matter (WHMT). ***Taken from Trabzuni et al (2012*** **[133].**
Bottom: The ratio of 4R-tau/total tau isoforms by H1 (blue) and H2 (green) haplotypes measured in frontal cortex (FC), temporal cortex (TC), Pons, cerebellum (CB), occipital lobe (OL) and globus pallidus (GP). ***Taken from Majounie et al (2012)*****[134].**

1.4.3 Tau phosphorylation

The biological activity of tau is regulated by its phosphorylation state [135]. The addition of phosphate groups to specific residues generally reduces the binding affinity of tau for tubulin and thus microtubules formed in the presence of phosphorylated tau (p-tau) are usually less stable than those formed with the unphosphorylated species [117, 129]. As with isoform production, the phosphorylation of tau is dynamic and developmentally regulated. Foetal tau is highly phosphorylated, significantly more so than the tau of the adult brain [136]. In the healthy adult brain, the level of tau phosphorylation is thought to decrease with age, though some evidence suggests this finding may be the result of de-phosphorylation post mortem [137]. In the tauopathy brain, however, the tau species forming the neurofibrillary tangles (NFTs) is hyperphosphorylated and resembles the phosphorylation pattern found in the foetal brain. [138].

The longest isoform of tau has 79 potential phosphorylation sites at serine and threonine residues, with phosphorylation confirmed to occur at over 50 of them [139-142]. Five tyrosine residues have also been shown to be phosphorylated [142]. These residues are phosphorylated by a number of different kinases, many of which have been implicated in neurodegenerative disease. These include the major serine-theronine kinases: glycogen synthase kinase 3β (GSK-3β), cyclin-dependent kinase 5 (cdk5), cyclic AMP-dependent protein kinase A (PKA), protein kinase N (PKN), microtubule affinity regulating kinase (MARK) and the mitogen-activated protein kinase (MAPK) and casein kinase 1 (CK1) families [129, 140, 142, 143]. Several tyrosine kinases have also been associated with tau pathology including Fyn, c-Abl and Syk [144-146]. The dual serine/threonine and tyrosine kinase, tau-tubulin kinase 1 (TTK1), has been implicated in Alzheimer's disease [140].

There is a large body of research concerned with the characterisation of tau phosphorylation and its role in neurodegeneration but as this is not directly relevant to this project it will not be discussed further here.

1.4.4 Tau aggregation

Tau aggregation is thought to occur following a change in the conformation of tau monomers that leads to hydrophobic sections of the protein becoming exposed. This allows contact between monomers at these hydrophobic sites, resulting in their association into oligomers, and eventually filaments [131]. These filaments (NFTs) are a feature of normal ageing, but occur with much greater frequency in the tauopathy brain [124]. The heat-shock protein 70 (Hsp70) family are a group of molecular chaperones that work to prevent abnormal tau aggregation. They bind to the exposed hydrophobic regions of tau and assist in its refolding. Voss and colleagues have demonstrated that Hsp70 directly inhibits the aggregation of all six tau isoforms, without affecting normal microtubule formation. They also showed that Hsp70 was more effective at inhibiting 3R-tau than 4R-tau isoforms and that the 2N3R isoform was inhibited at a lower Hsp70 concentration than the other 3R isoforms, 0N3R and 1N3R. All three 4R isoforms displayed inhibition at similar Hsp70 concentrations. Heat shock proteins are up-regulated in response to cellular and environmental stresses and Hsp70 expression has been shown to be increased in AD [147, 148].

1.5 The *MAPT* gene

1.5.1 Structure

Tau is encoded by the *MAPT* gene which is 134kb in size and located on chromosome 17q21.1 [122]. It consists of 16 exons, with exons 4A, 6 and 8 absent from most brain transcripts and exons -1 and 14 untranslated [149]. The alternative splicing of exons 2, 3 and 10 produces the six major isoforms of tau expressed in the adult brain (figure 1.9).

Exon 4A is the largest of the tau exons and, along with exon 6, is included in the tau isoform preferentially expressed in the retina, spinal cord and peripheral nervous system where it is referred to as high molecular weight tau (or ‘big tau’) due to its large size (approximately 110 kDa). Big tau is observed in the brain, albeit at much lower levels and in a different regional pattern to that of the major alternatively spliced exons 2, 3 and 10 [150]. Exon 6 contains three potential

splice sites and brain-specific changes in exon 6 splicing have been implicated in myotonic dystrophy type 1 (DM1), along with more minor changes in exon 2 and exon 10 splicing [150-152].

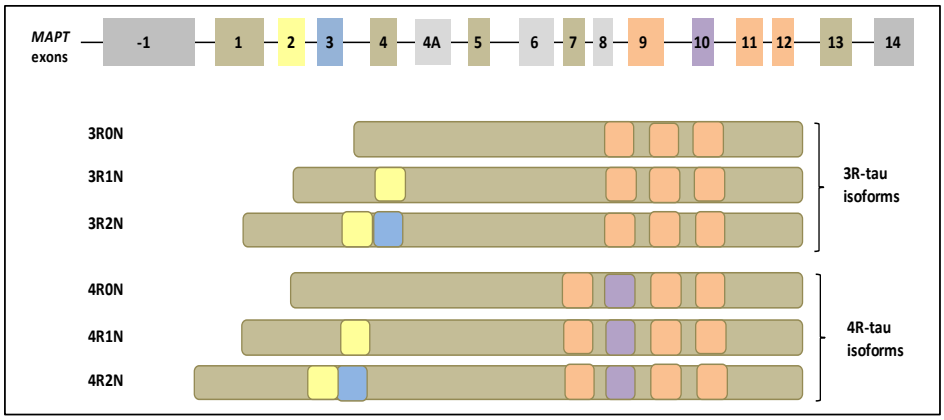


Figure 1.9 The structure of the *MAPT* gene and the six major tau isoforms.
Exons 4A, 6 and 8 are transcribed in very low abundance in the brain. Exon -1 constitutes the *MAPT* promoter region and is non-coding. Exon 14 constitutes the 3'UTR. Exons 9-12 each encode one of four microtubule binding domains. Exons 2, 3 and 10 are alternatively spliced giving rise to six isoforms which can be split into two groups based on the presence or absence of the binding domain encoded by exon 10.

1.5.2 Exons 2 and 3

Exons 2 and 3 each encode 29 amino acid N-terminal inserts that form part of the protein's acidic projection domain [153]. This domain is believed to interact with the plasma membrane and is also thought to regulate the spacing between microtubules, potentially acting as a polymer brush or spring to keep the microtubules apart. It has been suggested that failure of the polymer brush could lead to tau aggregation [154].

The alternative splicing of exons 2 and 3 produce three N-terminal isoforms denoted 0N (2-/3-), 1N (2+/3-) and 2N (2+/3+). These exons demonstrate so-called 'incremental combinatorial' splicing, which describes a situation where the downstream exon of an alternatively spliced pair is never present alone. This is a very rare occurrence and other reported instances include exons 7 and 8 of the amyloid precursor protein (APP) – the major pathological protein in AD – and the neuronal-specific N1 and N2 exons of the gene encoding the tyrosine-kinase Src.

This suggests that such splicing pairs have an important function in neuronal cells and that a shift in the balance of the expression of these exons may contribute to neurodegeneration.

In 1995, Andreadis and colleagues created splicing constructs that placed *MAPT* exons 2 and 3 between insulin exons 2 and/or 3. *In vitro* expression was driven by the simian virus (SV) promoter. They demonstrated that exon 2 behaves as a constitutive exon as it was always present in expressed mRNA regardless of whether the surrounding exons were from the tau or insulin genes or whether the *in vitro* cell line was neuronal or non-neuronal. Exon 3 was inefficiently incorporated into the mRNA transcripts unless either the insulin splicing signals were modified or exon 2 was pre-spliced downstream to exon 3. Exon 3 inclusion was detected, however, when exon 2 was physically absent from the construct, though inclusion significantly increased when exon 2 was present. In this instance the *in vivo* expression pattern was recapitulated and ex2-/ex3+ was not detected [155].

Although most studies have focussed on the splicing of exon 10, the inclusion/exclusion rates of the N-terminal exons has been shown to be altered in the brains of individuals with certain variants of the *MAPT* gene and this will be discussed in more detail in section 1.7.3.

1.5.3 Exon 10

Exon 10 encodes one of four microtubule-binding domains located in the C-terminal half of tau. Each domain is 31-32 amino acids in length and is encoded by one of four imperfect repeats constituting exons 9-12 [11]. Thus, inclusion of exon 10 produces tau protein isoforms with four microtubule-binding domains (4R-tau) and its exclusion produces isoforms with three (3R-tau). The two isoform groups appear to form distinct structures with complex intramolecular folding interactions, suggesting that they may have different functions [135]. This is consistent with the changes in tau isoform expression that occur during development (figure 1.6), with the exclusive expression of 0N3R in the foetal

brain suggesting that 3R-tau is important in neuronal plasticity [118].

Another piece of the *MAPT* puzzle was presented by Chen and colleagues, who investigated the effect of exon 10 inclusion on the expression level of *other* genes. Whole genome expression profiling of SH-SY5Y neuroblastoma cells over-expressing either 4R- or 3R-tau detected a decrease of transcripts involved in embryonic development when exon 10 was present, accompanied by a corresponding up-regulation of transcripts related to neurite outgrowth. The *Wnt* signalling pathway – which has been implicated in AD – was also shown to be differentially altered by the presence/absence of exon 10 [156].

Changes in exon 10 inclusion have been massively implicated in tauopathy pathogenesis and most of the *MAPT* mutations leading to FTDP-17 (section 1.6.1) are located in or around exon 10. The pathological consequences of changes to the rate of exon 10 inclusion are discussed below.

1.5.4 Tau pathology

The healthy adult brain has approximately equal amounts of 3R- and 4R-tau [153]; however in some tauopathy brains this ratio is disrupted with a predominance of one isoform group over the other. The different tauopathies are therefore classed according to the direction of the isoform ratio change, with PSP and CBD classed as 4R tauopathies due to their observed shift towards 4R-tau production. Conversely, Pick's disease is a 3R tauopathy. Over-expression of 4R-tau in transfected cell lines results in the displacement of 3R-tau from microtubules, indicating that a change in isoform regulation that favours 4R production is likely to impair both the properties of the microtubules and microtubule-dependent functions [122].

The splicing factor polypyrimidine tract-binding protein 2 (PTBP2) plays an important role in 3R-/4R-tau production, with knockdown in Neuro2a cells causing a significant reduction in 4R-tau levels and 4R:3R ratio. PTBP2 expression is decreased following binding of the microRNA miR-132 to its

3'UTR. Interestingly, miR-132 expression was found to be reduced in the brains of PSP patients compared to controls [157].

Alzheimer's disease (AD) – a secondary tauopathy – does not exhibit an overall shift in 3R/4R ratio. Recent evidence, however, suggests that tau isoform profiles in the hippocampal pyramidal neurons of AD brains are not homogeneous across the neuronal population and actually vary from one neuron to another. Double immunofluorolabelling of the 3R- and 4R-tau species has shown that each of the three tau isoform profiles (3R+/4R+, 3R+/4R- and 3R-/4R+) are present in the neuronal population and correlate with distinct neuronal morphologies. In line with previous observations [158-162], neurons exclusively expressing 4R-tau (3R-/4R+) resembled the so-called 'pre-tangle' stage, staining positively for diffuse cytoplasmic tau but without the neurofibrillary structures. Neurons exclusively expressing 3R-tau (3R+/4R-) had loosened, widely-spaced parallel fibres commonly termed 'ghost tangles'. Neurons expressing both 3R- and 4R-tau (3R+/4R+) exhibited the tight fibrillary structures of typical NFTs [163].

In PSP brains, cortical neurons displayed the 3R-/4R+ pre-tangle morphology, as expected. In the substantia nigra and midbrain, however, 3R-tau was detected in low abundance [164, 165], though double immunofluorolabelling showed it was only present alongside 4R-tau in NFTs (3R+/4R+) [163]. In CBD brains, 3R-tau reactivity was more widespread, including in a small number of structures in the cortex, but was still at much lower abundance than – and co-expressed with – 4R-tau [165].

These results together show that tau expression profiles differ not only between different regions of the brain and neuronal populations, but also between different tauopathies. It has been suggested that the 4R-selective cortical neurons in PSP and CBD may represent early tau deposition, with the superimposition of 3R-tau onto 4R-tau in nigral neurons indicating more advanced tau deposition. It would therefore follow that the presence of 3R-selective neurons in AD indicates further advancement of tau deposition and the absence of such neurons in PSP and CBD

may account for the lack of ghost tangles in these patients [163]. It is still unclear as to the exact contribution of each isoform to tau aggregation, though recent biochemical studies with synthetic tau have suggested that isoform recruitment to tau aggregates via a seeding model is dependent upon the initial composition of the seed. Pre-formed seeds of 3R-tau recruit both 3R and 4R isoforms to form larger aggregates; whereas aggregates seeded by 4R-tau exclusively recruit 4R isoforms [163].

These data provide an insight into the pathological consequences resulting from alterations to the inclusion rate of exon 10 in the tau protein. This project, however, focuses on elucidating the molecular mechanism behind these protein changes and to achieve this an understanding of the genetic complexities in and around the *MAPT* gene is required.

1.6 The genetics of *MAPT*

1.6.1 Genomic architecture

The tau gene, *MAPT*, falls in a structurally complex region on the long arm of chromosome 17, where a stretch of complete linkage disequilibrium (LD) spans approximately 2Mb and encompasses several genes [97]. LD describes a state in which genetic variants at differing locations are inherited together more often than would be expected under normal random inheritance. This usually applies to variants located within close proximity to one another and therefore the large region of LD on chromosome 17 is a somewhat unusual phenomenon. This complexity stems from an ancient 900kb inversion of the *MAPT* region – believed to have originated in European Caucasians approximately 18 to 45 thousand years ago [166, 167] – which resulted in the evolution of two completely separate haplotype clades, denoted H1 and H2. These haplotypes span the length of the *MAPT* gene and beyond [11, 99, 168] and there is a complete absence of recombination between them, with H2 remaining invariant [168, 169]. This haplotype is the rarer of the two, with a frequency of up to 30% in Caucasians, and is almost completely absent in East Asian, Native American and African

populations [97]. The H1 haplotype, on the other hand, is prevalent in all populations and represents the ancestral sequence. There is a normal pattern of variation and recombination within the H1 clade and several sub-haplotypes exist [11, 169].

As described earlier, mutations in *MAPT* have been shown to cause the tauopathy FTDP-17 [98, 170] and, in extremely rare cases, have been shown to cause a phenotype similar to that of PSP [105]. In 1999, Hasewaga and colleagues found that two FDTP-17-associated missense mutations located within exon 10 (N279K and S305N) affected tau function, not by altering the strength of the exon 10 microtubule-binding domain, but by modifying its inclusion rate [171]. This indicated a role for RNA processing in neurodegeneration, with mutations that lead to disturbances in protein isoform homeostasis sufficient to cause neurodegenerative disease. Since then a number of exon 10 mutations have been identified and shown to alter the ratio of 3R-/4R-tau expression. The key mutations and their pathological consequences were summarised by Niblock and Gallo in a recent review [172].

Additionally, a recent study has linked the V363I mutation in exon 12 of *MAPT* with frontotemporal dementia, but only in one individual that also carried the A/A genotype of the rs9897526 progranulin polymorphism and demonstrated homozygosity for the methionine amino acid at codon 129 of the prion protein. Individuals from the same family that carried the V363I mutation but not the two additional genotypes did not develop FTD by the time of the study [173].

These findings have confirmed that tau dysfunction plays an important role in neurodegenerative aetiology, whether directly or in combination with other disease modifying factors. In most cases, however, tauopathies are sporadic with no known causal mutations and genetic studies have instead focused on identifying polymorphisms that modify risk.

1.6.2 Common *MAPT* variation and PSP

1.6.2.1 Early association studies

The first published association of *MAPT* with PSP was of a dinucleotide repeat polymorphism in intron 9, the A0 allele of which was found to be over-represented in PSP cases [102]. This allele is found on the H1 background and after further analysis the association was expanded to include the entire H1 haplotype – an association that has been consistently replicated in Caucasian populations [99, 101, 104, 105, 174]. In addition, H2 was shown to be protective against PSP, perhaps explaining why this haplotype appears to be under positive selection in the Caucasian population [97]. Further associations of H1 were reported in CBD and AD populations and, surprisingly due to the lack of tau pathology, in several Parkinson's disease (PD) studies [11, 175-180]. These will be discussed in more detail in section 1.6.2.4. Subsequent high density LD mapping identified a number of H1 variant haplotypes and lead to the refinement of the PSP association to H1C, one of the more common H1 sub-haplotypes [169, 181, 182].

1.6.2.2 The H1C haplotype and rs242557

The H1C haplotype is tagged by the minor A-allele of a single nucleotide polymorphism (SNP) denoted rs242557. The frequency of this allele is around 30-40% in Caucasian populations, which is slightly higher than the frequency of the H1C haplotype (23.5%) [181]. This is due to the presence of the A-allele on several of the minor *MAPT* haplotypes, though H1C is the only common haplotype to carry this allele. H1C is one of only three major *MAPT* haplotypes that have a frequency greater than 5% (H1B and H1C =23.5%; H2 =17.6%) [181]. The A-allele of rs242557 appears to drive the association of H1C with PSP [11, 104, 181-183], with effect sizes of 1.8 and 2.4 reported in UK and US Caucasian populations, respectively [181]. Thus, individuals carrying the A-allele are approximately twice as likely to develop PSP than those carrying the G-allele.

The rs242557 polymorphism is located in a highly conserved region of exon -1, approximately 47kb downstream to the *MAPT* core promoter and 20kb upstream to the first coding exon. It is predicted to fall in or near to a *cis*-acting transcription regulatory domain and this therefore suggests that the influence of this allele on PSP risk stems from modifications to the regulation of *MAPT* transcription. This will be discussed further in section 1.7.

1.6.2.3 The PSP genome-wide association study

In 2011, a genome-wide association study (GWAS) [104] added to growing evidence supporting the association of rs242557 with PSP risk. The study looked at over half a million SNPs in 1,114 pathologically confirmed PSP cases and compared them to 3,287 healthy controls (Stage 1) followed by replication with an additional 1,051 clinically diagnosed PSP cases (Stage 2). The most strongly associated region of the genome was 17q21.31 where, in Stage 1, 56 SNPs covering a 1Mb region reached the genome-wide significance threshold of $p=5x10^{-8}$. This region contains the *MAPT* gene and the association was replicated in Stage 2. In the combined analysis the most strongly associated SNP in this region was rs8070723 ($p=1.5x10^{-118}$), a proxy for the H1/H2 inversion. More in-depth analysis of the region revealed that most of the associated SNPs mapped either directly or closely to the inversion itself. Therefore, as expected, most of the associated SNPs became non-significant when rs8070723 was used to control for inversion status. Some SNPs, however, did remain and of these rs242557 was the most strongly associated ($p=8.5x10^{-18}$), with the A-allele conferring a 1.4-fold increase in PSP risk. This confirmed that rs242557 makes an important contribution to PSP risk that cannot solely be accounted for by the H1/H2 inversion. In this, the most comprehensive study on the genetic risk factors of PSP completed to date, the H1 *MAPT* haplotype was shown to increase disease risk by 5.5-fold, making the magnitude of its effect equivalent to that of the ε3/ε4 *APOE* genotype in AD.

Three other genes were shown to confer increased risk of PSP in the GWAS: *MOBP, STX6* and *EIF2AK3;* though little progress has been made regarding their role in PSP risk.

1.6.2.4 *MAPT* haplotypes in other neurodegenerative disorders

In addition to its strong association with PSP, the rs242557-A allele has been associated with various other related neurodegenerative disorders including some outside of the tauopathy family. The strongest of these associations (outside PSP) is with corticobasal degeneration (CBD), a rare primary tauopathy that shares many similarities with PSP, including hallmark 4R-tau pathology [176, 181]. Accurate genetic studies of CBD are difficult due to the rarity of the disorder; however, one study calculated the effect size of the rs242557-A allele on CBD risk to be around 2.2 [181]. Further associations of rs242557-A have been reported with Guam amyotrophic lateral sclerosis (ALS-G), parkinsonism dementia complex (PDC-G) and dementia (GD), though the effect sizes are much smaller (1.03-1.5) despite each displaying significant tau pathology [184].

Investigation of the role of the *MAPT* haplotypes in Alzheimer's disease (AD) has produced mixed results, though many positive findings have been reported. An initial report found an association of the H1C haplotype with AD in two series of cases in which age of onset was >65 years [177] and this was independently replicated [185]. A recent finding has also associated the G-allele and G/G genotype of rs242557 with an increased risk of late-onset AD in a large Han Chinese population, with moderate effect sizes of 1.16 and 1.13, respectively [186]. This is puzzling as the A-allele is generally believed to be the risk allele of rs242557 and therefore this G-allele association may result from the absence of the H2 chromosome in Asian populations.

Several studies have failed to replicate these rs242557 associations with AD [187-189]. One study did, however, find a significant association of the combined genetic effects of the rs242557 A/A genotype and the T/T genotype of the rs2071746 polymorphism in the heme oxygenase-1 (HO-1) gene. In fact the

combined effects of these two genotypes conferred a 6.5-fold increase in AD risk for individuals who carry these markers compared to those who do not (p=0.037) [190]. The HO-1 gene is involved in the oxidative stress pathway, an intriguing finding as increases in oxidative stress have previously been shown to play a fundamental role in the aggregation of tau into NFTs and in AD risk [191, 192]. Another association study of the *MAPT* haplotypes in AD reported that the H1 haplotype was associated with reduced NFT pathology [193]. This was a surprising finding and is the opposite of what would be expected based on the disease associations of the haplotype identified to date. If this finding is replicated in independent studies, it would have important implications for our understanding of disease mechanisms in AD.

There is no robust evidence supporting an association of the H1C haplotype with sporadic Parkinson's disease (PD) [194, 195], although one association of rs242557 was detected in a Finish population [196]. The H1 haplotype, however, has been repeatedly shown to be over-represented in PD cases [189, 194-199]. The association of the *MAPT* gene with PD is surprising, as these patients do not exhibit tau pathology. The repeated association of the H1 haplotype strongly supports a role for this haplotype in PD; however, the apparent inability to refine this association to one of the H1 sub-haplotypes may indicate that the effect on PD risk arises as a consequence of the H1/H2 inversion affecting *MAPT* expression, rather than from *cis*-acting variation from within the gene [175, 183].

Interestingly, the H1 haplotype does not appear to be associated with frontotemporal dementia (FTD) [200], suggesting that common variation within *MAPT* does not affect risk and therefore the contribution of this gene to FTD arises solely from pathogenic mutations.

1.6.2.5 The effect of rs242557 on CSF tau levels

An increasing number of studies have investigated the effect of the rs242557 alleles on tau protein levels in the cerebrospinal fluid (CSF) of patients with neurodegenerative disease, again with mixed results. Two studies found an

association between rs242557-A and an increase in CSF tau in AD patients, though one added the caveat that this association was dependent on both the presence of dementia and low CSF β-amyloid levels [201]. This study also reported an association of the A-allele with increased phospho-tau levels. The second study provided significant evidence for the role of rs242557 in tau expression by demonstrating that the association of the A-allele was gene-dosage dependent, with an increasing copy number directly leading to proportional increases in CSF tau levels. Furthermore, sliding window analysis of the region pinpointed the causal variant to lie at or proximal to the rs242557 polymorphism [185]. This provides direct evidence that the increase in disease risk is actually conferred by rs242557-A, rather than a variant in LD with the polymorphism. That being said, one study did not find an association between this polymorphism and CSF tau in AD but did report associations with other polymorphisms from the *MAPT* gene [202].

The rs242557 polymorphism has also been shown to increase tau levels in CSF in both PSP/CBD and PD cases, though this role was played by opposing alleles depending on the disease. In PSP and CBD the A-allele, as expected, conferred the increase, whereas the G-allele increased CSF tau levels of PD patients [203]. The reasons for these differential associations are unclear but the fact that they were detected in the same study using the same methods suggests that they might be real. Replications of these findings have yet to be reported, however.

The above sections have provided overwhelming genetic evidence supporting a significant role for the rs242557 polymorphism in disease risk. The location of the polymorphism within a highly conserved region of the *MAPT* intron -1 suggests that risk may by conferred through changes to *MAPT* expression. Focus, therefore, has shifted to elucidating the functional mechanism behind the rs242557 association with neurodegenerative disease.

1.7 *MAPT* gene expression

1.7.1 *In vivo* allele-specific expression studies

In vivo studies of *MAPT* haplotypes – primarily allele-specific expression studies from post mortem brain tissue – have produced mixed results, with some reporting an increase in *MAPT* transcription from H1 chromosomes compared to H2 chromosomes [12, 204] and others not finding any difference at all [111]. One group [12] reported a further increase in expression specifically for H1C chromosomes compared to non-H1C chromosomes, though this has yet to be replicated [111, 204]. An interesting finding from these studies is that, where an increase in expression from H1 or H1C was observed, it was accompanied by an increase in the number of exon 10+ (i.e. 4R-tau) transcripts relative to exon 10- (i.e. 3R-tau) transcripts [12, 204]. This suggests that the H1/H1C association may actually exert its effect though two molecular processes – *MAPT* transcription and alternative splicing – and that these processes may be linked.

A recent eGWAS study – a genome-wide association study of quantitative trait loci (QTLs) – provided the most comprehensive assessment to date of the effect of genetic variation on gene expression in the brain [183]. The authors analysed gene expression in tissue samples from the cerebellum and temporal cortex of 400 autopsied patients with PSP, PD or AD. Overall, an enrichment of *cis*-acting SNPs (SNPs that alter expression of the gene in which they are located) was detected amongst disease-associated genes. Decreased *MAPT* expression was associated with a lower risk of AD, PD and PSP, with 78% of identified SNPs demonstrating concordant effect sizes between the cerebellum and temporal cortex. The rs242557-A allele conferred a significant increase in both PSP risk and *MAPT* expression (cerebellum: $p=9.78\times10^{-3}$ to 8.8×10^{-13}; temporal cortex: $p=1.1\times10^{-8}$), with the H2-defining rs3070723 minor allele conferring reductions in PSP risk and *MAPT* expression. This provides further confirmation of the detrimental and protective roles of the H1C and H2 haplotypes, respectively. The H2 polymorphism was also associated with reduced *MAPT* levels in PD, supporting a potential role for the H1/H2 dichotomy in this disease.

Two other genes previously identified in the PSP GWAS study (section 1.6.2.3) were also found to confer allelic differences in transcript expression levels in the brains of PSP patients. A reduction in *SLCO1A2* and an increase in *MOBP* expression were both associated with an increase in PSP risk. Thus, variation within the *MAPT*, *SLCO1A2* and *MOBP* genes has been shown to be associated with both an increase in PSP risk and changes in the expression of their respective genes in the PSP brain. This suggests that changes in brain expression patterns are an important component of PSP aetiology.

1.7.2 *In vitro* luciferase reporter gene studies

Several studies have examined the effect of rs242557 and the H1C haplotype on *in vitro MAPT* expression, using varying methodologies and with varying results. Two groups [11, 12], including our's, conducted *in vitro* luciferase reporter gene assays to study the allelic effect of rs242557 on the *MAPT* core promoter, with interesting results that differed on two major points. Firstly, Myers and colleagues (our group) [12] reported that the A-allele of rs242557 conferred significantly higher *MAPT* transcription than the G-allele, with the greatest difference observed between the A-allele on the H1 promoter background and the G-allele on H2. However, Rademakers et al [11] found that the G-allele of rs242557 conferred higher expression from the H1 promoter than the A-allele – in direct disagreement with the Myers study. Secondly, the results of the Myers study showed that when either allele of rs242557 was assayed in conjunction with the *MAPT* H1/H2 core promoters, the general level of transcription was lower relative to the core promoter alone. This suggests that the overall effect of the polymorphism is to repress transcription, with the G-allele conferring stronger repression and thus lower transcription compared to the A-allele. In contrast, the Rademakers study demonstrated that each allele of the polymorphism increased expression from both the *MAPT* H1 and SV40 control promoters, suggesting that rs242557 acts to enhance transcription and that the G-allele is a stronger enhancer than the A-allele.

These opposing results may potentially be explained by differences in methodology, with one major difference being that the Rademakers study cloned the region containing rs242557 upstream to the core promoter – altering the natural downstream position of the polymorphism which was used in the Myers study. It may, therefore, be possible that there are positional effects at play, though this phenomenon at present does not appear to have been investigated any further. Where both studies are in agreement, however, is that the region containing the rs242557 polymorphism appears to affect transcription from the *MAPT* core promoter and that there are allelic differences in this effect, providing support to the hypothesis that H1C and rs242557 have an important role to play in the regulation of *MAPT* transcription and that this may underlie the association with PSP.

1.7.3 N-terminal exons 2 and 3

In 2008, Caffrey and colleagues [205] conducted an allele-specific expression study to determine whether the H1 and H2 *MAPT* haplotypes confer differing rates of exon 2 and 3 inclusion. They took frontal cortex (FC) and globus pallidus (GP) tissue from the brains of 14 H1/H2 individuals and quantified the rate of exon inclusion from each chromosome. They found that H2 chromosomes expressed two-fold more 2N transcripts (2+3+) than H1 chromosomes in both the FC (H2:H1 ratio = 1.96; $p<0.0001$) and GP (H2:H1 ratio = 1.99; $p<0.0001$) tissues. There were no other transcripts in the FC or GP that demonstrated a biologically relevant allelic difference – that is a 20% or 1.2-fold difference – suggesting a role for exon 3 in the protection against PSP conferred by H2 carriers.

Interestingly, this increase in 2N isoform production from H2 chromosomes was observed independently of disease status. In addition, there was no significant difference in N-terminal transcripts between H1 chromosomes carrying the A-allele of rs242557 (the H1C sub-haplotype) and those carrying the G-allele. This suggests that the splicing of the N-terminal exons may be regulated separately from exon 10 and its role in PSP risk – if any – may thus be smaller.

In support of these results, a recent study analysed the largest collection of human brain samples in the most comprehensive study to date on the regional expression, splicing and regulation of *MAPT* [133]. Significant regional variation in *MAPT* mRNA expression and splicing was detected, with a 1.5-fold difference between the highest tau-expressing region (the frontal cortex) and the lowest (the white matter; $p=5.7x10^{-49}$). A relative reduction in exon 2 expression was detected in white matter compared to other brain regions and this corresponds to a reduction in 2N and 1N transcripts specifically in this region; the reason for which is unclear.

Overall, the regional tau mRNA expression levels were found to be highly correlated with total tau expression levels, though the relationship between mRNA and protein isoforms was not determined. At the genetic level, H2 chromosomes were found to express a significantly higher proportion of exon 3-containing transcripts in all brain regions and this was most significant in the frontal cortex (figure 1.10). No increase in exon 2 inclusion was observed in any brain region.

Figure 1.10 Expression of exon 3 from tau haplotypes in different brain regions.
Each brain region is presented in a different colour, with the H2 haplotype represented by the C/C genotype in the first lane. aveALL is the average across 10 brain regions: frontal cortex (FCTX); temporal cortex (TCTX); occipital cortex (OCTX); hippocampus (HIPP); thalamus (THAL); cerebellum (CRBL); substantia nigra (SNIG); putamen (PUTM); medulla (MEDU) and white matter (WHMT). ***Taken from Trabzuni et al (2012)*** **[133].**

1.8 Minigene studies of *MAPT* alternative splicing

Over the past decade, the *in vitro* expression of artificial *MAPT* constructs has provided important information regarding the expression of tau at the molecular

level. The value of these tools lies in both their small size and their focused investigation of small sections of DNA outside of their normal genomic context. It would generally be preferable – and more biologically relevant – to study the expression of the full-length *MAPT* gene but its large size makes this extremely difficult. Expression studies using smaller constructs do have some advantages, however, particularly for comparing the function of small sections of DNA, or variants of the same section, in the absence of variables such as differential *cis*- and *trans*-acting factors and the promoter driving expression. Indeed, most of the published *MAPT* constructs were created for investigation of splicing regulation of the alternative *MAPT* exons.

In 1993, unique cosmid constructs were created for exon trapping experiments designed to assess the inclusion rates of *MAPT* exons 3-9. This technique involved the insertion of the exon into the intronic section of a heterologous exon-intron-exon genomic fragment to see whether it was spliced in or out. This study was the first to demonstrate that *MAPT* exon 3 is spliced out in the absence of exon 2 [206]. Most *MAPT* constructs, however, were created for the purpose of assessing the effect on exon 10 splicing of the recently identified *MAPT* mutations shown to cause FTDP-17. These simple constructs typically comprised exon 9, 10 and 11 surrounded by small intronic segments with or without the mutation under investigation and transcription was driven by a control promoter such as the cytomegalovirus (CMV) or simian virus (SV) promoters [207-209].

Focus then shifted to the identification of splicing regulators that modulate the alternative splicing of exons 2, 3, 10 and even exon 6. These regulators may act in *cis* and were identified using a series of deletion constructs in which the size of the intronic segments around the alternative exon was gradually reduced until the sequence containing the regulator was excluded and the splicing pattern changed [210-214]. *Trans*-acting regulators were typically identified by over-expressing or knocking down the predicted protein factor in the presence of the *MAPT* construct and determining the effect on splicing pattern of the exon under investigation [211, 215-218]. An exon 10 minigene was also used to demonstrate that mRNA

secondary structure can affect the inclusion rate of this exon in mature transcripts [219].

Minigene constructs have also been created for the purpose of investigating the biochemistry of wildtype tau. Published studies include the use of tetracycline-inducible expression vectors to create cellular models expressing different tau isoforms that recapitulate the filamentous tau aggregation indicative of the tauopathy brain [220] and the use of 3R- and 4R- tau minigenes to determine a 20-fold increase in the binding of Fyn tyrosine kinase to 3R isoforms [221]. One study also used minigenes to assess the effect of exon 2 and 3 inclusion on the function of exon 6-containing N-terminal tau isoforms in microtubule assembly [222].

A final use of the *MAPT* construct involved the trialling of potential corrective treatments against aberrant exon 10 splicing. These included so-called *trans*-splicing, in which key sections of a *MAPT* minigene containing sequences responsible for aberrant splicing were replaced by the wildtype version [223, 224]. The use of antisense oligonucleotides to reduce the abundance of exon 10-containing transcripts was also trialled using a *MAPT* exon 10 minigene [225].

Each of the above splicing constructs contain only the short section of the *MAPT* gene under investigation and transcription was always driven by an exogenous control promoter to remove confounding by promoter-exon or promoter-intron interactions. One study, however, described the design of a minigene that expressed all six tau isoforms under the control of the *MAPT* core promoter element. All of the tau exons expressed in the brain were included and alternative exons 2, 3 and 10 were surrounded by 150-450bp of intronic sequence (figure 1.11). The authors used the minigene to create transgenic mouse models and showed that the tau N279K mutation confers increased exon 10 inclusion and recapitulates FTDP-17 pathology [226].

This minigene was the first of its kind created and its design potentially allows the study of mRNA isoform expression at the mRNA and protein levels. It was used here to study a specific *MAPT* mutation, but the basic design could be adapted to study the effect of common *MAPT* variation on the splicing of exons 2, 3 and 10. This minigene blueprint could therefore provide an ideal tool for the study of the role of the rs242557 polymorphism in the co-transcriptional regulation of *MAPT* alternative splicing.

Figure 1.11 A minigene to study alternative splicing at exons 2, 3 and 10 when expression is driven by the endogenous *MAPT* promoter.
Adapted from Dawson et al (2007)[226].

1.9 Project aims

This project investigated the molecular processes linking a common *MAPT* variant shown to be associated with increased PSP risk with the hallmark pathology observed in the brains of patients with this disease. Efforts were particularly focused on finding evidence of a co-transcriptional mechanism of alternative splicing regulation, which could potentially explain the connection between the rs242557 risk polymorphism located within the *MAPT* promoter region and aberrant downstream splicing events.

To this end, minigenes for the expression of all six tau isoforms and representing common *MAPT* variants were constructed and studied *in vitro* in human neuroblastoma cell lines. Investigations were conducted based on the hypothesis that the risk allele of the rs242557 polymorphism alters the physical interactions between the transcription and splicing machineries – either through conformational changes to the mRNA transcript or binding site abolition – which concurrently leads to an increase in transcription and a shift towards production of the more fibrillogenic exon 10-containing 4R-tau.

Accompanying this, two luciferase reporter gene studies aimed to identify the regions of the *MAPT* 5’ and 3’UTRs, and the genetic variants within them, that are critical for controlling mRNA transcription and stability respectively. This project is the first of its kind to study the co-transcriptional splicing of *MAPT* exons in order to elucidate the role of common promoter variation in neurodegenerative disease.

2 Methods and Materials

2.1 Methods

2.1.1 DNA/RNA sample extraction from tissue

DNA and RNA samples used in all cloning and genetic studies were previously extracted from flash frozen brain tissue by Proteinase K/phenol-choloroform extraction.

2.1.2 DNA/RNA quantification

The quantity and quality of DNA and RNA samples was determined by UV spectrophotometer absorption. The concentration (in ng/µl) of the sample is given by the absorption value at 260nm multiplied by 50 (the constant for DNA) or 40 (the constant for RNA). Sample purity is determined by the ratio of absorption at 260nm and 280nm, with a ratio greater than 1.8 and 2.0 indicating a pure DNA and RNA sample respectively.

2.1.3 Polymerase chain reaction

2.1.3.1 Standard PCR

The polymerase chain reaction (PCR) is a widely used *in vitro* method for amplifying defined sections of DNA from a known template sequence. It is a versatile and robust technique that is used in a wide variety of molecular biology protocols, including those for genotyping, mutation screening, DNA sequencing, cloning and gene expression quantifications. Genomic DNA is the most commonly used template, but sequences can also be amplified from plasmid DNA and cDNA reverse transcribed from RNA (section 2.1.3.2). The technique requires the design of two short oligonucleotides that anneal to the denatured DNA template at either end of the target region, creating short stretches of double-stranded sequence that act as primers for new DNA synthesis. Over repeated cycles of priming and synthesis, the concentration of the target sequence is exponentially and selectively amplified.

The specificity of the PCR product is determined by the design of the two oligonucleotide primers (approximately 15-25 nucleotides in length) that anneal exclusively to the target region. DNA synthesis occurs when the template and primers are subject to a series of heating and cooling steps, called thermal cycling. An initial heating step denatures the DNA, allowing the primers to access the single-stranded template. On cooling, the primers anneal to their complementary sequences at the ends of the target region. A final heating step results in primer extension and the synthesis of new DNA that is complementary and specific to the target sequence. DNA is synthesised by a heat stable DNA polymerase enzyme in the presence of a high concentration of the four deoxynucleoside triphosphate DNA components (dATP, dCTP, dGTP and dTTP). Repeated cycles of DNA synthesis – with the newly synthesised strands incorporated into the pool of template strands after each cycle – results in an exponential increase in the amount of product, with 25 cycles producing approximately 10^5 copies of the target sequence.

All primers were designed using the freely available PerlPrimer programme (http://perlprimer.sourceforge.net/). The AccuPrime™ *Taq* DNA High Fidelity Polymerase kit was used for all PCR reactions unless otherwise stated. Typical 25µl reactions comprised 2.5µl of Buffer I (10x), 0.2-0.4µM of forward and reverse primers, 2.5 units of *Taq* DNA polymerase and either 25ng of genomic DNA or 5-10ng of plasmid DNA. Additional magnesium chloride ($MgCl_2$; 1-2mM) and/or DMSO (5-10%) were added as necessary, depending on the structure of the primers and target sequence.

Thermal cycling was conducted using a Techne TC-Plus Thermal Cycler. After an initial four minute denaturation step at 94°C, 30-35 cycles of the following were completed: 30 seconds of denaturation at 94°C, 30 seconds of primer annealing at a temperature optimum for the primer pair (usually in the region of 55-65°C) and an extension step at 68°C of duration suitable to the size of the target sequence (45 seconds per 1 kb of target sequence). A final extension step at 68°C for 7 minutes completed the protocol.

2.1.3.2 Reverse transcription PCR

A variation of the PCR method, called reverse transcription PCR (or RT-PCR), was used for the selective amplification of RNA targets. This method involves an initial step in which the RNA template is converted into cDNA before specific target amplification by standard PCR. This conversion is facilitated by the reverse transcriptase enzyme, an RNA-dependent DNA polymerase that binds to primed RNA transcripts and synthesises complementary DNA copies (cDNA). Upon enzymatic degradation of the original RNA template, the reverse transcriptase acts as a DNA-dependent DNA polymerase and converts the single-stranded cDNA into double-stranded cDNA products.

There are three methods by which reverse transcription is achieved and these vary in the choice of oligonucleotide primer. The first method uses a primer that binds specifically to the target gene and therefore selectively amplifies transcripts expressed from this gene. The other two methods provide a means of reverse transcribing total RNA and therefore the pool of cDNA produced can be used for multiple analyses. With these methods priming is achieved with either random hexamers or oligo(dT) primers. Random hexamers are a pool of short oligonucleotides, each comprising a six nucleotide sequence generated at random. The short length and low specificity of the random hexamers result in the universal priming of total RNA transcripts. Oligo(dT) primers comprise stretches of 20 T-residues that exclusively bind to the poly-A tail of mature mRNA transcripts. With this method only processed transcripts that have been polyadenylated are reverse transcribed, with unprocessed nascent transcripts or those without a poly-A tail omitted from the reaction. All of the RT-PCR products in this project were produced using oligo(dT) priming of RNA samples extracted from transfected neuroblastoma cells.

A typical RT-PCR was conducted as follows: 1µg of total RNA was mixed with 1µl of oligo(dT) primers (50µM) and 1µl of dNTP mix (10µM) and adjusted to a total reaction volume of 10µl with RNase-free sterile water. The mixture was heated to 65°C for five minutes and immediately cooled on ice. A further 10µl of

a mastermix containing 2µl of RT buffer (10x), 4µl of $MgCl_2$ (25mM), 2µl of DTT (0.1M), 1µl of RNase OUT enzyme (40U/µl) and 1µl of SuperScript III reverse transcriptase (200U/µl) was added to the reaction mixture. The final 20µl reaction volume was heated to 42°C for 10 minutes, 53°C for 50 minutes, 85°C for 5 minutes and 10°C for 10 minutes; during which reverse transcription occurred. A volume of 1µl of *E.coli* RNase H (2U/µl) was added to the reverse transcribed cDNA to digest away the original RNA template. A final incubation at 37°C for 20 minutes completed the protocol. A volume of 1µl of the cDNA was used as the template in a standard PCR to amplify the target sequence.

2.1.3.3 Agarose gel electrophoresis

Agarose gel electrophoresis provides a method of visualising DNA products. It is most commonly used to check the size, specificity and quantity of products produced by PCR but can also be used to separate fragments of different size, such as a cloned target DNA fragment from its plasmid vector following digestion. Samples are loaded onto an agarose gel containing a nucleic acid stain that is visible under UV light. The application of an electrical current causes the DNA to migrate through the gel matrix and consequently become coated in the nucleic acid stain. Larger fragments migrate more slowly than smaller fragments, thus allowing size to be determined by comparing migration with that of a DNA ladder of known size.

To make a 1% w/vol gel, 1g of agarose powder was melted in 100ml of TAE buffer (1x). For visualisation of the DNA fragments under UV light, 5µl of the SYBRgreen nucleic acid stain, SafeView, was mixed into the melted agarose. The gel was cast around plastic combs to create wells for sample loading. Once set, the combs were removed and the gels were placed in the electrophoresis tank. TAE buffer (1x) was poured into the tank until the gel was completely submerged. DNA samples were mixed with 5x loading dye and loaded onto the gel alongside an appropriate size marker, either Hyperladder I or, for smaller products, Hyperladder IV. In the presence of an electrical current (typically 80-110 mV), the negatively charged DNA migrated through the gel towards the positive

electrode at a rate dependent on size. The duration of electrophoresis varied depending on the size and percentage of the gel, but typically lasted between 30 minutes and 1 hour. Bands were visualised under UV light using the MiniBis Pro (DNR Bio-Imaging Systems) and size was determined by comparison against the DNA ladder.

2.1.3.4 Polyacrylamide gel electrophoresis

Polyacrylamide gel electrophoresis (PAGE) is a variant of agarose gel electrophoresis and is used here to resolve PCR products of particularly small and/or similar size. PAGE is traditionally used in Western blotting to resolve protein samples or in Sanger sequencing protocols to separate DNA products that differ by a single nucleotide; neither of which are possible by agarose gel electrophoresis. PAGE is more sensitive than its agarose counterpart and allows the visualisation and quantification of products of very low concentration. The matrix of the polyacrylamide gel is much smaller than that of the agarose gel, with pore size determined by the relative concentrations of acrylamide and bis-acrylamide included in each gel. Another feature of PAGE is the vertical orientation of the electrophoresis tank, with the positive electrode located at the bottom of the tank.

Pre-cast polyacrylamide gels (in their plastic cases) were placed vertically in the electrophoresis tank with the sample wells at the top. The chamber was filled with TBE running buffer (1x) until the gel was completely submerged. PCR products were mixed with 5x loading dye and loaded into the vertical wells alongside a size marker. Following the application of an electrical current to the top and bottom of the gel, the DNA migrated downwards towards the positive electrode at the bottom of the tank. The length of the electrophoresis typically lasted between 50 minutes and 1 hour at 200v. The gel was then removed from its case and carefully placed into a small plastic container using clean forceps to prevent the gel from tearing. The gel was submerged in a 1:5000 dilution of Syto® 60 Nucleic Acid Stain in double distilled water and placed, in the dark, on a plate shaker at a low number of revolutions. After 30 minutes, the gel was twice washed with double

distilled water, returning each time to the plate shaker for 5 minutes. The DNA bands were visualised using the Odyssey Infrared Imaging System (LI-COR), with DNA bands stained with Styo® 60 visible using the 700nm channel.

2.1.4 Molecular biology: cloning

2.1.4.1 Purification of PCR products for use in cloning

PCR products amplified for ligation into plasmid vectors were purified using the QIAquick PCR Purification kit to remove leftover primers, nucleic acids and the DNA polymerase enzyme. The purification was conducted according to the manufacturer's instructions.

2.1.4.2 Purification of DNA products by agarose gel electrophoresis

In instances where two PCR products were produced in one reaction, or to separate cloned inserts from their plasmid vector, the DNA products were resolved by agarose gel electrophoresis to separate them by size. The desired band(s) were excised from the gel using a sterile scalpel and the DNA extracted and purified using the QIAquick Gel Extraction kit. This was done according to the manufacturer's protocol.

2.1.4.3 DNA ligation into plasmid vectors: pGEM-T Easy vector

Each target DNA fragment generated by PCR was routinely cloned into the pGEM-T Easy plasmid vector to produce a homogenous population for use in further cloning steps. All PCR products synthesised by the *Taq* DNA polymerase enzyme contain a single base (adenosine) overhang at the 5' end and this complements the 5' thymidine overhang of the linearised pGEM-T Easy vector, allowing the direct ligation of PCR products into the vector without the need for restriction digestion.

The ligation was facilitated by the T4 DNA ligase enzyme and typical 10µl reactions comprised: 50ng of linearised vector, 1-6µl of purified PCR product (at a 2-6 molar ratio), 1µl of T4 DNA ligase buffer (10x), 1µl of ATP (100mM) and 2

units of T4 DNA ligase (2-3U/µl). Ligation reactions were incubated at 4°C overnight before propagation in *E.coli* and plasmid harvest and purification (section 2.1.4.5). The amount of PCR insert included in each ligation reaction was calculated using the following formula:

$$\text{ng insert (purified PCR product)} = \frac{\text{ng vector} \times \text{kb insert}}{\text{kb vector}} \times \text{insert:vector ratio}$$

2.1.4.4 DNA ligation into plasmid vectors: Expression vectors

To create the expression constructs for *in vitro* analysis, the cloned DNA inserts were removed from the pGEM-T Easy vector by restriction enzyme digestion and purified by agarose gel electrophoresis. The purified fragment was then ligated into a similarly digested and purified expression vector (in this study pGL4.10 [*luc2*] or pMIR-REPORT) using the protocol described in section 2.1.4.3.

2.1.4.5 Propagation of plasmid constructs in *E.coli*

After ligation, the vector-insert constructs were transformed into *E.coli* cells. Bacterial cells take up plasmid DNA when subjected to a high temperature – or 'heat shocked'. Heating the cells to 42°C for a short interval (30 seconds to 1 minute) temporarily makes the bacterial cell wall porous, allowing the plasmid construct to pass through. Immediate cooling on ice closes the cell wall, trapping the plasmid inside where it is replicated as part of the cell division process. Thus, a large yield of homogenous plasmid construct is produced following transformation of one rapidly-dividing bacterial cell. Successfully transformed colonies are primarily identified by continued growth in the presence of a specific antibiotic, the resistance to which is conferred solely by the transformed plasmid construct.

Typical transformations comprised 50µl-100µl of High Efficiency JM109 or HB101 *E.coli* cells (thawed slowly on ice) and 5µl-10µl of ligation reaction (i.e. 10% of the cell volume). The transformation mix was incubated on ice for 15 minutes, "heat shocked" for 30 seconds in a 42°C water bath and returned to the

ice for 10 minutes. After the addition of 500µl or 1ml of L-broth (for low or high copy number vectors respectively), the cells were incubated at 37°C for 1 hour with horizontal agitation at 150rpm. Gentle centrifugation at 3,000 x g formed a cell pellet which, after removal of the supernatant, was re-suspended in 100µl of L-broth. The full cell suspension was spread on an LB-agar plate containing an appropriate selection antibiotic and incubated overnight at 37°C.

Single, well-defined colonies were individually picked using aseptic technique and cultured in 3ml of L-broth containing the selection antibiotic. Following overnight incubation at 37°C with vigorous horizontal agitation at 250rpm, the cultured cells were harvested by centrifugation at 17,000 x g. The cloned plasmid DNA was extracted from the bacterial cells using the QIAquick Spin Miniprep kit according to the manufacturer's protocol.

2.1.4.6 Blue-white screening of pGEM-T Easy clones

As described previously, pGEM-T Easy is a linearised vector that provides a single T-overhang for ligation with the target PCR products. The downside to the convenience of this method is that during ligation the T4 DNA ligase enzyme also catalyses the re-circularisation of empty pGEM-T Easy vectors. Thus, an additional method of selection is required to separate colonies that have been transformed with vector/insert constructs from those that have taken up re-circularised empty vectors.

The pGEM-T Easy vector contains the *lacZ* gene and therefore expresses the β-galactosidase enzyme. This enzyme, in the presence of IPTG, metabolises X-Gal into a blue product. Upon successful ligation of the insert into the pGEM-T easy vector, the *lacZ* gene is disrupted, β-galactosidase is not produced and the X-Gal remains unmetabolised. Thus, the addition of IPTG (0.1mM) and X-Gal (20µg/ml) to the LB-agar plate, allows the identification of colonies transformed with successfully ligated constructs by virtue of their white – not blue – colour.

2.1.4.7 Digestion by restriction enzyme

Restriction enzyme digestion was routinely used to remove DNA fragments from purified plasmid vectors following cloning. Restriction endonucleases are enzymes that cut double-stranded DNA molecules at a specific recognition sequence called a restriction site. Most restriction sites are 6bp in length and are palindromic, meaning they can be read in either direction. These sites occur naturally throughout the genome but can be artificially added onto the ends of target DNA sequences by PCR with primers containing the specific recognition sequence on their 5' ends. Digestion with a restriction enzyme can produce a blunt-ended fragment, but more commonly results in so-called 'sticky ends', in which a 5' or 3' single-stranded overhang is produced. Thus, two DNA fragments can be joined together using the complementary overhangs produced following digestion of each fragment with the same restriction enzyme. This technique is used extensively in cloning protocols, as the addition of two different sites onto the end of a target sequence facilitates its directional insertion into an expression vector containing the same two restriction sites.

In this project, restriction enzyme digestion was used for two main purposes: firstly to extract the cloned target DNA from the pGEM-T Easy vector for re-ligation with an expression vector; and secondly to simply confirm the presence of the target DNA insert in a cloned plasmid. For the former purpose, which required a high yield of extracted insert, digestions were conducted on a large scale (50µl) and comprised: 5-10µg of purified plasmid DNA, 5µl of bovine serum albumin (BSA; 10x), 5µl of enzyme-appropriate buffer (10x) and 50 units of restriction enzyme. The reactions were incubated overnight at a temperature suitable for optimum enzyme activity, which in most cases was 37°C. For the latter purpose – to simply confirm the presence of the target DNA – the reaction was scaled down to a total volume of 10µl and incubated for 1-2 hours.

In some instances a double digestion was required; for example when a directional insertion dictated that the DNA fragment was digested by two different restriction enzymes. If the two enzymes were sufficiently active in the same digestion buffer,

each was added in half the amount stated above (i.e. the total concentration of enzyme in the double digestion did not exceed that in the single digestion). If the two enzymes were not compatible with the same buffer, two single digestions were conducted with a purification step in between to remove all traces of the first enzyme and buffer. This purification was conducted using the QIAquick PCR Purification kit according to the manufacturer's protocol.

2.1.4.8 DNA sequencing

Final confirmation of cloned plasmid constructs was achieved by sequencing. All DNA sequencing was outsourced to a service provider, Source BioScience Ltd.

2.1.5 Cell Culture

2.1.5.1 Neuroblastoma cell culture

Two human neuroblastoma cell lines, SK-N-F1 and SH-SY5Y, were obtained from the European Collection of Cell Cultures (ECACC). SK-N-F1 cells were derived from the bone marrow metastasis of a male patient with neuroblastoma. These undifferentiated and slow-growing cells have a substrate-adherent, epithelial-like phenotype and an aneuploid karyotype. They secrete neuronal peptides in culture. SH-SY5Y is a sub-line of the SH-N-SH bone marrow biopsy-derived line. The parent line was extracted from the bone marrow metastasis of a four year old Caucasian female patient with neuroblastoma. SH-SY5Y cells are adherent with a neuroblast morphology and a diploid karyotype but have been shown to lose neuronal characteristics with increasing passage number. Both cell lines grow processes in culture and produce a neuronal phenotype following differentiation with retinoic acid.

Cell lines were expanded in F12-MEM cell culture medium supplemented with 10% foetal calf serum (FCS) and incubated at 37° with 100% relative humidity and 5% carbon dioxide. Medium was changed every two to three days and cells were passaged when approximately 90% confluent. Briefly, one flask of adherent cells was rinsed with 1x phosphate buffered saline (PBS) and detached from the

flask surface by incubation with trypsin-EDTA at 37°C for five minutes. The detached cells were collected, gently pelleted and resuspended in cell culture medium before being diluted and split among three to four new culture flasks. These new cultures were propagated as before.

Neuronal differentiation was achieved by replacing the cell culture medium with reduced FCS (1%) medium containing 10nM of retinoic acid. The addition of retinoic acid halts cell growth and induces the expression of neuronal markers. Differentiation medium was changed every two days and cells were fully differentiated after five days.

For long term storage of the neuroblastoma cell lines, cells were harvested by trypsinisation, pelleted by gentle centrifugation and resuspended in 1ml of freezing solution (F12-MEM medium with 10% FCS and 10% DMSO). Cells were transferred to cryotubes, slowly frozen to -80°C in an isopropanol bath and stored in liquid nitrogen.

2.1.5.2 Transfection of cells with plasmid DNA

Plasmid DNA constructs were expressed in SK-N-F1 and SH-SY5Y cells following lipid transfection. The TransFast transfection reagent is comprised of two lipids, a synthetic cationic lipid and L-dioleoyl phosphatidylethanolamine (DOPE), a neutral lipid. Plasmid DNA coated in lipid micelles are taken up by mammalian cells by endocytosis.

Cells were plated on a suitable cell culture plate and grown to 80% confluency in F12-MEM cell culture medium supplemented with 10% FCS. Approximately one hour before transfection, the medium was removed and replaced with serum-free culture medium. An appropriate amount of endotoxin-free plasmid DNA was mixed with TransFast reagent at a charge ratio of 1:1 in serum-free medium. The amount of plasmid DNA transfected varied depending on the size of both the plasmid and cell culture plate. The mixture was incubated for 15 minutes at room temperature with vigorous shaking to ensure the plasmid became fully coated with

the lipids. A small volume of fresh serum-free medium was added to the cells before addition of the DNA/TransFast mixture. Cells were incubated at room temperature for 10 minutes with gentle shaking and then returned to the 37°C incubator. After one hour, cells were topped up with 10% culture medium and allowed to recover for a further 48-72 hours before analysis.

2.1.5.3 The luciferase reporter gene assay

The luciferase reporter gene assay provides a simple and high through-put method of assessing the ability of short DNA sequences to initiate, regulate or differentially alter gene expression. Routinely used to ascertain the comparative strengths of different promoters or promoter variants, the assay may also be used to determine the regulatory effects on promoter activity of *cis*-acting elements or the effect of the 3'UTR on transcript stability. This is achieved by cloning the sequence of interest either upstream or downstream to a luciferase gene in which the promoter or the 3'UTR has been removed. When expressed *in vitro*, the level of luciferase produced is directly proportional to the transcriptional activity or stability conferred by the cloned sequence. The commonly used reporter gene is cloned from the firefly *Photinus pyralis* which, upon translation, gives rise to the luciferase enzyme. This enzyme catalyses the ATP-dependent oxidation of luciferin to oxyluciferin, a reaction that produces light at a rate proportional to the activity of the enzyme.

The Dual-Glo Luciferase Assay System was used for all reporter gene studies described here. This system was chosen due to its incorporation of an internal control plasmid, in this case the pRL-TK plasmid which expresses the *Renilla* luciferase gene from a tyrosine kinase control promoter. The *Renilla* luciferase reaction requires a different substrate to its firefly equivalent and therefore its independent quantification in the same well allows the normalisation of the firefly signal and compensation of well-to-well differences in transfection efficiency and/or cell density. Mammalian cells transfected with both the firefly and *Renilla* plasmids are analysed in a two-step assay that individually quantifies the luminescence produced by each plasmid. The first step quantifies the level of

firefly luciferase and involves the direct addition of the Dual-Glo luciferase reagent to the culture medium of the transfected cells. This reagent provides the substrate for the firefly luciferase enzyme, producing a stable luminescent signal that can be measured by a luminometer (Tecan GENios). The signal lasts for approximately two hours and must be measured within this time. The addition of a second reagent, the Dual-Glo Stop & Glo reagent, quenches the firefly signal and provides the substrate for the *Renilla* luciferase enzyme. This luminescent signal is similarly measured within a two-hour window.

Luciferase assays were conducted on neuroblastoma cells plated on opaque 96-well plates 48 hours post-transfection with the two luciferase plasmids. The cell culture medium was removed and replaced with 20µl of serum-free medium. A volume of 20µl of each reagent was sequentially added to each well, with the luminescent signal measured ten minutes after the addition of each reagent.

2.1.5.4 TRIzol® method for RNA extraction from cell culture

The isolation of RNA samples from transfected cells was conducted using the TRIzol® method. The TRIzol® Reagent is a monophasic solution of phenol and guanidine isothiocyanate developed specifically to maintain the integrity of large and small RNA species during cell lysis and cell component dissolution. In this project the method was used to extract total RNA from neuroblastoma cells transfected with the *MAPT* minigenes.

The first stage of the method involves cell harvesting and homogenisation. Cells were cultured in 6-well, 35mm dishes for 72 hours post transfection. To harvest the cells, the culture medium was removed from each well and replaced with 1ml of TRIzol® Reagent. Cells were lysed in the dish by vigorous pipetting before transferral to a 1.5ml centrifuge tube.

The second stage, phase separation, is essential for the separation of the DNA, RNA and protein species in the cell lysate. The homogenised cells were incubated at room temperature for 5 minutes to allow the complete dissociation of the

nucleoprotein complex. The addition of 0.2ml of chloroform followed by vigorous inversion and centrifugation at 12,000 x *g* for 20 minutes at 4°C, resulted in the separation of the mixture into three clearly visible phases. The upper aqueous phase contained the isolated RNA. The interphase and lower organic phase contained DNA and protein.

To isolate the RNA, the aqueous phase was carefully transferred to a separate 1.5ml centrifuge tube containing 0.5ml of 100% isopropanol. The RNA was precipitated during a 10 minute incubation at room temperature and collected by centrifugation at 12,000 x *g* for 15 minutes at 4°C. The supernatant was removed and the RNA pellet washed with 75% ethanol. After centrifugation at 7,500 x *g* for 10 minutes at 4°C, the wash was discarded and the pellet left to air dry for 5 minutes. The RNA was dissolved in 20µl of RNase-free water and left to resuspend for 2 hours at 4°C.

2.1.5.5 RNA purification

RNA samples extracted from cultured cells were treated with DNaseI to remove DNA contaminants. This enzyme selectively digests DNA, leaving the RNA molecules intact. This is an important step to ensure future analyses of reverse-transcribed transcripts are not compromised by amplification of genomic DNA. Each 20µl RNA sample was mixed with 1µl of DNaseI enzyme, 2.5µl of DNase buffer (10x) and 1.5µl of RNase-free water, and incubated at 37°C for 30 minutes. The sample was then purified using the RNeasy MinElute Cleanup kit according to the manufacturer's instructions.

2.1.5.6 DNA extraction from cell culture

When necessary, DNA was extracted from cultured cells using the CellsDirect Cell Resuspension and Lysis Buffers according to the manufacturer's protocol.

2.1.6 Minigene construction

2.1.6.1 Multisite Gateway® cloning by recombination

The *MAPT* minigenes were constructed using the Multisite Gateway® Pro Plus kit. This technology utilises recombination between two compatible sequences to transfer up to four target DNA fragments into one vector in a single step. This reduces both the need for multiple cloning steps and the likelihood of sequence errors being inserted, as often occurs during cloning in bacterial cells. This method is described in detail in chapter 4.

2.1.6.2 Creation of stable isogenic cell models

The Gateway® technology provides a method of integrating the expression constructs into the genome of mammalian cell lines at a pre-determined location. This adds reliability and versatility to expression studies by ensuring the contructs are not differentially affected by factors related to the insertion site. This method is also described in detail in chapter 4.

2.1.7 Cell Biology

2.1.7.1 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) provides a method of studying the association of certain DNA-binding proteins with specific sequences in the genome. Such proteins play important roles in numerous cellular processes such as gene expression, DNA repair and segregation, cell-cycle progression and epigenetic silencing. In this project, the ChIP assay was used to fulfil two project aims: firstly to confirm the association of proteins that were predicted to bind to specific sequences within the *MAPT* promoter; and secondly to ascertain whether single nucleotide polymorphisms within the predicted binding sites could lead to allelic differences in protein binding.

Immunoprecipitation of chromatin from neuroblastoma cells was achieved using the MAGnify™ ChIP System. Cultured cells were treated with formaldehyde to fix

DNA-protein and protein-protein associations by generating crosslinks between neighbouring molecules within the chromatin complex. Following cell lysis, chromatin was released from the cell nuclei and sheared by sonification to produce fragmented DNA of 200-500bp in size. The crosslinked protein of interest was selectively immunoprecipitated using a specific ChIP-grade antibody conjugated to specially-designed magnetic beads. This conjugation allowed the specific isolation of the crosslinked protein of interest from the rest of the nuclear chromatin extract. The crosslinking was reversed by heat treatment and the protein-associated DNA fragments purified. Target-specific PCR of the purified DNA confirmed the presence or absence of the target sequence in the pool of DNA fragments that were associated with the protein of interest. The specificity of the protocol was determined by the inclusion of the mouse IgG antibody as a negative control, as this antibody does not bind to human DNA. The reliability of the final PCR was confirmed by the inclusion of an 'input' control containing chromatin that had been purified but not subject to selective immunoprecipitation.

2.1.8 Genetics

2.1.8.1 Restriction fragment length polymorphism

A restriction fragment length polymorphism (RFLP) is a single nucleotide polymorphism that lies within a restriction enzyme recognition sequence. The two alleles can therefore be distinguished based on whether they complete or abolish this restriction sequence, as determined by the ability of the restriction enzyme to cut at this location. Thus, the presence of a specific allele in a given DNA sample is determined by PCR amplification of the surrounding region followed by restriction digestion. Resolution of the digestion products by agarose gel electrophoresis reveals a banding pattern unique to each allele. PCR product containing the allele that abolishes the restriction site will remain uncut, producing one solitary band. Product that contains the other allele, however, will be cut into two, with two smaller bands visible on the gel. When two variants of the target region are present – such as when genotyping human DNA samples – a

third banding pattern may be produced for individuals that are heterozygous. In these cases all three bands, the uncut and the two cut bands, are present together.

2.1.9 Statistics

2.1.9.1 The Student's *t*-test

The Student's *t*-test provides a method of determining whether the means of two normally distributed populations are significantly different to one another. The test is conducted under the null hypothesis, which states that the difference between the two means is zero. The mean, standard deviation and size of each population are all used to calculate the test statistic; a measure of the difference between the means. The null hypothesis is rejected when the probability of a detected difference occurring by chance is less than or equal to 5% ($p \leq 0.05$).

2.1.9.2 The Hardy-Weinberg Equilibrium

The Hardy-Weinberg Equilibrium (HWE) describes the stable inheritance of allele and genotype frequencies within a population. It is based on the assumption that the population undergoes random mating and is not subject to mutation, migration, selection or random drift. The Hardy-Weinberg equation provides a means of calculating the deviation of a given population from the genotype distributions expected under HWE. The equation ($p^2 + 2pq + q^2 = 1$) calculates the expected distribution of genotypes of a given polymorphism based on the population allele frequencies. Thus, if p is the frequency of the major allele (A), and q is the frequency of the minor allele (a), then p^2, $2pq$ and q^2 provide the population frequencies of the AA, Aa and aa genotypes, respectively. This expected distribution is compared with the actual distribution observed in the population, with a significant difference between the two detected using a Chi-square test (section 2.1.9.3). The population is said to be in HWE when $p>0.05$. In genetic studies, the Hardy-Weinberg equation is used to identify population stratification within a genotyped cohort and can also highlight potential problems with the accuracy of the genotyping assay.

2.1.9.3 Genetic association: The Chi-square test

The Chi-square test was used in single locus analyses of genotype and allele frequencies in case and control cohorts. It is a non-parametric test of independence and uses a contingency table of paired frequencies to calculate the Chi-square statistic. This statistic is generated by summing the normalised squared difference of each paired allele or genotype count. This, along with the degrees of freedom (the number of frequencies minus the number of parameters) is used to calculate a p-value, with $p \leq 0.05$ the threshold for significance. In instances where one of the values in the contingency table is below 5, Fisher's Exact test is used to calculate the significance value. The Chi-Square distribution only gives an approximation of statistical significance and this leads to inaccuracies when the sample size is too small. Fisher's Exact test is instead used to calculate an accurate significance value. In this project, genotype frequencies were tested using dominant and recessive models, with the heterozygote group added to one of the homozygote groups in each instance. This allowed the mode of inheritance of an associated allele to be determined by ascertaining whether one or two copies of the allele are required for association.

2.1.9.4 Genetic association: The odds ratio

The odds ratio is a measure of the effect size of an association. It is calculated as the probability of an event occurring in one group divided by the probability of it occurring in the other. An odds ratio of 1.0 indicates that the event is equally likely to occur in the two groups. An odds ratio above or below 1.0, however, indicates that the event is more likely to occur in one of the groups. For example in a case-control study, an associated genetic variant may be calculated as having an odds ratio of 2.4. This would mean that an individual with this particular genetic variant is 2.4-fold more likely to be in the case group than in the control group. An odds ratio of less than 1.0, however, would indicate that the individual is more likely to be in the control group and therefore the genetic variant confers protection. The further the odds ratio is from a value of 1.0, the greater the effect of the association.

2.1.10 Bioinformatics resources

2.1.10.1 NCBI

The National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) is a collection of publicly available databases that provides a valuable resource for molecular genetics studies. The databases facilitate the retrieval of information vital to the progression of such studies, including nucleotide sequences and polymorphism frequency data. The resource also provides web-based tools such as the basic local alignment search tool (BLAST) which can be used to identify unknown sequences through alignment with the known sequence database.

2.1.10.2 UCSC Genome Bioinformatics

The University of California Santa Cruz (UCSC; http://genome.ucsc.edu/) genome browser provides comprehensive and visual annotations of assembled reference genomes from, among other species, human, chimpanzee and mouse. Information available includes: the chromosomal location of genes, polymorphic variation, isoform composition, cross-species conservation and tissue-specific gene expression. The resource also has an *in silico* PCR function, which aligns potential primer pairs to a reference genome and generates the predicted sequence of the product(s), and a Blat tool, which aligns sequences of interest against an annotated reference genome. Both of these functions were used extensively throughout this project.

2.1.10.3 ClustalW2

The European Bioinformatics Institute (EBI) provides the web-based ClustalW2 programme (http://www.ebi.ac.uk/Tools/msa/clustalw2/) which aligns two or more protein or DNA sequences, clearly highlighting the differences between them. This tool is particularly useful for identifying the single nucleotide polymorphisms, deletions, insertions, and repeats in the sequences of genetic variants of the same gene or region. In this study, the ClustalW2 tool was used to

identify sequence differences between the haplotype variants of each *in vitro* expression construct.

2.2 Materials

2.2.1 PCR reagents

AccuPrime™ *Taq* DNA Polymerase High Fidelity (Invitrogen)
FastStart High Fidelity PCR System (Roche)
All primers and oligonucleotides were obtained from Sigma-Aldrich

2.2.2 Restriction enzymes

All restriction enzymes were obtained from New England Biolabs.

2.2.3 Molecular biology reagents

2.2.3.1 Gel electrophoresis reagents

Tris-acetate-EDTA (TAE) buffer (Qiagen)
Agarose (Sigma-Aldrich)
SafeView Nucleic Acid Stain (NBS Biologicals)
Hyperladder I and Hyperladder IV (Bioline)

2.2.3.2 DNA purification kits

QIAquick PCR Purification kit – genomic (Qiagen)
QIAquick Gel Extraction kit – genomic (Qiagen)
QIAquick Spin Miniprep kit – plasmid mini-preparation (Qiagen)
Endo-free Plasmid Maxi kit – plasmid maxi-preparation (Qiagen)

2.2.3.3 Plasmid vectors

pGEM-T Easy – sub-cloning plasmid (Promega)
pGL4.10 [*luc2*] – firefly luciferase plasmid (Promega)
pMIR-REPORT – firefly luciferase plasmid (Promega)
pRL-TK – *Renilla* luciferase plasmid (Promega)

2.2.3.4 Ligation reagents

T4 DNA ligase – single fragment ligation (1-2U/µl; Promega)
T4 DNA ligase high concentration – multi-fragment ligation (200U/µl; Promega)
T4 DNA ligase high concentration – splinkerette PCR (500U/µl; New England BioLabs)
ATP (100mM; Sigma-Aldrich)

2.2.3.5 Bacterial cells

JM109 High Efficiency competent cells (>10^8cfu/μg; Promega)
HB101 competent cells (Promega)

2.2.3.6 Cloning reagents

Luria-Bertani (LB) broth (Invitrogen):

- 20g/L LB broth dissolved in deionised water and sterilised by autoclaving at 121^{O}C for 20 minutes.

LB-agar (Invitrogen):

- 32g/L LB-agar dissolved in deionised water and sterilised by autoclaving at 121^{O}C for 20 minutes.

Ampicillin (Sigma-Aldrich)
Kanamycin (Sigma-Aldrich)
Hygromycin B (Invitrogen)
Zeocin (Invitrogen)
X-Gal (5-bromo-4-chloro-indolyl-β-D-galactopyranoside; Promega)
IPTG (isopropyl β-D-1-thiogalactopyranoside; Sigma-Aldrich)

2.2.4 Sequencing

All sequencing was conducted by Source BioScience.

2.2.5 Cell culture reagents

All cell culture reagents were supplied by Sigma-Aldrich unless stated otherwise.

Neuroblastoma cell culture growth medium (F12-MEM; 1-10%):

- 44% Ham's F12 nutrient mixture (F12)
- 44% Eagle's minimum essential medium (MEM)
- 1-10% foetal calf serum (FCS)
- 2mM L-glutamine
- 1% non-essential amino acids
- 20 units/ml penicillin
- 250ng/ml amphotericin B

Trypsin-EDTA solution
Cell freezing solution

- 10% dimethyl sulphoxide (DMSO)
- 90% neuroblastoma cell culture growth medium

Retinoic acid (10μM stock; Tocris)

2.2.6 Transfection reagents

TransFast transfection reagent (Promega)
Serum-free medium (all Sigma-Aldrich)

- 49% Ham's F12 nutrient mixture (F12)
- 49% Eagle's minimum essential medium (MEM)
- 2mM L-glutamine
- 1% non-essential amino acids

2.2.7 Luciferase assay reagents

Dual-Glo Luciferase Assay System (Promega)

2.2.8 DNA extraction (cells)

CellsDirection Cell Resuspension and Lysis Buffers (Invitrogen)

2.2.9 RNA extraction and purification (cells)

TRIzol® Reagent (Invitrogen)
Chloroform
Isopropanol
Ethanol
DNaseI (Invitrogen)
RNeasy MinElute Cleanup kit (Qiagen)

2.2.10 Minigene construction reagents

Multisite Gateway® Pro Plus Vector Module (Invitrogen)
Jump-In™ TI™ Platform Kit (Invitrogen)
Jump-In™ TI™ Gateway® Vector Kit (Invitrogen)

2.2.11 mRNA analysis reagents

SuperScript III First Strand Synthesis System for RT-PCR (Invitrogen)
4-12% Tris-borate-EDTA (TBE) polyacrylamide gels (Invitrogen)
10% TBE polyacrylamide gels (Invitrogen)
TBE running buffer (Invitrogen)
Syto® 60 DNA Nucleic Acid Stain (Invitrogen)

2.2.12 Chromatin Immunoprecipitation (ChIP) reagents

MAGnify™ Chromatin Immunoprecipitation System (Invitrogen)
Monoclonal ChIP-grade antibodies:

- Anti-Pol II (Covance)
- Anti-hnRNPU clone 366 (Millipore)
- Anti-β-actin (Abcam)
- Anti-mouse IgG (negative control; Millipore)

2.3 Suppliers

The suppliers of materials and services used throughout this project are listed below:

Abcam plc, 330 Cambridge Science Park, Cambridge CB4 0FL

Bioline Reagents Ltd, Unit 16 The Edge Business Centre, Humber Road, London NW2 6EW

Covance Inc, Compass House, Manor Royal, Crawley, West Sussex RH10 9PY

Invitrogen Ltd, 3 Fountain Drive, Ichinnan Business Park, Paisly, UK, PA4 9RF

Millipore (UK) Ltd, Suite 3 & 5, Building 6, Croxley Green Business Park, Watford WD18 8YH

NBS Biologicals Ltd, 14 Tower Square, Huntingdon, Cambridgeshire PE29 7DT

New England Biolabs UK Ltd, 73 Knowl Piece, Willbury Way, Hitchin, Hertfordshire SG4 0TY

Promega UK Ltd, Delta House, Chilworth Research Centre, Southamptom, SO16 7NS

Qiagen UK Ltd, Flemming Way, Crawly, West Sussex, RH10 9NQ

Roche Products Ltd, 6 Falcon Way, Shire Park, Hexagon Place, Welwyn Gardeen City, Hertfordshire AL7 1TW

Sigma-Aldrich Company Ltd, Fancy Road, Poole, Dorset, BH12 4QH

Source BioScience plc, 1 Orchard Place, Nottingham Business Park, Nottingham NG8 6PX

Tocris Bioscience, Tocris House, IO Centre, Moorend Farm Avenue, Bristol BS11 0QL

3 Luciferase reporter gene studies to investigate the effect on expression of genetic variation within the untranslated regions of the *MAPT* gene

3.1 Overview

The 5’ and 3’ regions of a gene, although non-coding, contain important genetic information required for the regulation of gene expression. The 5’ region is commonly referred to as the gene promoter (or intron -1). This region contains the core promoter (often denoted exon 0), which is responsible for initiating transcription, and numerous regulatory elements that modulate the rate of transcription through enhancement or repression. The 3’ region (often referred to as the 3’ untranslated region or 3’UTR) contains binding domains for microRNAs and RNA binding proteins along with signal sequences for the polyadenylation, degradation, localisation and stabilisation of mRNA transcripts. As a result, the 3’UTR plays a key role in mRNA stability, processing and translation. It therefores follows that genetic variation within these regions has the potential to have a profound effect on the normal regulation of gene expression.

The luciferase reporter gene assay is a useful tool in gene expression studies as it provides an easily detectable, high throughput method of comparing the regulatory potential of unknown DNA sequences. The assay is most commonly used to compare the ability of different promoters or promoter variants to drive expression of a promoterless luciferase gene *in vitro*. It can also be used to investigate the effect of variation within the 3’UTR on gene expression, achieved by cloning the 3’UTR variant downstream to a luciferase gene in which expression is driven by a control promoter.

In this study, a series of luciferase reporter gene constructs was designed for the in-depth investigation of two highly conserved regions of the *MAPT* promoter; the first containing the *MAPT* core promoter and the second comprising a predicted regulatory domain containing the PSP-associated rs242557 polymorphism. An additional series of luciferase constructs was created to determine whether genetic

variation within the *MAPT* 3'UTR can affect mRNA stability and thus gene expression levels. This was investigated in two ways: firstly by quantifying expression when the full 3'UTR is present, and secondly by splitting the 3'UTR into three overlapping fragments to pinpoint the sequences that are most critical for maintaining normal gene expression levels.

In both the promoter and 3'UTR studies three variants of each luciferase construct were created representing the genetic variation of the H1C, H1B and H2 *MAPT* haplotypes. This allowed for direct comparison between the 'risk', 'neutral' and 'protective' *MAPT* variants respectively.

3.2 Background

The A-allele of the rs242557 polymorphism is strongly associated with an increased risk of PSP [104, 181, 227] and defines the *MAPT* H1 sub-haplotype denoted H1C. This polymorphism is located within a highly conserved region of the promoter and has been shown to lie within or proximal to a transcription regulatory domain [11, 12]. There are, however, conflicting reports regarding the exact nature of the domain's effect on transcription and the first step towards unravelling the functional role of the rs242557 polymorphism is to clarify these conflicting reports.

As described earlier, two groups used the luciferase reporter gene assay to quantify the level of transcription conferred by the allelic variants of the rs242557 regulatory domain when cloned adjacent to an element containing the *MAPT* core promoter. Although both studies found allelic differences in the activity of the regulatory domain, one group reported an significant increase in transcription for the G-allele variant [11] while the other reported a significant increase for the A-allele variant [12]. Both studies quantified the transcriptional activity in human neuroblastoma cells and showed that the addition of the SNP domain causes a reduction in transcription (i.e. the domain functions as a repressor). The Rademakers study [11], however, additionally conducted the assay in mouse neuronal cells (N2a) and although the allelic differences in activity remained, in

these cells the domain acted as a transcription enhancer. This suggests that the cellular environment – including the components of the transcription machinery – has an effect on the interaction between the regulatory domain and core promoter. Enhancement was also observed when the SNP domain was assayed in conjunction with the SV40 immediate early promoter, rather than the *MAPT* H1 core promoter, indicating that promoter identity plays an important role in the functioning of the domain.

To clarify these findings, an *in vitro* luciferase reporter gene study was designed to incorporate aspects of the two published luciferase studies and to remove some of the technical differences that may account for the conflicting results; most notably the positioning of the regulatory domain relative to the core promoter and the size of the core promoter and rs242557 elements.

An additional aspect of this study includes an in-depth look at the *MAPT* core promoter region. Of particular interest is a highly conserved sequence located immediately downstream to the core promoter region. According to database annotations (figure 3.1), this 900bp sequence appears to contain a bi-directional promoter, with two non-coding transcripts originating from within this region, one transcribed in the sense and one in the antisense direction. The sense transcript, denoted *MAPT-IT1* (LOC100130148), is an intronic transcript of only 3016bp and its role is currently unclear. The antisense transcript, denoted *MAPT-AS1* (LOC100128977), may play a role in the regulation of *MAPT* transcription.

Natural antisense transcripts (NATs) are a group of RNA molecules that have sequence complementarity to other RNA transcripts. They can be coding or non-coding transcripts and form two separate groups depending on whether they are transcribed from the same (*cis*-NATs) or different (*trans*-NATs) genomic locations to their target. There are four proposed models for NAT regulation of target gene expression. The first is a knockdown model, in which the binding of the NAT to its complementary transcript converts it into a double-stranded molecule that is subsequently targeted for degradation, thereby reducing the

number of target transcripts available for translation. The second is an RNA masking model, in which duplex formation masks certain *cis*-acting regulatory elements residing in either of the transcripts and inhibits protein-RNA interactions such as those involved in splicing and mRNA transport. The third is an epigenetic model, in which the NAT aids the binding of methylation and/or histone-modifying complexes to the promoter region of the sense transcript, inhibiting its expression. This model is the least understood of the four. The final model is based on the observation that transcription cannot occur in two directions simultaneously due to collision of the two Pol II elongation complexes. Thus, the simple act of transcribing the NAT in the antisense direction impairs transcription in the sense direction, reducing the number of transcripts produced [228].

The *MAPT-AS1* gene (which expresses a *cis*-NAT) has only 2 exons but its extensive introns result in a DNA sequence spanning over 52kb, overlapping the core promoter, the upstream regions of the *MAPT* promoter and beyond the intronless presenilin homologue gene, *IMP5*. Using the luciferase reporter gene assay, the strength of the bi-directional promoter in both the sense and antisense directions, as well as its effect on transcription from the *MAPT* core promoter, was quantified in order to determine whether this region plays a role in the regulation of *MAPT* expression.

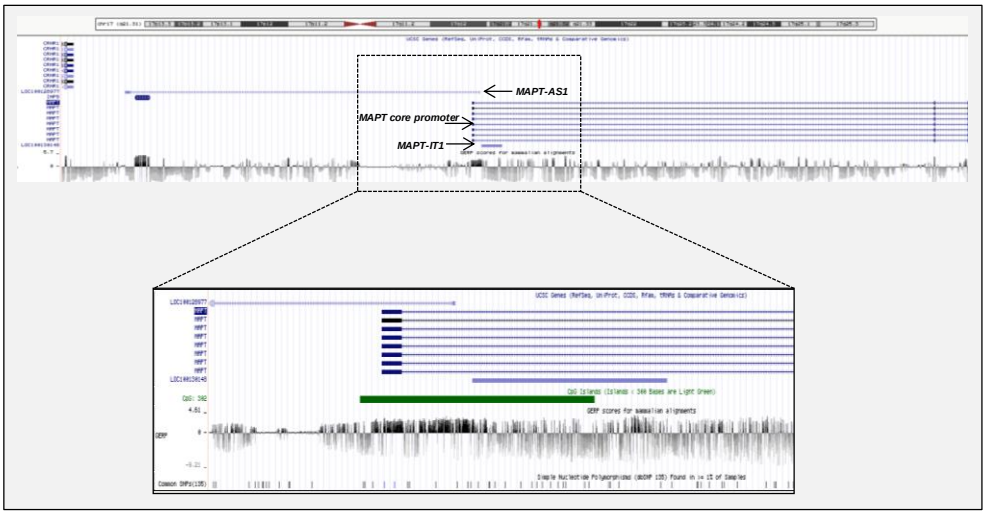


Figure 3.1 UCSC genome annotations of the region containing *MAPT* exon 0.
Two non-coding transcripts denoted *MAPT-IT1* and *MAPT-AS1* appear to be transcribed from the same region located immediately downstream to the core promoter at exon 0.

The second luciferase study investigated the effect of genetic variation within the *MAPT* 3'UTR on gene expression. As described previously, the 3'UTR is responsible for determining the half-life of the mRNA transcript and therefore plays an important role in mRNA stability. Stable mRNA transcripts are more highly expressed than unstable transcripts as they survive for longer before being degraded, producing higher steady-state levels. Thus, genetic variation within the 3'UTR that alters the stability of the transcript can also modify gene expression levels. Tau mRNA is a stable transcript and exhibits a relatively long half-life in neuronal cells [229].

The *MAPT* 3'UTR is approximately 4.4kb in length and contains several regions with a high level of sequence conservation. A series of luciferase constructs was created containing either the full-length 3'UTR or one of three overlapping fragments representing the 5', middle or 3' sections of the 3'UTR. The full-length constructs were created to determine whether genetic differences in the 3'UTRs of the H1B, H1C and H2 *MAPT* variants affected gene expression and, if so, the three deletion constructs would identify which regions are the most critical.

3.3 Patients

Three patients were identified from an existing cohort of pathologically confirmed PSP patients of European descent; one homozygous for the H1B haplotype, one homozygous for the H1C haplotype and one homozygous for the H2 haplotype. Haplotype and sub-haplotype status was previously confirmed by the genotyping of 5 tagging SNPs (rs1467967, rs242557, rs3785883, rs2471738, rs7521) [181] and the 238bp intron 9 deletion traditionally used to distinguish H2 from H1 (table 3.1).

Patient	rs1467967	rs242557	rs3785883	rs2471738	rs7521	intron9	Haplotype
P1	G/G	**G/G**	G/G	C/C	A/A	ins/ins	**H1B/H1B**
P2	A/A	**A/A**	G/G	T/T	G/G	ins/ins	**H1C/H1C**
P3	A/A	**G/G**	G/G	C/C	G/G	del/del	**H2/H2**

Table 3.1 *MAPT* haplotype determination of DNA samples.
The five tagging SNPs and one intron 9 deletion used to confirm haplotype status of the three PSP patients.

During the course of this project the PSP cohort was subject to a clinical review and the H2 patient was re-classified as a Parkinson's disease patient. The change in clinical diagnosis does not affect this project as *MAPT* haplotype status, rather than disease status, is of most importance.

3.4 DNA samples

Post mortem brain tissue was acquired from the Queen Square Brain Bank in London. DNA was previously extracted from frontal cortex tissue using standard molecular biology protocols. This study was approved by the Joint Ethics Committee of the Institute of Neurology and National Hospital for Neurology and Neurosurgery.

3.5 Luciferase reporter gene plasmids: Promoter constructs

3.5.1 Design

The first part of the study comprised an investigation into the influence of the rs242557 regulatory domain and its allelic variants on the regulation of *MAPT* transcription. The design of the study was based on that of Myers and colleagues [12], where 1.3kb of sequence surrounding the *MAPT* core promoter and 812bp of sequence surrounding the rs242557 polymorphism were cloned from three *MAPT* haplotype variants, H1B, H1C and H2. This study was more inclusive than that published by Rademakers *et al* [11], which incorporated only 182bp of sequence surrounding rs242557. In addition, the Rademakers study only looked at the effect of rs242557 alleles on the H1 core promoter variant and therefore potential H1/H2 differences were not investigated. The most intriguing difference between the studies, however, was the positioning of the rs242557 domain relative to the promoter element. The genomic location of this domain is approximately 46.8kb downstream to the *MAPT* core promoter in intron -1. In the Myers luciferase study the SNP domain was cloned immediately downstream to the core promoter element, in its more natural position. Rademakers and colleagues, however, chose to clone the domain immediately upstream to the promoter element.

To ascertain whether the positioning of the SNP domain relative to the core promoter could affect its function – and to clarify the conflicting results of the two previous studies – two luciferase constructs were created, one with the SNP domain cloned upstream to the core promoter element, and one with an identical domain situated downstream. Both the core promoter and SNP domain elements were identical to the ones described by Myers and colleagues [12].

The second part of the study involved an additional element comprising the 901bp sequence situated immediately downstream to the 1.3kb core promoter element and containing the bi-directional *MAPT-AS1* NAT promoter. Two luciferase constructs were created, with the element inserted alone in the forward and reverse orientations, in order to confirm the presence and strength of the bi-directional promoter. An additional luciferase construct was created, with the NAT promoter element cloned immediately downstream to the core promoter element in the forward direction – producing an extended promoter element encompassing the 2.2kb of highly conserved genomic sequence surrounding the major transcription start site at exon 0.

Three versions of each promoter construct were created, representing the genetic variation of the H1B, H1C and H2 *MAPT* haplotypes.

3.5.2 Element sequences and genetic variation

The three promoter elements – the core promoter (denoted 'CP'; 1.3kb), the SNP domain (denoted 'SD'; 812bp) and the NAT promoter region (denoted 'NP'; 901bp) – show a high degree of sequence conservation between human and mouse genomes (figures 3.2 and 3.3).

3.5.2.1 CP: the *MAPT* core promoter (chr17:43971166-43972505)

Exon 0 contains the sequences required for the initiation of *MAPT* transcription and falls between nucleotides 582 and 822 (240bp) in the CP element (figure 3.2). The H1B and H1C core promoter elements (the full 1.3kb) are identical, as confirmed by sequencing, and thus only one CP element representing the H1

haplotype was created for comparison with the H2 element. There are eight single nucleotide differences between the H1 and H2 CP elements, including two within exon 0 and one single nucleotide insertion/deletion polymorphism. There is also one five-nucleotide deletion (located between nucleotides 45 and 49) and one ten-nucleotide insertion (at position 78) on the H2 variant. At the 3' end of the element there is a TG dinucleotide repeat polymorphism that is predicted to form a binding domain for the RNA-binding protein TDP-43, a protein known to affect gene expression at multiple levels [230, 231]. A multiple sequence alignment (performed by ClustalW2) of the two CP element variants is given in Appendix A.

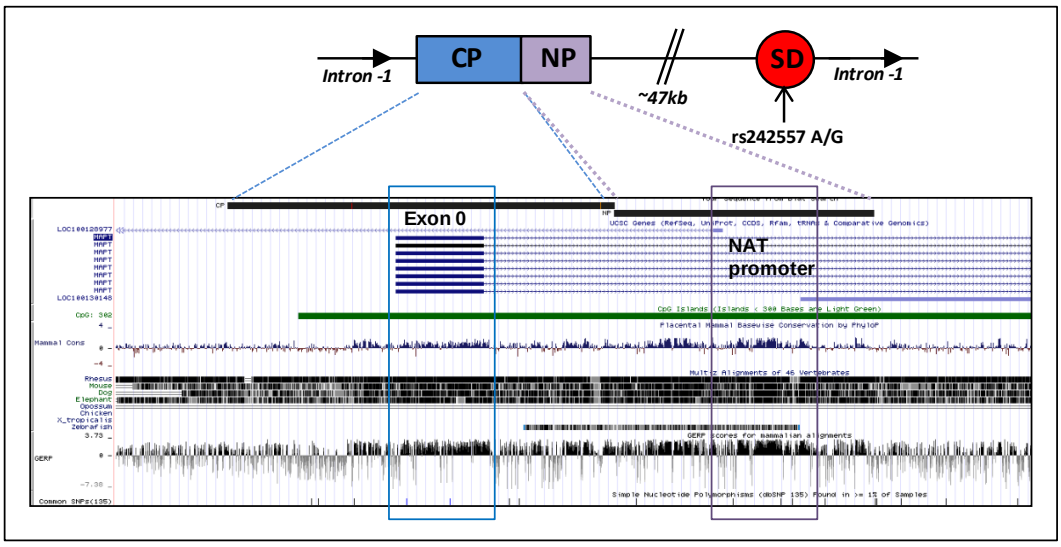


Figure 3.2 BLAT alignment of the CP and NP elements against *MAPT*
Alignments performed using the BLAT tool of the UCSC genome browser. Both elements are highly conserved and the NP element lies immediately downstream to the CP element.

3.5.2.2 The rs242557 'SD' SNP domain (chr17:44019339-44020150)

The rs242557 polymorphism is located at nucleotide 374 in the 812bp SD element. This is the only sequence difference between the H1B and H1C variants, with the G-allele present in H1B and the PSP risk-associated A-allele present in H1C. There are a further three single nucleotide differences unique to the H2 sequence, present alongside the rs242557 G-allele. A multiple sequence alignment of the three SD element variants is given in Appendix B and potential transcription factor binding sites within the SD are annotated in Appendix K.

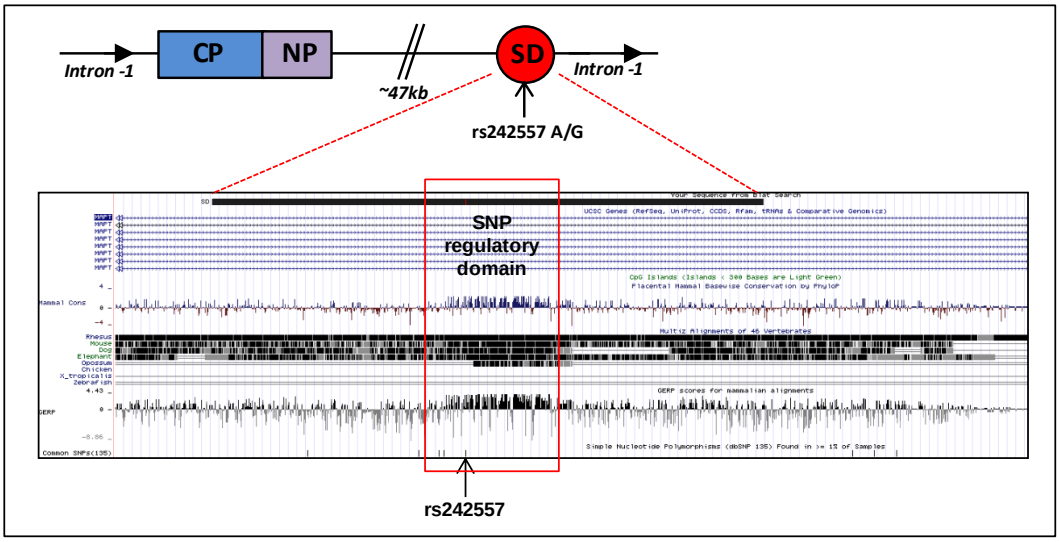


Figure 3.3 Blat alignment of the SD element against *MAPT*
Alignment performed using the Blat tool of the UCSC genome browser. The SD element is located approximately 47kb downstream to the CP. The putative transcription regulatory domain lies in the centre of the element in a highly conserved region and contains the rs242557 polymorphism.

3.5.2.3 The 'NP' NAT promoter region (chr17:43972506:43973404)

The predicted bi-directional promoter is located between nucleotides 340 and 374 (34bp) in the 901bp NP element. There is one nucleotide difference between the H1B and H1C elements – a known C/T polymorphism denoted rs3744457 – located just 36bp upstream to the antisense promoter at nucleotide 410. The C-allele is present on the H1B variant, with the H1C and H2 variants containing the T-allele. There are a further three single nucleotide differences unique to the H2 variant. A multiple sequence alignment of the three NP element variants is given in Appendix C.

3.5.3 Promoter element cloning: PCR

The promoter luciferase constructs were created using the pGL4.10 [*luc2*] luciferase reporter vector (Promega; figure 3.4A) and thus the cloning of the promoter elements was designed to facilitate their insertion into this vector. The pGL4.10 vector contains the promoterless *luc2* firefly luciferase gene located downstream to a number of unique restriction enzyme recognition sequences together comprising the multiple cloning site (MCS). The MCS was used to directionally insert each element into the vector upstream to the firefly luciferase

gene, thus requiring the attachment of specific restriction enzyme sites onto the end of each element. This was achieved by PCR, with the appropriate six-nucleotide recognition sequences added onto the 5' ends of the forward and reverse primers. The use of different sites at either end of the element allowed directional insertion into the pGL4.10 vector, as determined by the order in which the two sites appear in the MCS. The restriction sites incorporated into each element were selected based on four criteria: firstly, the sites were present in the MCS of the pGL4.10 vector; secondly, the sites did not occur naturally within the element itself; thirdly, the combination of sites would result in the insertion of the element into the MCS in the required orientation; fourthly, the chosen sites would allow, where required, the sequential and directional cloning of two different elements into the same pGL4.10 construct. Figure 3.4B presents schematics of the full set of promoter luciferase constructs created, with the combination of elements included in each one.

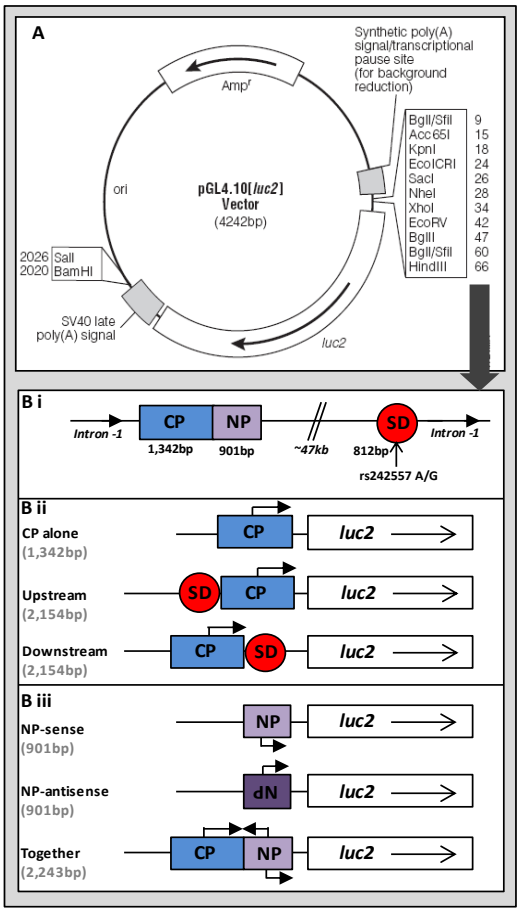


Figure 3.4 The pGL4.10 [*luc2*] promoter luciferase constructs
A: The pGL4.10 [*luc2*] vector used to construct each promoter luciferase construct (© Promega)
B: i) the genomic organisation of the three promoter elements; ii) the three constructs created to test the function of the SD domain on transcription from the CP; iii) the three constructs created to test the function of the NP domain

To create the upstream and downstream variants of the joint CP and SD constructs, two versions of the SD element were required with differing flanking restriction sites allowing insertion at either end of the CP element. The NP element – unlike the CP and SD elements – was cloned using a single restriction site (NheI) incorporated onto both ends. This removed the ability to control the direction of the insertion but enabled the creation of two separate pGL4.10

constructs in one cloning reaction; each with the same NP element inserted in either the forward or reverse orientations.

Genetic variants of each element were cloned from the genomic DNA of the three PSP patients carrying the H1B, H1C and H2 *MAPT* haplotype variants (see section 3.3). Each element was amplified by PCR using the primers and conditions detailed in table 3.2. The restriction enzyme recognition sequences attached to the 5' end of each primer are also included. Typical 25µl PCR reactions comprised: 50ng of genomic DNA, 1x FastStart High Fidelity reaction buffer with 1.8mM magnesium, dNTPs (each to a final concentration of 10mM), forward and reverse primers (each to a final concentration of 0.2µM) and 2.5 units of FastStart High Fidelity polymerase mix. PCR occurred during a series of heating and cooling steps comprising: an initial denaturation step at 94°C for 5 minutes, 35 cycles of denaturation at 94°C for 30 seconds, annealing at a temperature appropriate for the primer pair (table 3.2) for 30 seconds and extension at 72°C for 1-2 minutes, and a final extension at 72°C for 7 minutes. A 5µl aliquot of the product was resolved by agarose gel electrophoresis and successful amplification was confirmed by comparison against a size marker. Products from four replicate 25µl PCR reactions were pooled and purified using the QIAquick PCR Purification kit.

Element	Size (bp)	Primer (F/R)	5'site	Sequence (5'-3')	AT (°C)
CP	1,342	F	SacI	GAGCTC_CAAATGCTCTGCGATGTGTT	55
		R	NheI	GCTAGC_GGACAGCGGATTTCAGATTC	
SD upstream	812	F	KpnI	GGTACC_TGGGACAGATCCTCAGTG	60
		R	SacI	GAGCTC_GGCTGTCGATGAACCCTA	
SD downstream	812	F	NheI	GCTAGC_TGGGACAGATCCTCAGTG	60
		R	EcoRV	GATATC_GGCTGTCGATGAACCCTA	
NP	901	F	NheI	gaGCTAGC_TGCCGCTGTTCGCCATCAG	60
		R	NheI	gtGCTAGC_ACCCTCAGAATAAAAGCCAG	

Table 3.2 The primers used to amplify each promoter element from the genomic DNA of the three PSP patients.
Each primer contains a restriction enzyme site at the 5' end to allow the sequential ligation of each fragment into the multiple cloning site of the pGL4.10 luciferase reporter vector. AT: annealing temperature of the primer pair used during PCR.

3.5.4 Promoter element cloning: pGEM-T Easy

The purified PCR products were, in the first instance, cloned into the pGEM-T Easy vector (Promega; figure 3.5). This linearised plasmid vector is commonly used for the cloning of PCR products as it has a 3' T-nucleotide overhang that complements the 3' A-nucleotide overhang produced by most DNA polymerases during PCR. The pGEM-T Easy vector also contains the α-peptide of the β-galactosidase gene, allowing the easy identification of successful recombinants by blue/white screening (described in section 2.1.4.6).

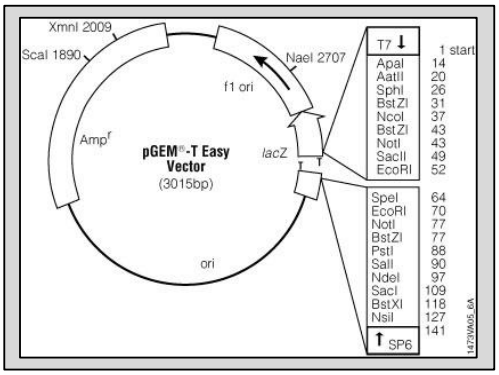


Figure 3.5 The pGEM-T Easy vector
This vector was used to clone PCR products (© Promega)

Ligation reactions (10µl) were incubated overnight at 4°C and comprised: 50ng of linearised pGEM-T Easy vector, 100-150ng of purified PCR product, 10mM ATP, 1x T4 DNA ligase buffer and 1-2 units of T4 DNA ligase enzyme. Half of the ligation mixture was transformed into 50µl of High Efficiency JM109 *E.coli* cells and incubated overnight at 37°C on LB-agar plates containing 50µg/ml of ampicillin, 0.1mM of IPTG and 20µg/ml of X-Gal. White colonies were manually picked and cultured overnight at 37°C in 3ml of L-broth containing 50µg/ml of ampicillin. Cultures were subject to continuous horizontal agitation at 250rpm during incubation. Plasmid DNA was extracted from the bacterial cells using the QIAprep Spin Miniprep kit. Positive clones were identified by sequencing with the SP6 and T7F primers that anneal at either side of the insertion site (SP6: TATTTAGGTGACACTATAG; T7F: TAATACGACTCACTATAGGG).

3.5.5 Promoter element cloning: pGL4.10 [*luc2*]

The cloned promoter elements were removed from the pGEM-T Easy vector by restriction enzyme digestion using the unique recognition sequences inserted onto the ends of each element. Each 50µl digestion comprised: 5µg of purified plasmid DNA, 1x digestion buffer, 1x bovine serum albumin (BSA) and 25 units of each restriction enzyme. An aliquot of the pGL4.10 vector was similarly prepared for

insertion of the promoter element by digestion with the same enzyme(s). The enzyme and buffer combinations for the digestion of each element are given in table 3.3.

Plasmid	Product size (bp)	5' Enzyme	3' Enzyme	Restriction buffer	Incubation temp (°C)
pGEM-T/CP	1,342	SacI	NheI	NEB1	37
pGEM-T/SD-up	812	KpnI	SacI	NEB1	37
pGEM-T/SD-down	812	NheI	EcoRV	NEB2	37
pGEM-T/NP	901	NheI	NheI	NEB2	37

Table 3.3 The restriction enzymes and digestion buffer used to remove the cloned promoter element from the pGEM-T Easy plasmid vector.
An aliquot of the pGL4.10 luciferase vector was similarly prepared for the insertion of each digested element.

The digestion products were resolved by agarose gel electrophoresis to separate the empty pGEM-T Easy vector from the newly liberated element. The latter was excised from the gel using a sharp, sterile scalpel and purified using the QIAquick Gel Extraction kit. The digested pGL4.10 vector was similarly purified.

The digested element was then ligated into the pGL4.10 vector by overnight incubation at 4°C with T4 DNA ligase. Each 10µl ligation reaction comprised: 50ng of digested pGL4.10 vector, 100-150ng of digested promoter element, 10mM of ATP, 1x T4 DNA ligase buffer and 1-2 units of T4 DNA ligase enzyme. JM109 *E.coli* cells were transformed with the full ligation mixture and selected on LB-agar plates containing 50µg/ml of ampicillin. Plasmid DNA was isolated from five to ten colonies as before using the QIAquick Spin Miniprep kit. A 1µl aliquot of the purified DNA was screened by digestion to ascertain the successful insertion of the promoter element. Final confirmation of the luciferase construct was achieved by sequencing with primers that anneal at either side of the insertion site:

RVp3: TAGCAAAATAGGCTGTCCCC

Luc-R: ATGTGCGTCGGTAAAGGCG

Some of the luciferase constructs required the insertion of a second promoter element immediately adjacent to the CP. To create these constructs, a second round of cloning was undertaken in which the pGL4.10/CP construct was digested with a pair of enzymes matching those on the ends of the second element. The additional element was inserted using the digestion-ligation method described above. The digestion components for the second round of cloning are given in table 3.4.

Plasmid	To insert	5' Enzyme	3' Enzyme	Restriction buffer	Incubation temp (°C)
pGL4.10/CP	SD-up	KpnI	SacI	NEB1	37
pGL4.10/CP	SD-down	NheI	EcoRV	NEB2	37
pGL4.10/CP	NP-sense	NheI	NheI	NEB2	37

Table 3.4 The restriction enzymes, digestion buffer and incubation temperature used to digest a second element for insertion into the CP luciferase construct.

3.6 Luciferase reporter gene plasmids: 3'UTR constructs

3.6.1 Design

The second part of the study investigated the role of genetic variation within the *MAPT* 3'UTR on gene expression. In addition to the full-length version, a set of deletion constructs were created to identify the regions of the 3'UTR that are most critical for maintaining the normal pattern of gene expression. It was not possible to amplify the full 3'UTR in one PCR reaction as the forward and reverse primers required to achieve this were incompatible. Instead, the 3'UTR was split into three separate fragments of 1,179bp, 1,828bp and 1,981bp in size, with the second fragment overlapping the first and third fragments by 312 and 314 nucleotides respectively. The three deletion constructs each comprised one of the overlapping fragments.

The full-length 3'UTR contains two naturally-occurring restriction sites that appear within the region only once. The deletion constructs were designed such that each overlapping region contained one of these unique recognition sequences. The AatII enzyme cuts in the overlap between fragments 1 and 2, with XbaI

cutting in the overlap between fragments 2 and 3. Thus, these two enzymes were used to ligate the three fragments together, creating the full-length construct. Figure 3.6B presents the design of the four 3'UTR constructs. As with the promoter plasmids, three sets of 3'UTR constructs were created to represent the genetic variation of the H1B, H1C and H2 *MAPT* variants.

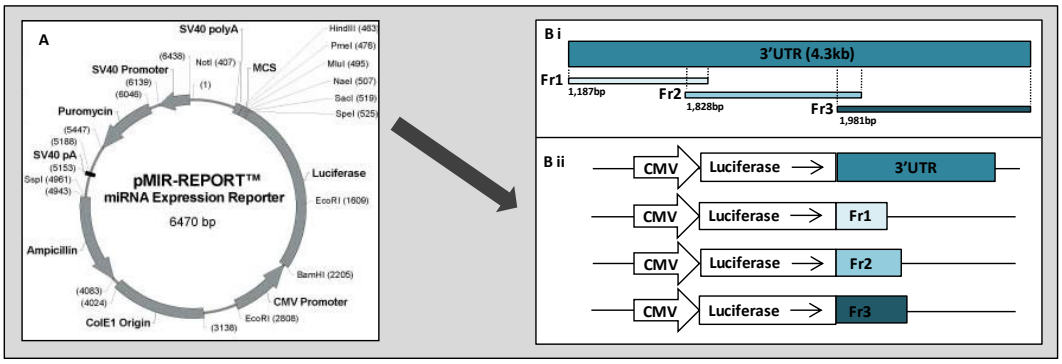


Figure 3.6 The pMIR-REPORT 3'UTR luciferase constructs
A: the pMIR-REPORT vector used to create each 3'UTR luciferase construct (© Promega). B: i) The section of the 3'UTR included in each deletion fragment; ii) The four luciferase constructs created to test the stability of the *MAPT* 3'UTR.

3.6.2 Fragment sequences and genetic variation

The full-length *MAPT* 3'UTR is 4,370bp in length and is located at chr17:44101545-44105914. The three deletion fragments are denoted Fr1, Fr2 and Fr3 and multiple sequence alignments (performed by ClustalW2) highlighting the genetic differences between the H1B, H1C and H2 variants of each fragment are given in appendices D, E and F.

3.6.2.1 Fragment 1 (Fr1)

The 5' end of the 3'UTR is contained in Fragment 1 (1,179bp). The three variants share 14 nucleotide differences; nine are H1/H2 differences including two insertion/deletion polymorphisms (one single nucleotide and one dinucleotide), four are H1B/H1C single nucleotide changes and one nucleotide at position 234 differs in all three variants (H1B-del, H1C-A, H2-T). Six of these sequence differences are in the region of Fr1 that overlaps with Fr2.

3.6.2.2 Fragment 2 (Fr2)

The middle section of the 3'UTR is contained in Fragment 2 (1,828bp). In addition to the six sequence differences in the region overlapping Fr1, there are a further 13 nucleotide variations between the three variants. The H1B and H1C variants differ by two nucleotides including one insertion/deletion polymorphism. The H2 variant differs from the H1 fragments by 11 nucleotides, including one dinucleotide and one trinucleotide deletion.

3.6.2.3 Fragment 3 (Fr3)

The 3' section of the 3'UTR is contained in Fragment 3 (1,981bp). The H1B and H1C variants have five nucleotide differences including one deletion. The H2 variant contains nine differences including two trinucleotide deletions. There are a total of 14 nucleotide differences between the three variants.

3.6.3 3'UTR fragment cloning: PCR

The 3'UTR constructs were created using the pMIR-REPORT luciferase vector (Promega; figure 3.6A), a commonly used plasmid vector containing the firefly luciferase gene under the control of the highly active cytomegalovirus (CMV) promoter. The 3'UTR of the luciferase gene has been removed and replaced with a multiple cloning site, allowing the insertion of the *MAPT* 3'UTR downstream to the luciferase gene. The three deletion fragments were individually cloned into pMIR-REPORT and these constructs were then used to create the full-length construct.

As with the promoter constructs, the 3'UTR fragments were inserted into the MCS of the pMIR-REPORT vector by restriction enzyme digestion and ligation. Thus, specific restriction enzyme recognition sequences were, again, inserted onto the ends of each fragment by PCR. As the full-length 3'UTR was formed using naturally-occurring restriction sites, the recognition sequences introduced onto the ends of the fragments were required solely for the creation of the individual deletion constructs. Thus, the same two sites, SacI and HindIII, were added onto 5' and 3' ends, respectively, of each fragment.

The three fragments were amplified from the same genomic DNA samples used in the promoter study (see section 3.3). PCR was conducted as described earlier, with the primer sequences and reaction conditions given in table 3.5. A 5µl aliquot of the PCR product was resolved by agarose gel electrophoresis and successful amplification was confirmed by comparison against a size marker. Products from four replicate PCR reactions were pooled and purified using the QIAquick PCR Purification kit.

Element	Primer (F/R)	5'site	Sequence (5'-3')	Mg (mM)	AT (°C)	Elongation time	PCR cycles	Size (bp)
Fr1	F	SacI	GAGCTC_CCTGGGGCGGTCAATAA	1.8	65	1.5 mins	35	1179
	R	HindIII	AAGCTT_AGGCAGTGATTGGGCTCTC					
Fr2	F	SacI	GAGCTC_GTAGGGGGCTGAGTTGAG	1.8	60	2 mins	35	1828
	R	HindIII	AAGCTT_ACCAGAAGTGGCAGAATTGG					
Fr3	F	SacI	GAGCTC_CAGACTGGGTTCCTCTCCAA	1.8	65	2 mins	35	1981
	R	HindIII	AAGCTT_GCCAGCATCACAAAGAAG					

Table 3.5 The primers, restriction sites, magnesium concentration (Mg), annealing temperature (AT) and number of PCR cycles used to amplify the three 3'UTR deletion fragments.

3.6.4 3'UTR fragment cloning: pGEM-T Easy

The Fr1, Fr2 and Fr3 PCR products, as with the promoter constructs, were first cloned into the pGEM-T Easy vector using the protocol described above in section 3.5.4.

3.6.5 3'UTR fragment cloning: pMIR-REPORT

To create the individual deletion constructs the cloned fragments were excised from their pGEM-T Easy vector in a double restriction digest comprising: 5µg of the plasmid DNA, 25 units each of SacI and HindIII enzymes, 1x NEB2 buffer and 1x BSA. The pMIR-REPORT vector was similarly prepared. The digestion mixture was incubated overnight at 37°C and the products resolved by agarose gel electrophoresis. The digested fragments were excised and purified using the QIAquick gel extraction kit, as before.

The fragments were individually ligated into the pMIR-REPORT vector by overnight incubation at 4°C with T4 DNA ligase. Ligation, transformation in JM109 *E.coli* cells, ampicillin selection, liquid culture and miniprep DNA purification were all conducted as described above. An aliquot of the purified

plasmid DNA was screened by digestion to confirm the presence of the promoter element. Final confirmation was achieved by sequencing with primers that anneal at either side of the insertion site (M13-F: TGTAAAACGACGGCCAGT; M13-R: AGGAAACAGCTATGACCAT).

To create the full-length construct, a further two rounds of cloning were required. The first inserted the Fr2 fragment into the Fr1 luciferase construct by digestion/ligation using the AatII (which cuts in the Fr1/Fr2 overlapping region) and HindIII (which cuts at the end of each fragment) enzymes. Fr3 was similarly inserted into the Fr1/Fr2 luciferase construct using the XbaI (which cuts in the overlap between Fr2 and Fr3) and HindIII enzymes. Final confirmation was, again, achieved by sequencing with the M13 F/R primers.

3.7 Cell lines

Each luciferase construct was assayed *in vitro* in two different human neuroblastoma cell lines, SH-SY5Y and SK-N-F1. These neuronal cell lines express tau endogenously, though only foetal tau (the shortest 0N3R isoform) is expressed when the cells are in an undifferentiated state. The addition of retinoic acid to the culture medium causes the cells to differentiate, producing a neuronal phenotype with the expression of all six adult tau isoforms that more closely resembles *in vivo* neuronal conditions. These particular cell lines were chosen for their *MAPT* haplotype status; the SH-SY5Y cells are H1 homozygous, whereas SK-N-F1 cells are H1/H2 heterozygous. In addition to genetic differences, the cell lines also show distinct morphologies in culture (figure 3.7). Comparison between the two lines will highlight any differences in luciferase activity that are due to endogenous differences in transcriptional regulation. Unfortunately there are no H2 homozygous cell lines currently available for comparison.

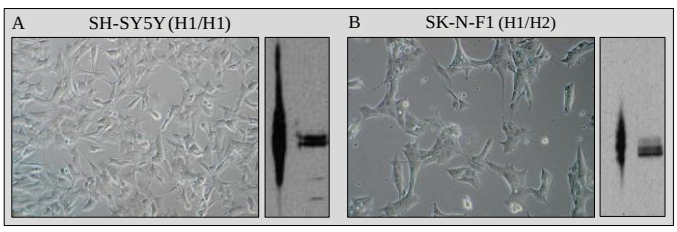


Figure 3.7 The two neuroblastoma cell lines. SH-SY5Y (A) and SK-N-F1 (B) cells have distinct morphologies in culture.

3.8 Transfection

Transfection was conducted 24 hours after the cells were transferred into a 96-well opaque cell culture plate, when approximately 80% confluent. Each promoter/3'UTR construct (containing the firefly luciferase gene) was transiently transfected into three replicate wells. The empty pMIR-REPORT luciferase vector was included in triplicate on all plates for normalisation (see section 3.10). Transfection in mammalian cells is inhibited by endotoxins commonly found in mini-preparations of plasmid DNA. Thus, an endotoxin-free maxi preparation of each luciferase construct was made specifically for transfection.

Each well was transfected with 200ng of the firefly construct and 50ng of a *Renilla* luciferase plasmid under the control of the tyrosine kinase promoter (pRL-TK). Co-transfection with the *Renilla* plasmid provides an internal control for the correction of differences in transformation efficiency. A volume of 1.5ul of TransFast transfection reagent (1mM) and 40ul of serum-free culture medium was added to each well (giving a 1:1 charge ratio of DNA to transfection reagent) and transfection occurred during a one hour incubation at 37°C.

3.9 Luciferase reporter assay

The luciferase assay was conducted 48 hours post-transfection using the Dual-Glo Luciferase Reporter Assay System (Promega). A volume of 20µl of Dual Glo Luciferase Reagent was added to 20µl of fresh serum-free culture medium in each well. This reagent induces a luciferase signal from the firefly luciferase reporter only (i.e. from the pGL4.10 and pMIR-REPORT constructs), which was quantified after a ten-minute room temperature incubation using the Tecan GENios luminometer and XFLUOR4 (version V 4.30) software. The luminescence reading was taken with an integration time of 1000ms and a gain setting of 150. A volume of 20µl of Dual Glo Stop & Glo Reagent was then added to each well. This reagent quenches the firefly luciferase signal and immediately induces the *Renilla* luciferase signal from the control plasmid. The *Renilla* signal was quantified with the same GENios settings after a ten-minute room temperature incubation.

3.10 Luciferase assay results

The relative luciferase activity is given by the ratio of firefly to *Renilla* signal emitted from each well and accounts for any changes in signal caused by well-to-well differences in cell density and/or transformation efficiency. The relative luciferase activity of each promoter construct was then normalised against the average relative luciferase activity of the three pMIR-REPORT positive control wells included on each plate. This allows direct comparison of luciferase activity from multiple plates and cell lines. The normalised results for each construct were averaged across the three replicate wells. Each construct was assayed in a minimum of three independent experiments and the mean relative luciferase activity across the replicates was calculated. A significant difference (defined as $p \leq 0.05$) in relative luciferase activity between two constructs was detected using a two-tailed Student's t-test.

3.11 The funtional effect of the rs242557 domain on transcription from the *MAPT* core promoter

3.11.1 'Upstream' vs 'Downstream' positioning of the rs242557 element affects its function

The first step in unravelling the functional role of the rs242557 polymorphism in PSP risk was to confirm the nature of the effect, if any, of the rs242557-containing regulatory domain (SD) on transcription from the core promoter (CP). In an attempt to clarify previous conflicting reports regarding the allelic effects of the polymorphism, two luciferase constructs were created in which the SD element was inserted either upstream or downstream to the CP element. These constructs, along with an additional construct containing the CP element alone, were assayed in undifferentiated SK-N-F1 (denoted 'F1') and SH-SY5Y ('SH') neuroblastoma cells. Comparative luciferase activity was initially assayed in neuronally differentiated cells treated with retinoic acid for five days but did not differ significantly from that quantified in undifferentiated cells (data not shown). The relative luciferase activities of each construct 48 hours post transfection in undifferentiated cells is given in figure 3.8. The results are presented by haplotype

set – H1B, H1C and H2 – with the error bars representing the standard error of the mean from three biological replicates. The key findings are summarised below.

3.11.1.1 The SD element functions as a repressor of transcription when inserted downstream to the CP

The addition of the SD element downstream to the CP produced the strongest and most consistent effect on transcription. For all three haplotype variants in both cell lines the downstream addition of the SD element significantly repressed transcription from the core promoter. This repression was strongest for the H2-G variant (F1: 6.7-fold reduction; SH: 11.8-fold; $p<0.0001$ for both) and weakest for the H1C-A variant (F1: 1.8-fold reduction, $p=0.0447$; SH: 3.7-fold, $p=0.0145$) (figure 3.8).

3.11.1.2 The function of the SD is determined by the cellular conditions when inserted upstream to the CP

The effect of the upstream addition of the SD element on transcription differed depending on the cell line in which it was assayed. In F1 cells, the element functioned as an enhancer, with a general trend of increased transcription observed for all three variants. This increase reached statistical significance for the H2-G variant (2.2-fold, $p=0.0007$) and trended towards significance for the H1B-G variant (1.6-fold, $p=0.0602$). The H1C-A allele of the SD did not significantly alter transcription when inserted upstream to the CP ($p=0.2284$).

In SH cells, however, the picture was very different. For the H2-G variant, the ability to modify transcription from the CP was lost when the SD was moved from the downstream to the upstream position, with no difference in activity observed between the CP and upstream constructs ($p=0.7981$). For the H1B-G and H1C-A variants, however, the SD element continued to function as a repressor, but only the H1C-A variant was close to achieving a statistically significant reduction in CP transcription (H1B-G: 1.5-fold, $p=0.3337$; H1C-A: 3.2-fold, $p=0.0687$) (figure 3.8).

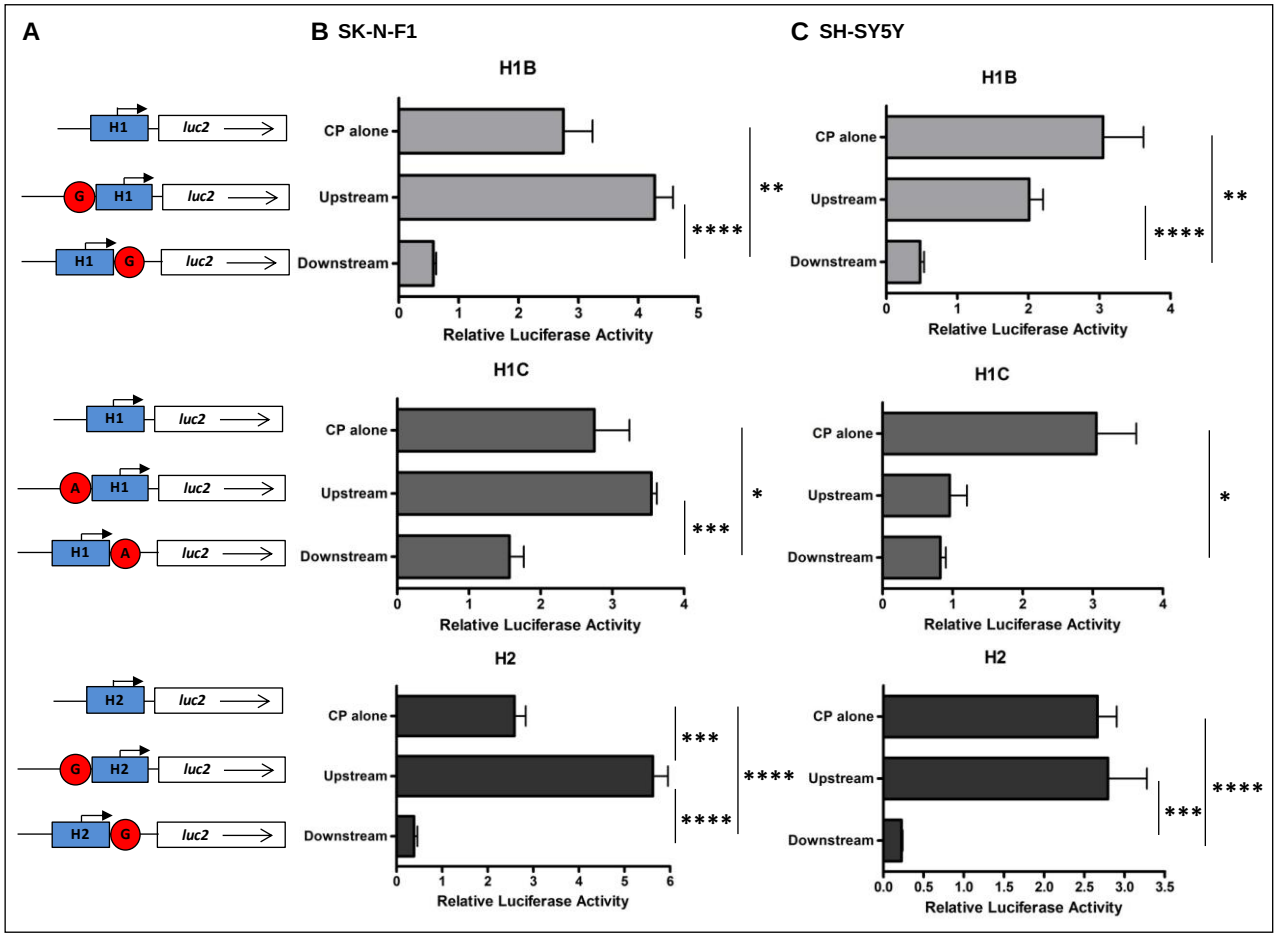


Figure 3.8 Promoter luciferase results 1

A: The luciferase construct variant assayed in each instance, separated by haplotype.

B: The relative luciferase activity of each construct in SK-N-F1 cells, separated by haplotype.

C: The relative luciferase activities in SH-SY5Y cells.

*	p≤0.05
**	p≤0.01
***	p≤0.001
****	p<0.0001

3.11.1.3 The function of the H1C-A SD variant is unaffected by a change in positioning in SH-SY5Y cells

In general, the relative luciferase activities of the upstream constructs were consistently and significantly higher than their downstream counterparts – a difference that can be solely attributed to the positioning of the SD. The only occurrence for which this was not true was with the H1C-A variant in SH cells, where the constructs conferred equal levels of transcriptional repression (p=0.5400). This suggests that the A-allele variant of the SD is not affected by its positioning relative to the H1 core promoter in this cell line.

3.11.2 The allelic variants of the rs242557 element differentially affect transcription from the core promoter

A simpler picture is produced when, rather than comparing positional variants within a haplotype set, the activity of the allelic variants of the same construct are considered. Figure 3.9 provides a different presentation of the results discussed above and given in figure 3.8. This time the results are split by construct, with the H1B, H1C and H2 variants presented on the same bar graph. Unlike the positional comparisons, the allelic differences in transcriptional activity were consistent in both cell lines. There was no significant difference in relative luciferase activity between the two CP variants (F1: p=0.7765; SH: p=0.5375), suggesting that unregulated *MAPT* transcription from H1 and H2 chromosomes is of equal strength. The H1C-A variant of the SD domain conferred a significantly different level of activity when added to the CP than the H1B-G and H2-G variants in both the upstream and downstream positions. When cloned upstream to the CP, the H1C-A variant conferred 1.2- to 2.1-fold lower transcriptional activity compared to H1B-G (F1: p=0.0805; SH: p=0.0282) and 1.5- to 2.9-fold lower compared to H2-G (F1: p=0.0037; SH: p=0.0278). In the downstream position the allelic differences were more pronounced, with a 1.7- to 2.7-fold and 3.6- to 4.1-fold *increase* in transcriptional activity conferred by the H1C-A variant compared to the H1B-G (F1: p=0.0140; SH: p=0.0060) and H2-G (F1: p=0.0006; SH: p<0.0001) variants, respectively.

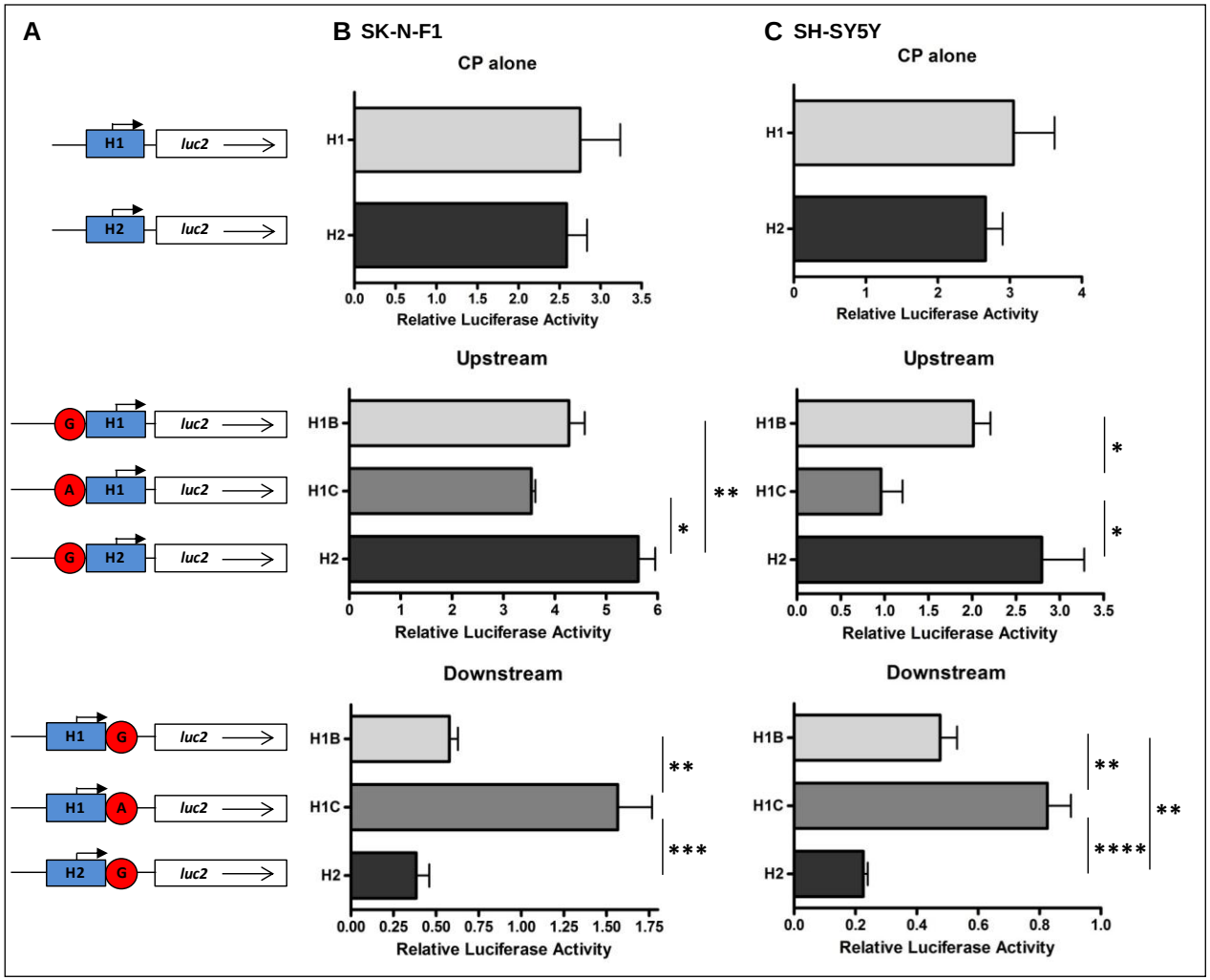


Figure 3.9 Promoter luciferase results 2

A: The luciferase construct variant assayed in each instance, separated by construct.

B: The relative luciferase activity of each haplotype variant in SK-N-F1 cells, separated by construct.

C: The relative luciferase activities in SH-SY5Y cells.

*	p≤0.05
**	p≤0.01
***	p≤0.001
****	p<0.0001

Thus, regardless of the positioning of the SD element, the H1C-A variant of the domain conferred significantly different activity than the H1B-G and H2-G variants, producing increased activity in the downstream and reduced activity in the upstream positions. As the positioning of the SD is the only difference between the upstream and downstream constructs, these results present a viable explanation for the discrepancies in the direction of the allelic effect of rs242557 reported by Myers and Rademakers; Myers and colleagues, who reported an increase in activity for the A-allele, cloned their element downstream to the CP element whereas Rademakers *et al* cloned theirs upstream and reported a decrease for the A-allele.

3.11.3 The relationship between the function of the SD and the strength of its interaction with the CP changes depending on the cell line

The differential behaviour of the SD constructs in each cell line is an intriguing finding, but perhaps becomes clearer when considered in conjunction with the comparative strengths of the different SD allelic variants. As described previously, the strongest and most consistent effect on transcription was observed when the SD was cloned in its more natural downstream position. This was the only finding that was consistent for all haplotype variants in both cell lines. There were, however, allelic differences in the strength of the repression and thus the SD variants can be classified based on the magnitude of their effect on the core promoter: H1C-A (weakest repression), H1B-G (moderate) and H2-G (strongest).

The behaviour of the SD element has also been shown to be affected both by changes in positioning and by variation in cell type. When these three things are taken together – the function of the SD, the strength of the allelic variant and the *in vitro* cellular conditions – an intriguing pattern begins to emerge which shows the importance of both the interaction between the CP and SD and the cellular environment.

In F1 cells, the strongest repressor in the downstream position – H2-G – becomes the strongest enhancer of transcription in the upstream position. Similarly, the

weakest downstream repressor – H1C-A – has no significant effect on transcription in the upstream position. The H1B-G moderate repressor becomes a moderate enhancer when moved upstream. This indicates that although the nature of the interaction between the SD and CP in F1 cells changes depending on the positioning of the two elements, the relative strengths of the allelic interactions do not significantly change (figure 3.10).

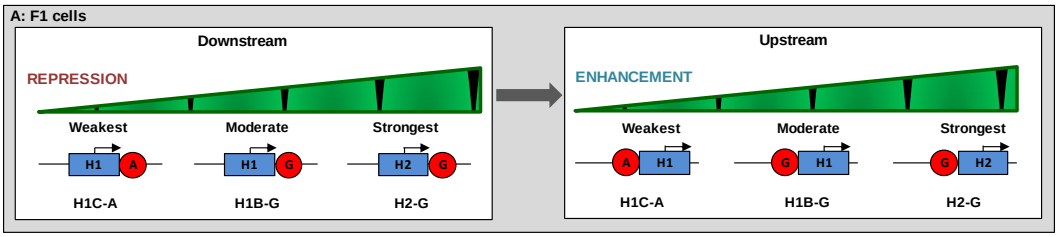


Figure 3.10 A schematic representation of the relationship between the positioning of the SD, haplotype-specific variation within it and SD function in SK-N-F1 cells. The strongest allelic SD repressor in the downstream position became the strongest enhancer in the upstream position.

In SH cells this relationship is inverted, with the H1C-A variant – the weakest downstream repressor – becoming the strongest upstream repressor. Similarly, the H2-G variant, which conferred the strongest downstream repression, was unable to exert influence on the CP from an upstream position. In therefore appears that, in these cells, it is the relative strength of the interaction, and not the nature of it, that is most affected by the change in positioning (figure 3.11).

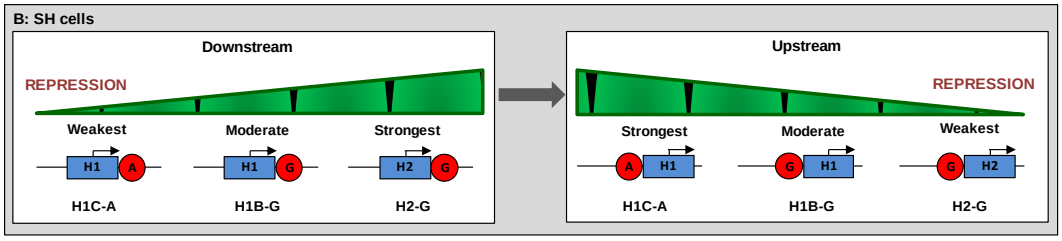


Figure 3.11 A schematic representation of the relationship between the positioning of the SD, haplotype-specific variation within it and SD function in SH-SY5Y cells. The strongest allelic SD repressor in the downstream position became the weakest repressor in the upstream position.

3.11.4 Evidence for an interaction between the *MAPT* core promoter and the rs242557 domain

During the construction of the promoter luciferase constructs, a CP clone was identified that had serendipitously mutated during the cloning process. This H1 CP construct – denoted H1X – had two single nucleotide errors inserted into the sequence during PCR or cloning in *E.coli* bacteria. The first error was an A to G transition at position 120 (A120G) at the 5' end of the CP element, over 460 nucleotides upstream to the start of exon 0 in a relatively unconserved region of the element. It was therefore unlikely that this error would affect transcription from the CP element. The second error, however, was a G to T transition at position 596 (G596T) and is located within exon 0 – the major transcription start site. The wildtype G nucleotide at this position is highly conserved. The A120G and G596T errors are highlighted in red in the ClustalW2 sequence alignment of H1 and H1X presented in Appendix G. To determine whether transcription rate was affected by these transitions, the H1X CP luciferase construct was assayed alongside the wildtype H1 version and the results are presented in figure 3.12. There was no difference in luciferase activity conferred by the two H1 CP variants in either of the cell lines (F1: p=0.8177; SH: p=0.9606), indicating that the G nucleotide at position 596 is not essential for the initiation and maintenance of transcription rate.

To determine whether this error altered the interaction between the CP and SD elements, the H1C-A SD element was inserted downstream to the H1X CP and assayed against the wildtype version (figure 3.12). The activity of the mutated construct (H1X-A) was 4- to 6-fold lower than the activity of the wildtype H1-A construct (F1: p=0.0026; SH: p=0.0002), reducing to a level equivalent to the activity of the H2-G wildtype variant. This is an intriguing finding and suggests, firstly, that a single nucleotide error in exon 0 can affect transcription rate, not by modulating the efficiency of transcription initiation at exon 0 but by altering the interaction between the SD regulatory domain and the core promoter; and secondly, that this altered interaction has a gain-of-function effect, serving to strengthen the normally reduced repression conferred by the H1C-A variant to

match that of the H2-G variant. The reason for this is unclear; however these findings provide evidence of a direct interaction between the core promoter and regulatory domain.

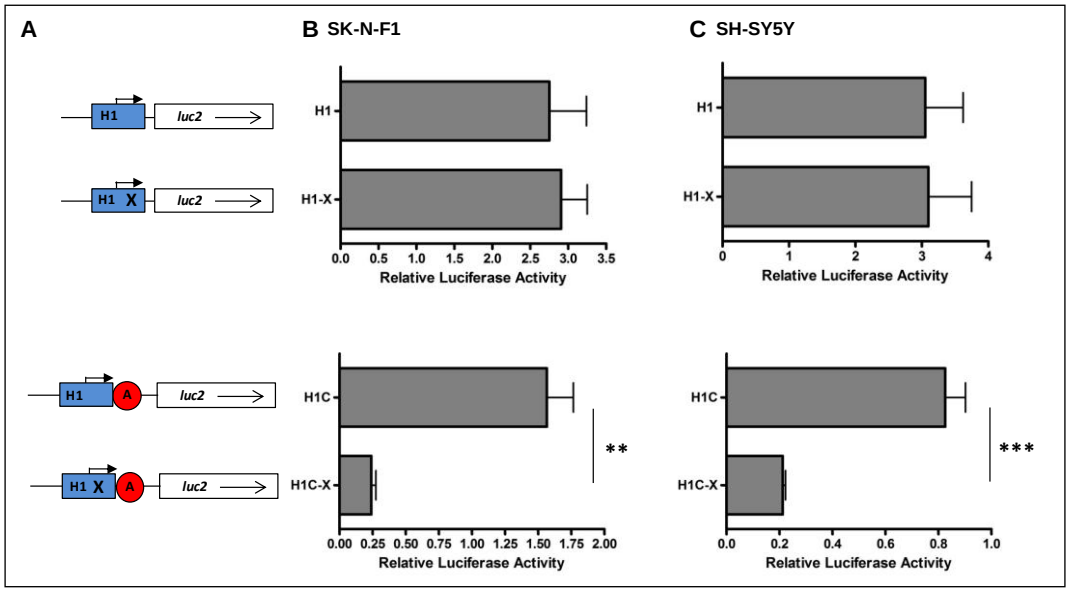


Figure 3.12 Promoter luciferase results 3
A: Schematics of the H1 and H1X versions of the CP luciferase construct and the H1-A and H1X-A versions of the downstream SD luciferase construct. B: The relative luciferase activity of each H1 variant in SK-N-F1 cells, separated by construct. C: The relative luciferase activities in SH-SY5Y cells. ** p≤0.01; *** p≤0.001.

3.11.5 Biological interpretation

This part of the study has confirmed previous reports that the rs242557 polymorphism falls within a *cis*-acting regulatory domain that is capable of modifying transcription from the *MAPT* core promoter. It has also shown that the position of the domain relative to the core promoter can potentially alter its function, thereby providing an explanation for the opposing results reported from similar studies by Myers and Rademakers. Most interesting of all, however, is the finding that the function of the regulatory domain can be differentially influenced, not only by genetic variation within the domain and its positioning, but by the cellular environment. This would suggest that the domain's influence on *MAPT* transcription results from a delicate balance between the proximity and orientation of its *cis*-acting signal, genetic variation within it and *trans*-acting binding factors expressed by the cell line. This relationship will be explored further in chapter 5.

The luciferase reporter assay is, of course, an artificial means of assessing the ability of short DNA sequences to initiate and/or modify transcription and therefore it is difficult to draw any biological conclusions from these results. In particular, these results have shown the importance of the positioning of the SD for both strength and function and therefore the removal of the intervening 47kb of intronic sequence between the two elements – as occurs in the genome – is highly likely to affect their interaction. It is hypothesised that the two elements interact through changes in confirmation that form a 'loop' structure and bring the regulatory domain into close proximity to the core promoter. Transcription factors bound to the regulatory domain may then interact with those bound to the core promoter, thus altering the rate of transcription. If this is the case, the distal location of the regulatory domain would be a vital factor in its normal functioning and thus the luciferase constructs described here would not be biologically representative.

That being said, this study has confirmed that, at the basic sequence level, the rs242557 polymorphism lies within a stretch of sequence that can regulate transcription from the *MAPT* core promoter in *cis*, even when located proximally. More importantly, it has consistently shown that the alleles of the rs242557 polymorphism differentially affect this regulation, regardless of positioning and *in vitro* cellular conditions.

3.12 Functional assessment of the NAT promoter region, individually and in conjunction with the *MAPT* core promoter

3.12.1 Sense vs antisense

There are well characterised examples of antisense-mediated transcriptional regulation occurring via each of the four models described in section 3.2 [228], though the effect on transcription of the overlapping sense gene varies depending on the model. Expression of the antisense transcript is often correlated – either positively or inversely – with expression of the sense gene, though this is not always the case. The relationship between sense and antisense transcription can

hint at the mechanism connecting the two, as well as provide clues as to the biological consequences of the non-coding natural antisense transcript (NAT).

The second part of this luciferase study takes a more in-depth look at the highly conserved region located immediately downstream to the core promoter element described above. The 900bp region includes a 34 nucleotide sequence that is predicted to act as a bi-directional promoter for two non-coding transcripts, one transcribed in the sense (*MAPT-IT1*) and one in the antisense (*MAPT-AT1*) direction. To test whether this region (the NP element) is capable of initiating transcription in the sense and antisense directions, two luciferase constructs were created, one with the NP element inserted upstream to the promoterless luciferase gene in the forward or 'sense' direction (NP-S) and the second in which the element has been 'flipped' and lies in the reverse or 'antisense' direction (NP-A). Three variants of each construct representing the H1B, H1C and H2 haplotypes were assayed in undifferentiated SK-N-F1 and SH-SY5Y cells and the results are presented in figure 3.13. The key findings are summarised below.

3.12.1.1 The NP element contains a promoter capable of initiating transcription in both the sense and antisense directions

A small level of transcription was detected when the NP element was cloned upstream to the luciferase gene in both the sense and antisense directions, thereby confirming the presence of a bi-directional promoter within this element. The H1 variants conferred a higher level of activity in the sense direction than in the antisense direction (approximately 1.7- to 2.5-fold higher) and this reached or trended towards statistical significance in both the F1 (H1B: $p=0.0039$; H1C: $p=0.0784$) and SH (H1B: $p=0.0437$; H1C: $p=0.0625$) cell lines. The consistent relationship between sense and antisense transcription from the NP indicates a positive correlation between the two. This is particularly evident when comparisons are made between the two cell lines, as an increased level of sense transcription in F1 cells was accompanied by a proportional increase in NP-antisense transcription.

The H2 variants, however, behaved differently depending on the cellular context. In the F1 line (which has one endogenous H2 chromosome), the comparison between the sense and antisense constructs resembled that of the H1 variants, with 2-fold greater activity in the sense direction. This did not reach statistical significance (p=0.1140); most likely as a result of the large standard error produced by the sense construct.

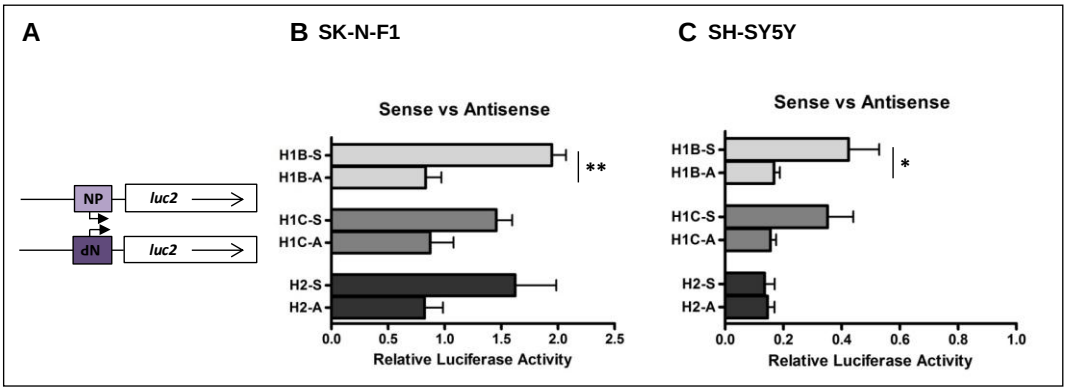


Figure 3.13 Promoter luciferase results 4
A: The NP-sense (top) and NP-antisense (bottom) luciferase constructs differ only by the orientation of the NP element. B: The relative luciferase activities of the three haplotype variants of the sense (NP-S) and antisense (NP-A) constructs in SK-N-F1 cells. C: The relative luciferase activities in SH-SY5Y cells. * p≤0.05; ** p≤0.01

In SH cells (which do not possess an endogenous H2 chromosome), however, the activity level of the H2-sense construct was reduced and equalled that of the H2-antisense construct (p=0.8156), which itself was barely above background levels. This suggests that there is something specifically produced by the endogenous H2 chromosome of the F1 cells that upregulates transcription of the H2 NP variant in the sense direction. The absence of an endogenous H2 chromosome in the SH cells appears to affect sense transcription but not antisense transcription, perhaps suggesting that they are regulated by different mechanisms. It is more likely, however, that the overall level of antisense transcription is too low in SH cells for any difference in H2 antisense expression to be detected.

3.12.1.2 The NP element modifies expression from the core promoter

Following confirmation that the NP element contains a second transcription start site, the effect of this element on transcription from the core promoter was

investigated. It was hypothesised that the addition of the NP element to the CP construct would cause either an increase – perhaps through the additive effects of transcription from two sense promoters – or a decrease – through antisense-mediated repression – in luciferase activity compared to the CP alone. To test this, the NP element was inserted downstream to the CP in the sense orientation; producing an extended promoter fragment covering the full 2.2kb of highly conserved sequence located between chr17:43971166 and chr17:43973404. The three haplotype variants were assayed alongside the CP alone and NP-sense luciferase constructs and the results are presented in figure 3.14.

The results show that the addition of the NP element to the core promoter does indeed alter significantly the activity level of the CP, though this was dependent on the cell line. In the F1 cell line, CP transcription rate was significantly increased following the addition of the NP element (H1B: p=0.0295; H1C: p=0.0469; H2: p=0.0478). In SH cells, however, the effect of the NP was to significantly decrease CP activity (H1B: p=0.0308; H1C: p=0.0237; H2: p=0.0001). This, once again, indicates the importance of *trans*-acting factors and endogenous cellular conditions in gene expression.

Indeed, the cellular context does appear to differentially alter the activity of the NP variants. When comparing the rate of transcription conferred by the NP in the sense direction with that of the core promoter element, there are clear differences in expression between the two cell lines which can only be attributed to differences in the endogenous conditions. Although transcription levels are low across the board, the activities of the NP elements are much higher in F1 cells, with NP-sense reaching 53-71% of the level of transcription from the core promoter – the major transcription start site. In fact, although NP-sense transcription is clearly lower than CP transcription, this difference does not quite reach statistical significance for any of the three variants in this cell line (H1B: p=0.2292; H1C: p=0.0795; H2: p=0.0708). In SH cells, however, the NP-sense elements can only manage 5-14% of the activity of the core promoter and this difference is statistically significant in all cases (H1B: p=0.0056; H1C: p=0.0047;

H2: p=0.0001). Thus, in F1 cells, the increase in activity observed when the NP-sense element is added to the CP may result from the combination of the two highly active sense promoters. In SH cells, however, although the level of NP-sense transcription was much lower than observed in F1 cells and, therefore as expected, the combined effects of the two sense promoters was also much lower, the overall repression of transcription from the CP was a surprising finding.

The differential activity of the joint CP and NP-sense constructs in the two cell lines is difficult to explain and is not concordant with any of the four models of antisense-mediated transcriptional regulation outlined in section 3.2. This is likely due to the bi-directional nature of the antisense promoter, with additional sense transcription from this promoter complicating the picture. It is likely that the significant differences in NP promoter activity observed between the two cell lines is an important part of the mechanism linking the activities of these three promoters. Indeed, at this stage, the only viable explanation for these opposing results may be that the reduced NP activity in SH cells causes the accumulation of RNA Pol II transcription complexes at this site that block upstream Pol II complexes elongating from the CP, thus reducing overall expression. This would be less of a problem in F1 cells, where NP activity is much higher; therefore additive transcription from the two sense promoters would increase overall expression. This cannot be proved using the luciferase assays described here as CP, NP-sense and NP-antisense transcription cannot be distinguished when expressed from a single luciferase construct.

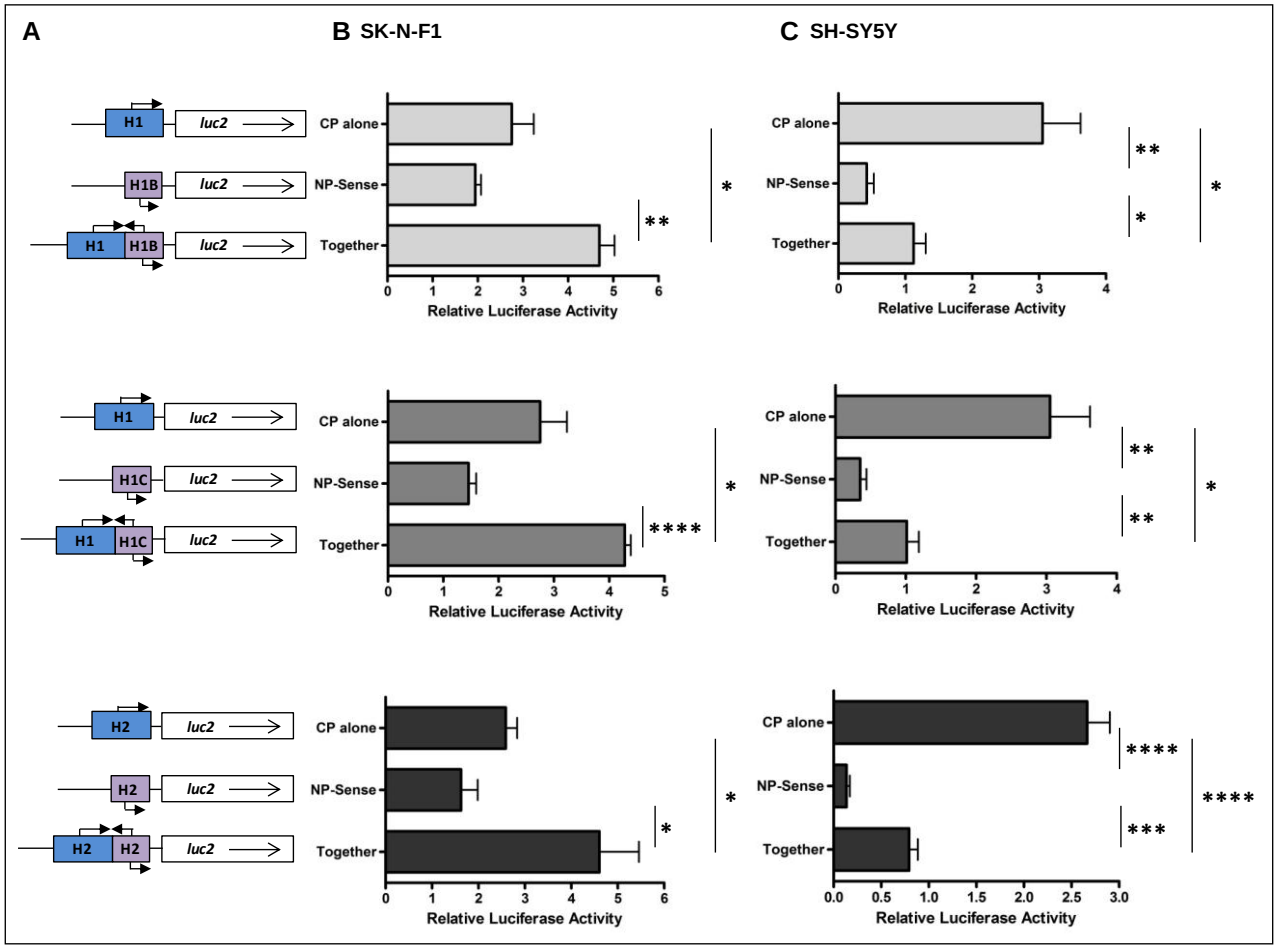


Figure 3.14 Promoter luciferase results 5

A: The luciferase construct variant assayed in each instance, separated by construct.

B: The relative luciferase activity of each construct variant in SK-N-F1 cells, separated by haplotype.

C: The relative luciferase activities in SH-SY5Y cells.

*	p≤0.05
**	p≤0.01
***	p≤0.001
****	p<0.0001

3.12.2 The effect of genetic variation on transcriptional regulation by the NP

As the main aim of this project was to look at the effect of genetic variation on tau gene expression, it was important to assess whether polymorphisms in the NP element could affect its function. Therefore, the relative luciferase activities of the three haplotype variants of the NP-sense, NP-antisense and CP+NP-sense combined constructs are presented together in figure 3.15. Only one significant difference was observed; between the H1B and H2 variants of the NP-sense construct in SH cells ($p=0.0313$). Thus, at first glance it does not appear that genetic variation within the NP domain affects either its independent function or its effect on transcription from the CP. However, a consistent, though non-significant, pattern emerged between the H1B and H1C variants of all three constructs in both cell lines.

First noticed with the NP-sense construct, the H1C variant conferred consistently lower activity than the H1B variant, as observed across numerous independent replications in both cell lines. This was intriguing as, if there was truly no difference between the two variants, the high variability of the luciferase technique would normally result in the H1C variant being slightly higher than the H1B variant in some of the replications. This was never the case; neither was such variability observed with the NP-antisense constructs, where the H1C variant was consistently higher than the H1B variant, at least in the F1 cell line (the expression level of this construct is too low to detect subtle haplotype differences in SH cells). When the same consistency in H1B/H1C expression occurred with a third NP construct – NP-sense together with the CP – a coincidence seemed unlikely. A comparison of the sequences of the two NP haplotype variants revealed a single nucleotide difference – a known C/T polymorphism denoted rs3744457. Furthermore, this polymorphism lies just 36bp downstream to the 3' end of the NP-sense promoter (and therefore 36bp upstream to the 5' end of the NP-antisense promoter), suggesting it may play a role in regulating NP transcription.

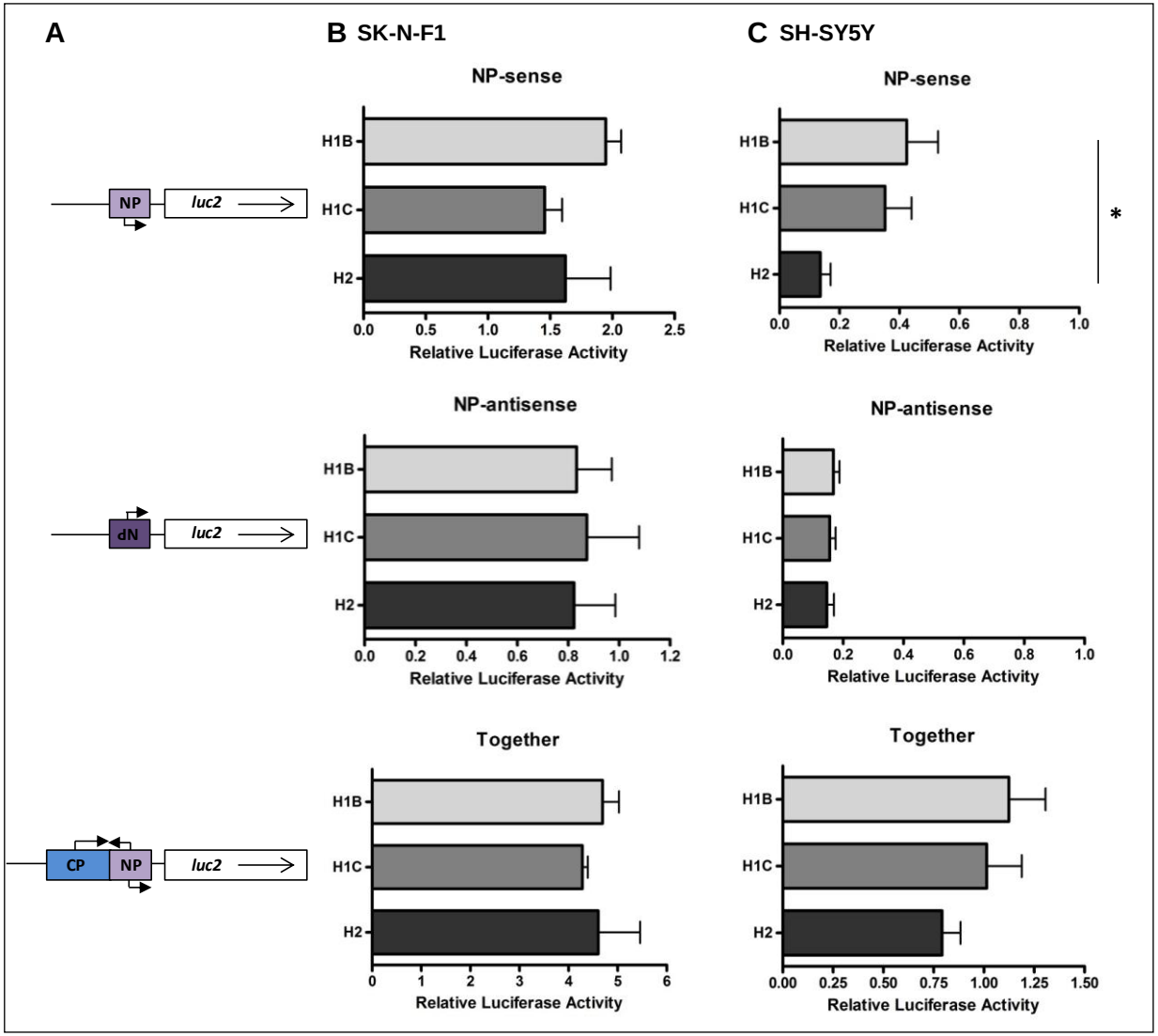


Figure 3.15 Promoter luciferase results 6

A: The luciferase construct variants assayed, separated by construct.

B: The relative luciferase activity of each haplotype variant in SK-N-F1 cells, separated by construct.

C: The relative luciferase activities in SH-SY5Y cells.

*	p≤0.05
**	p≤0.01
***	p≤0.001
****	p<0.0001

3.12.3 The C/C genotype of rs3744457 is over-represented in PSP

To determine whether rs3744457 plays a role in PSP risk, the polymorphism was genotyped in two DNA cohorts, one consisting of 125 clinically diagnosed PSP patients and one consisting of 127 neurologically normal control individuals. These cohorts were pre-existing and DNA was previously extracted from brain tissue using standard methods.

The sequences of the NP luciferase constructs show that the H1B haplotype contains the C-allele of rs3744457, with the H1C and H2 haplotypes carrying the T-allele. The global population frequency of the minor C-allele, as reported by the 1000 Genomes project, is 0.43, which is similar to that of the rs242557-A allele (0.42); however, these polymorphisms are not in LD with each other, nor with the H1/H2 inversion polymorphism (figure 3.16).

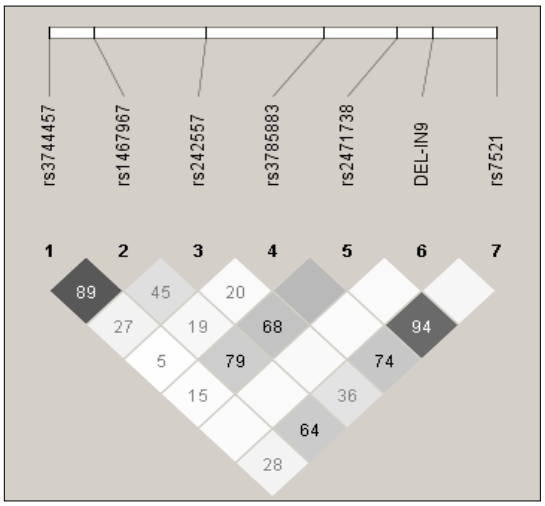


Figure 3.16 An LD plot of the six tagging SNPs commonly used to define the *MAPT* haplotypes and the rs3744457 polymorphism.
The numbers represent the R'squared measure of correlation between the polymorphisms; the higher the number the greater the correlation. The plot was created from the genetic information of 55 PSP patients.

The rs3744457 polymorphism is a restriction fragment length polymorphism (RFLP), with the C-allele abolishing one of two NlaIII recognition sites (CA**T**G) contained within the NP element (901bp). Thus, an element carrying the T-allele will be cut twice by the NlaIII enzyme (at nucleotides 412 and 690) whereas an element carrying the C-allele will be cut only once (at nucleotide 690).

The NP region was amplified from the DNA of each patient in the PSP and control cohorts using the method described in section 3.5.3. PCR products were incubated overnight at 37°C with 3 units of NlaIII enzyme, 1x NEB4 buffer and 1x BSA and the digestion products were resolved by agarose gel electrophoresis

using a 2% gel. Figure 3.17 gives examples of the banding pattern that identified each genotype.

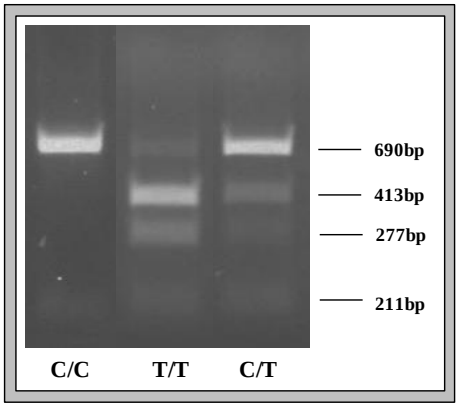


Figure 3.17 Genotyping of the rs3744457 polymorphism
The banding pattern produced by NlaIII digestion of NP PCR products amplified from the DNA of individuals carrying the three rs3744457 genotypes. Each cohort was independently genotyped on two separate occasions and genotypes were consistent with Hardy-Weinberg equilibrium.

As mentioned previously, during the course of this project the PSP cohort was subject to clinical review and 23 of the patients were re-classified with non-PSP diagnoses. These included PD (N=11), CBD (N=3), multiple system atrophy (N=6), parkinsonism (N=1) and PSP in conjunction with other neurodegenerative conditions (N=2). Thus two separate analyses were conducted, one using only patients classified as purely having PSP (N=102), and one using all patients diagnosed with a neurodegenerative condition (N=125). Table 3.6 gives details of the genotype and allele frequencies of each cohort. The frequency of the minor C-allele in the control population was lower than the global frequency reported by the 1000 Genomes project (0.26 v 0.43); however it more closely matched the 0.3 frequency reported by HapMap (NIH) specifically for populations of European descent.

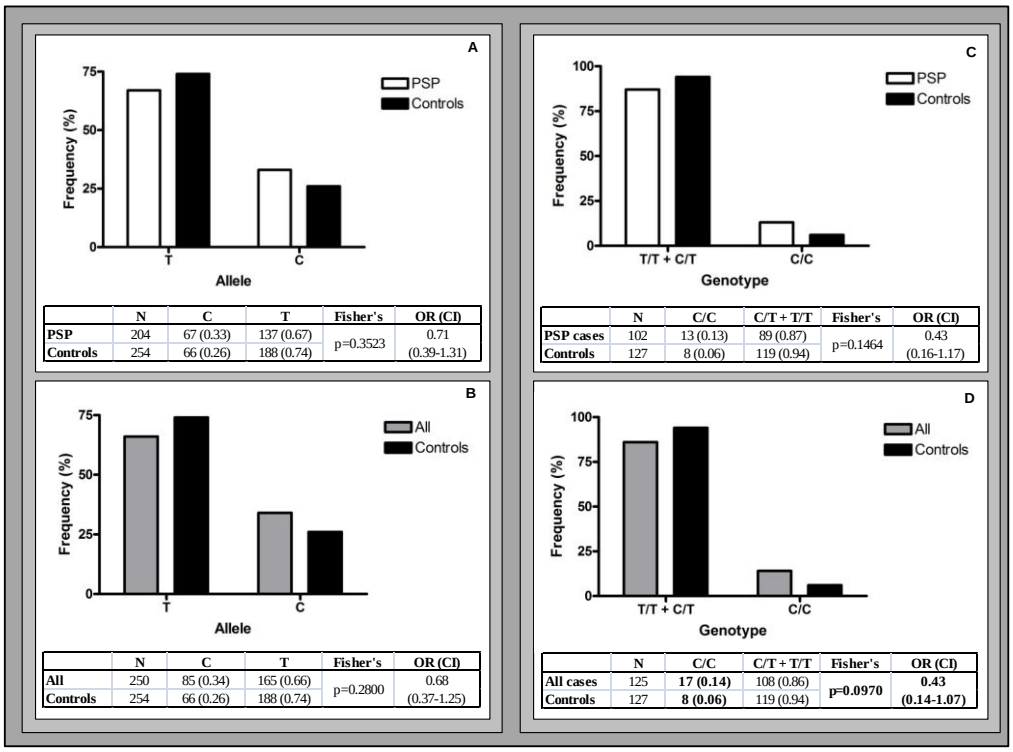
PSP cases	**N**	**Freq**
T/T	48	0.47
C/T	41	0.40
C/C	13	0.13
Total	102	1.00
T	137	0.67
C	67	0.33
Total	204	1.00
HWE	p=0.369862	

All cases	**N**	**Freq**
T/T	57	0.45
C/T	51	0.41
C/C	17	0.14
Total	**125**	**1.00**
T	165	0.66
C	85	0.34
Total	**250**	**1.00**
HWE	p=0.309442	

Controls	**N**	**Freq**
T/T	69	0.54
C/T	50	0.4
C/C	8	0.06
Total	**127**	**1.00**
T	188	0.74
C	66	0.26
Total	**254**	**1.00**
HWE	p=0.790851	

Table 3.6 Genotype and allele frequencies of the rs3744457 polymorphism.
Frequencies of the PSP only cohort, the PSP cohort including other neurodegenerative disorders ('All') and the neurologically normal control cohort following genotyping by RFLP. All genotyping was in Hardy-Weinberg equilibrium.

A statistically different distribution of allele or genotype frequency between the case and control cohorts was detected using a two-sided Fisher's Exact test and defined as $p \leq 0.05$. Comparisons of the C- and T-allele frequencies are given in figures 3.18A and 3.18B. There was a slight over-representation of the C-allele in both case cohorts versus the controls, but this did not reach statistical significance (PSP vs controls: p=0.3523, OR=0.71 [95% CI: 0.39-1.31]; All versus controls: p=0.2800, OR=0.68 [CI: 0.37-1.26]). This may be due to the small size of the cohorts producing insufficient statistical power to detect subtle shifts in frequency, as the p-value edged closer to significance with the 'All' cohort, which was slightly larger than the PSP cohort (N=125 vs N=102).



	N	C	T	Fisher's	OR (CI)
PSP	204	67 (0.33)	137 (0.67)	p=0.3523	0.71 (0.39-1.31)
Controls	254	66 (0.26)	188 (0.74)		

	N	C	T	Fisher's	OR (CI)
All	250	85 (0.34)	165 (0.66)	p=0.2800	0.68 (0.37-1.25)
Controls	254	66 (0.26)	188 (0.74)		

	N	C/C	C/T + T/T	Fisher's	OR (CI)
PSP cases	102	13 (0.13)	89 (0.87)	p=0.1464	0.43 (0.16-1.17)
Controls	127	8 (0.06)	119 (0.94)		

	N	C/C	C/T + T/T	Fisher's	OR (CI)
All cases	125	**17 (0.14)**	108 (0.86)	**p=0.0970**	**0.43 (0.14-1.07)**
Controls	127	**8 (0.06)**	119 (0.94)		

Figure 3.18 Results of the rs3744457 genotyping in PSP and control cohorts
The frequencies (%) of the C- and T-alleles in the PSP (A) and 'All' (B) cohorts versus controls; and the frequencies (%) of the C/C and C/T+T/T genotype groups in the PSP (C) and 'All' (D) cohorts versus controls. The p-value (Fisher's) and odds ratio (OR) produced from a Fisher's Exact of cases versus controls are given in each instance.

Testing a recessive mode of inheritance for the C-allele, a slight over-representation of the C/C genotype in the case cohorts was again observed (figures 3.18C and 3.18D) and this time the p-value for the 'All' comparison fell

below the p=0.10 threshold that indicates a trend towards significance (PSP vs controls: p=0.1464, OR=0.43 [CI: 0.16-1.17]; All vs controls: p=0.0970, OR=0.40 [CI: 0.14-1.01]).

These results are interesting and suggest that the rs3744457 polymorphism may be worth analysing in a larger PSP cohort. Due to the rarity of PSP it is difficult to collect enough samples to create a cohort large enough to allow confidence in the results. It is unfortunate that this polymorphism was not genotyped as part of the PSP genome-wide association study published by Hoglinger *et al* in 2010 [104], as this was the largest genetic study of PSP to date and pooled data from numerous independent cohorts.

It has been shown here, however, that the rs3744457 C/C genotype is slightly over-represented in PSP cases and this may therefore indicate an important role in tau gene expression and PSP risk for two factors that are yet to be investigated thoroughly. The first is the confirmed presence of a secondary sense promoter lying immediately downstream to the major core promoter at exon 0. This promoter is believed to express the *MAPT-IT1* transcript of which little is currently understood. The second is the role of antisense transcription and the *MAPT-AS1* transcript in the regulation of tau gene expression, again of which little is understood. It was shown earlier in section 3.12.1 that the cellular conditions can potentially alter the regulatory effect of this region on transcription from the core promoter. Can genetic variation also alter its activity? Does the rs3744457 C-allele alter the expression of either the sense or antisense transcripts and if so does this play a role in PSP risk? Are these transcripts capable of modulating either transcription rate and/or alternative splicing *in vivo*? These questions, however, are outside of the scope of this project.

3.13 The role of the 3'UTR in *MAPT* expression

The second luciferase study looked at the role of the *MAPT* 3'UTR in gene expression. This region of the gene is vital for determining mRNA transcript stability and has the potential to modulate protein expression by increasing or decreasing the half-life of the transcript. Thus, genetic variation within the 3'UTR that changes mRNA stability could directly influence tau gene expression and contribute to PSP risk. To test this hypothesis and to identify the most critical regions of the 3'UTR for determining stability, a set of luciferase constructs were created in which either the full-length 3'UTR or one of three overlapping fragments (the deletion constructs) were cloned immediately downstream to the luciferase gene. Luciferase expression driven by the highly active CMV promoter was quantified in undifferentiated SK-N-F1 and SH-SY5Y as described previously.

The first findings to be discussed concern the full-length 3'UTR (~4.4kb) and the effect of genetic variation on overall mRNA stability. This will be followed by an assessment of the deletion constructs (~1.2-2.0kb) and their individual contribution to gene expression. Finally, the three haplotype variants of each deletion fragment will be compared.

3.13.1 The *MAPT* 3'UTR increases stability of the luciferase transcripts

The relative luciferase activity of the three haplotype variants of the full-length 3'UTR did not differ significantly in either cell line (figure 3.19); indicating that haplotype variation within this region does not contribute to the increased risk of PSP conferred by the H1C haplotype. In general, the activities of the 3'UTR constructs were much higher than their promoter counterparts, though there were sizeable differences in overall expression between the two cell lines. In F1 cells, the normalised luciferase activities of the three variants were 4- to 5-fold higher than the activities of the CP constructs in this cell line. In SH cells, however, activities were 8- to 9-fold higher from the H1 variants and 14-fold higher from the H2 variant than their CP counterparts. This, again, highlights the importance of the cellular context in these types of studies.

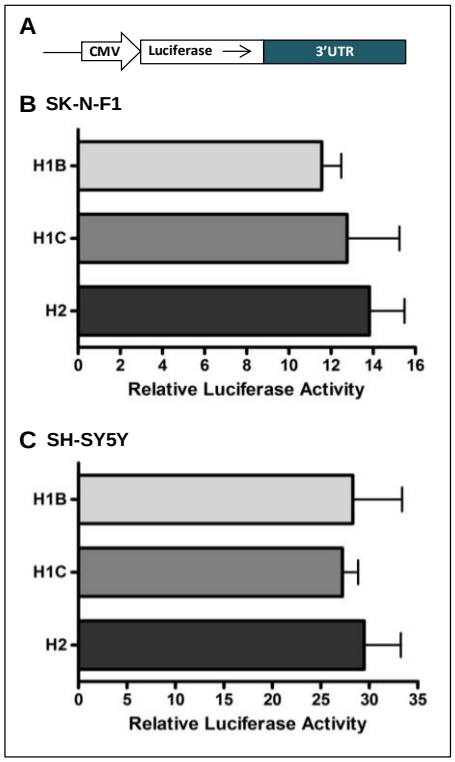


Figure 3.19 3'UTR luciferase results 1
The full-length 3'UTR construct (A) and the relative luciferase activities of the three haplotype variants in SK-N-F1 (B) and SH-SY5Y cells (C).

All results – both promoter and 3'UTR – were normalised against the activity of an empty CMV-driven pMIR-REPORT vector and thus this increase cannot be attributed to the higher activity of the viral promoter in SH cells. Thus, although at the basic sequence level genetic variation within the 3'UTR does not differentially affect tau gene expression, the difference in stability observed between the two cell lines opens the possibility for allelic differences in the interaction of the 3'UTR with differentially expressed endogenous factors.

It has been shown that actively transcribed genes adopt a loop formation in which factors within the polyadenylation complex at the 3' end of the transcribed gene interact with promoter-associated transcription factors to reduce aberrant transcription and promote mRNA transcription in the sense direction [232]. In particular, the poly(A)-associated factor, Ssu72, has been shown to play an important role in this, with mutations preventing loop formation in the *FMP27* gene. Interestingly, these Ssu72 mutations also caused an increase in Pol II density at and antisense transcription from the *FMP72* promoter. This would therefore suggest that the 3'UTR plays a role in determining the activity and directionality of the promoter. Thus, it would be interesting to combine the promoter and 3'UTR luciferase studies to determine the effect of the full-length *MAPT* 3'UTR on transcription from both the CP and NP elements. This may add an additional layer of complexity to the regulation of *MAPT* transcription but could also potentially reveal haplotype differences that are dependent on the interaction of the 3'UTR with the promoter and therefore were undetectable when

the 3'UTR was assayed alone. Due to time constraints, this unfortunately falls outside the scope of this project.

3.13.2 H1/H2 differences in determining transcript stability

The relative luciferase activities of the three deletion constructs are given in figure 3.20 alongside that of the full-length constructs. Results are separated by haplotype and cell line and appear to reveal H1/H2 differences in the determination of overall luciferase expression. For the H1 variants, the Fr1 fragment conferred a marked increase in expression compared to its full-length counterpart; an increase that reached or trended towards statistical significance in both the F1 (H1B: p=0.0083; H1C: p=0.0083) and SH (H1B: p=0.0058; H1C: p=0.0634) cell lines.

The reason for this increase may lie with the utilisation of the three polyadenylation signals (AATAAA) present in the *MAPT* 3'UTR. Polyadenylation is one of the final steps in the production of the mature mRNA transcript. The addition of a string of A-residues to the 5' end of the transcript initiates transcription termination, facilitates export of the mature transcript from the nucleus and subsequent subcellular localisation, prevents transcript degradation in the cytoplasm, and is required for translation of the mRNA transcript into protein. Each process, however, is not solely dependent on polyadenylation and requires additional regulation from RNA binding factors such as microRNAs (miRNAs).

The first *MAPT* poly(A) site is present at the 5' end of Fr1 and its usage results in mature transcripts with a short 3'UTR of around 220 nucleotides. The other two signals are located at the 3' end of Fr3, producing transcripts containing almost the full-length 3'UTR (~4130 and 4280 nucleotides respectively). It is therefore likely that these sites have different roles to play in the 3'UTR-mediated regulation of transcript expression. Thus, the absence of the Fr3 polyadenylation signals results in transcripts that exclusively contain the shorter 3'UTR and this may therefore lead to an increase in overall expression through the loss of the Fr3-

mediated regulation of some or all of the polyA-dependent processes listed above. For the H2 Fr1 variant, although a relatively small increase in expression was observed, it did not quite reach statistical significance in either cell line, though came close in F1 cells (F1: p=0.0841; SH: p=0.3137; figure 3.20).

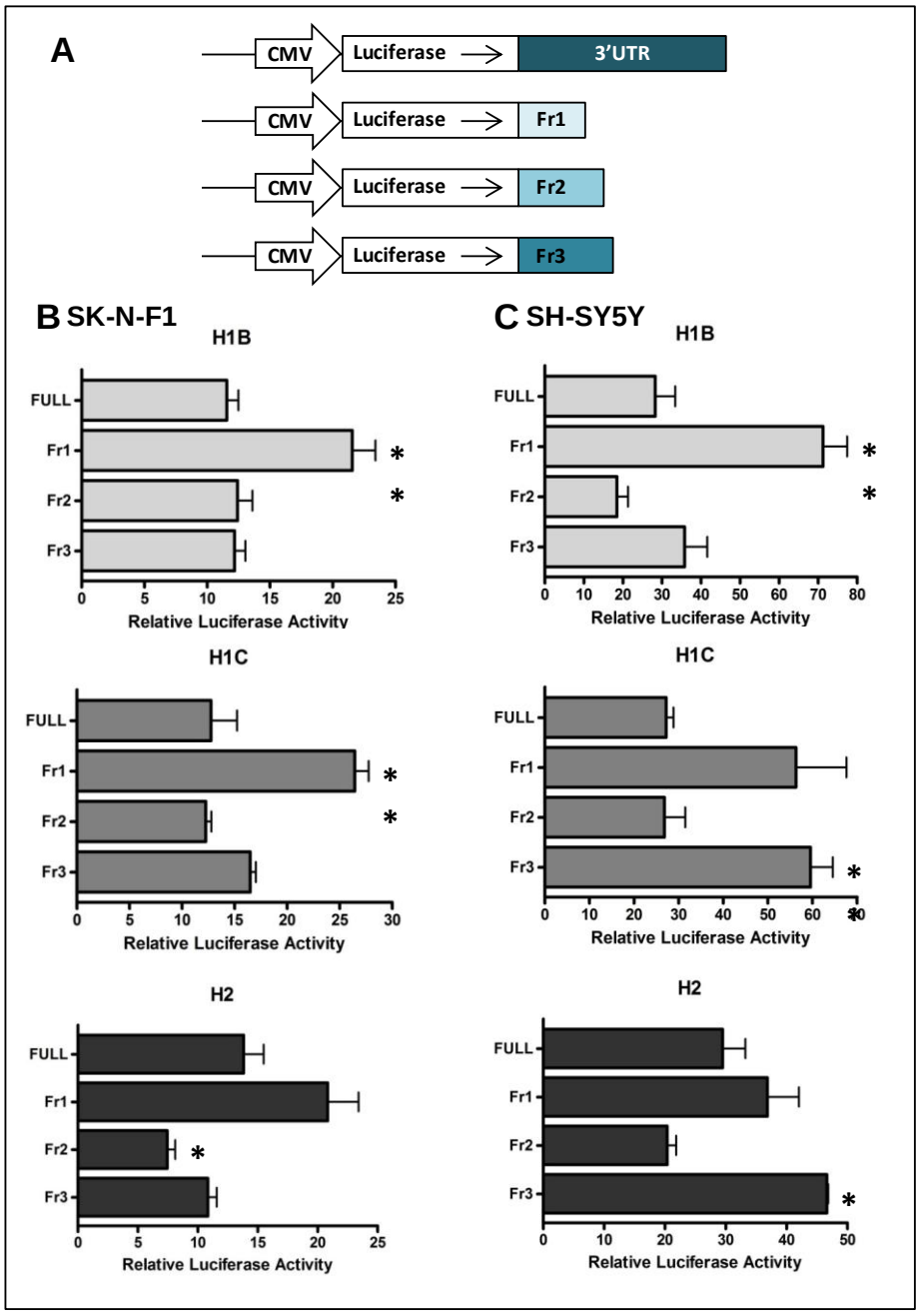


Figure 3.20 3'UTR luciferase results 2
A: The luciferase construct variant assayed in each instance; B: The relative luciferase activity of each construct in SK-N-F1 cells, separated by haplotype; C: The relative luciferase activities in SH-SY5Y cells. A two-tailed Student's t-test compared the relative activity of each deletion construct to that of the full-length construct (* p≤0.05).

As Fr1 expression is the most consistently altered compared to that of the full-length construct, it is likely that the key sequences for determining the overall level of expression are contained on Fr2, Fr3 or both. Again there appears to be H1/H2 differences in how this is achieved. For the H2 variants, the overall stability of the mRNA transcripts appears to result from a balance between the stabilising and de-stabilising functions of the three individual fragments, with no single fragment consistently matching the expression of the full-length construct and therefore appearing dominant. For the H1 variants, however, Fr2 could be considered the dominant fragment, as the expression levels of the Fr2 H1 deletion constructs conferred expression that was consistently similar to their full-length counterparts in both the F1 (H1B: p=0.6004; and H1C: p=0.8939) and SH (H1B: p=0.1691; and H1C: p=0.9397) cell lines. Fr2 expression from the H2 variant was reduced compared to full-length expression and this reached or trended towards significance in both cell lines (F1: p=0.0238; SH: p=0.0877).

Fr3 expression was the most variable, behaving differently depending on the cell line. In F1 cells, none of the three haplotype variants conferred expression levels that were significantly different from that of their full-length counterparts (H1B: p=0.6477; H1C: p=0.2170; H2: p=0.1758). In SH cells, however, expression significantly increased for the H1C and H2 variants, but not for the H1B variant (H1B: p=0.3855; H1C: p=0.0035; H2: p=0.0104).

As mentioned earlier, Fr3 contains two of the three *MAPT* poly(A) sites and therefore the differential behaviour of this fragment in the two cell lines may result from differences in the recognition of the sites by the endogenous polyadenylation machinery. A useful tool for determining poly(A) site usage is 3' RACE, in which priming at the poly(A) tail of mature transcripts, followed by extension and sequencing, allows the identification of the sequence lying immediately upstream to the poly(A) tail. Thus the relative usage of the two Fr3 poly(A) sites – and the Fr1 site – in each cell line could be determined, though this is, again, outside the scope of this project.

Taken together, these results suggest that the Fr2 deletion fragment contains the key sequences for determining the expression level of H1 *MAPT* transcripts. This was the only fragment for which expression did not significantly deviate from that of the full-length 3'UTR, regardless of H1 sub-haplotype status and cell line. Fr2 comprises the middle section of the 3'UTR (between nucleotides 1179-3007) and contains a string of A-residues that, when translated into mRNA forms a U-rich binding domain for the ELAV-like protein, HuD. It has been shown that HuD regulates the stability of some tau mRNA transcripts by anchoring them to microtubules and protecting them from decay [233]. This not only leads to an increase in stability, but ensures the correct subcellular localisation of the mature transcripts, a process that requires association with functioning microtubules. This is particularly important in neuronal cells, where local translation of *MAPT* transcripts occurs at large distances from the cell body and is vital for maintaining cellular polarity, which, in turn, is important for generating synapses and for neuronal plasticity during development.

In fact, HuD expression has been shown to increase during development – first appearing following cessation of the cell cycle – and therefore regulates the stability of tau mRNA during neuronal differentiation. Inhibition of HuD expression in PC12 cells results in a decrease in the number of tau transcripts and a failure to respond to neuronal differentiation [233]. Interestingly, the H2 Fr2 variant contains a triplet deletion (GAA) that reduces the A-rich stretch from 24 nucleotides, as occurs in the H1 variants, to 21 nucleotides, likely weakening the HuD binding site and accounting for the lower stability conferred by the H2 Fr2 variant.

In consideration of the changes in HuD expression during development, it would have been desirable to compare the 3'UTR constructs in neuronally differentiated cell lines, but unfortunately the luciferase reporter assay was not a suitable technique in this instance. As the increase in expression from the 3'UTR constructs was so high – particularly in SH cells – the luciferase signal was close to the maximum limit of the Tecan GENios plate reader. Thus if, as expected,

expression increased further upon differentiation of the cells, the GENios limit for differentiating the signals would be exceeded.

3.13.2.1 The H1C variant of Fr3 confers significantly increased expression compared to the H1B and H2 variants

Although there was no difference in overall expression between the three full-length haplotype variants, the variability in individual function of the three deletion constructs may suggest a role for genetic variation. Thus, the 3'UTR luciferase results described above and in figure 3.20 by haplotype are presented by construct in figure 3.21. The main influence of genetic variation was on the expression of Fr3. In both cell lines, the H1C variant conferred significantly increased expression compared to the H1B variant (F1: p=0.0135; SH: p=0.0360) and reached or trended towards a significant increase compared to the H2 variant (F1: p=0.0034; SH: p=0.0600). A significant H1/H2 difference was also observed for Fr2 in F1 cells, with the H2 variant conferring significantly lower expression than the H1B (p=0.0220) and H1C (p=0.0051) variants, likely due to the triplet deletion in the HuD binding site.

It may therefore be that the H1C variant of the *MAPT* 3'UTR can affect gene expression by modulating Fr3-mediated processes, perhaps through alterations to poly(A) site preference. This is, of course, pure speculation and the H1C difference does disappear when the Fr3 region is assayed as part of the full-length 3'UTR construct. The genetic difference, however, may come to prominence when assayed in conjunction with the *MAPT* promoter, allowing the formation of a gene loop that apparently plays an important role in promoting transcriptional activity in the sense direction (as described in section 3.13.1) [232].

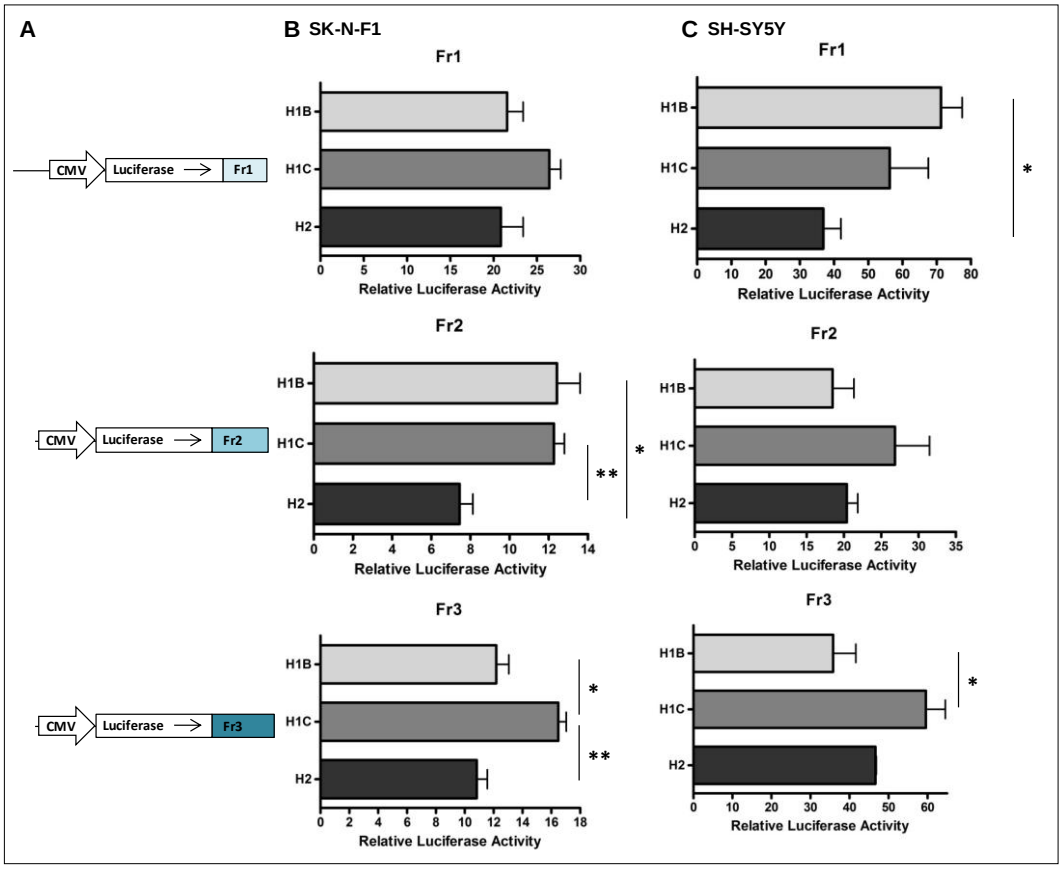


Figure 3.21 3'UTR luciferase results 3
A: The three luciferase deletion construct variants; B: The relative luciferase activity of each construct in SK-N-F1 cells, separated by construct; C: The relative luciferase activities in SH-SY5Y cells. * p≤0.05; ** p≤0.01.

3.14 Discussion

This chapter has described three luciferase reporter gene studies that investigated the effect of genetic variation within the 5' and 3' UTRs on transcription and mRNA stability. The most important finding described here was the consistently different regulation of transcription conferred by a *cis*-acting distal transcription regulatory domain containing the A-allele of the rs242557 polymorphism; the variant strongly associated with an increase in PSP-risk. It was also shown that this differential regulation of transcription by the allelic variants of rs242557 remained regardless of the positioning of the domain relative to the core promoter or differences in endogenous cellular conditions. These factors are, however, important in determining the strength and function of the interaction between the domain and the core promoter and therefore the genomic location of this domain and polymorphism is likely key to its regulation of tau transcription *in vivo*. The

H1 mutant constructs demonstrated that a single nucleotide mutation within exon 0 was sufficient to abolish the differential effect of the A-allele variant of the regulatory domain. This provided evidence of a physical interaction between the two elements, which again signifies the importance of the distal location of the domain.

The second luciferase study confirmed the presence of a secondary bi-directional promoter located immediately downstream to the core promoter. Transcriptional activity in both the sense and antisense directions was differentially affected by the endogenous cellular conditions, as was the effect of this additional activity on transcription from the core promoter. Not only was overall transcription from the bi-directional promoter increased in SK-N-F1 cells compared to SH-SY5Y cells in both directions, H2 transcription in the sense direction was apparently affected by the lack of a H2 chromosome in the SH line; failing to produce the 2-fold greater level of activity compared with that in the antisense direction that was observed in F1 cells. It is therefore likely that transcription from this promoter is dependent upon factors that are specific to the H1/H2 polymorphism and the cell type.

Although there were no significant differences in either sense or antisense transcription conferred from the three haplotype variants, a polymorphism was identified that appeared to have a subtle but consistent allelic effect on activity. Following genotyping in PSP and control cohorts, the C-allele and C/C genotype were found to be slightly over-represented in PSP patients, though statistical significance was not reached. Investigation by more sensitive methods – such as allele-specific quantitative RT-PCR of endogenous *MAPT-IT1* and *MAPT-AS1* expression from heterozygous cell lines – would allow differential activity of the allelic variants to be detected. This polymorphism is certainly worth investigating in larger PSP cohorts and its potential functional consequences bring into play non-coding RNAs, antisense transcription and chromatin modifications in *MAPT* expression and PSP risk.

The final luciferase study aimed to identify genetic variation within the 3'UTR of the *MAPT* gene that differentially affected mRNA stability and thus gene expression. Although no such genetic variation was detected when the full-length 3'UTR was assayed, when split into three individual fragments the expression level conferred by the 3' end of the 3'UTR (Fr3) was shown to be increased for the H1C variant. Fr3 expression was also the most affected by a change in cellular conditions, suggesting that this fragment may play an important role in regulating expression of the tau transcripts during development. It may therefore be that the increase in expression conferred by the H1C variant of this fragment is only of significance in neuronally differentiated cells, or when assayed in conjunction with *MAPT* promoter elements. As the luciferase assay was not suitable to investigate this hypothesis, more sensitive *in vivo* methods such RNA-FISH may prove to be more valuable in this instance. This method could also be used to quantify alternative poly(A) site usage and the effect on sub-cellular localisation of *MAPT* transcripts.

In conclusion, the investigations described here have identified three H1B/H1C genetic differences that could account for the increase in *in vivo* expression reported for H1C chromosomes [12] and thus contribute to the increased risk of PSP conferred by this *MAPT* haplotype variant. The first – and strongest – is the rs242557 A-allele which was shown to confer increased transcriptional activity by weakening repression of the *MAPT* core promoter when cloned in its natural downstream position. The second is the rs3744457 C-allele and/or C/C genotype which may have a subtle effect on the transcription of two non-coding RNAs. The regulation of transcription by non-coding RNAs, particularly natural antisense transcripts, is a growing field of investigation but has been shown to play an important role in gene transcription. Thus, determining the function of these transcripts and the allelic effects of rs3744457 on their expression may further unlock the role of common genetic variation in *MAPT* transcription and PSP risk. The final genetic finding was the discovery that the H1C variant of Fr3 (the 3' end) of the 3'UTR conferred significantly increased luciferase expression compared to the H1B and H2 variants. Although this increase was lost in the full-

length constructs, the variability of Fr3 expression in the two cell lines suggests this region may play an important role in differential tau expression during development.

After confirming the first part of this project's hypothesis: that common genetic variation can affect the transcriptional activity of the tau promoter; the next stage was to determine the effect on the alternative splicing of downstream exons, and this is described in chapter 4.

4 Design, construction and validation of *MAPT* minigenes for the investigation of the effect of the rs242557 polymorphism on *MAPT* transcription and alternative splicing

4.1 Overview

To further investigate the link between the rs242557 polymorphism and the regulation of *MAPT* expression, a set of *MAPT* minigenes were created. The minigene blueprint comprised the 11 protein coding exons expressed in the adult brain, the key intronic sequences surrounding the alternatively spliced exons and the 3'untranslated region (3'UTR). Interchangeable promoter elements allowed the comparison of *MAPT* expression when driven by the *MAPT* core promoter with or without the addition of the rs242557 regulatory domain. Different versions of the minigenes were created representing the genetic variation of the common *MAPT* haplotypes. Each haplotype set included the minigene variant under the control of the *MAPT* core promoter alone, the *MAPT* core promoter in conjunction with the rs242557 regulatory domain, and the cytomegalovirus (CMV) promoter as an independent control.

Each minigene variant was, in the first instance, transiently transfected into two neuroblastoma cell lines and the *in vitro* expression profiles determined. Platform cell lines were also created for the integration of the minigenes into the genome of each cell line to create stably-expressing isogenic cell models. This chapter details the design, construction and validation of these minigene cell models.

4.2 Background

Minigenes are artificially constructed versions of a gene in which most of the non-essential sequences have been removed. They are traditionally used in splicing studies to identify the key *cis*- and *trans*-acting factors responsible for regulating the splicing of constitutive and alternative exons [234]. Minigenes are significantly smaller than their full-length counterparts, making them easier to manipulate in mutagenesis studies and allowing transfection and study in *in vitro* cell models.

To date, most tau minigenes have been created for the specific purpose of studying the splicing of exon 10 and as such have consisted of part or all of the region spanning exons 9-11 under the control of a either the *MAPT* core promoter or an independent control promoter [11, 12, 207, 209]. These minigenes allow the study of exon 10 splicing at the mRNA level but not at the protein level and not in conjunction with the N-terminal splicing events at exons 2 and 3.

In 2007, Dawson *et* al [226] created a transgenic mouse model to investigate the effect of the N279K *MAPT* mutation – previously associated with FTDP-17 – on exon 10 inclusion (figure 4.1 (A)). Using a mixture of genomic and cDNA fragments of human tau, they created a chimeric minigene (denoted T-279) that contained all 14 tau exons and expressed mRNA that was alternatively spliced at exons 2, 3 and 10 (figure 4.1 (B)). The promoter consisted of a 5,049bp fragment of the human tau promoter and the N279K mutation was introduced into exon 10 by site-directed mutagenesis. Two control models were created, one wildtype version without the mutation (T-WT) and one in which the tau promoter was swapped for the CMV promoter (C-279). In adult mice, the T-WT wildtype minigene expressed approximately equal amounts of 3R and 4R human tau mRNA, as observed in the healthy human adult brain (figure 4.1 (C)). The T-279 and C-279 versions, however, almost exclusively expressed mRNA containing exon 10, showing a direct effect of the N279K mutation on exon 10 splicing.

Interestingly, in foetal mice the T-279 model replicated the wildtype expression pattern of an equal 3R- and 4R-tau ratio, whereas the C-279 model demonstrated the same 4R-tau exclusivity seen in the adult mice. In a separate analysis, all three minigene models were found to constitutively include exons 2 and 3 (2N tau) in adult mice, with the 1N and 0N tau isoforms completely absent. The reason for this remains unclear. At the protein level, expression of human tau from the minigenes was found to be extremely low (approximately 1-2% of endogenous murine tau expression) and therefore the expression of exons 2, 3 and 10 in tau protein were not quantified here. Despite the low level of human protein expression, pathological aggregates resembling those found in the tauopathy brain were detected in the T-279 – but not the C-279 or T-WT – mouse brain post

mortem. Together, these findings revealed four things: firstly, that the tau promoter plays a role in the regulation of exon 10 inclusion; secondly, that this regulation changes during development; thirdly, that the N279K mutation causes a change in exon 10 splicing and fourthly, that an increase in tau expression is not sufficient to cause tau pathology unless there is an accompanying shift in 3R- and 4R-tau ratio.

Figure 4.1 A published minigene study of tbe *MAPT* N279K mutation
***Taken from Dawson et al. J Neurosci (2007)*. A: The tau minigene used to create transgenic mouse models for the study of the effect of the N279K exon 10 mutation on tau expression. B: The mRNA transcribed from the minigene. The numbered arrows represent the location of primers used to analyse exon 2 and 3 (primers 1 and 2) and exon 10 (primers 3 and 4) inclusion. C: RT-PCR results of minigene cDNA extracted from mouse brain. 'non-TG' = non-transgenic control mouse; Human = human tau cDNA from healthy adult brain; C-279 = the mutated Tg mouse with the CMV promoter; T-279 = the mutated Tg mouse with the tau promoter; T-WT = the wildtype Tg mouse with the tau promoter.**

This study by Dawson *et al* was the first to show that the tau promoter can influence splicing at exon 10 and has demonstrated a direct effect of a known disease-causing mutation on this process. It has yet to be shown, however, whether common variation that is known to increase risk, rather than cause disease, can have a similar effect on tau splicing. In addition, it has yet to be determined whether the increase in tau transcription reported for the PSP-associated H1/H1C haplotypes can actually cause the pathology-associated

alteration in tau splicing ratio directly, rather than the two phenomena simply occurring concurrently. Thus, in this study the basic design of the Dawson minigene was adapted to create *in vitro* mammalian cell models to determine whether the rs242557 promoter polymorphism can directly affect the splicing events at exons 2, 3 and 10. Their construction is described here.

4.3 Multisite Gateway® Pro Technology

Multisite Gateway® Technology (Invitrogen) presents a highly efficient method for the simultaneous transfer of several heterologous DNA sequences into a chosen vector system in a defined order and orientation [235-237]. The technology is typically used to bring together separate elements of a gene for expression analysis *in vitro*. DNA transfer is facilitated using a modified version [238] of the bacteriophage lambda site-specific recombination system [239]. Briefly, lambda utilises specific recombination sequences – called attachment or *att* sites – to integrate itself into the genome of its *E.coli* host and to switch between its lytic and lysogenic pathways [240]. Recombination reactions are catalysed by a mixture of enzymes that bind to the *att* sites, bring the two target sequences together, and facilitate DNA strand exchange by cleavage and covalent re-attachment. Recombination is reversible, with different sets of proteins catalysing the lytic and lysogenic pathways, and conservative, as there is no gain or loss of nucleotides and DNA synthesis is not required. Although strand exchange occurs within a core region common to all *att* sites, the new site formed post-recombination is a hybrid that combines the differing flanking sequences donated by the two parental sites [239].

The basic Gateway® technology utilises four *att* recombination sites, modified from the wildtype lambda site to improve efficiency and specificity, to transfer the target DNA in a two-step reaction. In step one, *att*B sites recombine with *att*P sites to produce the hybrid *att*L and *att*R sites in a reaction termed the ‘BP’ reaction. This reaction is used for the transfer of the target DNA sequence (typically a PCR product flanked by *att*B sequences) into a ‘donor’ vector (pDONR) containing *att*P sequences to produce ‘entry clones’ (figure 4.2A). In

step two, the 'LR' reaction catalyses a reversal of the BP reaction, with recombination between *att*L and *att*R sites, giving rise to *att*B and *att*P sites. This reaction is used to transfer the target DNA sequence from the entry clone (now flanked by the newly formed *att*L sequences) into the 'destination' vector (pDEST) containing *att*R sequences. This 'expression clone' is now ready for *in vitro* expression analysis (figure 4.2B). The BP and LR reactions are catalysed by proprietary mixtures of different recombination proteins, ensuring each reaction is unidirectional.

Figure 4.2 The two-step recombination process using Gateway® technology
A. The BP reaction transfers the target DNA fragment into a pDONR vector to produce an 'entry clone'; B. The LR reaction transfers the target DNA from the entry clone into a destination vector to produce the final 'expression clone'. ***Taken from the Invitrogen Gateway® Technology with Clonase™ II manual.***

A modified version of the Gateway® protocol, called Multisite Gateway®**,** allows the simultaneous and directional transfer of up to four different target sequences into one destination vector. This is made possible by specific modifications to the *att*B and *att*P sequences that increase specificity and give the sites an orientation. Up to five modified *att*B sites (*att*B1-B5) in two orientations are used in the Multisite Gateway® BP reaction, depending on the number of DNA sequences to be transferred. The *att* sites are not palindromic and therefore their orientation relative to the target DNA sequence determines the type of hybrid site produced following recombination. When the orientation of the *att*B site (illustrated by the direction of the arrowhead in figure 4.3) points *towards* the target DNA sequence, the modified sites are denoted *att*B1-B5 and recombination with donor vectors containing similarly modified *att*P1-P5 sites results in the production of entry

clones with *att*L sites (*att*L1-L5). Specificity among the *att*B variants is maintained as *att*B1 sites will only recombine with *att*P1 sites to produce *att*L1 sites, *att*B2 with *att*P2 to produce *att*L2 etc. Conversely, when the orientation of the *att*B site points *away* from the target DNA sequence the sites are denoted with an 'r' (*att*B1r-B5r) and recombination with *att*P1r-P5r sites produces *att*R sites (*att*R1-R5). Thus, up to four different entry clones can be created, each containing a target DNA sequence flanked by different *att*L and *att*R variants.

Figure 4.3 The Multisite Gateway® process
Four target DNA sequences are combined in one expression clone. ***Taken from Invitrogen's 'Multisite Gateway Pro' manual.***

In the Multisite Gateway® LR reaction, recombination occurs simultaneously between the four pDONR entry clones. Specificity is again maintained as *att*L1 sites will only recombine with *att*R1 sites etc. Thus, by flanking the initial target DNA sequences with specific combinations of the *att*B/*att*Br variants, the order and orientation of the multiple fragments in the final expression clone can be controlled. Figure 4.3 outlines the experimental process and the *att* site combinations required for the simultaneous transfer of the maximum four DNA sequences, as required here for construction of the *MAPT* minigenes. The

simultaneous transfer of two or three fragments is also possible using this system, but the combination of *att*B sites on the ends of each fragment must be altered.

4.4 Jump-In™ TI™ (Targeted Integration) Gateway® System

The Jump-In TI System (Invitrogen) presents a method by which expression clones generated using the Gateway® or Multisite Gateway® Technology can be irreversibly inserted into specific locations in the mammalian genome, creating stably expressing isogenic cell lines. The technology uses the PhiC31 and R4 integrase enzymes to stably insert target DNA sequences into a specific, predetermined location in the genome of mammalian cells. The 'targeted integration' process involves, in step one, creating a platform cell line by inserting the unique R4 *att*P sequence into the genome of the chosen cell line and determining the site of integration. In step two, the expression clone is integrated into the platform cell line at the predetermined genomic locus in a process called 'retargeting' (Figure 4.4).

Platform creation involves the PhiC31-mediated integration of a 'platform' vector into the genome of the chosen cell line. This is possible, firstly, due to the presence of naturally occurring PhiC31 'psuedo-*att*P' sites in the mammalian genome and, secondly, the ability of the PhiC31 integrase enzyme to catalyse recombination between two non-identical sites. Thus, in the presence of PhiC31 integrase, recombination occurs between PhiC31 *att*B sequences located on the platform vector and the endogenous pseudo-*att*P sites, resulting in the insertion of the platform vector into the genome of the cell line. PhiC31 integrase lacks a corresponding excisionase enzyme, making this insertion unidirectional.

The platform vector also contains the unique *att*P target sequence of the R4 integrase. R4 target sites do not occur naturally in the mammalian genome and will therefore only be present at the insertion site of the platform vector. The hygromycin resistance gene and a promoterless zeocin, blasticidin or neomycin resistance gene are also included in the platform vector and are required for the

selection of successful recombinants during the two-step targeted integration process described below.

In step one, cells successfully transformed with the R4 platform vector are selected by their resistance to the antibiotic Hygromycin B. Each resistant cell colony is isolated and expanded to produce new monoclonal cell lines, each with the R4 platform vector inserted into the genome at one or more of the naturally occurring pseudo-*att*P sites. The site(s) of integration of each new cell line is determined by cell harvest and DNA extraction followed by Splinkerette PCR (see section 4.10.2). Cell lines with more than one integration site are immediately discarded, as are those in which the platform vector has been inserted into a critical region of the genome, for example within a gene. In this instance, the insertion may disrupt normal cell function and make subsequent expression analyses unreliable and misleading. The use of a platform vector allows the effect of the insertion on normal cell functioning to be monitored in the absence of any influence conferred by the expression clone and provides a single, unique *att*P integration site for the subsequent retargeting step.

In step two, the Gateway® expression clone is integrated into the predetermined genomic locus by virtue of the newly inserted R4 *att*P target site in the platform cell line. The destination vector used to create the expression clone (see section 4.3) contains the *att*B target sequence for the R4 integrase and an independent promoter element (EF1α) that is required for antibiotic selection following integration. In the presence of R4 integrase, recombination occurs between the R4 *att*B site in the expression clone and the R4 *att*P site in the platform cell line, inserting the expression clone into the genome of the cell line at the predetermined site of the platform vector. This integration event results in the insertion of the EF1α promoter element upstream to the promoterless zeocin, blasticidin or neomycin resistance gene present in the platform vector, giving successful recombinants resistance to the zeocin, blasticidin or neomycin antibiotic as appropriate and providing a new agent for selection.

This two-step targeted integration process leads to the production of an isogenic cell model in which DNA sequences of interest are stably expressed from a predetermined genomic location and can be differentiated from the platform or wildtype cell line by virtue of their antibiotic resistance. The main advantage of this system, however, is that once a platform cell line has been selected, it can be used for the integration of all subsequent expression clones generated using Gateway® Technology, creating a series of isogenic cell models in which the integration of the expression clone is always into the same genomic locus. This removes the possibility of gene expression being differentially influenced by the insertion site of the expression clone, increasing reliability and reproducibility of subsequent expression studies.

Figure 4.4 The Gateway® two-step targeted integration process
The Gateway® expression clone is integrated into the genome of a chosen mammalian cell line to create stably-expressing cell models. ***Taken from Invitrogen's 'Jump-In TI' manual.***

4.5 Cell lines

Two human cell lines were chosen to create the stable cell models. The SK-N-F1 and SH-SY5Y cell lines are both derived from neuroblastomas and were described in sections 2.1.5.1 and 3.7. As the two cell lines are morphologically different, the effect of cell type – particularly the difference in endogenous *MAPT* haplotype status – on minigene expression can be investigated.

4.6 *MAPT* minigenes: design

4.6.1 The minigene blueprint

The *MAPT* minigenes were created for the specific purpose of studying the effect of the rs242557 polymorphism on the co-regulation of *MAPT* transcription and alternative splicing. A unique objective of this project was to investigate the effect at both the transcript and protein levels and to allow this a number of elements had to be incorporated into the minigene design.

To ensure expression of the full tau protein *in vitro*, the minigene had to include: a promoter with a start site for the initiation of transcription, the Kozak sequence ((gcc)gccRccAUGG) for the initiation of translation, all *MAPT* protein-coding exons to produce full-length tau protein, the splicing signals at the 5' and 3' ends of introns 1, 2, 3, 9 and 10 to ensure the expression of all six tau isoforms, and the 3' untranslated region (UTR) which plays a major role in the post-transcriptional processing of tau mRNA. Exons 4A, 6 and 8 were not included as they are not widely expressed in the adult brain [149]. To minimise the size of the minigene, introns 4-8 and 11-12 were completely excluded as their neighbouring exons are constitutively included in tau mRNA and are therefore not subject to alternative splicing. It was necessary to have a method of distinguishing the tau mRNA and protein produced by the minigene from the species produced endogenously by the cell lines. For this reason a 27 nucleotide tagging sequence – unique to the minigene and recognised by the non-native FLAG antibody – was inserted downstream to the final coding exon, upstream to the stop codon. Panel A in figure 4.5 presents a schematic of the basic *MAPT* blueprint.

4.6.2 The minigene promoter elements

The rs242557 polymorphism falls within a transcription regulatory domain and therefore three minigenes, each under the control of a different promoter but otherwise identical, were necessary to fully ascertain the contribution of rs242557 to the regulation of minigene expression:

1. The first promoter comprised the *MAPT* core promoter (the 'CP' element described in chapter 3) with the rs242557 regulatory domain (the 'SD' element) cloned immediately downstream. By comparing minigene expression conferred by the H1B-G, H1C-A and H2-G variants of this promoter construct, the differential effect of each rs242557 allele on alternative splicing could be determined. The downstream positioning of the SD domain in this promoter construct was preferred over the upstream version (see chapter 3) as it more closely resembles the endogenous genomic organisation.

2. The second promoter acted as a control from within the *MAPT* gene and simply comprised the core promoter alone. By comparing minigene expression conferred with and without the addition of the SD, the contribution of the rs242557 domain to both transcription rate and alternative splicing could be determined.

3. The third promoter was cloned from the cytomegalovirus (CMV) and provided an independent control. Comparison of the minigene expression profiles of the three haplotype variants when driven by the same CMV promoter would highlight any changes in expression that were due solely to genetic variation elsewhere in the minigene and not to the promoter itself. Further comparison with the *MAPT* promoter-driven minigenes would confirm whether sequences specific to the native *MAPT* core promoter play a role in regulating the pattern of *MAPT* alternative splicing.

The minigenes were created using the Multisite Gateway® technology (see section 4.3), with four entry clones containing different sections of the minigene

recombining in the LR reaction to produce the complete construct. Thus, by isolating the promoter element on the first of the four entry clones, different versions could be swapped into the minigene simply by exchanging the promoter entry clone in the final LR reaction. With the other three entry clones comprising the body and the 3'UTR of the minigene and remaining the same each time, three separate minigenes were created for each haplotype variant in which only the promoter element varied.

4.6.3 Adaptations for the Multisite Gateway® protocol

To fulfil the requirements for the Multisite Gateway® system all of the components of the minigene must be contained within four DNA fragments. Due to its size, large sections of the *MAPT* gene - including most of the introns and some of the exons – were excluded from the minigene to facilitate its study *in vitro*. This meant that the minigene was made up of numerous smaller elements distributed across the gene – typically at large distances from each other – making it impossible to clone all of the required elements in just four PCR reactions. Figure 4.5B details the nine separate elements (not including the promoter and 3'UTR) that made up the body of the minigene, with each element amplified in one PCR reaction from either genomic DNA or reverse transcribed cDNA. A series of cloning steps joined the individual minigene elements together to produce two larger fragments that were compatible with the Multisite Gateway® protocol.

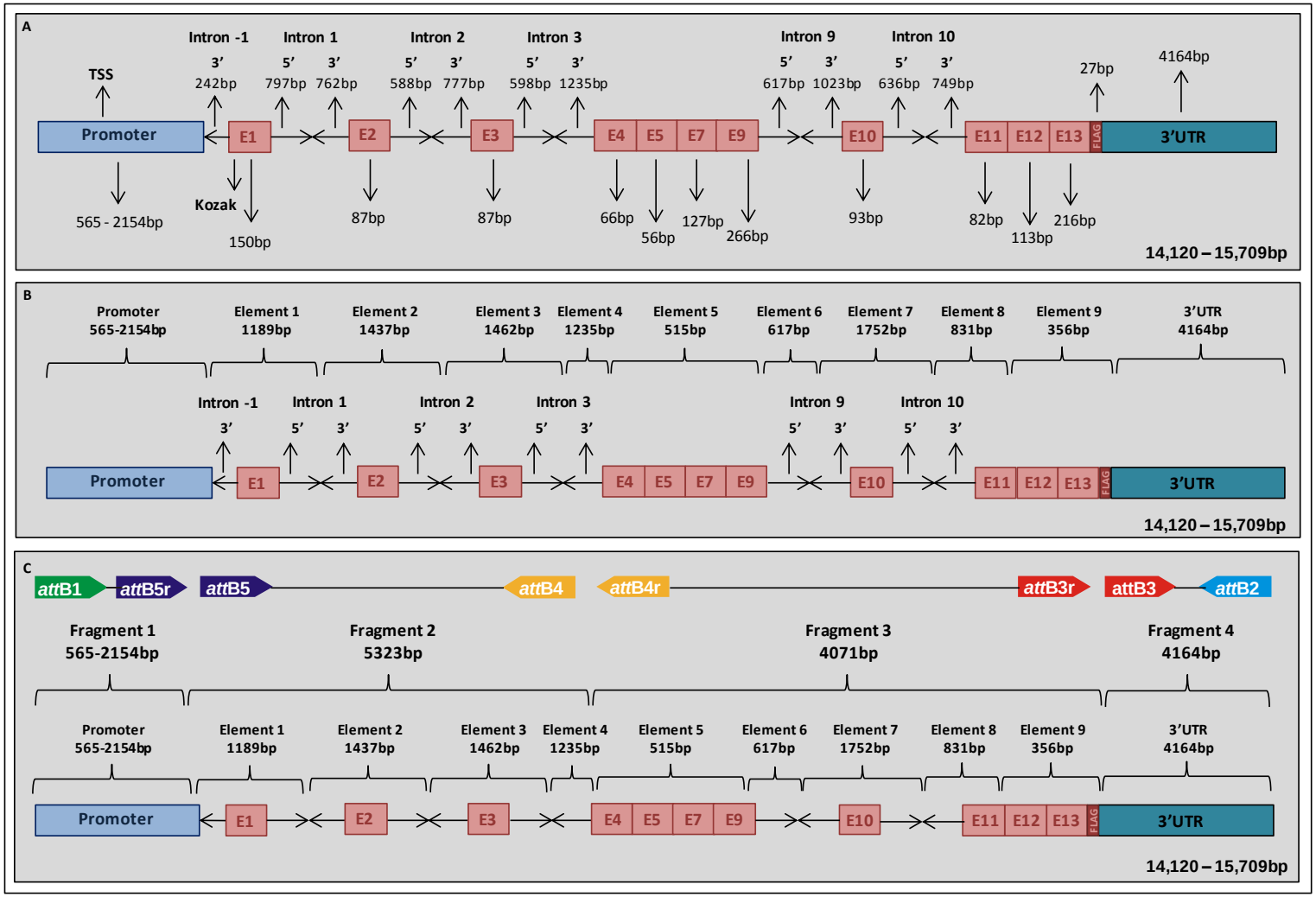


Figure 4.5 The minigene blueprint

A: The basic design

B: The separate PCR elements required to construct the minigene. Elements 5 and 9 were amplified from cDNA, the rest from genomic DNA

C: The PCR elements were grouped into four larger fragments to meet the requirements of the Multisite Gateway® protocol (Invitrogen)

The composition of the final minigene fragments was carefully considered to take advantage of the modular design of the Gateway® system. As described previously, one of the main advantages of this system is the flexibility it provides in the study of genetic variation, as variants of one of the DNA fragments can be swapped in while the other three remain the same. It is also possible – with some adjustments to the protocol – to leave one or even two fragments out completely, allowing a more focused investigation of certain sections of the minigene. Another consideration was the final size of each fragment, as the closer the four fragments are in size, the more efficient the final recombination reaction will be. For these reasons, and in consideration of the hypothesis under investigation, the first fragment comprised the promoter element alone to allow the straightforward swap-in of the three different promoters. The second fragment contained the alternatively spliced exons 2 and 3 and the third contained the 3R/4R-determining exon 10. Placing exons 2 and 3 on a separate fragment to exon 10 creates the possibility of studying the N-terminal (exons 2 and 3) and C-terminal (exon 10) alternative splicing events separately if necessary at a later date. The fourth fragment contained the full-length 3'UTR. Figure 4.5C details the minigene elements included in each fragment.

4.7 *MAPT* minigenes: Target DNA fragment construction

4.7.1 Fragment 1 (F1): the promoter elements

4.7.1.1 Promoter 1 (F1-242): The *MAPT* core promoter in conjunction with the rs242557 regulatory domain

The first promoter element comprised the *MAPT* H1 or H2 core promoter (CP) in conjunction with the distal regulatory domain containing the rs242557 polymorphism (the SD element). The construction of this promoter was described in section 3.5, when it was cloned into the pGL4.10 [*luc2*] vector as part of the luciferase reporter gene study of the *MAPT* promoter elements. To recap, the H1B and H1C variants comprised the H1 core promoter with either the rs242557 G-allele (H1B-G) or A-allele (H1C-A) variant of the SD cloned immediately

downstream. The rs242557 polymorphism in the SD represented the only sequence difference between these two constructs, allowing a direct examination of the effect of this polymorphism on minigene expression. The H2 variant comprised the H2 core promoter with the G-allele variant of the SD domain, with several sequence differences in both the CP and SD separating this variant from its H1 counterparts. The luciferase reporter study revealed that the H1C-A variant conferred an approximate 2-fold increase in transcriptional activity in comparison to the H1B-G variant, with a corresponding 4-fold increase over the H2-G variant observed. Thus, the study of this promoter variant in the wider context of the minigene allowed the determination of the effect of the increased transcription rate of H1C-A on the inclusion/exclusion rate of exons 2, 3 and 10.

To make these constructs – denoted from here onwards as F1-242 – compatible with the Gateway system, *att*B1 and *att*B5r recombination sequences (the combination required for fragment 1) were introduced onto the 5' and 3' ends respectively. This was achieved by PCR, with the pGL4.10 luciferase construct used as the template to amplify the full element in one reaction. The appropriate *att*B sequences were added onto the 5' end of the forward (*att*B1) and reverse (*att*B5r) primers, producing a 2,218bp PCR product containing the promoter element flanked by the *att*B1 site at the 5' end and the *att*B5r site at the 3' end. A full description of the *att*B PCR protocol is given in section 4.7.4.

4.7.1.2 Promoter 2 (F1-CP): The *MAPT* core promoter alone

The second promoter was the 1,342bp core promoter (CP) element containing the major *MAPT* transcription start site (exon 0) and is described in section 3.5. The H1 variant doubled as the promoter element for both the H1B and H1C minigene variants, whereas the H2 variant was included on the H2 minigene only. The inclusion of this promoter element in the investigation allowed the effects of the rs242557 domain of the F1-242 variant to be separated from the effects conferred by the core promoter alone. It also allowed the examination of basic differences between the unregulated H1 and H2 core promoters independently of the rs242557 regulatory domain and, as the H1B and H1C core promoters are

identical, provided a means of detecting any changes in alternative splicing pattern that resulted from genetic variation elsewhere in the minigene.

As with the F1-242 constructs, the required *att*B1 and *att*B5r sequences were introduced onto the ends of the H1 and H2 CP elements (denoted from here onwards as F1-CP) by PCR, using the pGL4.10 luciferase constructs described in section 3.5.5 as the template. The total size of the F1-CP *att*B PCR product was 1,400bp.

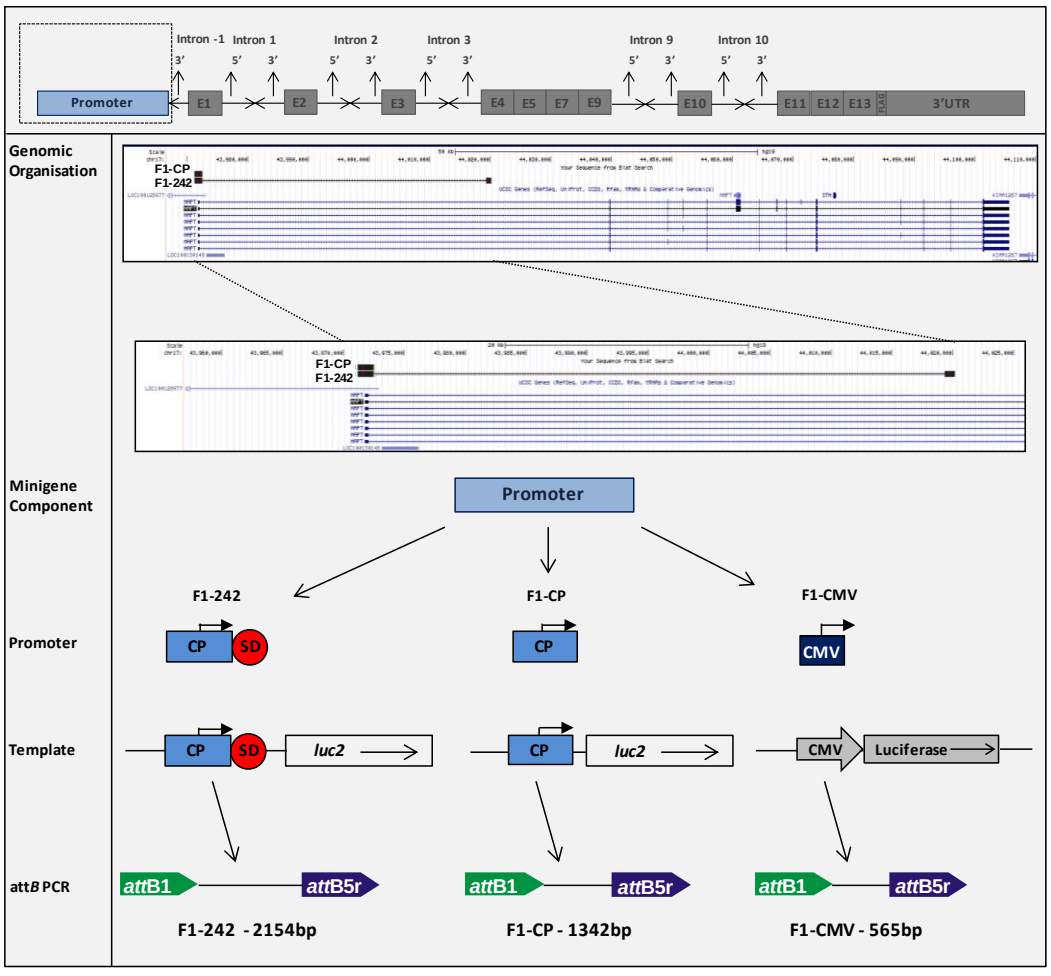


Figure 4.6 The cloning process for the construction of the three minigene promoters. CP = the H1 *MAPT* core promoter; SD = the rs242557 regulatory domain; CMV = the cytomegalovirus immediate early promoter.

4.7.1.3 Promoter 3 (F1-CMV): The cytomegalovirus promoter

The third promoter element was an extrinsic, non-mammalian control promoter originating from the cytomegalovirus (CMV). Placing the minigene variants under the control of the same CMV promoter allowed the detection of differences in alternative splicing and mRNA processing due solely to the genetic variation in the body of the minigene and 3'UTR (i.e. in Fragments 2, 3 and 4 only). Of greater interest, however, is the assessment of the role of promoter identity in the regulation of tau alternative splicing, achieved by determining whether a viral promoter can recapitulate the pattern of expression conferred by the intrinsic *MAPT* promoter. If differences are observed, this would indicate that elements specific to the *MAPT* promoter region play a role in alternative splicing, pointing towards co-transcriptional regulation as a likely mechanism for the control of tau isoform expression.

The CMV promoter is highly active – significantly higher than the *MAPT* promoter – and is commonly used in expression studies as either a control promoter or in instances where the promoter is not under investigation, for example in many miRNA studies. There are therefore many plasmid vectors available containing the CMV immediate early promoter element, one of which is the pMIR-REPORT vector used in the 3'UTR luciferase reporter gene study and described in section 3.6.3. Thus, the CMV promoter element was amplified by PCR using the pMIR-REPORT vector as the template. The full sequence of the CMV immediate early promoter element is given in Appendix H.

As with the F1-242 and F1-CP constructs, the *att*B1 and *att*B5r sequences were added onto the 5' and 3' ends of the construct (denoted from here onwards as F1-CMV) respectively by PCR. The total size of the F1-CMV *att*B PCR product was 623bp. The processes involved in producing each of the three F1 promoter variants are presented in Figure 4.6.

4.7.2 Fragment 2 and Fragment 3

4.7.2.1 Fragment 2 (F2) composition

Fragment 2 comprised protein-coding exons 1, 2 and 3 with surrounding 5' and 3' segments from introns 1, 2 and 3. The intronic sequences are necessary as they contain vital signals required for the regulation of alternative splicing events. The minimum sequence lengths of these and all introns present in the minigene were based upon the findings of Yu *et al*, who demonstrated that 569bp of the 5' sequence and 725bp of the 3' sequence of intron 10 was required to maintain the correct pattern of splicing in their exon 10 construct [210]. In the absence of similar information for the other *MAPT* introns, these sizes were set as a minimum inclusion length for intronic segments throughout the minigene. This marks an important departure from the Dawson minigene, where as little as 172bp of intronic sequence was included.

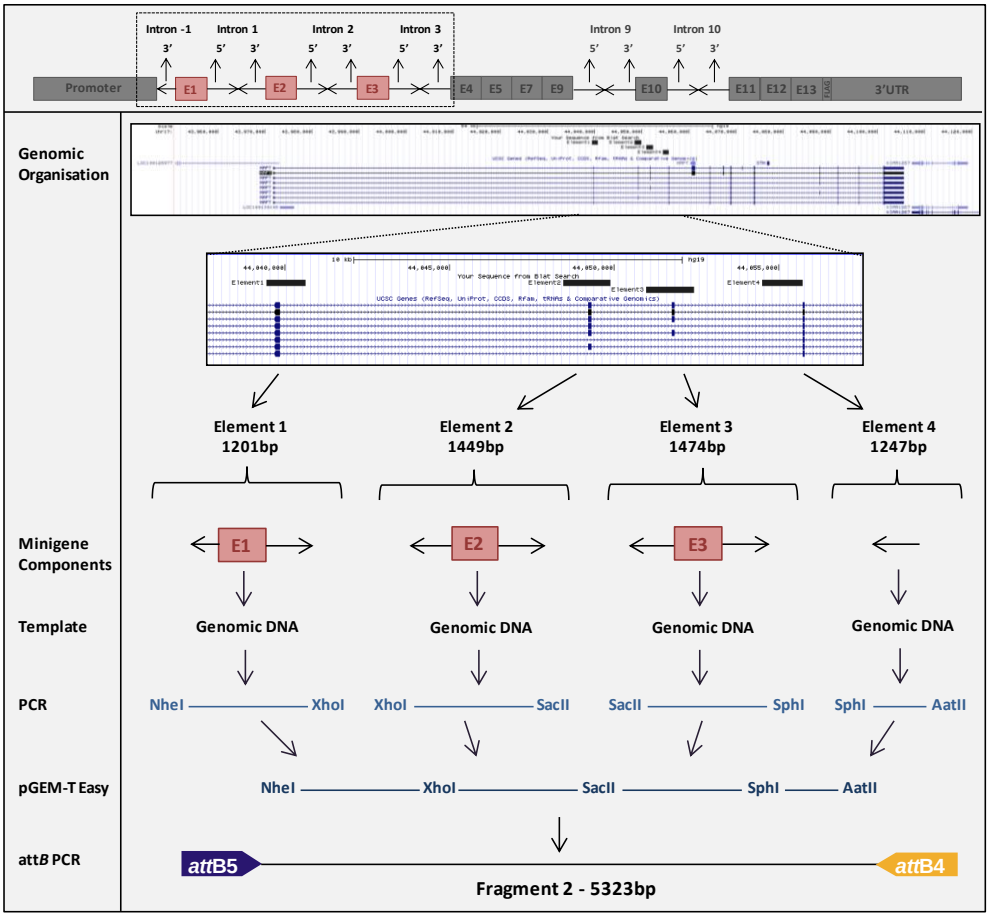


Figure 4.7 The cloning process for the construction of Fragment 2

Exons 1, 2 and 3 were contained within three separate elements (denoted 1-3), with each element amplified from genomic DNA by PCR. One set of primers per element was sufficient to capture the 3' segment of the upstream intron, the exon and the 5' segment of the downstream intron. An additional fourth element contained only the 3' segment of intron 3. The composition of each element in Fragment 2 is presented in figure 4.7.

4.7.2.2 Fragment 3 (F3) composition

Fragment 3 contained the remaining protein-coding exons and the intron 9 and 10 segments surrounding exon 10. As exons 4, 5, 7, 9, 11, 12 and 13 are constitutively present in *MAPT* mRNA, the surrounding intronic segments containing the splicing signals were deemed unnecessary. Thus, to reduce the size of the minigene – and therefore increase transformation efficiency – introns 4, 5, 7, 11 and 12 were excluded in their entirety.

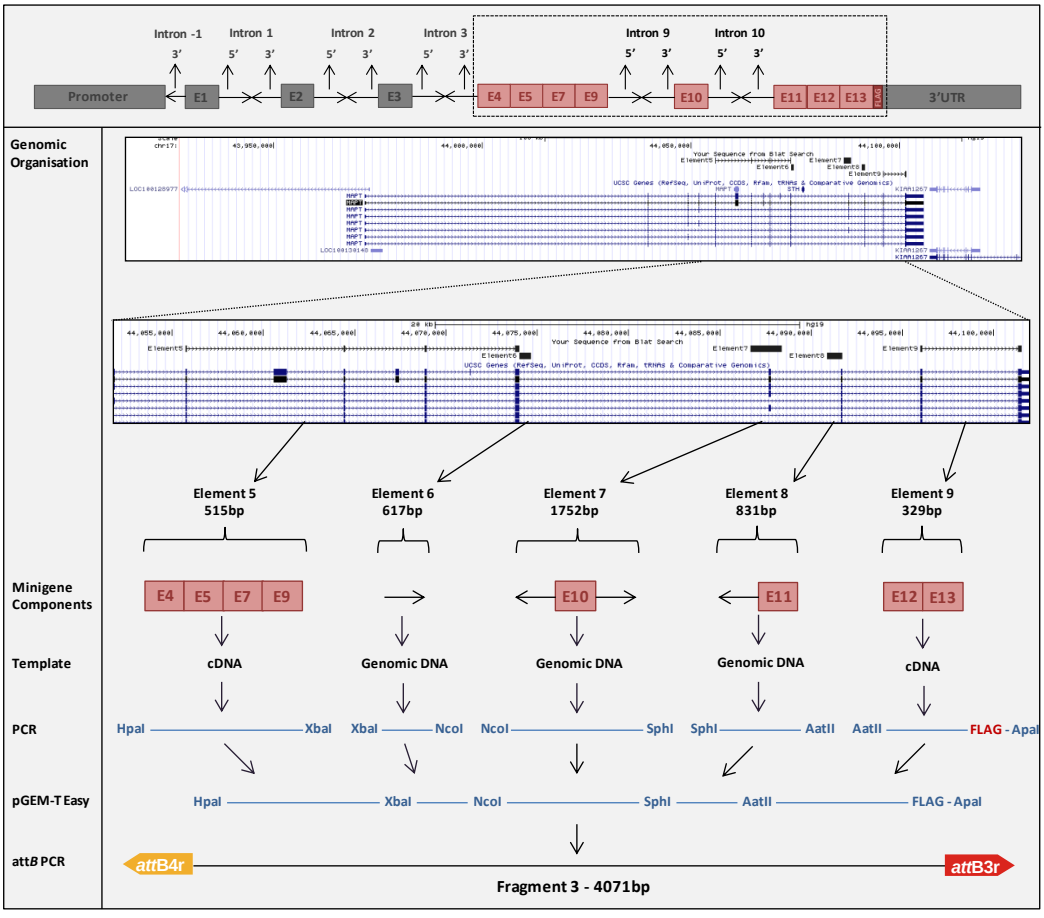


Figure 4.8 The cloning process for the construction of Fragment 3

The exons were amplified in two groups by PCR using cDNA reverse transcribed from RNA as the template. The RNA samples were extracted from the brain tissue of the same three patients from whom the genomic DNA elements have been cloned, thus encompassing the same genetic variation. Exons 4, 5, 6 and 9 were amplified in one PCR to form element 5. Element 9 was initially designed to similarly include exons 11, 12 and 13; however the primers required to achieve this in one PCR reaction were incompatible and therefore only exons 12 and 13 were incorporated into this element. Exon 11 was instead amplified from genomic DNA in a PCR reaction that also incorporated the 3' segment of intron 10. Exon 10 and the remaining intronic segments were amplified from genomic DNA to form the final two elements. The exact composition of elements 5-9 are presented in figure 4.8.

4.7.2.3 Fragment 2 and 3 construction

As described above, elements 1-9 were created by PCR from a starting template of either genomic DNA or reverse transcribed cDNA. To create Fragment 2, elements 1-4 were ligated together, as were elements 5-9 to create Fragment 3. To facilitate this, specific restriction enzyme recognition sequences were introduced onto the 5' and 3' ends of each element during PCR. Fragments 2 and 3 were each constructed in the pGEM-T Easy vector (section 3.5.4) and therefore the combinations of restriction sites were chosen to allow the sequential and directional insertion of each element into the multiple cloning site (MCS) of this vector (figures 4.7 and 4.8). Each six-nucleotide restriction sequence was attached onto the 5' end of the forward or reverse primer and was thus incorporated onto the appropriate end of the element during PCR. The primer sequences, restriction sites and PCR conditions required for the amplification of each element are given in table 4.1.

The element 9 reverse primer contained the FLAG-tag motif in addition to the required ApaI restriction site. The FLAG-tag motif is 27 nucleotides in length and provides a method of distinguishing the minigene tau protein from the endogenous species through its reactivity with the FLAG antibody. The

positioning of the motif between the restriction site and the element 9 target sequence at the 5' end of the reverse primer resulted in its incorporation onto the 3' end of element 9, immediately downstream to the final coding exon. The stop codon at the end of exon 13 was moved to the end of the FLAG motif to ensure the tag is transcribed and translated. The sequence of the FLAG-tag motif is: GATTACAAGGATGACGACGATAAGTAA.

The PCR products were purified using the QIAquick PCR Purification kit and individually ligated into the pGEM-T Easy vector using the cloning method described in section 3.5.4 Positive clones were identified by blue/white *β-galactosidase* screening and confirmed by sequencing.

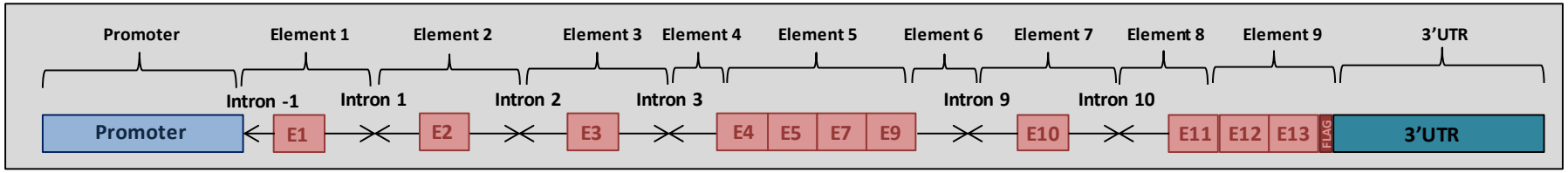
To construct Fragment 2, component elements 2, 3 and 4 were cut out of their pGEM-T Easy vectors and inserted into the vector containing element 1. This was done by digestion of each clone with the appropriate pair of enzymes (table 4.1), followed by a single multi-fragment ligation reaction between the linearised pGEM-T/element 1 vector and the three digested elements. A 6:1 ratio of insert:vector was used for this reaction, with optimal efficiency achieved by calculating the exact amount of each insert required to provide an equimolar ratio of DNA ends. This calculation takes into account the size of each element and uses the following formula to convert μg of DNA to pmol of ends:

$$\mu g\ DNA \times \frac{pmol}{660pg} \times \frac{10^6 pg}{1\mu g} \times \frac{1}{N} \times 2 \times \frac{kb\ element}{1000bp} = pmol\ DNA\ ends$$

Where: N is the number of nucleotides (in kb), 660pg/pmol is the average molecular weight of a single nucleotide pair, 2 is the number of ends in a linear DNA molecule, and kb/1000bp is a conversion factor for kilobases to base pairs (Promega BioMath calculator, "Linear DNA: Micrograms to Picomoles of DNA Ends").

A total of 25ng of digested pGEM-T/element 1 vector (0.018 pmol ends) and 52ng, 23ng and 45ng of digested elements 2, 3 and 4 (each 0.108 pmol ends) were included in the 20µl ligation reaction, which was incubated overnight at 16°C with 40 units of T4 DNA ligase, 1x T4 DNA ligase buffer and 10mM ATP. Half of the ligation mix was transformed into 100µl of JM109 *E.coli* cells and incubated overnight on LB-agar plates containing 50µg/ml of ampicillin. Positive clones were screened by digestion and confirmed by sequencing.

To construct Fragment 3, elements 6, 7, 8 and 9 were similarly inserted into the vector containing element 5. The enzymes used to digest each element are given in table 4.1. The large discrepancy in size between element 7 (1,764bp) and the other four elements (368-843bp) meant a multi-fragment ligation was unsuitable in this instance, as optimal efficiency occurs when each element is similar in size. Thus, the elements were inserted one at a time over four rounds of cloning to produce the final Fragment 3. Each single-fragment ligation comprised 150ng of digested element, 50ng of linearised pGEM-T/element(s) vector, 2 units of T4 DNA ligase, 1x T4 DNA ligase buffer and 10mM ATP and was incubated overnight at 4°C. A volume of 50µl of JM109 cells were transformed with 5µl of ligation mix and incubated overnight on LB-agar plates containing 50µg/ml of ampicillin. Positive clones were screened by digestion and confirmed by sequencing following each round of cloning.



	DNA template	Primer	RE site	Sequence (5'-3')	Length (bp)	AT (°C)	Mg (mM)	Size (bp)	Double digestion, buffer
Element 1	Genomic	Forward	NheI	GCTAGCCCTGGTGGTGTTGAATATGA	26	63	2.5	1201	XhoI/AatII, NEB4
		Reverse	XhoI	CTCGAGAAGAGGCGAAGTCAATTTGG	26				
Element 2	Genomic	Forward	XhoI	CTCGAGCACAGGGAAGGGACAATTCA	26	65	1.8	1449	XhoI/SacII, NEB4
		Reverse	SacII	CCGCGGCTTGACTGACACAGATGGGA	26				
Element 3	Genomic	Forward	SacII	CCGCGGAAAGCCTTCAAAGCTGACAA	26	65	1.8	1474	SacII/SphI, NEB4
		Reverse	SphI	GCATGCGCCCTGTCTGATTGATTCCC	26				
Element 4	Genomic	Forward	SphI	GCATGCCCCGTGAGCCCATTG	21	65	1.8	1247	SphI/AatII, NEB4
		Reverse	AatII	GACGTCCTGGTGTATGTGTCAGCAAA	26				
Element 5	cDNA	Forward	HpaI	GTTAACCTGAAGAAGCAGGCATTGGA	26	60	2.5	527	XbaI/ApaI, NEB4
		Reverse	XbaI	TCTAGACTTCCCGCCTCCCGGCT	26				
Element 6	Genomic	Forward	XbaI	TCTAGAGTGAGAGTGGCTGGCTG	23	65	1.8	629	XbaI/NcoI, NEB4
		Reverse	NcoI	CCATGGTAACGCACCCAGACGA	22				
Element 7	Genomic	Forward	NcoI	CCATGGAAGACGTTCTCACTGATCTGG	27	63	2.5	1764	NcoI/SphI, NEB2
		Reverse	SphI	GCATGCCACTTTGGTTTGGCTCTTTG	26				
Element 8	Genomic	Forward	SphI	GCATGCCTCGAGCTTACTGAGACACTA	27	65	1.8	843	SphI/AatII, NEB4
		Reverse	AatII	GACGTCCTGGTTTATGATGGATGTTGCCTA	30				
Element 9	cDNA	Forward	AatII	gtcattacatattGACGTCGAGGTGGCCAGGTG	33	60	1.8	368	AatII/ApaI, NEB4
		Reverse	ApaI	GGGCCCTTACTTATCGTCGTCATCCTTGTA ATCCAAACCCTGCTTGG	47				

Table 4.1 The primers, restriction enzyme sites, PCR conditions and digestion conditions for the cloning and ligation of each minigene element. PCR products were cloned into the pGEM-T Easy vector to product Fragments 2 (elements 1-4) and 3 (elements 5-9).

4.7.2.4 Gateway modifications

Once Fragments 2 and 3 had been constructed in the pGEM-T Easy vector, it was necessary to add the appropriate *att*B sequences onto the 5' and 3' ends. As with Fragment 1, this was done by PCR with the *att*B sequences added onto the 5' end of the forward and reverse primers. Fragment 2 was flanked by *att*B5 and *att*B4 sequences, with *att*B4r and *att*B3r sites incorporated onto the ends of Fragment 3. The total size of the F2 and F3 *att*B PCR products was 5,398bp and 4,086bp respectively.

Full schematics of the processes involved in the creation of Fragments 2 and 3 are given in figures 4.7 and 4.8, respectively.

4.7.3 Fragment 4 (F4): the 3'UTR

Fragment 4 constituted the *MAPT* 3'UTR which was previously cloned in the pMIR-REPORT vector for investigation in the luciferase reporter gene study described in section 3.12. As such, the full-length construct was used as the template in the *att*B PCR reaction to introduce the *att*B3 and *att*B2 sequences onto the 5' and 3' ends of the fragment, respectively. The total size of the F4 *att*B PCR product was 4,428bp. A schematic of the preparation of Fragment 4 for inclusion in the Multisite Gateway® protocol is given in figure 4.9.

4.7.4 *att*B PCR

The standard PCR protocol had to be modified to take into account the addition of 27-31 nucleotides of *att*B sequence – which does not anneal to the target DNA in the initial PCR cycles – to the primers and the large size (up to 5.4kb) of the target fragments. The AccuPrime High Fidelity polymerase enzyme blend (Invitrogen) is designed for the amplification of large products from plasmid DNA and provides increased specificity for PCR conducted with suboptimal primers. Each 25µl PCR reaction comprised: 100ng of purified plasmid construct, 0.25µl of AccuPrime Taq, 1x AccuPrime Buffer I, dNTPs (each to a final concentration of 10mM), the forward and reverse primers (each to a final concentration of 0.2µM), with magnesium and DMSO added as necessary. Each reaction was heated to

94°C for 4 minutes followed by 10 cycles of 94°C for 30 seconds, a suitable annealing temperature for 30 seconds and 68°C for 1-5 minutes depending on the product size. A further 20 cycles of 94°C for 30 seconds and 68°C for 1.5-5.5 minutes was followed by a final extension at 68°C for 7 minutes. The primer sequences and PCR conditions used for the amplification of each fragment are given in table 4.2. The *att*B PCR products were resolved by agarose gel electrophoresis and purified by QIAquick Gel Extraction kit.

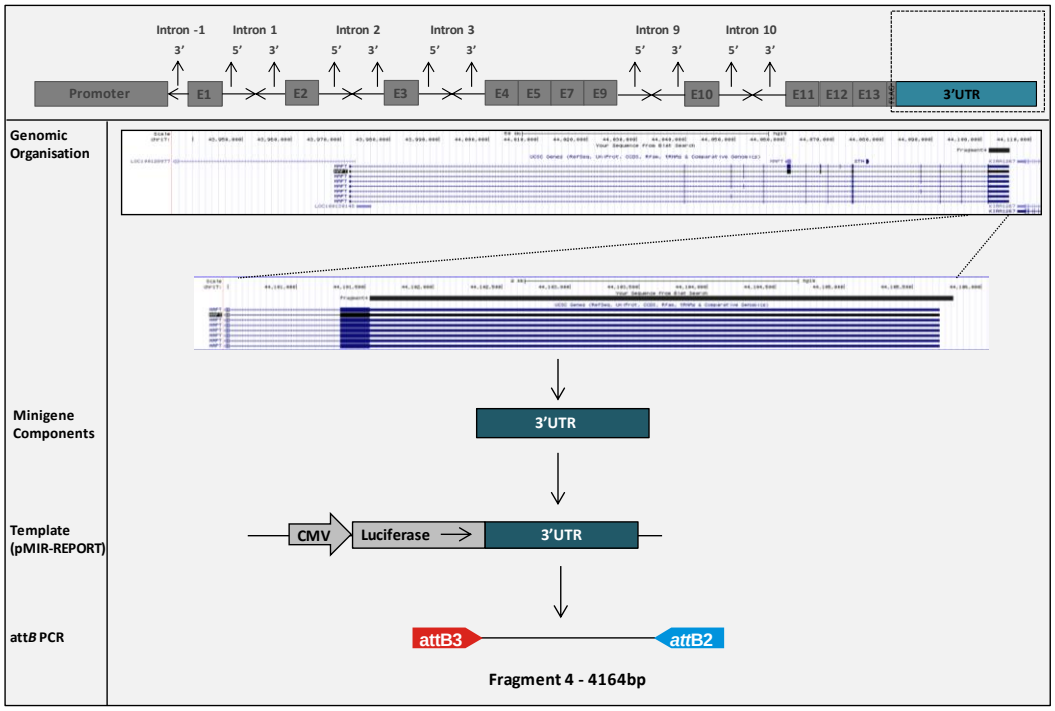


Figure 4.9 The cloning process for the construction of Fragment 4 containing the *MAPT* 3'UTR

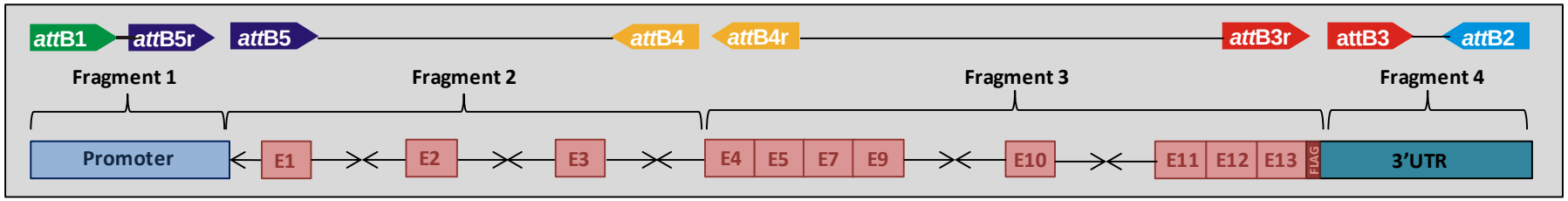
4.7.5 Entry clone creation and the BP reaction

The first stage of the Gateway® process involves the insertion of each fragment into one of four pDONR vectors to create an 'entry clone'. This requires the proprietary BP Clonase™ II enzyme which catalyses recombination between the *att*B-flanked PCR fragments and the pDONR vector carrying the corresponding *att*P sites. There are several versions of the pDONR vector that differ only by the *att*P site variants they carry (*att*P1-5/*att*P1r-P5r). All pDONR vectors carry the kanamycin resistance gene for the selection of positive recombinant clones and

are approximately 4,770bp in size. A schematic of the original pDONR 221 vector is given in figure 4.11A.

Each 10µl 'BP' reaction comprised: 2µl of BP Clonase II, 150ng of pDONR vector and 150ng of purified *att*B PCR product. BP reactions were incubated overnight at 25°C and then at 37°C for 10 minutes with an additional 1µl of Proteinase K. The full recombination mixture was transformed in 100µl of HB101 *E.coli* cells and positive recombinants were selected on LB agar plates containing kanamycin. The pDONR vector contains the *ccd*B cassette situated between the two *att*P sites. This cassette produces a protein that interferes with *E.coli* DNA gyrase, thereby suppressing the growth of most strains of *E.coli*, including HB101 cells. On successful recombination with the *att*B fragments, the *ccd*B cassette is removed from the entry clone and therefore only cells that have taken up a recombinant vector will grow, providing a second means of selection in addition to kanamycin resistance. Bacterial strains that contain the F' episome, including the JM109 cells used previously, are resistant to the effects of the *ccd*B cassette and are thus unsuitable in this instance.

Successful entry clone formation was confirmed by sequencing with the M13 forward and reverse primers that anneal at either side of the recombination site (F: GTAAAACGACGGCCAG; R: CAGGAAACAGCTATGAC). A midi preparation of the final entry clones produced the larger yields required for the next stage of the Gateway® protocol.



	Primer	*att* B variant	*att* B sequence (5'-3')	Target-specific sequence (5'-3')	Length (bp)	AT (°C)	Mg (mM)	DMSO (%)	Size (bp)
F1-242	F	*att* B1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTC	CAAATGCTCTGCGATGTGTT	51	56	1.5	5	2218
	R	*att* B5r	GGGGACAACTTTTGTATACAAAGTTGT	GGCTGTCGATGAACCCTA	45				
F1-CP	F	*att* B1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTC	CAAATGCTCTGCGATGTGTT	51	55	1.5	5	1400
	R	*att* B5r	GGGGACAACTTTTGTATACAAAGTTGT	GGACAGCGGATTTCAGATTC	47				
F1-CMV	F	*att* B1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTC	CTCTGCTTATATAGACCTC	50	51	1.5	-	623
	R	*att* B5r	GGGGACAACTTTTGTATACAAAGTTGT	AGTTATTAATAGTAATCAATTACGGGG	53				
F2	F	*att* B5	GGGGACAACTTTGTATACAAAAGTTGTC	CCTGGTGGTGTTGAATATGA	48	60	3.5	-	5398
	R	*att* B4	GGGGACAACTTTGTATAGAAAAGTTGGGTG	CTGGTGTATGTGTCAGCAAA	50				
F3	F	*att* B4r	GGGGACAACTTTTCTATACAAAGTTGTCCAG	CTGAAGAAGCAGGCATTGGA	51	57	1.5	-	4086
	R	*att* B3r	GGGGACAACTTTATTATACAAAGTTGT	TCACTTATCGTCGTCATCCTTGTAATC	54				
F4	F	*attB3r*	GGGGACAACTTTGTATAATAAAGTTGTC	CCTGGGGCGGTCAATAA	45	60	1.5	-	4428
	R	*att* B2	GGGGACCACTTTGTACAAGAAAGCTGGGTA	GCCAGCATCACAAAGAAG	48				

Table 4.2 The primer sequences and PCR conditions for the introduction of the *att*B sequence variants onto the 5' and 3' ends of the four large minigene fragments (F1-F4).

Figure 4.10 contains schematics of each BP reaction, detailing the pDONR vector used to create each clone, the combination of *att*B and *att*P sites involved and the *att*R and *att*L sites formed post-recombination. Three entry clones were produced for each fragment, representing the three *MAPT* haplotype variants.

Figure 4.10 The creation of entry clones using the BP reaction
The four entry clones are created by individual recombination reactions between the *att*B PCR products and specific pDONR vectors. *Modified from Invitrogen's 'Mulitsite Gateway Pro' manual.*

4.8 Final minigene construction

4.8.1 The LR reaction

The four fragments were joined together to produce the complete minigene in a final one-step recombination reaction denoted 'LR'. The four entry clones each contain one of the minigene fragments flanked by a different combination of *att*R1-5 and *att*L1-5 variant sites and this specificity is vital in ensuring the fragments are incorporated into the minigene in the correct order and orientation.

The final minigene is formed by the simultaneous transfer of the four fragments into a 'destination' vector (pDEST) to produce the expression clone. There are different versions of the pDEST vector available but only the R4 pDEST vector (figure 4.11B) allows the integration of the expression clone into the genome of a platform cell line, as desirable in this investigation.

Figure 4.11 The basic blueprint of the Gateway® vectors
A: Variants of the pDONR™221 vector were used to create the Gateway® entry clones. The *att*P sites recombine with the *att*B sites at the ends of the minigene PCR fragments to produce the individual entry clones required to create the final expression clone. B: The R4 destination vector used to create the final minigene or 'expression clone'. ***Modified from Invitrogen's 'Mulitsite Gateway Pro' manual.***

The proprietary LR Clonase™ II enzyme simultaneously catalyses recombination between the *att*R and *att*L sites present in the four entry clones and the R4 pDEST vector. To achieve the highest efficiency in the 'LR' reaction, an equimolar ratio of each entry clone was required. This was calculated based on the total size of the clone and ensured the same number of DNA molecules were present for each. Ten femtomoles (fmoles) of each entry clone was required for the LR reaction and the ng conversion for each fragment is given in table 4.3. The formula for converting femtomoles (fmoles) of DNA into nanograms (ng) of DNA is:

$$ng = x \text{ fmoles} \times N \times \frac{660fg}{\text{fmoles}} \times \frac{1ng}{10^6 fg}.$$

where *x* is the number of fmoles and N is the size of the DNA in bp.

Entry clone	CP + rs242557		CP alone		CMV	
	Size (bp)	10fmol (ng)	Size (bp)	10fmol (ng)	Size (bp)	10fmol (ng)
F1-242	6,919	**14.2**	-	-	-	-
F1-CP	-	-	6,107	**8.9**	-	-
F1-CMV	-	-	-	-	5,330	**3.7**
F2	10,114	**35.2**	10,114	**35.2**	10,114	**35.2**
F3	8,751	**26.6**	8,751	**26.6**	8,751	**26.6**
F4	9,144	**28.8**	9,144	**28.8**	9,144	**28.8**

Table 4.3 The size of each entry clone and the amount (in ng) required to ensure an equimolar ratio (10fmol) of each component in the final LR reaction.
Sizes include the pDONR vector.

The DNA mixtures were incubated overnight at 25°C with 20fmoles of R4 pDEST vector and 2µl of LR Clonase™ II proprietary enzyme mix (adjusted to a final volume of 10µl with 1x TE buffer, pH 8.0). HB101 *E.coli* cells were transformed with half of the recombination mixture and selected on LB-agar plates containing ampicillin. Resistance to this selection antibiotic is only conferred by the bacterial cells that have taken up the pDEST vector and therefore those that have only taken up the component entry clones will not grow. Additional selection was provided by the presence/absence of the *ccd*B cassette (described in section 4.7.5) in the pDEST vector. For a four fragment LR reaction, approximately 50-100 colonies were produced each time. Ten clones per LR reaction were purified by miniprep and successful minigene formation was confirmed as described in the following section.

4.8.2 Confirmation of final minigene expression clones

The presence of the completed minigene in the expression clone was determined by restriction enzyme digestion of the purified plasmid DNA with the SphI enzyme. This enzyme cuts at multiple sites in the minigene and in the pDEST vector, producing a specific banding pattern that only occurs when all four fragments are present in the correct order and orientation (figure 4.12). Final endotoxin-free maxi preparations of expression clones exhibiting the correct SphI banding pattern were fully sequenced using a set of primers that annealed at approximate 800bp intervals along the minigene. The full sequences of the H1B and H1C minigenes are given in appendices I (CP variants) and J (CP+rs242557

variants), with a multiple sequence alignment (ClustalW2) highlighting the genetic differences between them.

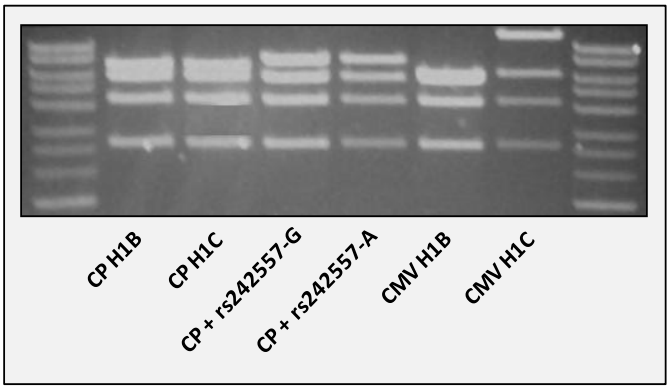


Figure 4.12 Confirmation of the successful minigene construction.
Digestion with restriction enzyme SphI produces a banding pattern that only occurs when all four fragments have been transferred into the R4 pDEST vector in the correct order and orientation.

4.8.3 The H2 minigene variants

Due to problems with the cloning of Fragment 3, it was not possible to produce H2 variants of the three minigenes. For unknown reasons, multiple and varied attempts to insert element 7 into Fragment 3 at the pGEM-T Easy stage all resulted in failure. It is unclear why this would occur for the H2 variant only, as the cloning processes for the H1B and H1C variants were completed successfully and efficiently. It was thus deemed more appropriate to focus on the H1 minigenes. It is also questionable as to whether the H2 variant would provide a fully informative model in this instance, as the design of the system would result in the H2 minigenes eventually being inserted into the genome of the cell lines in the same orientation as the H1 variants. As described previously, the genomic location containing the *MAPT* gene was subject to an ancient inversion, with the H2 variant subsequently lying in the opposite orientation to the H1 variants. This makes it highly likely that the endogenous H2 *MAPT* gene is subject to different positional effects and chromatin modifications than the H1 variant. Thus, although it would be interesting to determine expression differences purely at the sequence level, comparison between the H1 and H2 minigene variants may not be fully informative in this model. The H2 versions of Fragments 1, 2 and 4 were all completed and may by ‘swapped in’ to the H1 minigenes at a later date if deemed appropriate.

4.9 Transient expression of the minigene variants

4.9.1 Transfection

To confirm that the minigenes were capable of correctly expressing tau *in vitro*, each minigene was transiently transfected into the SK-N-F1 and SH-SY5Y cell lines. Cells were plated onto a 6-well culture plate and allowed to grow until approximately 80% confluent. An 18µl volume of TransFast transfection reagent was used to transfect 6µg of plasmid DNA with a TransFast:DNA charge ratio of 1:1. Cells were incubated post-transfection for 72 hours before harvest.

4.9.2 mRNA analysis

4.9.2.1 Reverse transcription-PCR

Total RNA was extracted from transfected cells using the TRIzol reagent (Invitrogen) and following the manufacturer's protocol. The RNA sample was incubated with DNase I for 30mins at 37°C to remove DNA contaminants before purification using the RNeasy MinElute Cleanup kit (Qiagen). One microgram of total RNA was reverse transcribed into cDNA using the Superscript III transcriptase and oligo(dT) primers. These primers comprise a string of T-nucleotide residues, ensuring only polyadenylated mRNA transcripts are reverse transcribed. The reverse transcription reactions were incubated at 42°C for 10 minutes, then at 53°C for 50 minutes. The transcriptase was inactivated during a 5 minute incubation at 85°C, followed by a final cooling step at 4°C for 10 minutes.

4.9.2.2 Exon 10 inclusion

The presence of exon 10 in the minigene mRNA was determined by PCR of the reverse transcribed cDNA. The forward primer annealed to exon 9 of *MAPT*, with a FLAG-tag-specific reverse primer ensuring only the minigene cDNA – and not endogenous tau – was amplified in the PCR reaction (F: AAGATCGGCTCCACTGAGAA and R: TTACTTATCGTCGTCATCCTTG). An amplicon of 585bp was produced when exon 10 was present in the minigene

mRNA and this represented the 4R-tau isoform. An amplicon of 492bp represented exon 10 exclusion and 3R-tau.

A volume of 1µl of cDNA template was amplified by 35 cycles of PCR using standard conditions and an annealing temperature of 55°C. The PCR products were resolved by polyacrylamide gel electrophoresis, with 6µl of product loaded onto a pre-cast 4-12% gradient TBE polyacrylamide gel and run at 200v for 50 minutes. To visualise the bands, the gel was stained with a 1:5000 dilution of SYTO® 60 red fluorescent DNA stain (Invitrogen) for thirty minutes, followed by two 5 minute washes with double distilled water. The stained gel was then visualised with the LI-COR Odyssey scanner, using the 700nm channel, the DNA gel setting, a 0.8mm focus offset and an intensity setting of either 1.5 or 3.0. The ratio of 4R-tau product (upper band) to 3R-tau product (lower band) was quantified using the ImageJ software (NIH).

The final outcome measure was an internal ratio between the 4R- and 3R-tau products and it was therefore vital that the PCR reaction did not reach the point of saturation. When saturation is reached, one or both of the amplification products stop increasing at an exponential rate, potentially altering the ratio between the two. To determine the saturation point for the minigene cDNA templates, a PCR was conducted as described above, with 37 cycles of denaturation, annealing and extension. A 5µl aliquot of product was removed after 28, 30, 32, 34, 36 and 37 cycles and resolved by polyacrylamide gel electrophoresis as described above. The 4R:3R-tau ratio of aliquots from the exponential stage of the PCR should not differ significantly. A significant reduction or increase in ratio indicates that the PCR saturation point has been reached. This must also be considered when scanning the polyacrylamide gel using the LI-COR system, as a high intensity setting could similarly saturate the SYTO®60 fluorescent signal and significantly alter the tau ratio. To optimise the Odyssey intensity settings, the gel was scanned twice at intensities of 1.5 and 3.0 and the tau ratios compared.

The PCR was performed on cDNA templates from SK-N-F1 (figure 4.13) and SH-SY5Y cells transfected with the H1C-CP and H1C-CMV minigene variants. The results show that, in all cases, 34 to 36 cycles of PCR was sufficient to allow accurate quantification without reaching the saturation point. The optimal Odyssey intensity setting, however, differed depending on the minigene variant. The CP-H1C ratios were consistent at both intensity settings. The 4R-tau signal of the CMV-H1C minigene, however, appeared to reach saturation at the 3.0 intensity level, resulting in a general decline of the 4R:3R-tau ratio. This is likely due to the marked increase in expression conferred by the CMV promoter in comparison to the CP promoter. Thus, the RT-PCR results from the CMV minigene variants were always quantified with the 1.5 intensity setting.

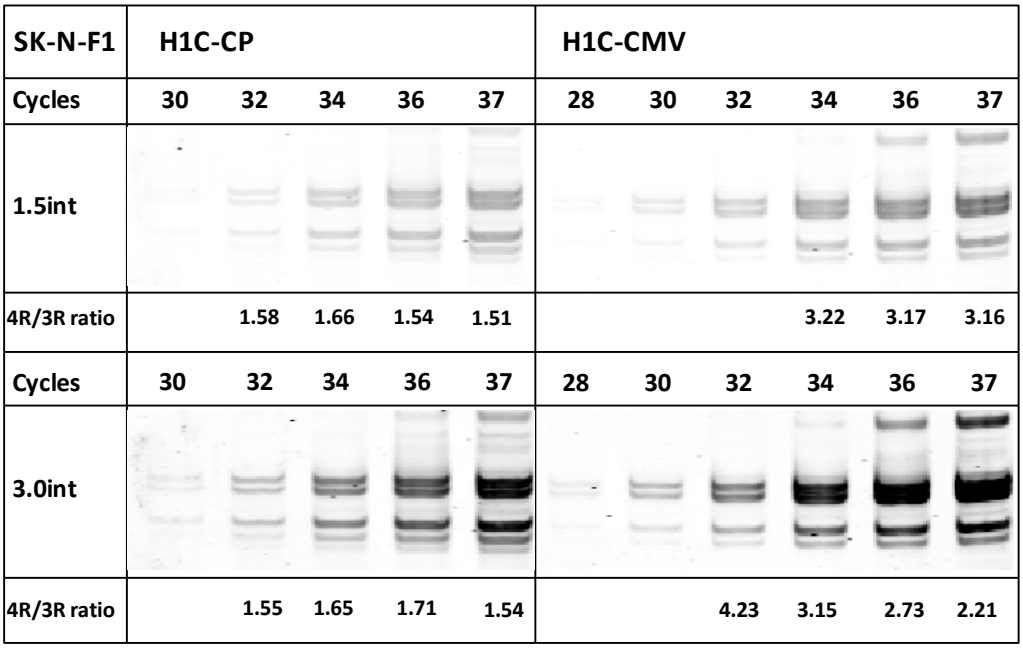


Figure 4.13 Polyacrylamide gel images of the exon 10 PCR optimisation.
cDNA was reverse-transcribed from undifferentiated SK-N-F1 cells transfected with the H1C variants of the CP and CMV minigenes. Aliquots of the PCR reaction were taken at 30, 32, 34, 36 and 37 cycles. The 4R/3R tau ratio (upper band divided by lower band) was quantified using the ImageJ software.

4.9.2.3 Mis-splicing events at exon 9

The exon 10 PCR optimisation revealed that both the 4R and 3R PCR products resolve as doublets and not, as expected, as single bands (figure 4.13). To determine the reason for this, the 4R and 3R PCR bands were purified and cloned into the pGEM-T Easy vector using standard protocols described previously in

section 3.5.4. Random selections of 36 clones were sequenced and the results show the occurrence of a mis-splicing event at the exon 9/intron 9 border in 23 of the 36 clones (64%). In these clones, the splice site at the exon/intron boundary was skipped, with splicing occurring 24bp downstream in intron 9. This caused the insertion of 26bp of intron 9 into the mRNA sequence between exons 9 and 10. All of the clones that were spliced correctly were 4R isoforms and accounted for 52% of the total number of 4R clones sequenced (N=23). None of the 3R clones sequenced were spliced correctly (N=12).

The reason for this mis-splicing event appears to lie in the design of the minigene. Exons 4, 5, 7 and 9 were cloned together in one element amplified from cDNA (element 5, figure 4.5). This was done purposely to remove the intervening intronic elements and reduce the size and transfection efficiency of the minigene. The 5' region of intron 9 was amplified from genomic DNA (element 6 in figure 4.5) and attached to the 3' end of exon 9 via restriction enzyme digestion and ligation. This resulted in the introduction of the XbaI sequence (TCTAGA) at the exon/intron boundary and appears to have weakened the splicing signal at this site. The full signal consists of the AG dinucleotide located at the 3' end of exon 9 and the GTG triplet at the 5' end of intron 9. The AGGTG motif is repeated 21bp downstream at the site of the mis-splicing event, suggesting that intron 9 contains a second, cryptic splice site (figure 4.14B). The purpose of this second site is currently unknown but comparative analysis of cDNA from untransfected cells suggests that the use of this site, if at all, is extremely rare in endogenous tau.

4.9.2.4 Exon 2 and 3 inclusion

The inclusion of exons 2 and 3 was determined in three separate PCR reactions using forward primers annealing in exons 1, 2 and 3. As with exon 10, the reverse primer in each PCR annealed to the minigene-specific FLAG-tag motif. The exon 1 PCR should amplify all six tau isoforms, though the six products resolve as four separate bands due to the small difference in size – 6bp – of the 1N3R and 0N4R, and 2N3R and 1N4R products. The exon 2 PCR should similarly produce four products visible as three bands, with 2N3R and 1N4R again resolving together.

The 2N3R and 2N4R products should be easily distinguishable by the exon 3 PCR. The expected product sizes for each PCR are given in table 4.4.

The PCRs were performed as described previously for the exon 10 PCR, with an annealing temperature of 55°C and 35 cycles of denaturation, annealing and extension. An aliquot of 6µl of PCR product was resolved by polyacrylamide gel electrophoresis, stained with SYTO®60 fluorescent stain and visualised by LI-COR Odyssey, as described previously. The results are presented in figure 4.15.

When performed on minigene cDNA, PCR product was detected for all three sets of primers, confirming the presence of exons 1, 2 and 3. The exon 1 and exon 2 PCRs, however, produced fewer bands than expected, with the largest band representing the 2N4R product notably absent in each case. The 2N4R isoform is the rarest of the six tau isoforms and therefore its absence is not necessarily surprising. The exon 3 PCR, however, revealed that this isoform is present, albeit in low abundance.

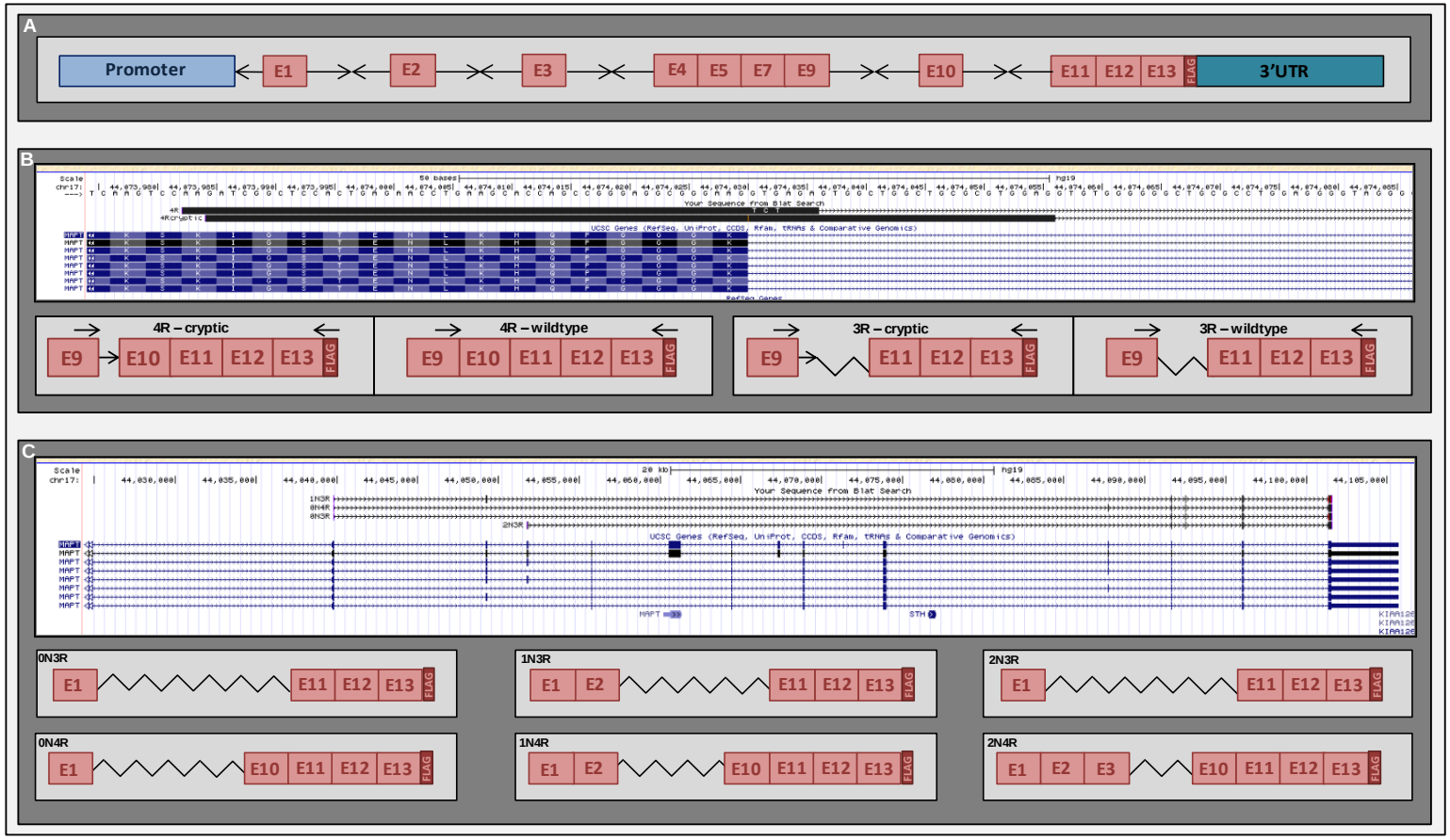


Figure 4.14 The minigene transcript mis-splicing events.
A: The basic blueprint of the *MAPT* minigene; B: The insertion of a restriction site into the exon 9/intron 9 5'splice site causes the preferential usage of a cryptic splice site located 21 nucleotides downstream; C: The skipping of the exon4-9 element results in the transcripts representing the six tau isoforms being shortened.

	Primer	Sequence (5'-3')	AT (°C)	0N3R	1N3R	2N3R	0N4R	1N4R	2N4R
Exon 1	Ex1-F	CATGCACCAAGACCAAGA	55	1002/487	1089/574	1176/661	1095/580	1182/667	1269/754
	FLAG-R	ATCCTTGTAATCCAAACCCTG							
Exon 2	Ex2-F	CTCTGAAACCTCTGATGCTAAG	55	-	999/484	1086/571	-	1092/577	1179/664
	FLAG-R	ATCCTTGTAATCCAAACCCTG							
Exon 3	Ex3-F	AGCACCCTTAGTGGATGAG	55	-	-	1038/523	-	-	1131/616
	FLAG-R	ATCCTTGTAATCCAAACCCTG							

Table 4.4 The primer sequences, PCR annealing temperature and expected product sizes of the Exon 1, Exon 2 and Exon 3 PCRs with the FLAG reverse primer.
The expected product sizes without element 5 (exons 4-9) are given in red.

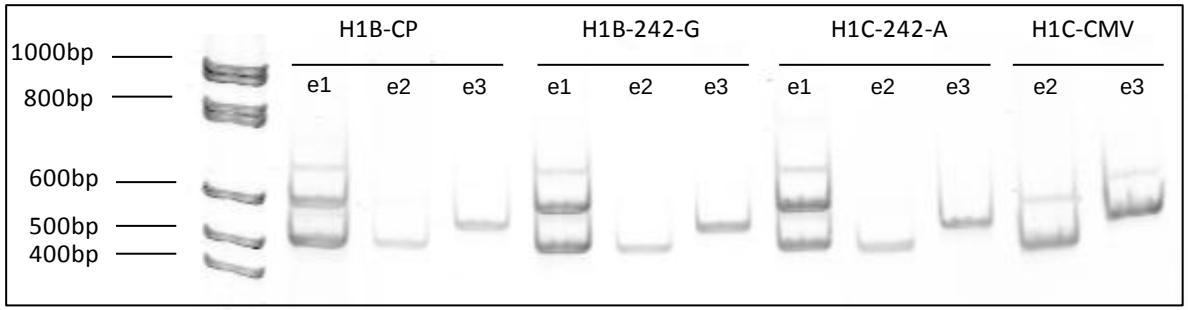


Figure 4.15 Optimisation of the exon1, exon 2 and exon 3 PCRs.
PCR with forward primers annealing in exon 1 (ex1), exon 2 (ex2) and exon 3(ex3), produces products ~500bp smaller than expected, as visualised by PAGE.

An important and intriguing finding, however, concerns the sizes of the products of each minigene N-terminal PCR, as all were approximately 500bp smaller than expected. This is roughly the same size as the exon 4-9 element of the minigene, suggesting that the splice site at the intron 3/exon 4 boundary has been skipped and exons 4, 5, 7 and 9 have been completely spliced out of the minigene mRNA. This was confirmed by the purification, cloning and sequencing of the PCR products (figure 4.14C). The complete absence of exons 4-9 in these transcripts was somewhat puzzling, as the exon 10 PCR described above – with a forward primer annealing in exon 9 – yielded product, confirming that exon 9 must be present in some of the transcripts even though there were no bands representative of this in the exon 1, 2 and 3 PCR products. In addition, the cloning and sequencing of these products appear to show that the bands observed for the exon 1 and 2 PCRs almost exclusively represent the 3R-tau isoforms, suggesting that the mis-splicing event at exons 4-9 favours the exclusion of exon 10. The reason for this is unclear; however, further analyses of the N-terminal splicing events conferred by the minigenes will be described in chapter 5.

4.9.3 Protein analysis

The mis-splicing events described above, specifically the use of the cryptic splice site in intron 9, causes a shift in opening reading frame and thus a change in protein translation. A truncated form of the tau protein may still be expressed but the absence of the FLAG-tag at the C-terminal end prevents its detection using the minigene-specific FLAG antibody. It was therefore not possible to study minigene expression at the protein level in this study.

4.9.4 Suitability of the minigenes for use in this project

Despite the mis-splicing events at the intron 3/exon 4 and exon 9/intron 9 boundaries, the minigene *is* expressed at the mRNA level *in vitro* and is alternatively spliced at exons 2, 3 and 10 in a consistent and highly replicable pattern. The mis-splicing events are common to all six variants of the minigene, with each affected in the same way. The aim of this project is, firstly, to link the promoter and transcription to the regulation of tau alternative splicing and,

secondly, to assess the effect of the rs242557 polymorphism on the splicing of exon 10. The transient transfections have shown that minigene mRNA is alternatively spliced at exon 10 and therefore the effect of the promoter and the genetic variation within it on exon 10 splicing can still be determined. Thus, the core themes of this project can still be tested at the mRNA level by determining whether small changes in sequence can affect the splicing pattern. Any biological interpretation of the results, however, must be made with extreme caution.

4.10 Cell models

4.10.1 Generating the R4 platform cell line

Transient expression is highly variable, with factors such as the passage and growth rate of the cells, the large size of the minigene plasmid (~20kb) and variance in transfection efficiency having an effect on the overall level of minigene expression. As described in section 4.4, the use of the Gateway® system to create the minigenes provides a means of integrating them into the genome of chosen cell lines to produce stably-expressing cell models. This should remove most of the variability, increase minigene expression and produce more consistent results.

The platform cell lines were created by co-transfecting the pJTI/Zeo platform vector with a vector containing the PhiC31 integrase gene (pJTI/PhiC31). The pJTI/Zeo vector contains the Hygromycin B resistance gene for initial selection, the R4 *att*B sequence and a promoterless zeocin resistance gene for selection during the retargeting stage (see figure 4.4). The PhiC31 integrase is required for integration of the platform vector at pseudo-PhiC31 *att*P sites that occur naturally in the mammalian genome. A 6µg amount of each plasmid was co-transfected into SK-N-F1 and SH-SY5Y cells plated at approximately 80% confluency onto 6-well culture plates. After recovering from transfection, cells were transferred to a 100mm culture dish and recombinant colonies were selected using 50µg/ml of the antibiotic Hygromycin B. Well-defined colonies were visible after 1-2 weeks of

selection. These colonies were manually picked and individually expanded under selection to form new, isogenic platform cell lines.

Genomic DNA was isolated from each platform cell line using the CellsDirect Resuspension and Lysis Buffers according to the manufacturer's protocol. Briefly, 10,000 to 30,000 cells were pelleted and washed with 500µl of PBS before resuspension in a mixture of 20µl of Resuspension Buffer and 2µl of Lysis Buffer. Following incubation at 70°C for 10 minutes, 3µl of cell lysate was analysed by nested PCR using primers specific to the pJTI/Zeo platform vector. A final product size of 397bp indicated successful integration.

PCR1 F: GCCAGACCCTGAATTTGTGT
R: GTTCTTTCCTGCGTTATCCC
PCR2: F: CCCAAAGCGATACCACTTG
R: AAGTTCGTGGACACGAC

4.10.2 Splinkerette PCR

The site of integration was determined by splinkerette PCR (spPCR). This method was initially developed by Potter and Luo [241] to identify unknown genomic DNA sequences located between a known restriction site and a target gene and is commonly used to map the integration site of viral DNA sequences in the mouse genome. The process requires the attachment of unique, double-stranded 'splinkerette' linkers onto the ends of genomic DNA fragments digested with a specific restriction enzyme. The fragment containing the integrated target DNA is then amplified by PCR using a forward primer that anneals to the target gene and a reverse primer specific to the splinkerette linker. Sequencing of the PCR product and subsequent genome mapping allows the site of integration to be identified. A schematic of the process involved and the oligonucleotide sequences of the splinkerette linker and PCR primers used to determine the integration of the pJTI/Zeo platform vector are given in figure 4.16.

SpPCR was performed following 'Splinkerette Protocol S1' published by Potter and Luo [241] and using the BstYI restriction enzyme. This enzyme cuts at either side of the PhiC31 *att*B site in the pJTI/Zeo vector and at regular intervals throughout the human genome. Digestion of genomic DNA with this enzyme produces a four-nucleotide 3' overhang (GATC) which anneals to a compatible 5' overhang on the splinkerette linker. The splinkerette overhang is unphosphorylated; increasing specificity by ensuring only the 3' recessed end on the bottom strand is capable of ligation. The presence of a stable hairpin loop prevents the splinkerette from binding to genomic DNA at anywhere other than the compatible sticky ends and reduces the background produced by non-specific and end-repair priming.

The genomic DNA isolated from each platform cell line (described in section 4.10.1) was purified by ethanol precipitation to remove all traces of the CellsDirect buffers. One microgram of purified DNA was digested with BstYI in a total reaction volume of 35µl, including 3.5µl of NEB buffer 4, 3.5µl of BSA (10x) and 10 units of enzyme. Following overnight incubation at 60°C, the BstYI enzyme was inactivated by heating to 80°C for 20 minutes. The full digestion volume was added to 6µl of double-stranded splinkerette linker, 600 units of T4 DNA ligase (NEB), 5µl of ligase buffer (10x) and 2µl of water. The double-stranded splinkerette linkers were created by heating the two oligonucleotides (figure 4.16B) to 95°C for 3 minutes in the presence of NEB buffer 2 and allowing natural cooling to room temperature. The final ligation reaction volume of 50µl was incubated at room temperature for 3 hours to facilitate the annealing of the splinkerette linkers to the genomic DNA.

The nested PCR reactions were optimised using Phusion Taq polymerase (Finnzymes) and the reaction mixtures are detailed below in table 4.5. The primer sequences are given in figure 4.16C.

Round 1 PCR	Volume (µl)	Round 2 PCR	Volume (µl)
Ligation reaction	10	Round 1 PCR	0.5
Water	1.5	Water	11
Phusion MasterMix (2x)	12.5	Phusion MasterMix (2x)	12.5
p1 (10µM)	0.5	p2 (10µM)	0.5
s1 (10µm)	0.5	s2 (10µm)	0.5
Total	**25**	**Total**	**25**

Table 4.5 The composition of the nested PCR reactions conducted to confirm the presence of the pJTI/Zeo platform vector in the genome of the cell line.

The round 1 reaction was heated to 98°C for 75 seconds, followed by two cycles of 98°C for 20 seconds and 64°C for 15 seconds. A further 30 cycles of 98°C for 20 seconds, 68°C for 15 seconds and 72°C for 2 minutes was followed by a final extension at 72°C for 7 minutes.

The round 2 reaction was heated to 98°C for 75 seconds, followed by 30 cycles of 98°C for 20 seconds, 68°C for 15 seconds and 72°C for 90 seconds. Final extension occurred at 72°C for 7 minutes. A 5µl aliquot of the round 2 PCR product was resolved by agarose electrophoresis to confirm the presence of a single band (Figure 4.16D). The remaining PCR product was purified using the QIAquick PCR purification kit and sequenced using a primer that anneals to the pJTI vector (F: TCCCGTGCTCACCGTGACCAC). The resulting sequence was mapped to the human genome using the Blat tool (UCSC).

Two platform lines were identified from each of the SK-N-F1 and SH-SY5Y parental lines, with details of the integration sites given in table 4.6. Each parental line produced one platform line with an integration site that did not disrupt a gene and one that did. The undisrupted lines were favoured as their sites of integration are less likely to disrupt normal cell functioning; however, the positioning outside of an active area of the genome may cause low minigene expression levels. Comparison between the two platform lines will highlight any minigene expression differences caused by the insertion site.

Figure 4.16 The splinkerette PCR method
A: The process involved in identifying the insertion site of the pJTI/Zeo platform vector into the genome of mammalian cell lines; B: The sequence and structure of the 'splinkerette' linker; C: The two sets of primers used in a nested PCR to confirm the presence of the platform vector in the genome of the cell line; D: An agarose gel image of the nested PCR products. The ringed cell clones contained single insertions of the platform vector. ***Adapted from Potter et al (2010) [241].***

Clone	Cell line	Match (bp)	Chr	Location	Gene	Insertion	Function	Expressed in brain?
F4	SK-N-F1	21	4	45328896	-	-	-	-
F2	SK-N-F1	36	19	44564015	ZNF223 Zinc finger protein 223	Intron 1	Zinc finger protein	No (thymus and ovary)
S20	SH-SY5Y	596	3	65179876	-	-	-	-
S19	SH-SY5Y	119	18	40535992	RIT2 Ras-like without CAAX 2	Intron 2	RAS family of GTPases	Yes

Table 4.6 The Gateway® TI platform cell lines.
Four platform cell lines were identified; two each from the SK-N-F1 and SH-SY5Y parent lines.

4.10.3 Retargeting

The final stage in the creation of the cell models involved the insertion of each minigene into the chosen platform cell lines in a process called 'retargeting' (see section 4.4.). Due to the mis-splicing events, however, the decision was taken to delay the creation of the stable models. Simple alterations to the minigenes at exons 4-9 should correct the mis-splicing events occurring in this region (see

section 4.11 below) and it was therefore considered preferable to make these corrections before integrating the minigenes and creating the stable cell lines. This, unfortunately, falls outside the timeline of this project and therefore the final stage in the creation of the stable cell models was not completed.

4.11 Discussion

This chapter has described the design, assembly and validation of minigenes to study the role of promoter variation in tau alternative splicing. The use of Invitrogen's Gateway® technology provided the flexibility to study the specificity of the *MAPT* promoter – and in particular the effect of the rs242557 polymorphism – on the inclusion rate of exon 10. Six minigenes were created, three representing the genetic variation of the *MAPT* H1B haplotype and three similarly representing H1C. The three minigenes in each haplotype set differed only by their promoter, with expression controlled by either the *MAPT* core promoter alone, the core promoter in conjunction with an allelic variant of the rs242557 regulatory domain or the CMV promoter.

The minigenes were created using a mixture of genomic DNA and cDNA fragments cloned from two PSP patients confirmed as having the H1B and H1C variants of the *MAPT* gene. A series of intricate and varied cloning strategies were used to join together multiple individual elements into complete minigenes of 14.1-15.7kb in size. Following transient transfection in SH-SY5Y and SK-N-F1 cells, *in vitro* expression and alternative splicing of minigene mRNA transcripts was confirmed and found to be consistent between all minigene variants in multiple independent transfections.

Two mis-splicing events, however, were identified and can be traced to the inclusion of a cDNA element in the minigene design. The decision to amplify constitutive exons 4, 5, 7 and 9 from cDNA in a single element reduced the size of the minigene by excluding unnecessary intronic sequences. This ultimately reduced the number of cloning steps required to construct the minigene – reducing the number of opportunities for sequence errors to be introduced – and increased

the efficiency of *in vitro* transfection – thereby increasing the number of cells expressing the minigene. It seems, however, that the complete removal of introns 4, 5, 6, 7 and 8 altered the splicing signals at the intron 3/exon 4 and exon 9/intron 9 boundaries. Exons often contain splicing regulatory elements, called exonic splicing enhancers (ESEs) or repressors (ESRs). Grouping together the four exons may have produced competition between the exonic signals, leading to confusion and failure of the splicing machinery to recognise the exon/intron boundaries. This was certainly not helped by the unavoidable insertion of an *att*B sequence at the intron 3/exon 4 boundary, although the sequence is designed such that the AG exon recognition sequence is reformed at the boundary following recombination. The *att*B insertion may, however, disrupt the intronic polypyrimidine tract or increase its distance from the AG exon recognition sequence, therefore weakening the splicing signal. The signal at the exon 9/intron 9 boundary was similarly weakened by the unavoidable insertion of a restriction site into the exon/intron recognition motif.

The minigene design – particularly the inclusion of cDNA elements – was based on that described by Dawson *et al* in the creation of their transgenic mouse models. Although the authors did not report any mis-splicing issues with their minigene, they did observe extremely low protein expression, quantified at approximately 1-2% of the endogenous murine tau level. This was unexpected and suggests that a mis-splicing event leading to a shift in reading frame may have occurred during mRNA processing, thereby affecting tau protein expression. In addition, the N-terminal exons of their minigene mRNA were not alternatively spliced, with exons 2 and 3 constitutively present. This may be due to the inclusion of shorter intronic segments (approximately 200bp) surrounding these exons, confirming that a minimum length of intronic sequence is required to produce the correct pattern of exon 2 and 3 splicing. Unlike in this study, Dawson and colleagues did not require the presence of a C-terminal FLAG-tag motif to distinguish the minigene human tau from the endogenous murine tau. They therefore had more flexibility in the placement of their PCR primers and did not need to amplify the full-length mRNA in order to quantify the inclusion rate of

exons 2 and 3. Thus the inclusion of these exons in the context of exon 10 splicing was not reported and any mis-splicing events at the exon 4-9 cDNA segment may not have been detected.

A few changes to the minigene should correct the mis-splicing events at exons 4-9 and facilitate tau protein expression. Ideally, the exon 4-9 element would be replaced with four separate elements each comprising one of the exons surrounded by approximately 600bp of upstream and downstream intronic sequence. This would move the restriction sites and *att*B sequence to the intronic junction between two elements, away from the critical exon/intron boundaries. The restriction enzyme-based construction of the minigene, however, makes this correction unfeasible. The sheer number of unique restriction enzymes required – with each not cutting internally in the new inserts but cutting at exactly the right points in the current minigene to remove and replace the exon 4-9 element – is simply too high. Even when going back several cloning steps to insert the new elements into the individual component Fragments 2 and 3 (figures 4.7 and 4.8), this approach is not possible and the resulting increase in overall minigene size would likely reduce transfection efficiency too far. A more feasible approach, however, may be to insert only two of the new elements; the ones containing exon 4 and exon 9. This will keep the size of the final minigene down and ensure the two critical boundaries – the intron 3/exon 4 and exon 9/intron 9 boundaries – are preserved. Although exons 5 and 7 would be entirely excluded, this should not affect the inclusion rate of exons 2, 3 and 10 at both the mRNA and protein level and the size of the minigene should not increase by too much. A schematic of how this correction may be achieved is given in figure 4.17.

The second aspect of this project was the creation of stable cell models by integrating each minigene into the same location in the genome of the SH-SY5Y and SK-N-F1 cell lines. Although two platform cell lines were created per cell type and the sites of integration verified, the decision was taken to delay the creation of the integrated cell models until the mis-splicing events of the

minigenes have been corrected. This will increase the value and versatility of the models for use in future investigations of *MAPT* expression.

Despite the mis-splicing events, the minigene variants still express tau mRNA *in vitro* and demonstrate a highly replicable pattern of splicing at alternative exons 2, 3 and 10. Thus, they are a valid tool for studying the role of the tau promoter and the rs242557 allelic variants on the splicing pattern of these exons. This investigation will be described in detail in Chapter 5.

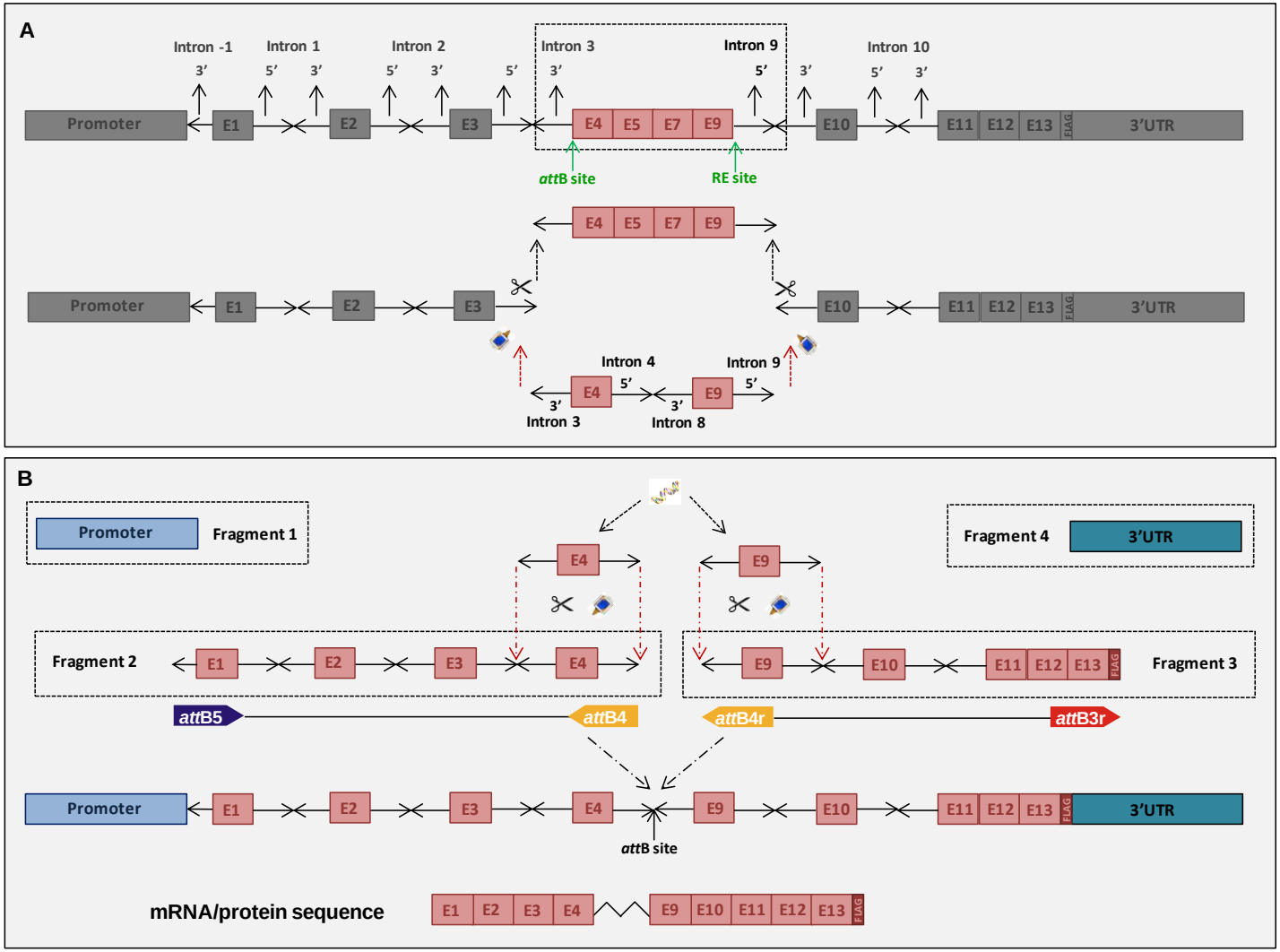


Figure 4.17 Correction of the mis-splicing events at exons 4-9.

A: Removal of the exon 4-9 element would allow the insertion two new elements containing exons 4 and 9 with surrounding intronic segments.

B: The new exon 4 element could be inserted into the 3' end of Fragment 2, with the new exon 9 element inserted into the 5' end of Fragment 3. This would result in the problematic *attB* sequence falling outside of the intron/exon boundary.

5 The role of the tau promoter and rs242557 polymorphism in the alternative splicing of *MAPT* exons 2, 3 and 10

5.1 Overview

It is evident that transcription and alternative splicing are not independent processes and an increasing number of genes have been shown to demonstrate co-regulation [1-5]. The luciferase reporter gene study of the *MAPT* promoter region (chapter 3) confirmed that the rs242557 polymorphism lies within a transcription regulatory domain and that its two alleles differentially alter the strength of this regulation, regardless of the positioning of the domain relative to the core promoter and the endogenous cellular environment. Thus, if *MAPT* transcription and alternative splicing are co-regulated, it would follow that the allelic variants of the rs242557 regulatory domain could also differentially affect the inclusion rate of the alternatively spliced *MAPT* exons.

To test this hypothesis and determine the influence of the promoter and the rs242557 polymorphism on *MAPT* alternative splicing, the six *MAPT* minigenes were expressed *in vitro* in neuroblastoma cells, as described in chapter 4. To recap, two minigenes were created representing the genetic variation of the H1B and H1C *MAPT* haplotypes. Three versions of each minigene were produced, differing only by the promoter element driving expression. The three promoter elements comprised: the *MAPT* H1 core promoter, the CMV promoter and the H1 core promoter in conjunction with the allelic variants of the rs242557-containing regulatory domain. Each minigene was expressed in SK-N-F1 and SH-SY5Y neuroblastoma cells in undifferentiated and neuronally differentiated states. The relative inclusion rates of exon 10 and of exons 2 and 3 in minigene mRNA transcripts were quantified, as was the differential binding of specific transcription and splicing factors to the alleles of the rs242557 polymorphism.

5.2 Background

There are two main models supporting the co-transcriptional regulation of alternative splicing: physical coupling and kinetic coupling (section 1.2.3).

Physical coupling describes the physical interactions between components of the transcription and splicing machineries and in this scenario genetic variation that disrupts the normal functioning of one could directly modulate the functioning of the other. We have shown that the two allelic variants of the rs242557-containing regulatory domain differ in the strength of their effect on transcription from the *MAPT* core promoter. This may result from a change in mRNA conformation that changes the relative proximity of the regulatory domain to the core promoter, restricting or increasing the interaction between the two. Transcriptional changes may also result from an allelic modification to a transcription factor binding site that increases or reduces binding affinity. In either case, these differences would likely change the composition of the transcription complex, which in turn would alter its physical interaction with the splicing machinery.

Kinetic coupling refers to the specific linking of transcription rate to splice site recognition, with a faster rate of transcription increasing the likelihood that a weak alternative splice site is outcompeted for spliceosome assembly by a stronger downstream constitutive splice site. Thus, the rate of alternative exon skipping would increase in this scenario. This model is unlikely to provide the mechanism linking the rs242557-A allele with an increase in PSP risk, as this allele (to the best of our knowledge) confers an increase in *MAPT* transcription rate. The kinetic model would therefore suggest that rs242557-A confers an increase in *MAPT* exon 10 skipping (i.e. an increase in 3R-tau), which is not consistent with the increase in 4R-tau observed in the PSP brain.

Although the allelic differences in rs242557-mediated transcriptional regulation may also affect the inclusion rates of exons 2 and 3, previous evidence suggests that the N-terminal and C-terminal alternative exon splicing events are regulated by different mechanisms (section 1.7.2) [205]. Thus, although we would expect to see differences in the splicing of exons 2 and 3 between the *MAPT* and CMV promoter-driven minigenes, the rs242557-regulatory domain – and its allelic variants – is not expected to contribute to splicing in this region.

5.3 *In vitro* expression of the minigene variants

5.3.1 Overview

The six minigene variants were expressed in SK-N-F1 (F1) and SH-SY5Y (SH) neuroblastoma cells as described in section 4.9. This section also describes the methods of mRNA analysis used to determine the inclusion rate of exon 10 and of exons 2 and 3 in the minigene transcripts. Unless otherwise stated, these methods were used in all comparative analyses of minigene mRNA transcripts described in this chapter.

5.3.2 Neuronal differentiation

The addition of retinoic acid to low-serum cell culture medium suspends cell growth and induces morphological changes including the formation of long neuronal processes (figure 5.1). Most significant, however, are the changes in gene expression that result as a consequence of differentiation, with numerous proteins and neuronal markers upregulated. In particular, tau expression is significantly altered in differentiated cells with all six isoforms expressed, concurrent with a shift from exclusive 3R expression to an approximately equal ratio of 4R:3R transcripts. Thus, the changes in endogenous cellular conditions achieved through differentiation would likely influence minigene transcript processing.

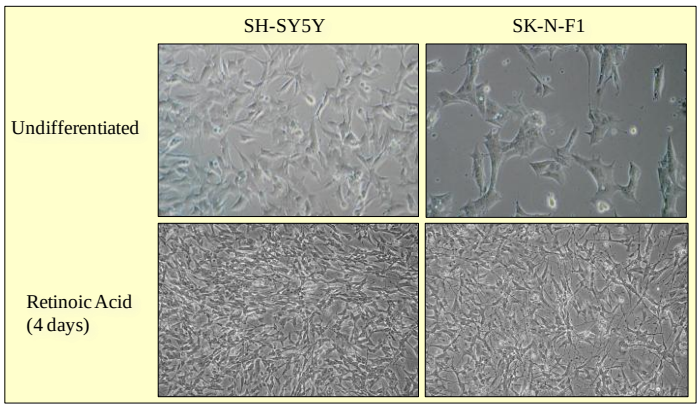


Figure 5.1 SH-SY5Y and SK-N-F1 cells undergo morphological changes after treatment with retinoic acid.

Cells were cultured for 24 hours in 10% cell culture medium (see section 2.2.5) and then rinsed with 1x PBS to remove all traces of the serum. The PBS was then replaced with cell culture medium containing 1% foetal calf serum (1% medium) and supplemented with 10nM of retinoic acid. Medium was changed every 2 days and minigene transfection was conducted after five days of treatment. Cells were allowed to recover from transfection with 10% cell culture medium for 24 hours before retinoic acid treatment was resumed for a further 48 hours.

5.4 Minigene quantification of exon 10 inclusion

5.4.1 mRNA analysis

The six minigene variants were transfected in differentiated and undifferentiated F1 and SH cells as described previously. After 72 hours, total RNA was extracted, reverse transcribed and the rate of exon 10 inclusion determined by PCR with primers annealing to exon 9 and the 3' FLAG-tag motif, again as described previously in chapter 4. Exon 10 inclusion was measured as an internal ratio and defined as the number of exon 10+ transcripts divided by the number of exon 10- transcripts (i.e. the 4R:3R-tau mRNA ratio); values that were determined by the relative intensities of the RT-PCR products when visualised by polyacrylamide gel electrophoresis (PAGE; section 2.1.3.4). Quantification of band intensity was conducted using the ImageJ software (NIH), which calculated the area under the intensity peak produced by each band (figure 5.2C). The 4R:3R ratios conferred by the minigene variants are presented on a bar graph with results averaged from 4-8 independent transfections in 2-4 biological replicates. The error bars represent the standard deviation from the mean and a significant difference in 4R:3R ratio was detected by a Student's t-test and defined as $p \leq 0.05$ (section 2.1.9.1).

5.4.2 Exon 10 splicing in undifferentiated cells is heavily influenced by the *in vitro* cell model

Figure 5.2 presents the 4R:3R-tau ratios of the six minigenes when expressed in undifferentiated F1 cells. Each of the promoter variants (CP, CP+rs242557 and CMV) demonstrated significantly different rates of exon 10 inclusion,

independently of haplotype status, and this indicates the importance of promoter identity in the regulation of exon 10 splicing in this cell line. The CMV promoter variants produced the highest 4R:3R ratios, with four times as many 4R transcripts than 3R transcripts expressed (H1B ratio = 4.05; H1C = 4.21). The CP minigenes produced the lowest ratios but still expressed twice as many 4R transcripts than 3R transcripts (H1B = 2.13; H1C = 2.18). This is surprising as undifferentiated cells generally express only 3R-tau endogenously and even in differentiated cells the ratio of 4R:3R-tau is approximately 1.0. Thus, the preference of the minigenes for 4R expression suggests that additional *cis*-acting factors involved in the regulation of exon 10 splicing are absent from the minigene construct, likely as a consequence of the exclusion of large sections of the *MAPT* promoter region.

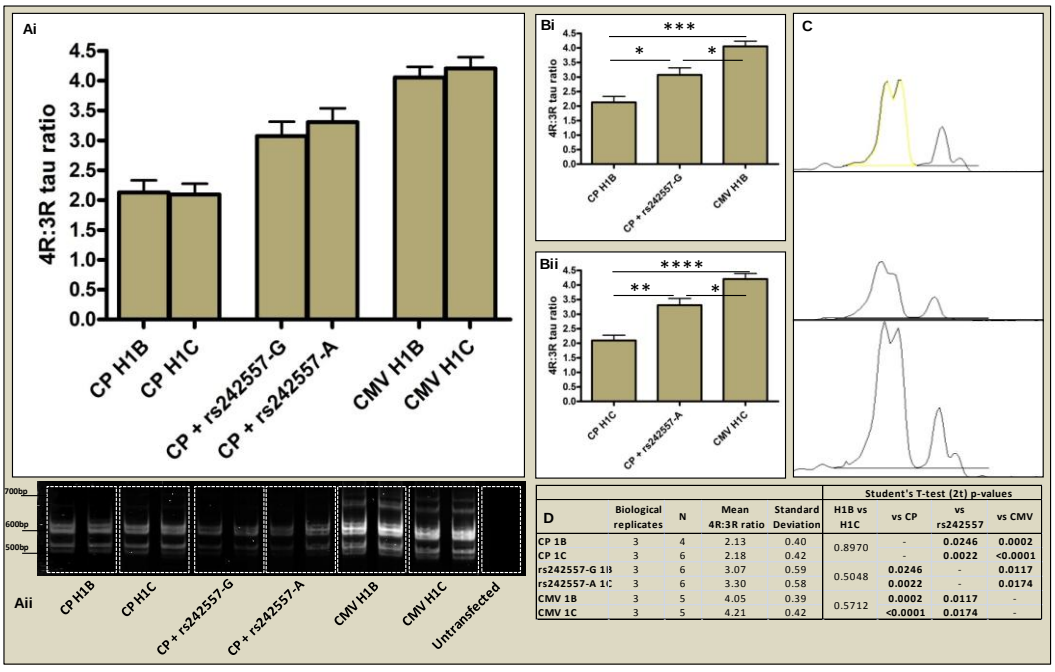
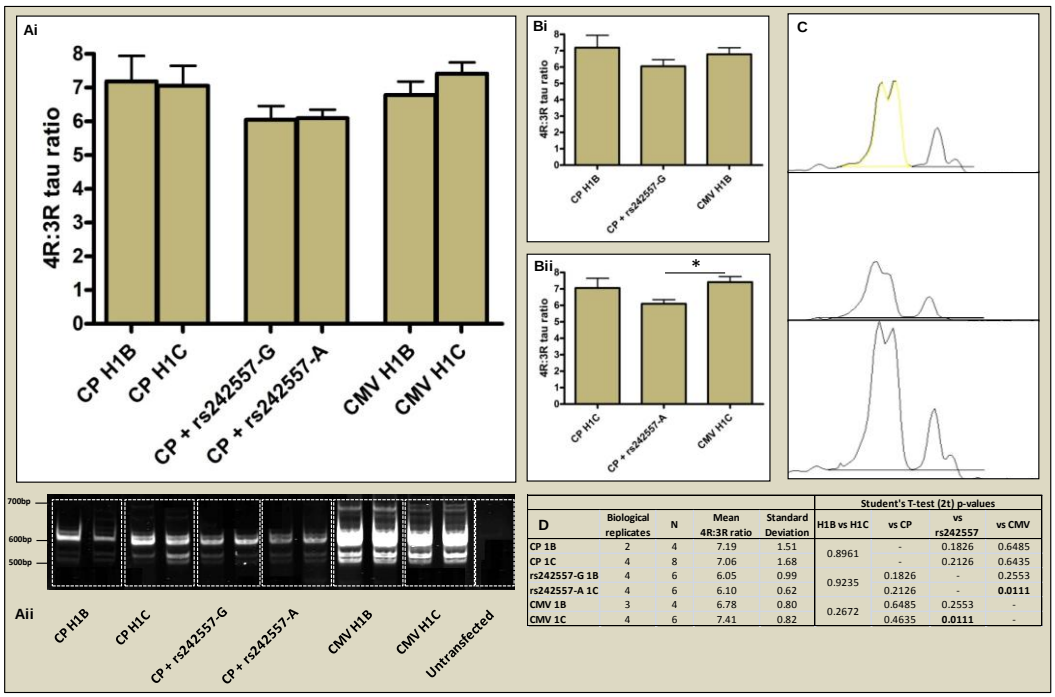


Figure 5.2 The exon 10 quantification of minigene transcripts expressed in undifferentiated SK-N-F1 cells.
Ai: The 4R:3R mRNA ratio produced by each minigene variant; Aii: An example of a polyacrylamide gel image used to quantify the 4R and 3R RT-PCR products; B: The individual results of the H1B (Bi) and H1C (Bii) haplotypes; C: The quantification was achieved by calculating the area under the intensity peak using the ImageJ software (NIH); D: Significant differences in ratio were detected by Student's t-test. * p≤0.05; ** p≤0.01; *** p≤0.001; **** p<0.0001

The CP+rs242557 minigene variants produced three times as many 4R than 3R transcripts (H1B-G = 3.07; H1C-A = 3.30). Although there was no significant difference in 4R:3R ratio between the two rs242557 allelic variants, these constructs produced a significantly higher ratio than their CP counterparts (H1B: p=0.0246; H1C: p=0.0022) and a significantly lower ratio than their CMV counterparts (H1B: p=0.0117; H1C: p=0.0174), as presented in figure 5.2B. This confirms that the rs242557 domain has the ability to regulate exon 10 splicing, acting to increase the proportion of 4R transcripts expressed by the core promoter in this cell line. Figure 5.3 presents the 4R:3R-tau ratios in undifferentiated SH cells and in this instance the picture was very different. In general, the proportion of 4R transcripts produced by each minigene was much higher in this cell line and there was little difference in 4R:3R ratio between either the promoter or haplotype variants.



D	Biological replicates	N	Mean 4R:3R ratio	Standard Deviation	Student's T-test (2t) p-values: H1B vs H1C	vs CP	vs rs242557	vs CMV
CP 1B	2	4	7.19	1.51	0.8961	-	0.1826	0.6485
CP 1C	4	8	7.06	1.68		-	0.2126	0.6435
rs242557-G 1B	4	6	6.05	0.99	0.9235	0.1826	-	0.2553
rs242557-A 1C	4	6	6.10	0.62		0.2126	-	**0.0111**
CMV 1B	3	4	6.78	0.80	0.2672	0.6485	0.2553	-
CMV 1C	4	6	7.41	0.82		0.4635	**0.0111**	-

Figure 5.3 The exon 10 quantification of minigene transcripts expressed in undifferentiated SH-SY5Y cells.
Ai: The 4R:3R mRNA ratio produced by each minigene variant; Aii: An example of a polyacrylamide gel image used to quantify the 4R and 3R RT-PCR products; B: The individual results of the H1B (Bi) and H1C (Bii) haplotypes; C: The quantification was achieved by calculating the area under the intensity peak using the ImageJ software (NIH); D: Significant differences in ratio were detected by Student's t-test. * p≤0.05.

Indeed, only one significant difference in ratio was detected between the CP+rs242557-A and CMV versions of the H1C minigene, with the former variant producing a significantly smaller proportion of 4R transcripts (Figure 5.3 Bii: p=0.011).

An interesting finding, however, concerns the behaviour of the CP variants in the two cell lines. As mentioned earlier, the 4R:3R ratio was notably increased across the board in SH cells compared to F1 cells. For the CP variants, however, this increase was greater (3.2- to 3.4-fold) than for the CP+rs242557 and CMV variants (1.7- to 2.0-fold), which the results of the promoter luciferase study in chapter 3 shows cannot be attributed to differences in CP transcriptional activity in the two cell lines (figure 3.8 in chapter 3). The reason for this is unclear, but may, once again, highlight the important contribution of endogenously expressed *trans*-acting factors in the regulation of tau gene expression.

5.4.3 Neuronal differentiation induces allelic differences in the contribution of rs242557 to splicing regulation of exon 10

The significance of the cellular context in tau isoform expression is exemplified *in vivo* by the well-established changes in exon 10 splicing that take place during development. Foetal tau consists exclusively of the 0N3R isoform, with exon 10 constitutively spliced out. In the adult brain, however, exon 10 skipping is downregulated and the overall expression of 3R- and 4R-tau isoforms is approximately even. To determine whether minigene splicing was similarly affected by developmental changes in the cellular environment, comparative analyses were conducted in F1 and SH cells that had been neuronally differentiated by treatment with retinoic acid (section 5.3.2).

Figure 5.4 presents the 4R:3R-tau ratios produced by each minigene variant when expressed in differentiated F1 cells. The most important finding is the emergence of an allelic difference in 4R:3R ratio between the two CP+rs242557 variants, with the H1C-A variant producing a significantly higher proportion of 4R transcripts than the H1B-G variant (4.88 versus 2.81 (1.7-fold increase); p=0.0103). Comparison of the allelic variants to their CP minigene counterparts

(figure 5.4B) also reveals interesting allelic differences. The addition of the H1B-G allele variant of the rs242557 domain conferred a reduction in the proportion of 4R transcripts produced, thereby significantly reducing the 4R:3R ratio (p=0.0366). The H1C-A allele variant, however, appears to increase the proportion of 4R transcripts produced and thus confers an overall increase in 4R:3R ratio compared to the minigene containing the CP alone (p=0.0295) – concordant with the findings from the analyses in undifferentiated cells.

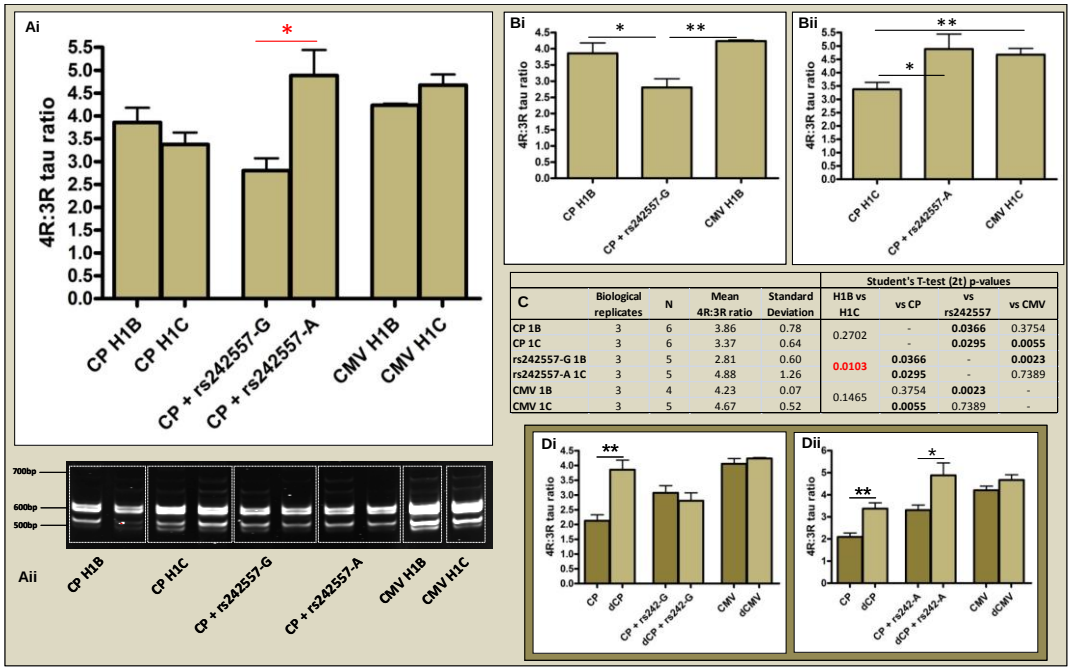


Figure 5.4 The exon 10 quantification of minigene transcripts expressed in differentiated SK-N-F1 cells treated with retinoic acid for 5 days.
Ai: The 4R:3R mRNA ratio produced by each minigene variant; Aii: An example of a polyacrylamide gel image used to quantify the 4R and 3R RT-PCR products; B: The individual results of the H1B (Bi) and H1C (Bii) haplotypes; C: Significant differences in ratio between minigene variants were detected by Student's t-test; D: Comparison of 4R:3R-tau ratios in undifferentiated and differentiated (prefixed with 'd') cells for H1B (Di) and H1C (Dii) haplotype variants. * p≤0.05; ** p≤0.01.

Panel D in figure 5.4 presents individual comparisons of the H1B (Di) and H1C (Dii) promoter variants in undifferentiated versus differentiated F1 cells. As expected, the 4R:3R ratio produced by the CMV variant does not change following differentiation. The CMV promoter is a viral promoter and is therefore unlikely to be as sensitive to changes in the endogenous environment of human cells as the tau promoter variants. Minigene expression driven by the intrinsic

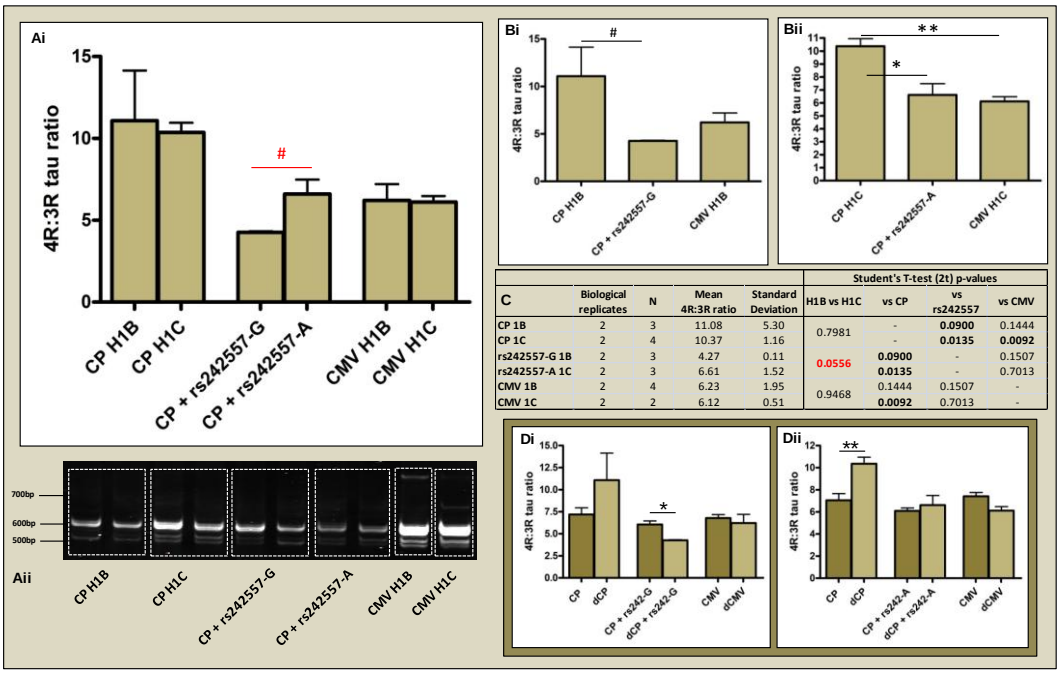
MAPT CP promoter, by contrast, exhibited a significant increase in the proportion of mRNA transcripts containing exon 10; a finding which was also expected as this is consistent with the changes in endogenous tau expression produced following differentiation. This was also the case with the minigene containing the A-allele variant of the rs242557 domain.

The H1B-G minigene, however, did not behave as expected, with the 4R:3R mRNA ratio unaltered following neuronal differentiation. This was surprising, as the PSP-associated H1C-A variant would be expected to be the one to deviate from the general trend. This may be explained by considering the potential function of the rs242557 domain in exon 10 splicing regulation. If the role of the domain is to suppress exon 10 inclusion following the upregulation of tau expression in differentiated neurons, then the H1C-A variant would be the one that does not respond to retinoic acid treatment, failing to regulate the increase in exon 10 inclusion conferred by the CP following differentiation. In section 3.11 it was shown that the A-allele variant of the rs242557 domain conferred significantly weaker repression of transcription from the CP compared to its G-allele counterpart. Together these results suggest an overall loss-of-function for the A-allele variant of the regulatory domain, causing a weakening of its repressive effect on the CP in terms of both transcription rate and exon 10 inclusion.

Figure 5.5 presents comparative 4R:3R-tau ratios produced by the minigenes when expressed in differentiated SH cells. The rs242557 allelic difference in ratio detected in F1 cells was replicated in this second cell line (figure 5.5.Ai), with a 1.5-fold increase in ratio conferred by the H1C-A variant. Although statistical significance was not quite reached, it was extremely close to the threshold (6.61 versus 4.27; p=0.0556).

Another important finding was the emergence of a significant difference in exon 10 splicing between the two tau promoter types following differentiation (figure 5.5Bi and Bii) – a finding that was absent in undifferentiated SH cells. Both allelic

variants of the rs242557 domain acted to reduce the 4R:3R ratio produced by the CP alone, though the H1C-A variant was notably less efficient at this (H1B: 2.6-fold reduction, p=0.0900; H1C: 1.6-fold reduction, p=0.0135).



C	Biological replicates	N	Mean 4R:3R ratio	Standard Deviation	Student's T-test (2t) p-values: H1B vs H1C	vs CP	vs rs242557	vs CMV
CP 1B	2	3	11.08	5.30	0.7981	-	**0.0900**	0.1444
CP 1C	2	4	10.37	1.16		-	**0.0135**	**0.0092**
rs242557-G 1B	2	3	4.27	0.11	0.0556	**0.0900**	-	0.1507
rs242557-A 1C	2	3	6.61	1.52		**0.0135**	-	0.7013
CMV 1B	2	4	6.23	1.95	0.9468	0.1444	0.1507	-
CMV 1C	2	2	6.12	0.51		**0.0092**	0.7013	-

Figure 5.5 The exon 10 quantification of minigene transcripts expressed in differentiated SH-SY5Y cells treated with retinoic acid for 5 days.
Ai: The 4R:3R mRNA ratio produced by each minigene variant; Aii: An example of a polyacrylamide gel image used to quantify the 4R and 3R RT-PCR products; B: The individual results of the H1B (Bi) and H1C (Bii) haplotypes; C: Significant differences in ratio between minigene variants were detected by Student's t-test; D: Comparison of the 4R:3R-tau ratios in undifferentiated and differentiated (prefixed with 'd') cells for the H1B (Di) and H1C (Dii) haplotype variants. # p<0.09; * p≤0.05; ** p≤0.01.

Comparison of minigene expression in undifferentiated and differentiated SH cells (figure 5.5D) again shows that 4R expression is upregulated from the CP minigenes following differentiation, as occurs *in vivo*. For the rs242557 variants, however, the 4R:3R ratio conferred by the H1B-G variant is significantly reduced following differentiation (p=0.0195), while the ratio conferred by the H1C-A variant does not change (p=0.4781). It would, again, appear that the allelic difference in 4R:3R ratio stems from the inability of the H1C-A variant to fully respond to retinoic acid treatment, demonstrating a weaker ability to suppress exon 10 inclusion compared to the H1B-G variant in differentiated cells. The reason for this is unclear, but one hypothesis may be that the A-allele of rs242557

causes a change in the conformation of the mRNA transcript, blocking access of neuronal-specific *trans*-acting protein factors. The weakening or abolition of the binding site of such a factor by the presence of the A-allele may be another potential hypothesis.

5.5 Quantification of minigene exon 2 and exon 3 alternative splicing

5.5.1 Distinguishing the 0N, 1N and 2N isoforms

In section 4.9.2.4, a method was described for analysing the alternative splicing events at exons 2 and 3 in combination with splicing at exon 10. This method proved unsatisfactory, as the mis-splicing out of the exon 4-9 minigene element produced a PCR bias that resulted in the selective amplification of the shorter, mis-spliced transcripts. Correctly spliced transcripts containing the exon 4-9 element – known to be present due to the successful use of an exon 9 primer in the exon 10 quantification – were not detectable using this method. An additional issue with this method was the inability to separate the six tau isoforms due to small size differences (6bp) between some of the transcripts. Thus, although it is desirable to study the alternative splicing events along the full length of the minigene transcripts – as would have been possible by Western blot if minigene-expressed protein analysis was viable – it was instead decided to focus on the splicing of exons 2 and 3 independently of exon 10.

This approach was also not without problems, as the ability to distinguish minigene tau transcripts from those expressed endogenously was reliant on the use of a reverse primer annealing to the FLAG-tag motif at the 3' end of the transcript – downstream to exon 10. To selectively amplify the N-terminal section of correctly spliced transcripts and exclude those missing the exon 4-9 element, a reverse primer annealing to exon 4 was required. To enable this, a nested PCR was performed on the minigene cDNA, with the first round of amplification using primers annealing to exon 1 and the FLAG-tag motif, as described previously in section 4.9.2.4. The only change in conditions was the reduction in the number of PCR cycles from 35 to 29 to reduce the level of background after the second

amplification. A 1µl aliquot of the exon 1-FLAG PCR product was used to set up a second PCR, using the same forward primer in exon 1 (F: CATGCACCAAGACCAAGA) and a reverse primer in exon 4 (R: TCCAATGCCTGCTTCTTC). PCR was conducted during 30 cycles of denaturation, annealing at 55°C and elongation for 1 minute.

To ensure that the nested PCR selectively amplified transcripts expressed from the minigene – particularly in the absence of a FLAG-tag primer in the second PCR – a control PCR was conducted on reverse-transcribed endogenous tau transcripts from untransfected cells. As endogenous tau does not contain a FLAG-tag motif, the first amplification of the nested PCR was conducted with a reverse primer annealing in exon 12 (R: GTCCAGGGACCCAATCTTCGA) at an annealing temperature of 55°C. The rest of the PCR conditions and the second PCR was the same as for the minigene cDNA. In undifferentiated cells, this control PCR produced only one product representing 0N transcripts, as this is the only tau isoform expressed endogenously in this state. The minigene PCR products consistently deviated from this banding pattern, indicating that the nested method was indeed specifically amplifying the minigene transcripts.

All nested PCR products were resolved by agarose gel electrophoresis (section 2.1.3.3) and visualised bands were individually quantified using the ImageJ software. In all instances results were pooled from four independent transfections in four biological replicates.

5.5.2 Promoter identity affects N-terminal splicing in differentiated F1 cells

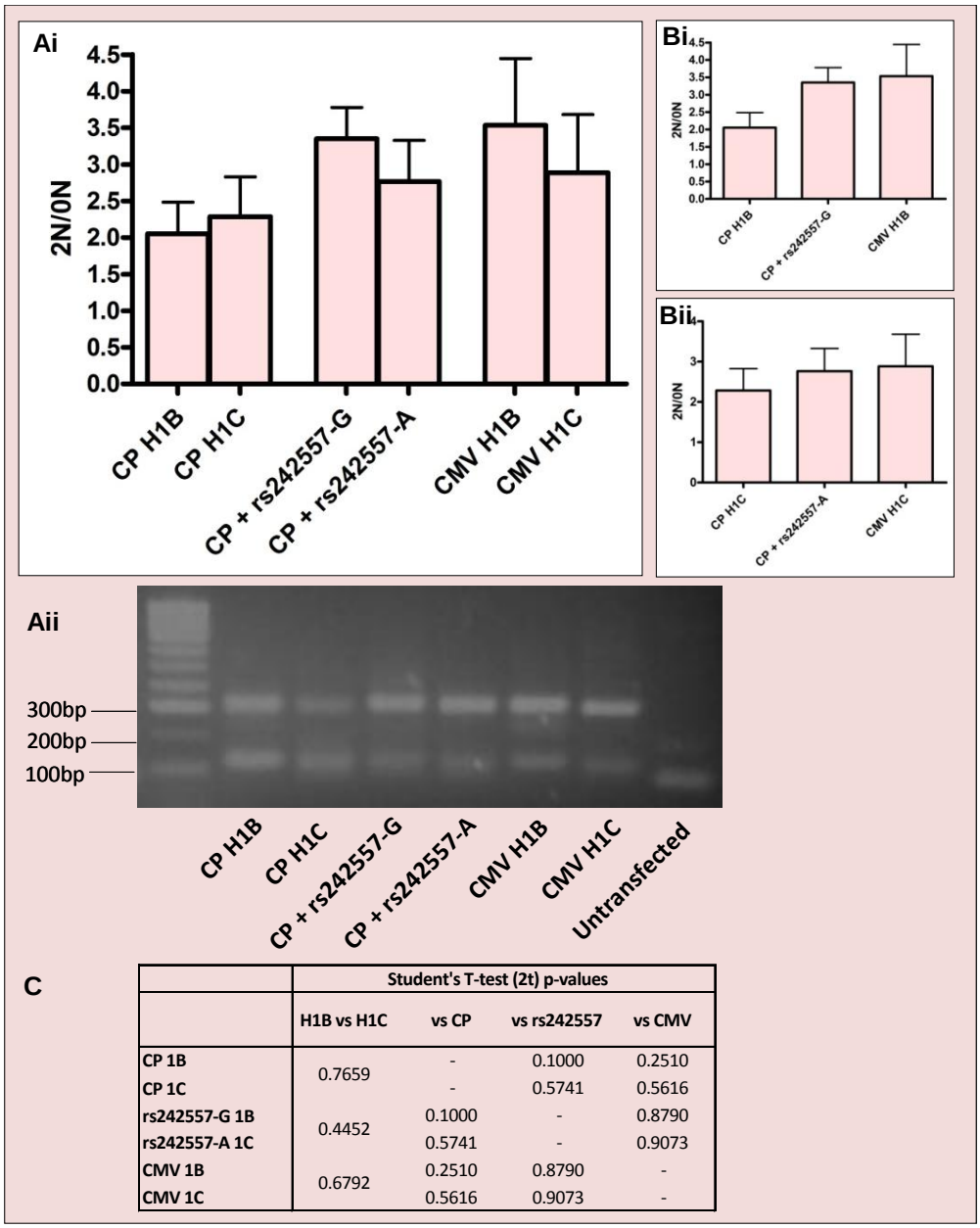
The minigene nested PCR products were slightly larger than expected, as observed by comparison of the 0N isoform band (~100bp) with that produced from the endogenous tau transcripts (~60bp; figure 5.6). As described in section 4.9.2.4, one of the Gateway *att*B sequences used to create the minigene lies within the 3'splice site at the intron 3/exon 4 boundary, and it therefore appears that the *att*B insertion causes the intronic sequences of the splicing motif to shift approximately 40bp upstream to the exon boundary. Thus all nested PCR products

amplified from minigene transcripts were approximately 40bp larger than expected.

The most notable finding from the results of the nested PCR was the preferential expression of 2N mRNA containing both exons 2 and 3 (278bp with *att*B insertion). The 0N mRNA was also detected (ex2-/3-; 104bp); however, although an appropriately-sized 1N band (ex2+/3-; 191bp) was detected, it was of too low abundance to be accurately quantified. As mentioned previously, 2N isoforms are the least abundant in the adult brain and therefore its over-expression here may result from the splicing issues at exon 4, with exon 3 inclusion increasing the likelihood of the transcript being correctly spliced at exons 4-9. Although far from a perfect model, the preference for 2N and 0N expression – and the *att*B insertion – was the same for all six minigene variants. Thus, a preliminary quantification of promoter and genetic influences on the ratio of 2N/0N isoform expression could still be determined. Figures 5.6 and 5.7 present these ratios from minigenes expressed in undifferentiated and differentiated F1 cells, respectively. Error bars represent the standard deviation from the mean and significance was detected using the Student's t-test, as previously.

In undifferentiated F1 cells, there were no significant differences in 2N/0N tau ratio between any of the minigene variants (figure 5.6), suggesting that the rs242557 domain does not play a role in regulating N-terminal splicing events when this cell line is in the undifferentiated state.

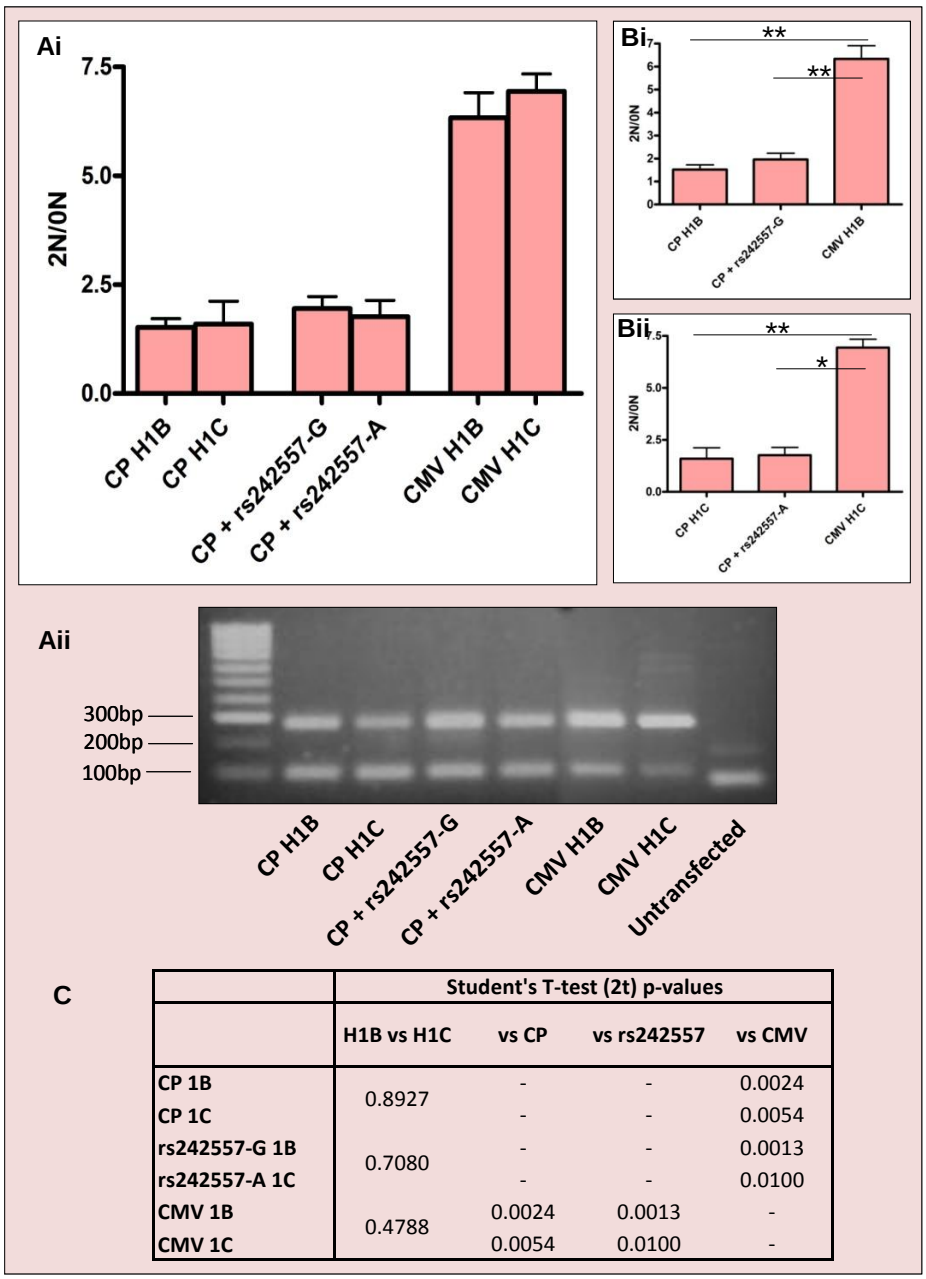
Promoter identity had a greater influence on N-terminal exon splicing when F1 cells were neuronally differentiated by treatment with retinoic acid (figure 5.7), with the *MAPT* promoter variants (CP and CP+rs242557) conferring a significantly lower 2N/0N ratio than their CMV counterparts. There were no differences in 2N/0N ratio between the H1B/H1C haplotype and rs242557 allelic variants, nor between the two *MAPT* promoter types.



C

	Student's T-test (2t) p-values			
	H1B vs H1C	vs CP	vs rs242557	vs CMV
CP 1B	0.7659	-	0.1000	0.2510
CP 1C		-	0.5741	0.5616
rs242557-G 1B	0.4452	0.1000	-	0.8790
rs242557-A 1C		0.5741	-	0.9073
CMV 1B	0.6792	0.2510	0.8790	-
CMV 1C		0.5616	0.9073	-

Figure 5.6 The 2N/0N mRNA ratio of N-terminal tau isoform expression in undifferentiated F1 cells.
A: Comparison of the 2N/0N ratio between H1B and H1C haplotype variants presented i: by bar graph; ii: by an agarose gel image of resolved nested PCR products; B: Comparison of the promoter variants of i: H1B and ii: H1C minigene variants; C: Significant differences in ratio were detected by Student's t-test.



C

	Student's T-test (2t) p-values			
	H1B vs H1C	vs CP	vs rs242557	vs CMV
CP 1B	0.8927	-	-	0.0024
CP 1C		-	-	0.0054
rs242557-G 1B	0.7080	-	-	0.0013
rs242557-A 1C		-	-	0.0100
CMV 1B	0.4788	0.0024	0.0013	-
CMV 1C		0.0054	0.0100	-

Figure 5.7 The 2N/0N mRNA ratio of N-terminal tau isoform expression in differentiated F1 cells.
A: Comparison of the 2N/0N ratio between H1B and H1C haplotype variants presented i: by bar graph; ii: by an agarose gel image of resolved nested PCR products; B: Comparison of the promoter variants of i: H1B and ii: H1C minigene variants; C: Significant differences in ratio were detected by Student's t-test. * p≤0.05; ** p≤0.01

The results of the 2N/0N mRNA quantifications in F1 cells suggest that tau promoter identity is important in the splicing of N-terminal exons – but only when cells are differentiated. As the splicing pattern produced by the minigenes does not resemble that observed *in vivo*, it is difficult to make any biological interpretations relating to the upregulation or downregulation of specific isoforms. It can, however, be determined that the rs242557 regulatory domain does not play a role in N-terminal splicing regulation in this cell line, as the 2N/0N ratio produced by the CP+rs242557 variants did not differ from that of the CP variants in either undifferentiated or differentiated state.

5.5.3 The N-terminal exon splicing events conferred by the minigenes in SH cells

Analysis of the N-terminal exon 2 and 3 splicing events from minigenes expressed in SH cells produced some surprising and intriguing results. In undifferentiated SH cells, differences were detected between the promoter variants, with the highest 2N/0N ratio conferred by the CMV minigene variants and the CP+rs242557 allelic variants conferring the lowest (figure 5.8A). There was also an allelic difference in ratio between the CP+rs242557 H1B-G and H1C-A variants that almost reached statistical significance (p=0.0787), with H1C-A conferring a lower ratio than its G-allele counterpart. There were no differences between the CP (p=0.1326) and CMV (p=0.5055) haplotype variants, suggesting that the rs242557 allelic difference is directly due to the promoter element and not genetic variation elsewhere in the minigene.

There was, however, evidence of a role for promoter identity, with the two tau promoter types conferring significantly lower 2N/0N ratios than their CMV counterparts (figure 5.8B; CP H1B: p=0.0306; CP H1C: p=0.0492; rs242557-A: p=0.0043; vs CMV). There was also a significant difference between the two *MAPT* promoter types, with the addition of the rs242557 domain reducing the expression of 2N tau conferred by the unregulated CP minigene (figure 5.8B; CP vs CP+rs242557; H1B: p=0.0445; H1C: p=0.0500).

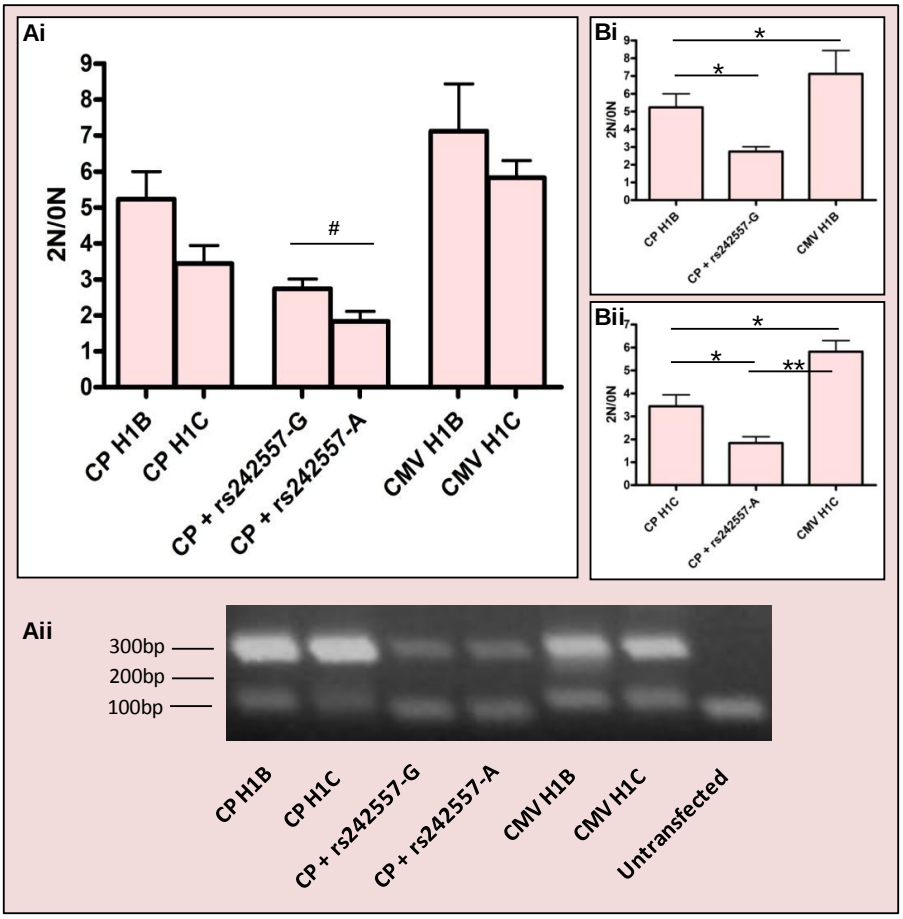


Figure 5.8 The 2N/0N mRNA ratio of N-terminal tau isoform expression in undifferentiated SH cells.
A: Comparison of the 2N/0N ratio between H1B and H1C haplotype variants presented i: by bar graph; ii: by an agarose gel image of resolved nested PCR products; B: Comparison of the promoter variants of i: H1B and ii: H1C minigene variants. # p<0.08; * p≤0.05; ** p≤0.01

In differentiated SH cells a clear haplotype difference was observed between all three promoter variants, with the H1C minigenes, firstly expressing a much lower abundance of transcripts overall, which, secondly, consisted almost exclusively of the 2N mRNA (figure 5.9). The reason for this is unclear, but the H1C-specific shift in N-terminal splicing suggests that the H1C minigene backbone contains an element that influences this ratio in SH cells, regardless of the nature of the promoter element. As 0N mRNA was undetectable for the H1C minigenes, internal ratios were not calculated in this instance. It is fair to say that quantification of the minigene N-terminal exon splicing events was erratic in the

SH cell line and these results are difficult to fathom. Further replications must be undertaken before anything can be read into these contradictory findings.

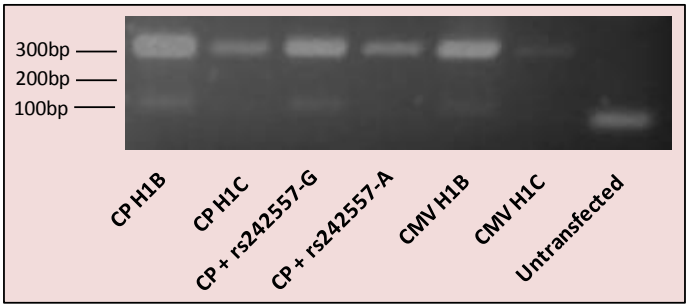


Figure 5.9 The nested PCR products of N-terminal exon 2 and 3 splicing events in differentiated SH cells.

Overall, however, these quantifications have shown that *MAPT* promoter identify plays a role in the regulation of N-terminal exon splicing in F1 cells when in the differentiated – but not undifferentiated – state; perhaps indicating its importance in regulating the significant changes in exon 2 and 3 splicing that take place during development.

5.6 Differential binding of factors to the rs242557 allelic variants

5.6.1 Overview

Chromatin immunoprecipitation (ChIP) provides a method of determining whether specific proteins bind to a known DNA sequence (section 2.1.7.1). It was used here to determine whether the allelic variants of the rs242557 domain differentially bind specific transcription and splicing factors. These factors were chosen for investigation based on experimental evidence produced previously in the laboratory [JF Anaya, PhD thesis, UCL 2012, [242]].

This work included electrophoretic mobility shift assays (EMSAs) and DNA affinity purification, which were used to identify a number of proteins that bind to the rs242557 region and to determine the comparative strength of their binding to the two alleles. The top candidates for differential binding – that is the proteins that demonstrated the largest difference in binding strength to the two alleles – were further analysed *in vitro*. Expression of each candidate protein was individually knocked down by siRNA treatment of SH cells and the effect on expression of the SD-downstream promoter luciferase constructs (described in

section 3.5) was determined. Out of the candidates analysed, knockdown of hnRNP U had the greatest differential effect on the activity of the rs242557 allelic variants, with the H1C-A construct conferring a significant reduction in activity compared to the H1B-G construct following hnRNP U knockdown (figure 5.10).

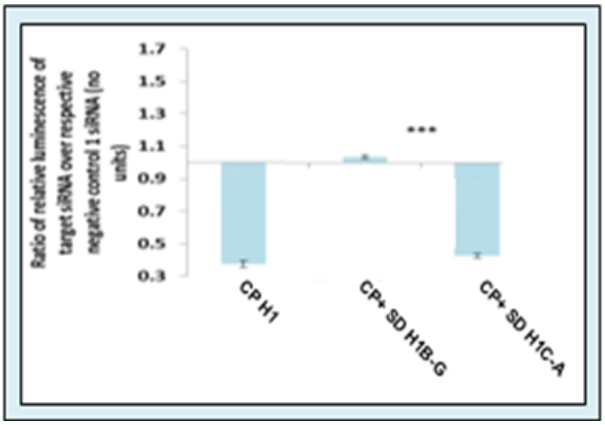


Figure 5.10 The effect of siRNA knockdown on promoter luciferase activity.
The relative luciferase activity of the CP H1 and CP+SD allelic variants in SH-SY5Y cells treated with an siRNA targeted against hnRNP U. Results were normalised against the activity of each construct in cells treated with a control (scrambled) siRNA. *JFAnaya, PhD thesis, UCL 2012*. *** p≤0.001

Heterogeneous ribonucleoproteins (hnRNPs) are a family of RNA-binding proteins that are expressed ubiquitously and have been implicated at all levels of gene expression, including in mRNA splicing, stability, transport and translation. The hnRNP U protein is a component of ribonucleoprotein particles and is thought to have an additional role in the regulation of transcription through its association with histone acetylase and the transcriptional activator CBP/p300 [243]. This is supported by its apparent physical association with the phosphorylated CTD of RNA Pol II (see section 1.2.2.2) [243, 244]. It has also been suggested that hnRNP U forms a complex with β-actin to regulate Pol II-mediated transcription during the initial activation phases. Indeed, antibodies against hnRNP U and β-actin have been shown to block transcription of Pol II-transcribed genes [243]. The association of hnRNP U to the rs242557 domain – a transcription regulatory domain – would therefore further support the hypothesis that *MAPT* transcription and splicing processes are co-regulated.

The allelic minigene variants were not suitable for use in these ChIP experiments, as the minigene rs242557 domain could not be separated from its endogenous

counterpart. Thus, ChIP was performed on endogenous chromatin extracted from untransfected cells of determined genotype.

5.6.2 rs242557 genotyping

The rs242557 genotype status of five human cell lines was determined: SK-N-F1, SH-SY5Y, BE(2)-M17, SK-N-MC and HEK293. This polymorphism is an RFLP (section 2.1.8.1), with the A-allele abolishing a restriction site of the ApaLI enzyme. Thus, the rs242557 genotype of each cell line was determined by the PCR amplification of a 384bp fragment containing the polymorphism followed by digestion with ApaLI.

DNA was extracted from cultured cells as described in section 2.1.5.6. A volume of 3µl formed the template in a PCR with specially-designed primers (F: ACAGAGAAAGCCCCTGTTGG; R: ATGCTGGGAAGCAAAAGAAA). PCR was performed as described previously using the FastStart High Fidelity PCR System and comprised 35 cycles of denaturation, annealing at 60°C and elongation for 1 minute. PCR products were digested overnight at 37°C with 25 units of ApaLI enzyme, NEB4 buffer (1x) and BSA (1x) in a total reaction volume of 50µl. Digestion products were visualised by agarose gel electrophoresis and genotypes were called based on the banding pattern observed. Cell lines homozygous for the A-allele produced one band of 384bp, whereas those homozygous for the G-allele produced two bands of 188 and 196bp that resolved together as one band. Heterozygous cell lines were identified by the appearance of two bands of 384 and ~190bp. The banding patterns produced are presented in figure 5.11.

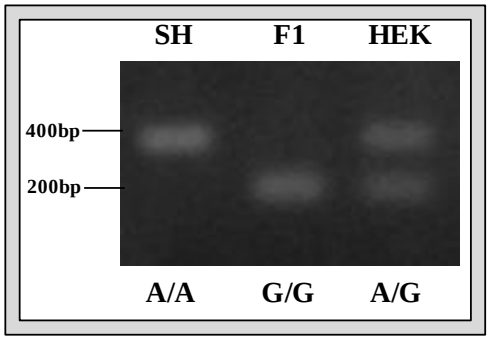


Figure 5.11 Genotyping of the rs242557 polymorphism.
The banding pattern produced by the ApaL1 digestion of rs242557-containing PCR amplicons of 384bp. SH: SH-SY5Y; F1: SK-N-F1; HEK: HEK293.

Three of the cell lines were homozygous for the G-allele, with SH-SY5Y homozygous for the A-allele. Although these genotypes are informative, a heterozygous line (A/G) was desirable as it would have avoided confounding due to general differences between the cell lines. Unfortunately, the only heterozygous line identified was HEK293, which is not a neuronal line and therefore was not deemed suitable in this instance. Thus, the SH-SY5Y (A/A) and SK-N-F1 (G/G) neuroblastoma cell lines were chosen for ChIP as, in addition to carrying opposing rs242557 genotypes, they have also been used throughout this project and therefore results produced here would further inform the transcription and splicing findings described thus far.

5.6.3 Chromatin immunoprecipitation (ChIP)

ChIP was performed using the MAGnify™ System (Invitrogen). Chromatin was extracted from cultured SH and F1 cells at four time points during retinoic acid-induced differentiation: before treatment (day 0) and after 1, 3 and 5 days of treatment. Comparison over this time course provided a more detailed analysis of the changes in protein binding during the different stages of neuronal differentiation.

Endogenous chromatin was fixed, extracted and sonicated into 100-500bp fragments according to the manufacturer's protocol. Each IP was conducted with chromatin from approximately 200,000 cells and samples were diluted accordingly in the presence of protease inhibitors (1x). Three investigative IPs were preformed on each chromatin sample using antibodies against RNA Pol II, hnRNP U and β-actin. There are various commercially-available antibodies that react against Pol II in specific phosphorylation states. Although hnRNP U is believed to bind to phosphorylated Pol II, the evidence for this is not absolute and it is not clear which residues of the Pol II CTD must be phosphorylated to enable hnRNP U binding. Thus, a phospho-independent Pol II antibody was used in these ChIP experiments to detect all Pol II molecules bound to the rs242557 domain, regardless of phosphorylation state. A final IP using the mouse IgG antibody was included as a negative control as this should not react against human proteins.

Each IP was conducted with 1µg of antibody, with the exception of the β-actin IP, for which 2µg was used. The IP, reverse cross-linking and DNA purification steps were conducted according to the manufacturer's protocol. For each IP, a positive 'input' control was included, in which the chromatin sample was subject to the same reverse cross-linking and DNA purification steps but did not undergo the antibody IP.

The binding of the three factors to the rs242557 domain was determined by PCR using 3µl of each purified IP product. The primers and PCR conditions used to genotype the rs242557 polymorphism in the cell lines were also used here and are described in section 5.6.2. A product of size 384bp indicated the binding of the factor to the rs242557 domain. A second PCR was conducted on each IP product using primers annealing to the GAPDH gene (AT=60°C; F: TACTAGCGGTTTTACGGGCG; R: TCGAACAGGAGGAGCAGAGAGCGA)

GAPDH is a highly active housekeeping gene that is constitutively expressed in most cell and tissue types [245]. Thus, this PCR acted as a measure of the efficiency of the IP, as transcription-associated proteins should bind to this gene regardless of differentiation state or cell type. For each PCR, two positive controls were included consisting of genomic DNA extracted from untransfected F1 and SH cells. All PCR products were resolved by agarose gel electrophoresis using a 2% gel. The results of the rs242557 and GAPDH PCRs are presented in figures 5.12 and 5.13 respectively.

5.6.4 RNA Pol II and hnRNP U associate with the rs242557 domain in a manner dependent on differentiation state

The ChIP results reveal that RNA Pol II and hnRNP U are detectable at the rs242557 domain (figure 5.12). This is a significant finding as the association of these two factors at the PSP-associated rs242557 domain has not been investigated before. If hnRNP U only binds to phosphorylated – and therefore elongating – Pol II, these findings may shed light on both the mechanism by

which the rs242557 domain influences transcription and the observed allelic differences in this influence.

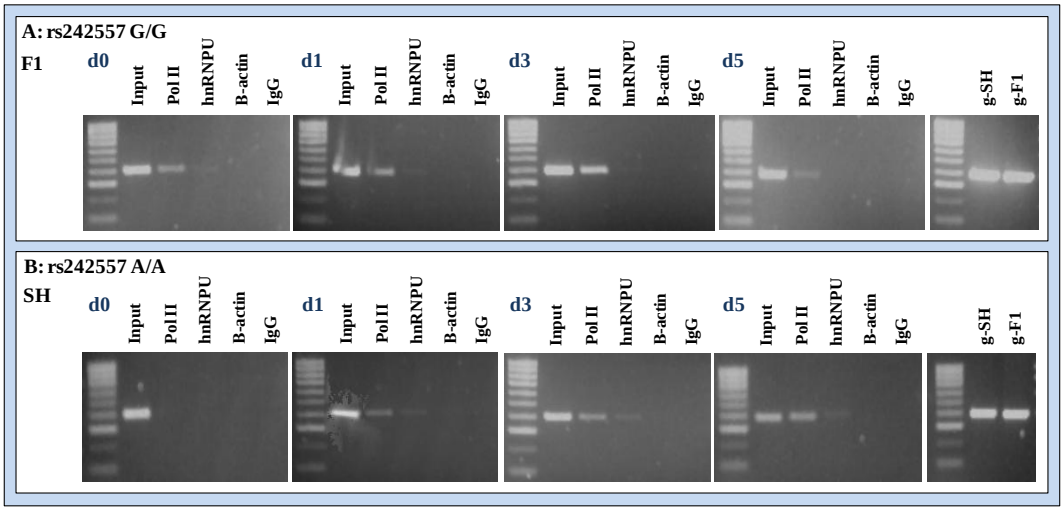


Figure 5.12 ChIP results of rs242557 binding.
ChIP was performed on cultured F1 (A) and SH (B) cells harvested following 0, 1, 3 and 5 days of retinoic acid treatment. Four IPs were performed on each chromatin sample using antibodies against: RNA Pol II, hnRNP U, β-actin and mouse IgG. The positive input control completed the ChIP protocol but skipped the IP stage. Two PCR controls comprised genomic DNA from SH (g-SH) and F1 (g-F1) cells.

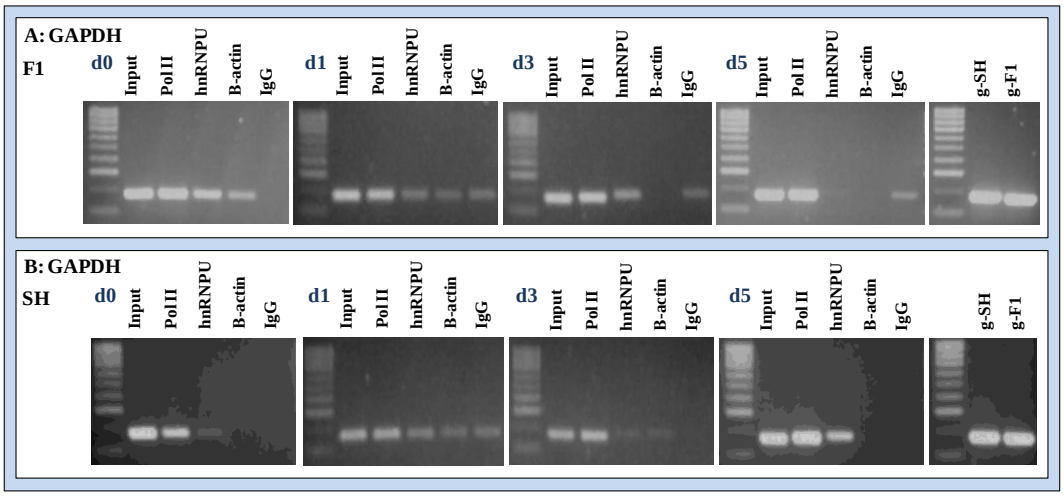


Figure 5.13 Comparative ChIP results of GAPDH binding.
ChIP was performed on cultured F1 (A) and SH (B) cells harvested following 0, 1, 3 and 5 days of retinoic acid treatment.

The rs242557 ChIP results from the F1 cells (carrying the G/G genotype) show that Pol II is associated at all points of the differentiation process, increasing steadily from day 0 to a maximum concentration at day 3 (figure 5.12A).

Increasing Pol II binding is concurrent with a decrease in hnRNP U binding, leading to almost complete dissociation of the hnRNP U factor by day 3. Thus, if hnRNP U is an indicator of active transcription, this would suggest that transcriptional pausing and Pol II accumulation is enhanced at the rs242557 domain during the later stages of differentiation. It may be, therefore, that the rs242557 domain represses *MAPT* transcription by inducing Pol II pausing.

In SH cells (carrying the A/A genotype) the pattern of Pol II and hnRNP U binding is slightly different (figure 5.12B). Neither factor was associated with the region at day 0, suggesting transcriptional pausing does not occur at this region in undifferentiated cells. This may be due to low basal tau levels when cells are in their undifferentiated form, with Pol II resident at the CP. Following induction of tau expression during differentiation, transcription complexes may proceed to the rs242557 domain. Indeed, as with F1 cells, Pol II binding at the rs242557 domain increased over time but in SH cells, the association of hnRNPU remained, albeit at lower relative levels. Thus, it would appear that the weaker transcriptional repression conferred by the A-allele of rs242557 (figure 3.9) results from a reduced ability to induce transcriptional pausing and therefore higher levels of elongating Pol II are detected at this allelic variant.

β-actin does not appear to be associated with either of the rs242557 allelic variants, Pol II or hnRNP U. It must be emphasised, however, that these results are preliminary and technical difficulties have currently prevented independent replication. In particular the β-actin IP needs further optimisation as the hit-and-miss GAPDH results (figure 5.13) suggest the lack of product from the rs242557 PCR may be a consequence of failure of the IP.

5.6.5 GAPDH binding confirms Pol II association with hnRNP U

Pol II and β-actin were highly associated with GAPDH in both cell lines, though hnRNP U binding decreased over the course of the differentiation process (figure 5.13). This appears to confirm the previous reports that hnRNP U associates with elongating Pol II. Accumulation of Pol II indicates transcriptional pausing and the

results presented here describe an inverse relationship between Pol II accumulation and hnRNP U binding, suggesting the latter factor has a low affinity, if any, for inactive Pol II. Reactivity with anti-mouse IgG was detected in four of the GAPDH PCRs. This could indicate contamination; however, sporadic positive results from this IP have been widely reported and are generally considered technical artefacts.

5.7 The ability of the rs242557 regulatory domain to initiate transcription in undifferentiated and neuronally differentiated cells

Following the finding that Pol II accumulates at the rs242557 domain, it was important to determine whether transcription could originate from the rs242557 domain itself. In other words, does the rs242557 domain contain sequences that are capable of initiating transcription independently of the core promoter? To answer this question, two pGL4.10 luciferase constructs were created, each containing one of the allelic variants of the rs242557 domain (the 'SD' H1B and H1C elements from chapter 3). The constructs were created using the NheI- and EcoRV-flanked SD elements and the cloning techniques described in chapter 3. Constructs were transfected into F1 and SH cells and luciferase activity was quantified as described previously.

As the ChIP results revealed that the association of Pol II and hnRNP U to the rs242557 domain changes during the course of differentiation, the SD constructs were assayed in both undifferentiated and differentiated (5 days) cells to see if transcriptional activity – if any – changed with the differentiation state. The CP H1 construct described in section 3.5.5 was included for comparison. Each assay was conducted in triplicate and the results are presented in figure 5.14. The error bars represent the standard deviation from the mean. Significant differences in activity between the two differentiation states were detected by Student's t-test, as previously.

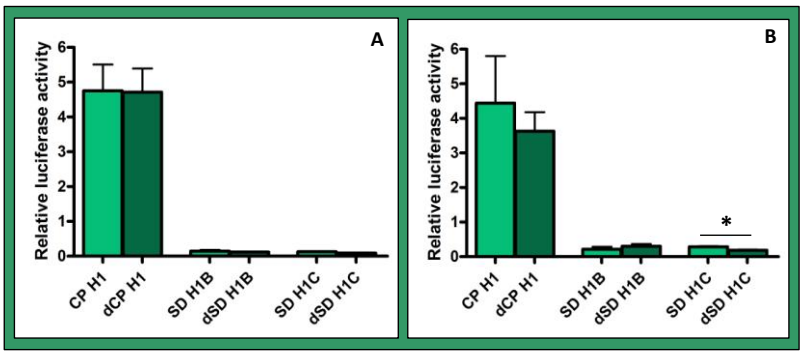


Figure 5.14 Promoter luciferase results 7
The relative luciferase activity conferred by the H1 core promoter (CP) and the allelic variants of the SD in undifferentiated and neuronally differentiated (dark green; prefixed with 'd') F1 (A) and SH (B) cells. * p≤0.05

The rs242557 domain did not confer transcription in F1 cells in either differentiation state (figure 5.14A), suggesting the active Pol II associated with this domain in the ChIP experiments originated from upstream regions of the *MAPT* gene and not the domain itself. Activity levels did appear slightly higher in SH cells (figure 5.15B), which may suggest this region is capable of initiating a low level of transcriptional activity, but only in certain cellular conditions. The activity of the H1 core promoter did not increase following differentiation of either cell line, presumably due to the over-expression of the construct required by the luciferase technique.

As a final investigation, the CP+rs242557-A and CP+rs242557-G luciferase constructs – previously assayed in undifferentiated cells in chapter 3 – were assayed again, this time in F1 and SH cells in both undifferentiated and differentiated states. The results were intriguing and are presented in figure 5.15. In F1 cells, differentiation did not significantly alter the activity of either of the rs242557 allelic variants; again presumably due to construct over-expression potentially masking subtle differences. In SH cells, however, the activity of the CP+rs242557-A variant, significantly reduced following differentiation ($p=0.0044$), suggesting the difference in ability of the allelic variants to modulate transcription rate, though still present ($p=0.0854$), is muted in differentiated cells. It must be noted, however, that these analyses were conducted on results gained

from one single experiment, with at least two further biological replicates needed before firm conclusions can be drawn.

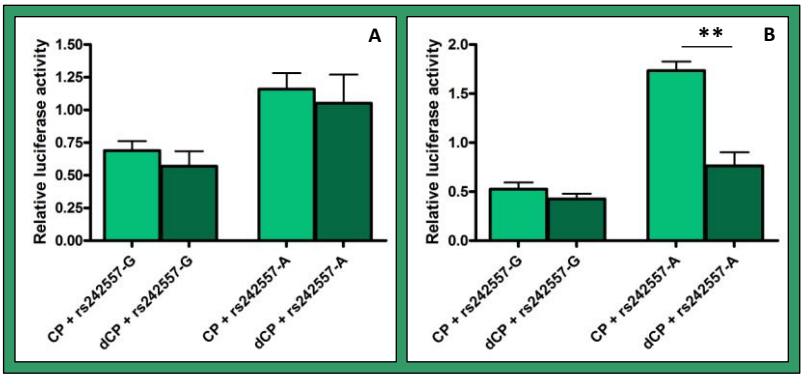


Figure 5.15 Promoter luciferase results 8
The relative luciferase activity of the allelic variants of the CP+rs242557 construct in undifferentiated and neuronally differentiated (dark green; prefixed with 'd') F1 (A) and SH (B) cells. ** p≤0.01

5.8 Discussion

This chapter has described an investigation into the role of the rs242557 polymorphism in the regulation of *MAPT* alternative splicing. The *MAPT* minigene variants were expressed *in vitro* in order to quantify the rate of exon 2, 3 and 10 inclusion in mRNA transcripts produced from different promoter and haplotype variants.

The rs242557 polymorphism was not found to play a role in the N-terminal exon splicing events in F1 cells. In undifferentiated cells, the ratio of 2N and 0N expression did not differ significantly between the six minigene variants. In differentiated F1 cells, however, promoter specificity did play a role, with a reduction in the proportion of 2N isoforms expressed from the tau promoter-driven minigenes compared to the CMV minigenes. The addition of the rs242557 domain to the CP did not affect 2N/0N ratio and this was in concordance with previously reported evidence suggesting N-terminal splicing events are regulated by a different mechanism to that at exon 10 [133, 205]. This picture was complicated in SH cells, where significant differences in 2N/0N ratio between the three promoter variants, and between the two rs242557 variants were detected. It is difficult to take any biological interpretations from these results and further

replications are required before conclusions can be drawn. If these cell line differences prove to be real, it would indicate that cell-specific *trans*-acting factors have a significant influence on the inclusion rate of exons 2 and 3. Recent evidence has assigned a protective role for exon 3 against neurodegeneration and therefore these findings may help to explain why certain subgroups of neurons are vulnerable to tau aggregation, where others are not.

Despite the over-expression of transcripts containing exon 10, the quantification of the 4R:3R ratio from minigene mRNA revealed that the rs242557 polymorphism plays a role in exon 10 splicing in differentiated cells, with the rs242557-A allele conferring increased exon 10 inclusion in both cell lines. This was accompanied by the ChIP results, which revealed an association of the hnRNP U factor with the A-allele domain in differentiated SH cells that was absent with the G-allele domain of differentiated F1 cells. When combined, these results provide some intriguing insights into the potential mechanism behind the association of rs242557 with the over-expression of 4R-tau in PSP. Putting all of the evidence together, the role of the rs242557 domain in exon 10 splicing may be characterised as follows:

1. Early in the differentiation process, as *MAPT* transcription is upregulated, the abundance of phosphorylated Pol II molecules increases
2. hnRNP U associates with elongating phospho-Pol II and recruits/interacts with the splicing machinery to promote exon 10 inclusion.
3. As differentiation progresses, the rs242557 domain induces increased Pol II pausing, leading to a decrease in transcription rate and Pol II accumulation.
4. hnRNP U dissociates from accumulated, and therefore paused, Pol II, altering the recruitment/interaction with the spliceosome and leading to a decrease in exon 10 inclusion.

The evidence presented by the ChIP results suggests that the A-allele of the rs242557 domain is inefficient at inducing Pol II pausing during differentiation,

with a proportion of the Pol II molecules remaining active and associated to hnRNP U, promoting higher levels of exon 10 inclusion. This may therefore link the increase in transcriptional activity conferred by the CP+rs242557-A luciferase construct with the relative increase in exon 10 inclusion produced by the CP+rs242557-A minigene. These results, along with the other evidence reported in this project, will be discussed in more detail in chapter 6.

6 Discussion

6.1 Summary of results

The role of common haplotype variation in *MAPT* expression was investigated at multiple levels: firstly, by quantifying the level of transcription conferred by *MAPT* promoter elements containing the genetic variation of the H1B, H1C and H2 haplotypes; secondly, by determining the rate of alternative exon inclusion conferred by minigenes representing the H1B and H1C haplotypes; and thirdly, by identifying protein factors that differentially bind to the key polymorphism that differentiates H1B from H1C.

The project began with three luciferase reporter gene studies designed to investigate the effect of genetic variation within the 5' and 3' UTRs on gene expression. Particular emphasis was placed on the role of a highly conserved distal domain containing the rs242557 polymorphism that has been strongly associated with PSP. The allelic variants of this domain were repeatedly shown, using reporter gene vectors in cell culture, to differentially alter the transcriptional activity of the *MAPT* core promoter, regardless of their relative positioning and the *in vitro* cell model in which they were expressed. The nature of the effect did, however, change depending upon whether the rs242557 domain was placed upstream or downstream to the core promoter, with the former position resulting in a significantly increased transcription rate compared to the latter and an inversion in the direction of the allelic effect on domain function. Thus, the genomic organisation of the two elements is likely to be a key factor in the functioning of this domain *in vivo*.

This was further exemplified by the comparative luciferase activity of the H1 mutant constructs containing a single nucleotide error inserted into the transcription start site (exon 0) of the core promoter element. The results provided evidence of a physical interaction between the two elements by demonstrating that a single alteration to the conserved sequence of one of the elements affects the functioning of the other.

The second luciferase study investigated the function of a bi-directional promoter located immediately downstream to the *MAPT* core promoter element. The results confirmed that this promoter was capable of initiating transcription in both the forward (sense) and reverse (antisense) directions and revealed that this transcriptional activity was up to 2.5-fold higher in the sense direction than the antisense direction. The transcriptional activities of the H1 haplotype variants did not differ significantly from one another in either direction and the fold-difference between sense and antisense transcription was consistent in both cell lines. The H2 variant, however, behaved differently in the two cell lines, with sense transcription equalling antisense transcription in SH cells.

The study of an extended promoter fragment, containing both the core promoter and bi-directional promoter in their natural genomic orientation, revealed that the addition of the bi-directional promoter modifies the transcriptional activity of the core promoter. Expression in F1 and SH cell lines produced opposing results, with the bi-directional promoter conferring increased activity in F1 and decreased activity in SH cells compared to the CP alone. The reason for this significant difference, however, could not be determined using the luciferase reporter gene assay.

During the course of the investigation into the bi-directional promoter, a polymorphism was identified that appeared to have a subtle but consistent allelic effect on activity. The C-allele – which was found to tag the H1B haplotype – conferred marginally increased sense and marginally decreased antisense transcription compared to the T-allele of the H1C haplotype. The C/C genotype of this polymorphism – denoted rs3744457 – was found to be slightly over-represented in PSP patients compared to control individuals but this did not quite reach statistical significance in these small cohorts.

The final luciferase study identified the region of the *MAPT* 3'UTR that was most influenced by genetic variation. The insertion of the full-length *MAPT* 3'UTR downstream to a CMV promoter-driven luciferase gene conferred a significant

increase in luciferase expression compared to the empty luciferase vector, presumably due to increased stability of the luciferase transcripts. Comparison of the H1B, H1C and H2 variants of the full-length 3'UTR did not reveal genetic differences in the regulation of luciferase expression. When split into three overlapping fragments, however, genetic variation in the 2kb section at the 3' end of the 3'UTR was shown to differentially affect luciferase expression, with the H1C variant conferring significantly increased expression compared to its H1B and H2 counterparts.

Focus then shifted to studying the effect of allelic promoter variation on *MAPT* alternative splicing. This was achieved through the construction of *MAPT* minigenes containing all of the exons, intronic segments and regulatory elements required to produce the six tau isoforms expressed in the human adult brain. Although the minigenes were far from perfect – exhibiting aberrant splicing at two important sites – and due to time constraints the planned isogenic cellular models could not be completed, these minigenes proved a valuable tool for determining the role of promoter identity and common genetic variation on the regulation of *MAPT* alternative splicing events.

In neuronally differentiated cells, the quantification of exon 10 splicing events revealed, firstly, that promoter identity plays a role in exon 10 splicing, with the two *MAPT* promoter variants (the core promoter (CP) alone and the core promoter in conjunction with the rs242557 domain) conferring significantly reduced exon 10 inclusion (4R-tau) compared to their CMV promoter-driven counterparts. An allelic difference between the two rs242557 domain variants was also observed, with the H1C minigene containing the A-allele variant producing a significantly higher proportion of 4R (exon 10+) transcripts than the H1B minigene containing the G-allele variant. The absence of differences between the haplotype variants of the CP and CMV minigenes confirmed that this difference in exon 10 inclusion was driven by the rs242557 polymorphism. Thus, the polymorphism that drives the strong association of the H1C haplotype with increased PSP risk – and has been shown to differentially modulate *MAPT* transcription rate – was here shown

to additionally drive a change in *MAPT* alternative splicing towards preferential 4R-tau expression following neuronal differentiation.

Quantification of the splicing events at the N-terminus of the minigene transcripts was less reliable, but did reveal a potential influence of promoter identity in the regulation of exon 2 and 3 inclusion. For unknown reasons, the minigenes produced a bias towards 2N tau (ex2+/ex3+) expression, with 1N tau (ex2+/ex3-) virtually undetectable. Biological interpretations relating to individual N-terminal isoforms could not, therefore, be made; however, the effect of promoter variation on overall N-terminal splicing – here defined as the 2N/0N ratio – could be determined and significant differences between *MAPT* promoter-driven minigenes and CMV promoter-driven minigenes were detected in differentiated F1 cells. This was not accompanied by rs242557-mediated allelic differences, which suggests that N-terminal and C-terminal splicing events are regulated by different mechanisms.

The final investigation of this project identified differential binding of two major protein factors to the alleles of the rs242557 polymorphism. Chromatin immunoprecipitation (ChIP) was performed on extracts from cells homozygous for the A-allele (SH) and G-allele (F1) of rs242557 and reactivity of antibodies against phospho-independent RNA Pol II and RNA-binding factor hnRNP U was determined at four stages of neuronal differentiation. These results revealed that RNA Pol II accumulates at the rs242557 domain in increasing concentration as differentiation progresses, with hnRNP U found to exhibit an inverse relationship with Pol II concentration. Pol II accumulation at the A-allele variant of the rs242557 domain appeared to be lower than at the G-allele variant, with a corresponding increase in hnRNP U reactivity.

6.2 General discussion

6.2.1 The role of antisense transcription in *MAPT* gene expression

The luciferase reporter gene assay was used to confirm the presence of a bi-directional promoter located downstream and proximal to the *MAPT* core promoter. In general, the level of expression conferred from this promoter in the sense direction was significantly higher than that in the antisense direction in both cell lines. Overall, however, it exhibited significantly higher activity in F1 cells than in SH cells, as demonstrated by comparing the sense transcription conferred by the bi-directional promoter alone with that from the *MAPT* core promoter alone. As described previously, there is no difference in core promoter activity in the two cell lines; however, for the bi-directional promoter the level of activity in F1 cells was equal to 71% of the activity of the core promoter, significantly higher than the 14% observed in SH cells. This cellular difference in relative promoter strength is likely to be behind the difference in the bi-directional transcription-mediated regulation of core promoter expression observed in the two cell lines.

The hypothesis that a low level of activity from the bi-directional promoter blocks elongating transcription complexes from the core promoter is an intriguing one and preliminary evidence from the ChIP experiments described in section 5.6.3 may support this theory. Figure 6.1 presents a PCR analysis of the bi-directional promoter region (~150bp) using the DNA products from the Pol II, hnRNP U and β-actin IPs conducted in chapter 5. Both Pol II and hnRNP U were detected at the bi-directional promoter in F1 cells, again demonstrating an inverse relationship. This indicates that Pol II accumulation does occur at this secondary promoter and may therefore cause the transcriptional arrest of Pol II complexes originating from the core promoter. The accumulation of Pol II is fairly low, however, in F1 cells and the additional activity of the highly active secondary sense promoter presumably masked this transcriptional arrest when quantification was conducted using the luciferase assay, leading to the relative increase in luciferase expression observed in F1 cells. Due to technical issues stemming from the difficulty in amplifying this GC-rich region by PCR, comparative results from SH cells are not

yet available; Pol II accumulation would, however, be expected to be much higher in this cell line to account for the repressive effect on CP expression observed in the luciferase study.

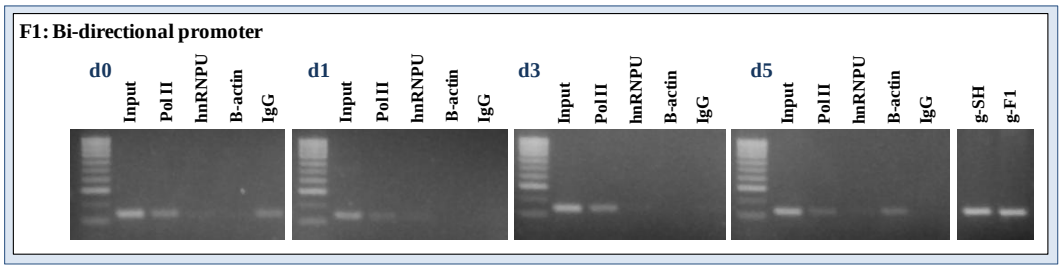


Figure 6.1 ChIP results of the binding of RNA Pol II, hnRNP U, β-actin and mouse IgG to the bi-directional promoter.
ChIP was performed on cultured F1 cells harvested following 0, 1, 3 and 5 days of retinoic acid treatment. The positive input control completed the ChIP protocol but skipped the IP stage. Two PCR controls used SH and F1 genomic DNA (g-SH/g-F1)

The presence of a bi-directional promoter – transcribing non-coding transcripts in both the sense and antisense directions – immediately downstream to the *MAPT* core promoter element indicates that the regulation of expression of this gene is very complex. The methods used in this project did not allow a detailed investigation into the effect of either of the non-coding transcripts on core promoter expression. There are, however, well-characterised examples of antisense-mediated transcriptional regulation that allow us to speculate on how antisense transcription may play a role in *MAPT* expression.

There are three different kinds of naturally occurring sense-antisense transcript pairs: a head-to-head model in which the 5' ends of the transcripts overlap, a tail-to-tail model in which the 3' ends of the transcripts overlap and a complete overlap model in which one gene is completely overlapped by the other (figure 6.2) [228]. The *MAPT* promoter region demonstrates a head-to-head organisation (figure 6.2A), with the antisense promoter lying approximately 1kb downstream to the core promoter. The picture is, however, slightly more complicated than this due to the bi-directional nature of the antisense promoter and further regulation by the sense-transcribed non-coding transcript is likely to play an additional – but as yet unknown – role in the regulation of transcription in either direction.

Figure 6.2 The three kinds of antisense pairs A: head-to-head; B: tail-to-tail; C: completely overlapping. *Taken unchanged from Lapidot et al (2006) [228].*

In humans, 15% of protein-coding genes have an associated antisense transcript and approximately 22% of all transcripts genome-wide are involved in transcriptional overlap – which is significantly higher than observed in species such as rat (5%), chicken (5%) and nematode (0.5%) – and this suggests that these transcripts contribute to the great complexity of human gene expression [246]. The main models proposed for this role were touched upon briefly in section 3.7 and include: transcriptional interference and Pol II collision, duplex formation leading to RNA editing or RNAi-mediated degradation, and chromatin re-modelling to induce gene silencing [228]. The most interesting model, however, that may provide a mechanism for the antisense-mediated regulation of *MAPT* expression involves RNA masking.

RNA masking describes a scenario in which the overlapping sections of the two transcripts form a double-stranded duplex that masks a *cis*-acting regulatory sequence located on the sense transcript. If the masked sequence is, for example, a splice site or splicing regulator, duplex formation would cause a shift in splicing ratio towards the preferential expression of one isoform over another and therefore antisense transcription would be positively correlated with one splice variant and inversely correlated with the other. A real example of this is given by the gene encoding the α-thyroid hormone receptor (*erbA*) which is overlapped in the antisense direction by the *RevErb* gene. Expression of *RevErb* strongly correlates with an increase in the expression ratio of the erbAα1/erbAα2 splice variants [247]. The *MAPT* bi-directional promoter element (the ‘NP’ element) was not included in the minigene models, and therefore its potential effect on *MAPT* splicing was not determined.

The creation of such *MAPT* minigenes would allow this determination, and a specific motif located within the overlapping region of the CP and NP transcripts makes this approach appealing. The 3' end of the core promoter element contains a TG dinucleotide repeat polymorphism – a motif at which the RNA-binding factor TDP-43 is known to bind [230]. TAR DNA-binding protein 43 (TDP-43) belongs to the hnRNP class of proteins and binds to both DNA and RNA. This protein has multiple functions in transcriptional repression, pre-mRNA splicing and translational repression [230, 248, 249] and is well characterised in neurological disease. So-called TDP-43 proteinopathies exhibit major neuronal and glial inclusions of TDP-43 and include frontotemporal lobar degeneration (FTLD-TDP), FTLD with motor neuron disease (FTLD-MND) and amyotrophic lateral sclerosis (ALS) [250, 251]. TDP-43 inclusions have also been detected in the limbic system of some PSP patients [252].

Given the functions of TDP-43 in transcription and splicing and its association with neurodegeneration, the location of a potential binding site between the major core promoter and secondary antisense promoter in the *MAPT* gene is intriguing. Preliminary data from our group has shown that TDP-43 does indeed bind to this motif, with the strength of binding determined by the length of the TG repeat [Roberto Simone; personal communication]. Knockdown of the protein by siRNA causes a significant increase in expression from the CP luciferase construct [Roberto Simone; personal communication]. Thus, if the NP-antisense transcript (commonly denoted *MAPT-AS1*) forms a duplex with the CP-expressed *MAPT* transcript, the repressive TDP-43 motif would be masked (figure 6.3). This could potentially lead to an increase in *MAPT* transcription which, in turn could alter the co-regulated splicing ratio of *MAPT* alternate exons.

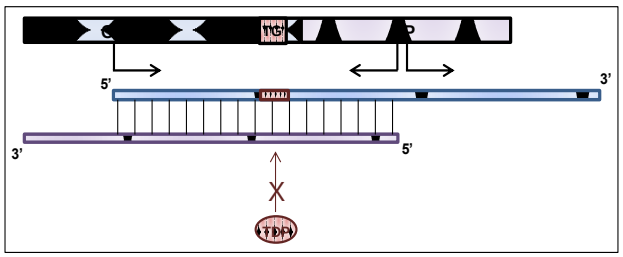


Figure 6.3 Potential masking of *MAPT* transcripts
Antisense transcripts expressed by the bi-directional promoter (NP) may form a duplex with transcripts expressed from *MAPT* core promoter (CP). The TG dinucleotide motif (TG; red box) would therefore be masked and transcriptional repression by TDP-43 would be inhibited.

It has to be said, however, that a role for TDP-43 in *MAPT* splicing has yet to be determined, with a recent study failing to show an effect on exon 2, 3 and 10 splicing following the siRNA knockdown of TDP-43. This is concordant with initial results gained from the *MAPT* minigenes described here, which detected an overall increase in minigene transcription following TDP-43 siRNA knockdown, but not an accompanying shift in exon 10 splicing ratio (data not shown).

It may therefore be unlikely that antisense-mediated transcriptional regulation plays a role in the regulation of *MAPT* exon 10 splicing, though a more detailed investigation into the function of this antisense transcript – and the non-coding sense transcript, *MAPT-IT1* – will further inform our understanding of the complex mechanisms regulating *MAPT* transcription.

6.2.2 The ability of the *MAPT* 3'UTR to regulate gene expression

The pMIR-REPORT luciferase study conducted in undifferentiated F1 and SH cells did not identify differences in luciferase expression conferred by genetic variants of the *MAPT* 3'UTR. Recent work by Tan-Wong and colleagues has shown that the 3'UTR region of a gene can influence transcription by directly interacting with the promoter. This is dependent upon factors that associate to both the 5' and 3' ends of the gene, with the formation of a 'gene loop' conformation bringing the two ends together (figure 6.4).

Figure 6.4 Ssu72 enables interaction of the 3'UTR and promoter through the adoption of a gene loop conformation.
Adapted from Tan-Wong et al (2012)[232].

This mechanism was found to determine the directionality of transcription from bi-directional promoters in which non-coding transcripts were produced in the antisense direction and protein-coding mRNA was produced in the sense

direction. The adoption of a gene loop formation enforced transcription in the sense direction and reduced aberrant transcription of non-coding antisense transcripts. The key protein involved in this regulation was the polyadenylation factor, Ssu72, which associates to both the 5' and 3' ends of genes. Mutation of Ssu72 was shown to prevent gene loop formation across the *FMP27* gene and was concurrent with an increase in promoter-associated antisense transcripts and Pol II density [232].

To date, gene loop formation has only been demonstrated with certain bi-directional promoters, a description that does not include the uni-directional *MAPT* core promoter. Loss of the 3' polyadenylation site from a mammalian gene has, however, been shown to directly influence the recruitment of transcription factors, with a subsequent reduction in gene expression [253]. The pMIR-REPORT vector used in the *MAPT* 3'UTR luciferase study expresses the firefly luciferase gene under the control of the CMV promoter. Thus, it is unlikely that gene loop formation would occur between a human 3'UTR and a viral promoter and therefore potential effects on expression of *MAPT* promoter-3'UTR interactions are presumably absent in these assays. Genetic variation within the *MAPT* 3'UTR that differentially affects its interaction with the *MAPT* promoter region cannot, therefore, be ruled out.

Another finding from this luciferase study was the significantly increased expression conferred by the H1C variant of the Fr3 fragment when compared to its H1B and H2 counterparts. The multiple sequence alignment of the three Fr3 variants (Appendix F) reveals a H1C-specific T/A polymorphism lying just 8bp upstream to a putative polyadenylation motif (ATAAAA; in green below). If the T to A transition on the H1C haplotype strengthens the signal from this putative site – leading to its recognition by the polyadenylation machinery – the 3'UTR of transcripts produced from this variant would be shortened by approximately 442bp. As transcripts with shorter 3'UTRs are generally more stable than those with longer 3'UTRs, this may account for the increased expression conferred by the Fr3 luciferase construct in both cell lines.

Fr3 Multiple Sequence alignment (1433/39-1492/99)

```
H1B CGTGTCCCATCTACAGACCTGCAGCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAA 1499
H1C CGTGTCCCATCTACAGACCAGCGGCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAA 1498
H2  CGTGTCCCATCTACAGACCTGCGGCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAA 1492
    ******************:**.************************************
```

6.2.3 The *MAPT* minigenes

The main component of this project was the design, construction and *in vitro* investigation of *MAPT* minigenes representing the genetic variation of two common *MAPT* haplotypes: H1B and the PSP risk-associated H1C. The completed minigenes conferred two transcript mis-splicing events – one major and one relatively minor. The minor splicing event occurred at the exon 9/intron 9 boundary and was caused by a weakening of the 5' splicing signal due to the insertion of a restriction site necessary for minigene construction. This resulted in the preferential use of a secondary, intronic splice site – located 26bp upstream to the native site – in a portion of the transcripts.

It is difficult to tell what effect this had on the exon 10 splicing ratio, as both 4R and 3R mRNA transcripts were produced despite the change in splice site utilisation. It may be that this cryptic splice site is somehow involved in exon 10 splicing and its over-use in transcripts expressed by the minigenes may be responsible for the over-expression of exon 10-containing 4R transcripts that was a feature of all six of the minigene variants regardless of the differentiation status of the *in vitro* cell lines – an unexpected finding given the exclusive 3R expression observed endogenously in undifferentiated cells. This would seem unlikely, however, as the use of this secondary splice site results in the insertion of 26 nucleotides of intronic sequence into the RNA message, which causes a shift in the open reading frame during protein translation. It is for this reason that protein analyses could not be undertaken using the *MAPT* minigenes produced here.

The major mis-splicing event resulted in the complete removal of exons 4-9 from the minigene transcripts and stemmed from the original minigene design. Although the *att*B sequence at the intron 3/exon 4 boundary was designed to re-

capitulate the 3' splice site, the increased distance between the exon boundary and the intronic splicing elements (such as the Py tract) appear to have resulted in inefficient splice site recognition.

The importance of intronic sequences in exon recognition was highlighted in a study by Dewey and colleagues, who showed that the length of the intron between two exons determines the strength of the splicing signal by dictating the number of splicing enhancer elements contained within the exon. Longer introns require a higher number of enhancers in order to maintain the splicing signal over a greater distance. Thus, the decision to completely remove the introns between exons 4, 5, 7 and 9 likely caused an accumulation of exonic splicing signals within a short stretch of sequence [256], presumably confusing the splicing machinery and resulting in its failure to recognise the element as an exon.

Although the Dawson study did not report a similar problem with their minigenes, they were not able to conduct protein analyses due to the low level of minigene-expressed tau protein. Their study was conducted on a murine tau background, awarding greater flexibility in the analysis of the minigene mRNA as they did not have to rely on a FLAG-tag motif to separate human minigene tau from endogenous murine tau. As a result, their mRNA analyses were much clearer. This is encouraging in terms of the stable cellular models that were planned for this project, as stable integration into the genome of the cell line should increase yield and add confidence to the splicing studies – particularly those at the N-terminal exons. Before this, however, the problem of the exon 4-9 element and the intron 9-mediated frameshift must be corrected and an outline of how this could be achieved was discussed in section 4.11.

One important observation regarding the *in vitro* expression of the minigenes was that the mis-splicing events were common to all six minigenes, with the consequences on expression the same and highly replicable in each instance. It was therefore agreed that these minigenes could still fulfil the purpose for which

they were initially designed – to study the effect of promoter identity and genetic variation of the alternative splicing events at exons 2, 3 and 10.

6.2.4 The role of promoter identity in the regulation of *MAPT* N-terminal splicing events

The design of the *MAPT* minigenes made the quantification of the N-terminal splicing events difficult. As aberrant splicing events inhibited full-length protein expression – preventing Western blot analysis that would have allowed the adequate separation of the six tau isoforms formed by exon 2, 3 and 10 alternative splicing – quantification was only possible at the mRNA level. The shortcomings of the mRNA analysis method meant the N-terminal exon splicing ratio was quantified independently of exon 10 inclusion; however, the aberrant splicing event that resulted in the exclusion of exons 4-9 seemed to occur preferentially in 1N transcripts. Therefore the refinement of the analysis to detect only the transcripts that were spliced correctly produced an over-representation of the 2N isoforms – the isoform that is actually the least expressed *in vivo*. In the absence of quantifiable 1N isoforms, the 2N/0N ratio was used to investigate a potential role for promoter identify and the rs242557 domain on N-terminal splicing regulation.

A role for promoter identity was observed in differentiated F1 cells, though additional regulation by the rs242557 domain was not apparent. This is concordant with the results of a previous study, which failed to find an association between rs242557 and exon 2 and 3 inclusion [205]. This was largely concordant in undifferentiated SH cells, though the behaviour of the variants in differentiated SH cells – in which all three H1C variants puzzlingly exhibited constitutive exon 3 inclusion – does not inspire confidence in the accuracy of the N-terminal splicing ratios quantified from these cells.

There have been reports of an association between the *MAPT* H2 haplotype and increased exon 3 inclusion, leading to suggestions that this exon contributes to the protective role attributed to H2 against PSP [133]. Indeed, one study reported a 2-

fold increase in the number of H2 transcripts containing exon 3 compared with H1 transcripts. There were also suggestions of an additional increase in 1N isoforms of the H2 transcript, though this was not deemed biologically relevant. Exactly how exons 2 and 3 may confer protection against neurodegeneration is currently unclear, though some studies have indicated potential mechanisms.

Investigations into the effect of the N-terminal exons on tau protein folding and aggregation suggests that tau forms a paperclip-like conformation in solution, which brings together the N- and C-termini as the C-terminus is associating with the microtubule binding repeat domains. As this conformation is similar to the aberrant tau epitope that is detected in early-stage AD, it has been suggested that the stabilisation of the paperclip confirmation may be pathologically significant [258]. The N-terminal may also be key to maintaining tau solubility, as N-terminal fragments have been shown to inhibit the polymerisation of tau into insoluble aggregates [259]. Increased exon 3 inclusion (and thus exon 2 inclusion due to their incremental relationship) in H2 *MAPT* transcripts may protect against neurodegeneration by altering the conformation of the tau protein and preventing its aggregation into insoluble filaments. Further clarification of the role of the N-terminal inserts and the effect of genetic variation on their inclusion rate is likely to be achieved in the near future as focus increasingly shifts from the study of exon 10 to exon 2 and 3 splicing.

Unfortunately the absence of H2 versions of the minigenes prevented a H1/H2 comparison of N-terminal exon splicing events that would be highly informative and is currently of great interest in the field. Thus, although these minigene quantifications reveal a role for the *MAPT* promoter in the regulation of exon 2 and 3 alternative splicing, little else pertaining to the mechanism and the relationship with exon 10 splicing can be determined here.

6.2.5 Evidence for a role of rs242557 in the co-regulation of *MAPT* transcription and exon 10 splicing

6.2.5.1 rs242557 and transcription

This project has provided evidence of a role for the rs242557 polymorphism in the co-regulation of *MAPT* transcription and exon 10 splicing, gained from analysis using multiple methods. The first method was the luciferase reporter gene assay, which clarified previous reports into the differential regulatory effect of the rs242557 alleles on *MAPT* transcription by demonstrating that the function of the domain and the direction of the allelic effect is dependent upon both the positioning of the rs242557 domain relative to the core promoter (CP) and the cellular environment in which the luciferase construct was assayed. This, combined with analyses of mutant CP constructs, indicated that the genomic positioning of the rs242557 domain – approximately 47kb downstream to the *MAPT* core promoter – is vital to its regulatory function *in vivo* and hinted at a physical interaction between the two domains.

This physical interaction may result from the formation of a loop structure which brings the core promoter and rs242557 domain into close proximity. This would allow proteins bound to either element to interact and for the rs242557 domain to modulate the activity of the transcription machinery assembled at the core promoter – similar to the 'gene loop' mechanism described for 3'UTR-mediated transcriptional regulation (section 6.2.2; figure 6.4). Although the direction of the regulation by the rs242557 domain changed depending on both its positioning relative to the core promoter and the cell model, the A-allele demonstrated consistently weaker regulation of transcription than its G-allele counterpart regardless of these factors. Furthermore, a single nucleotide error at position 596 of the core promoter element (in exon 0) – in which the wildtype G nucleotide was substituted for a T nucleotide – compensated for the altered regulation by the A-allele domain variant and caused a strengthening of domain function to match that of the G-allele variant.

This would suggest that the A-allele exerts a gain-of-function effect on the domain, which is abolished by the exon 0 mutation. One hypothesis could be that the A-allele of the rs242557 polymorphism forms (or strengthens) a binding site for an unknown protein factor that, in turn, recruits an additional unknown protein factor to exon 0 and together these proteins weaken the regulatory signal from the rs242557 domain (figure 6.5).

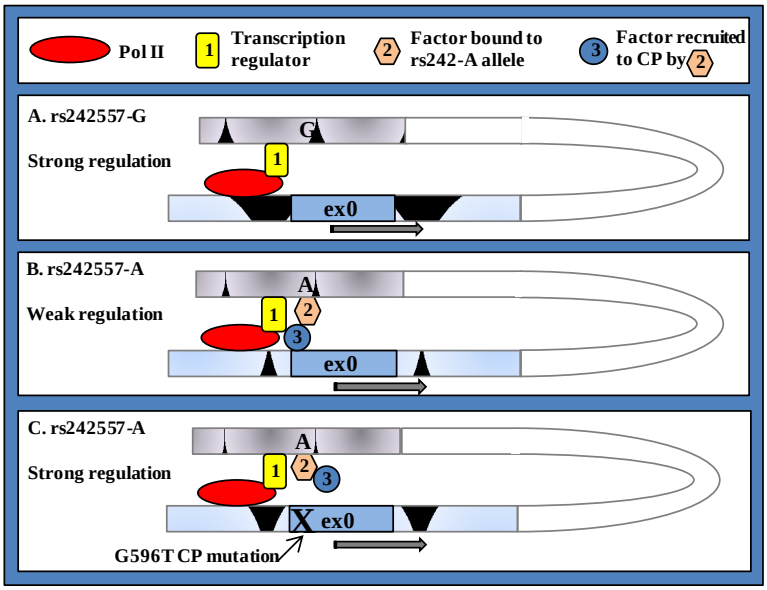


Figure 6.5 A potential mechanism for the differential regulation of CP expression by the rs242557-G and rs242557-A domain variants.
A: The rs242557 domain carries out its regulatory effect on CP transcription by adopting a loop confirmation that brings the two elements into close proximity. Proteins bound to the rs242557 domain (factor 1; yellow rectangle) interact with components of the transcription machinery assembled at the core promoter (such as RNA Pol II, red oval) to regulate transcription rate;
B: The A-allele of rs242557 completes a binding site for an additional DNA binding factor (factor 2; orange hexagon). This factor recruits a third factor to the CP (factor 3; blue circle). Factors 1, 2 and 3, interact and weaken the regulatory function of the rs242557 domain;
C: A single G to T transition within exon 0 of the core promoter element abolishes the binding site of factor 2, restoring the strength of rs242557-mediated regulation.

In a bid to identify factors that may differentially bind to the A- and G-allele variants of the rs242557 domain – and therefore explain the differences in transcriptional regulation conferred by these variants – ChIP was performed on extracts from A/A homozygous (SH) and G/G homozygous (F1) cell lines using antibodies against phospho-independent Pol II and hnRNP U epitopes. Both proteins were found to associate with the region containing rs242557 and

demonstrated an inverse relationship in regards to one another. The hnRNP U factor was chosen for investigation as EMSA and siRNA knockdown-mediated luciferase reporter gene studies had previously indicated that the two alleles of the rs242557 polymorphism differentially associate with this protein [JF Anaya, PhD thesis, UCL 2012; [242]]. The ChIP experiments supported these results and showed for the first time that hnRNP U – a known splicing factor – associates with Pol II at the *MAPT* promoter region.

It has been shown in other genes that hnRNP U binds to the phosphorylated CTD (carboxy terminal domain) of Pol II [243, 244] and could therefore potentially be used as an indicator of active and elongating transcription complexes. This would explain the apparent inverse relationship between Pol II and hnRNP U concentration that was observed over the differentiation time course. Pol II abundance increased during differentiation, with the consequential decrease in hnRNP U binding indicating transcriptional pausing and Pol II accumulation at the rs242557 domain.

In general, there was a higher level of Pol II accumulation at the G-allele variant compared to the A-allele variant of the rs242557 domain. Furthermore, by day 3 of differentiation, hnRNP U was undetectable at the G-allele variant, suggesting a lowering of transcription rate that is concurrent with the general repressive effect of the rs242557 domain that occurs when it is cloned downstream to the core promoter, as demonstrated by the promoter luciferase reporter gene study. The A-allele variant of the domain remained associated with hnRNP U throughout differentiation, although the abundance of the RNA-binding factor reduced as Pol II accumulation increased. This indicates that whatever it is that induces transcriptional pausing at the rs242557 domain (*cis*- or *trans*-acting factors) is weakened by the presence of the A-allele and therefore a higher proportion of Pol II complexes fail to pause at the rs242557 domain and continue to elongate. This is, again, concordant with the promoter luciferase assay results that reported a reduced capacity of this domain variant to repress transcription from the core promoter.

In summary, therefore, the combination of the luciferase reporter gene studies and the ChIP experiments have revealed two potential mechanisms that may explain both the role of the rs242557 domain in regulating transcription from the core promoter and the allelic differences in this regulation. Although the luciferase quantification of the effect of the rs242557 domain on expression could not determine the specific function of the domain, as either enhancement or repression was observed depending on its relative positioning and the *in vitro* cellular environment, the ChIP results support a repressive role for this domain and is in agreement with the highly consistent effect on luciferase expression produced in both cell lines when the domain was cloned in its more natural position downstream to the core promoter. Thus, if the rs242557 domain does, indeed, function as a repressor of transcription and the A-allele variant weakens this function, then this raises the question of its role in the alterations in exon 10 splicing ratio produced from the *MAPT* minigenes containing this domain variant.

6.2.5.2 rs242557 and alternative splicing: model 1

Figure 6.6A presents a possible mechanism in which transcriptional pausing at the rs242557 domain facilitates splicing factor recruitment and spliceosome assembly. In this model, *cis*- and/or *trans*-acting factors (orange hexagon in figure 6.6A) at the rs242557 domain (purple oblong) blocks the progression of the transcription complex (represented by Pol II; red oval). This causes the Pol II accumulation and dissociation of hnRNP U (blue triangle) observed in the ChIP experiments, followed by splicing factor recruitment and assembly of the spliceosome (green oval) on the nascent transcript. As elongation resumes, the nascent transcript is spliced as it emerges from the transcription machinery. The appropriate inclusion of alternate exons 2, 3 and 10 may be dependent on the recruitment of specific components to the spliceosome, which itself may be dependent on the length of transcriptional pausing.

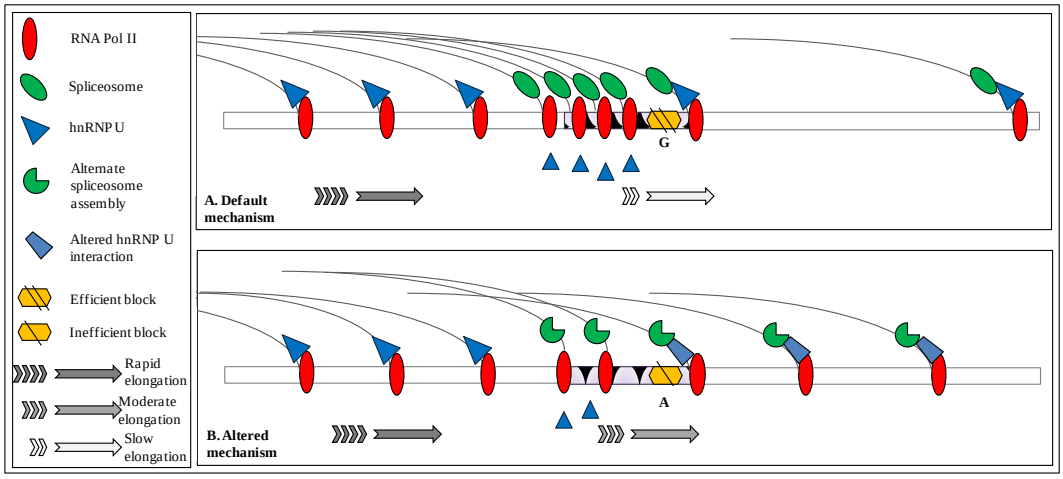


Figure 6.6 A potential mechanism for the repressive effect of the rs242557 domain (purple oblong) on transcription.
A: In the default mechanism, the G-allele domain variant may act as a 'speed bump', causing a reduction in transcription rate and accumulation of Pol II (red oval). The reduction in transcription rate could lead to hnRNP U (blue triangle) to dissociate from the transcription complex, facilitating spliceosome assembly (green oval). B: In the altered mechanism, the A-allele domain variant may be inefficient at reducing transcription rate, resulting in sub-optimal spliceosome assembly (cut out green circle) and altered interactions between the spliceosome and hnRNP U (blue flattened triangle).

One problem with this mechanism is that spliceosome formation occurs during transcription elongation and not when the transcription complex is stationary. As it is currently unclear as to what stage of elongation hnRNP U associates with Pol II, it may be that transcription is not completely blocked; rather it is simply slowed down, with the rs242557 acting as a 'speed bump'. This reduction in transcription rate may be necessary for the specific recruitment and assembly of the spliceosome on the nascent transcript, which may, in turn, determine the pattern of alternate exon inclusion by ensuring vital interactions between the splicing and transcription components can take place.

Under this hypothesis, the rs242557-A domain would appear to be inefficient at reducing the elongation rate of the transcription complex, with a proportion of complexes remaining associated with hnRNP U and passing straight through the domain (figure 6.6B). This may result in sub-optimal splicing factor recruitment and spliceosome assembly, presumably due to the transcription complex being at

the domain for a shorter amount of time and/or a modifying signal caused by the continued association of hnRNP U.

Matching this with endogenous exon 10 splicing patterns, Pol II accumulation would appear to cause a shift in isoform expression from default constitutive exon 10 exclusion (3R-tau), as observed in undifferentiated cells, towards increased exon 10 inclusion (4R-tau), as observed in neuronally differentiated cells. It is difficult, however, to reconcile a reduction in transcription rate with an increase in exon 10 inclusion, as the kinetic model of co-transcriptional splicing (section 1.2.3.3) dictates that a lower rate of transcription elicits greater exon *exclusion* – in this case increased 3R expression – due to competition from stronger downstream splice sites.

6.2.5.3 rs242557 and alternative splicing: model 2

A second potential mechanism centres on the secondary structure of the nascent pre-mRNA transcript. *In silico* evidence suggests that a single change from a G nucleotide to an A nucleotide at rs242557 can potentially cause a significant change in mRNA conformation. Figure 6.7 shows the strikingly different RNA secondary structures predicted to form when the G-allele (6.7A) or the A-allele (6.7B) is present. These predictions were created using the RNAfold web server (University of Vienna) and the full 812bp sequences of the rs242557 SD elements (Appendix B).

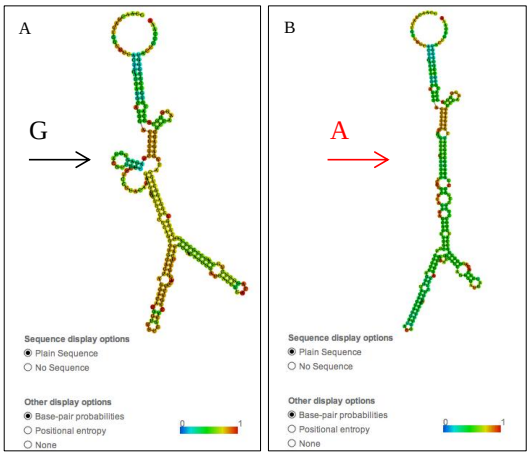


Figure 6.7 *In silico* predictions of differences in rs242557 RNA conformation.
The G-allele (A) and A-allele (B) variants of the 812bp rs242557 domain may confer significantly different secondary RNA structures, as predicted by the RNAfold web server.

These predictions are, of course, artificial and change significantly when the input sequences are extended by just a few nucleotides. They do show, however, that one nucleotide change can significantly alter RNA conformation and this may be important when considering co-transcriptional mechanisms of alternative splicing.

Figure 6.8 presents a mechanism by which the folding of the nascent transcript as it emerges from the transcription complex facilitates the recruitment and assembly of the splicing machinery (panel A). In this model, the alterations to the secondary structure caused by the A-allele of rs242557 could affect the assembly of the spliceosome, perhaps by masking – or exposing – binding sites for certain splicing factors (panel B). In addition to this, the reduced ability of the A-allele to lower transcription elongation rate may contribute to the production of differential RNA secondary structures, as a faster elongation rate would imply a smaller window for RNA folding before assembly of the spliceosome.

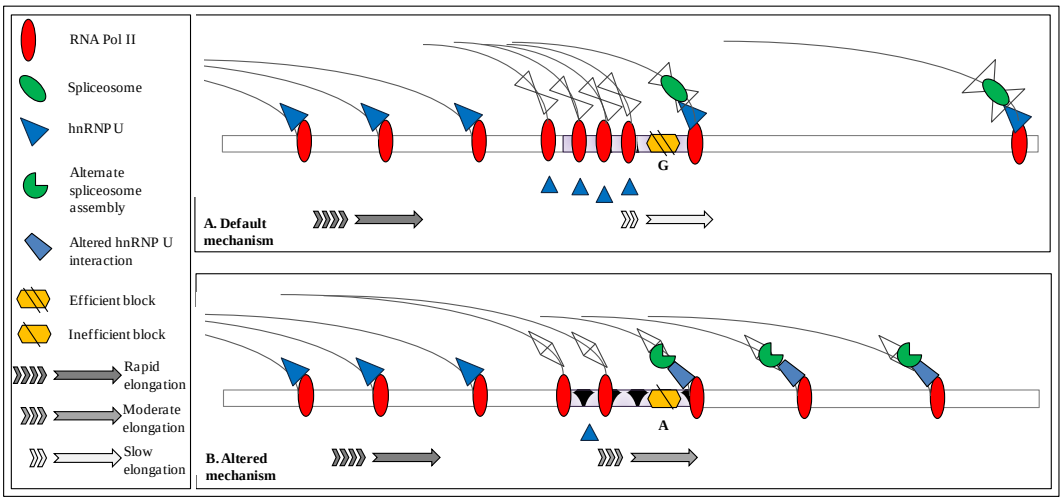


Figure 6.8 Folding of the nascent pre-mRNA transcript into a specific secondary structure may influence splicing factor recruitment and spliceosome assembly.
A: The default secondary structure produced in the presence of the G-allele of rs242557, combined with a reduced elongation rate, facilitates optimal spliceosome assembly; B: The A-allele of rs242557, combined with an increased elongation rate, alters the secondary structure of the nascent transcript and results in sub-optimal splicing factor recruitment. The continued association of hnRNP U to Pol II as a result of the increased elongation rate may further alter the composition and functioning of the spliceosome.

6.2.5.4 rs242557 and the co-transcriptional regulation of alternative splicing

The above sections have outlined partial mechanisms that could potentially explain some of the findings described here relating to the role of rs242557 in *MAPT* transcription and alternative splicing. Figure 6.9 brings together some of these ideas and speculates as to how these individual mechanisms may impact upon each other.

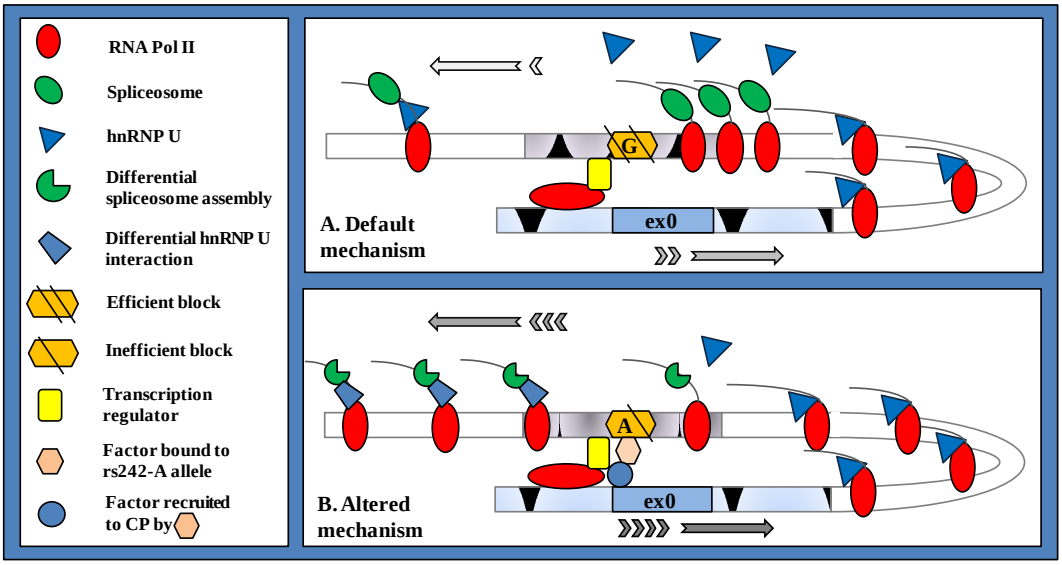


Figure 6.9 A potential mechanism of *MAPT* co-transcriptional splicing.
This mechanisms brings together the individual mechanisms proposed in this chapter. The G-allele (A) and A-allele (B) have multiple effects on transcription that together influence splicesome assembly. A thorough description of this figure is given in the text of this section.

Starting with the gene loop theory from section 6.2.4.1 – in which proteins bound to the rs242557 regulatory domain physically interact with proteins bound to the core promoter to reduce the overall rate of transcription – the binding of an additional transcription factor to rs242557-A may contribute to the reduced Pol II accumulation observed at this domain variant. The binding of an extra protein – and its physical interactions with the core promoter – presumably alters the conformation of this domain, making it easier for the transcription complex to pass through. Thus, the insufficient reduction in elongation rate of the transcription complex – combined with competing signals from the additional transcription factor and the continued association of hnRNP U with Pol II – would

likely alter the recruitment and assembly of the spliceosome on nascent transcripts. This, in turn, may modify the ability of the spliceosome to recognise certain splicing signals at intron-exon boundaries and cause a shift in the inclusion rate of certain alternate exons.

This mechanism is, of course, purely speculative and infers several relationships that are currently unproven. It does, however, bring together the individual data described in this project and provides an initial hypothesis which future studies can investigate. As research continues into the regulation of *MAPT* expression, this hypothesis will be proved or disproved, embellished or fragmented. Either way, the results presented here have provided valuable insight into the molecular consequences of the PSP risk allele, rs242557-A, and has, for the first time provided evidence that *MAPT* transcription and alternative splicing are co-regulated and that promoter identity plays a vital role in determining the outcome of this co-regulation. By establishing the regulatory potential of the rs242557 domain at the basic sequence level, a platform is provided for further investigation.

6.2.6 Cellular differences in gene expression

One overarching theme of this project has been the differential behaviour of many of the luciferase and minigene constructs when assayed in the two neuroblastoma cell lines, SK-N-F1 and SH-SY5Y. This may be because cancer cells, by their very nature, are abnormal and the two cell lines likely have different abnormalities that lead to differences in the way gene expression is regulated. There is evidence, however, that suggests that these two cell lines may be at different points in the differentiation pathway, despite the common belief that both are in the undifferentiated state. Several findings have indicated that, while the SH cells appear truly undifferentiated, the F1 cell line may in fact be at least one day into the differentiation pathway. The evidence for this includes:

1. N-terminal exon splicing analysis on RNA extracts from untransfected F1 cells shows that the 1N isoform is expressed in this cell line (figure

5.8Aii) in addition to the 0N isoform that was present alone in SH cells (figure 5.9). It is widely accepted that 0N3R is the only isoform expressed when cells are in their undifferentiated state [132].

2. Exon 10 splicing quantifications in undifferentiated F1 cells showed that promoter identity was already beginning to influence minigene expression, with the CP+rs242557 minigenes showing significantly different 4R:3R tau mRNA ratios than their CP and CMV counterparts. In SH cells promoter specificity was not detected until cells were in their differentiated state.

3. Expression of the promoter luciferase and minigene constructs was generally higher in F1 cells than in SH cells, consistent with the upregulation of *MAPT* expression that occurs during differentiation. All assays, particularly minigene quantifications, demonstrated greater consistency in F1 cells, and this suggests that gene expression is more tightly regulated in this cell line than in SH cells, where replicate assays were more inconsistent.

4. The rs242557 ChIP experiments performed on SH cell extracts show that neither Pol II nor hnRNP U is detectable at day 0 and first appear at day 1, with Pol II reaching a maximum concentration at day 5 (figure 5.12B). In F1 cells, however, Pol II and hnRNP U are already present at ‘day 0’ and Pol II accumulation reaches a maximum at day 3 (figure 5.12A), indicating that the cells have reached the end of the differentiation pathway. This would suggest that F1 cells are actually 1-2 days into the differentiation pathway in the so-called ‘undifferentiated’ state.

If ‘undifferentiated’ F1 cells are, indeed, one day into the neuronal differentiation pathway, this may account for the differences in expression observed in some of the assays and this must be taken into account when considering the results. If

nothing else, these differences show the significance of the endogenous cellular environment on gene expression and demonstrate the important of choosing the right cellular model for *in vitro* expression studies.

6.3 Conclusions

This project has presented evidence that *MAPT* transcription and splicing processes are co-regulated and has confirmed a role for the rs242557 regulatory domain in these processes. The A-allele of this polymorphism is highly associated with the tauopathies, in particular with PSP and CDB, and it has been shown here that this allele – at the basic sequence level – has the ability to modify both the rate of transcription conferred by the core promoter and the inclusion rate of the alternatively spliced disease-associated exon 10.

Initial chromatin immunoprecipitation experiments have provided a potential link between the *MAPT* transcription rate and splicing and have shown that RNA Pol II differentially accumulates at the allelic variants of the rs242557 domain. The apparently reduced Pol II accumulation at the A-allele presumably contributes to – or is a consequence of – the increased transcription rate conferred by this domain variant. An inverse relationship between Pol II accumulation and hnRNP U – a known splicing factor – and the increased association of this factor at the A-allele variant compared to the G-allele variant is likely to contribute to the differences in exon 10 inclusion detected in transcripts expressed from the two minigene variants.

To conclude, although this project has encountered several problems that have restricted the interpretation of the data produced, a minigene model has been created that has, for the first time, linked rs242557 to exon 10 splicing, and found evidence of co-transcriptional regulation of *MAPT* alternate exon inclusion. Perhaps most importantly, by bringing together elements that have previously been individually associated with PSP risk – the rs242557 allelic variants, an increase in *MAPT* expression and increased exon 10 expression – many exciting new avenues for future research have been opened.

6.4 Future directions

6.4.1 Correction of the *MAPT* minigenes

The first thing that must be done is to correct the mis-splicing events displayed by the *MAPT* minigenes. Section 4.11 describes a potential way of achieving this, though the resultant minigenes would be without exons 5 and 7. This is more preferable than the current model, however, and should allow improved mRNA quantifications and additional protein analyses.

Following these corrections, the creation of the isogenic cell models will increase assay yield and replicability. In chapter 4, the creation of platform cell lines was described, with two monoclonal lines selected from each of the F1 and SH cell types for minigene integration: one in which integration occurred within an active gene and one in which integration occurred in an non-critical, though inactive, region of the genome. These integrated models have two major advantages: firstly, that each minigene will be inserted into the same place in the genome and, secondly, that minigene expression should be significantly higher than observed with transient transfection. Such integration should remove confounding factors such as well-to-well differences in cell density and transfection rate and positional effects due to differential insertion sites. A comprehensive description of the method for creating these isogenic cell models is given in sections 4.4 and 4.10 and will not be repeated here.

6.4.2 Further analyses using the *MAPT* minigenes

6.4.2.1 H2 minigenes

One frustrating aspect of this project was the failure to complete the H2 minigene variants. As described in section 4.8.3, this was due to problems with the cloning of minigene fragment 3. This minigene variant would, however, add significant value to the project as H1/H2 differences in alternative exon splicing and tau protein expression, if present, are likely to be more pronounced than those detected between the two H1 sub-haplotypes. Such analyses are particularly

desirable following recent evidence describing the contribution of increased exon 3 inclusion to the protective role of H2 against PSP and neurodegeneration [133, 205].

As three of the four H2 minigene fragments – including the promoter variants – have been completed, a potential solution (if completion of the final fragment continues to be problematic) could be to create H2 minigenes using one of the H1 variants of fragment 3. This is obviously less desirable than a complete H2 minigene – particularly as fragment 3 contains exon 10 – but as the focus of this project was to investigate the effect of genetic variation within the *promoter* region, such hybrid minigenes should still prove informative and add significant value to the findings already drawn from this project.

Completion of the H2 isogenic cell models as planned would allow comparison with the H1 variants at the basic sequence level but, as described in section 4.8.3, would not take into account the positional effects conferred by the inversion polymorphism at the *MAPT* genomic location. There may be some value, therefore, in altering the design of the H2 minigene to allow its insertion into the platform cell line in the opposite orientation to the H1 models. This could be achieved by altering the combination of *att*B sequences inserted onto the ends of each minigene fragment, to reverse the orientation of the final minigene in the integrated cell models. Comparison between two H2 models, in which the H2 minigene has been inserted in opposite directions, may shed some light on the impact of the inversion, independently of H2-specific sequences, on *MAPT* expression.

6.4.2.2 Alternative mRNA analysis

In addition to the mRNA analysis methods used in this project, there are alternate methods of mRNA quantification that may prove valuable in the analysis of transcripts expressed from the *MAPT* minigenes. Real-time quantitative PCR (qPCR) in conjunction with reverse transcription provides an accurate method of quantifying the abundance of a specific transcript. Furthermore, the relative

abundance of two different isoforms of a transcript can be determined in one PCR by designing two probes, labelled with different colour fluorophores, that each bind specifically to one isoform. Thus, to quantify the proportion of minigene transcripts containing, for example, exon 10, two forward primers could be designed: one overlapping the exon 9/exon 10 boundary to detect 4R isoforms and the other overlapping the exon 9/exon 11 boundary to detect 3R-tau isoforms.

This method could also be used to quantify the splicing at the N-terminal exons – though only after the appropriate corrections to the minigenes have been made. Indeed, commercial TaqMan probes are available for the quantification of individual tau transcript isoforms.

6.4.2.3 Protein analyses

Quantification of tau protein expression from the minigene variants is a vital investigation. Not only will it determine whether the exon 10 inclusion rate in tau mRNA directly translates into 4R/3R tau protein isoform expression – as has been suggested to be the case [133] – but Western blotting will allow the analysis of all six protein isoforms in one assay; something which cannot be achieved by mRNA analysis.

A primary antibody targeting the FLAG-tag motif would detect the presence – and relative abundances – of all of the protein isoforms correctly expressed by the minigene variants and would distinguish them from endogenously expressed tau protein. This would determine whether the allelic differences in exon 10 inclusion detected in minigene mRNA translate into allelic differences in minigene protein isoform expression, not to mention the significant clarification the resolution of all six isoforms would give to the N-terminal exon splicing investigation.

Commercial antibodies are available that target specific phosphorylation sites on the tau protein, with reactivity only detected when the site is phosphorylated. Such antibodies will allow information to be gained on the phosphorylation status of the minigene-expressed protein isoforms, potentially bringing together the two major

areas of tau research – molecular processing and protein phosphorylation. In this instance, however, care must be taken to normalise the minigene tau blots against equivalent endogenous tau blots, as these phospho-specific antibodies will not distinguish between the two.

6.4.2.4 Alternative promoters

Although the modular nature of the *MAPT* minigenes has proved unsatisfactory in certain respects, one significant advantage to their design is the ability to easily swap in different promoters or promoter variants. Thus, the minigene blueprint can be used to investigate the role of other regions of the *MAPT* promoter, such as predicted regulatory elements or highly conserved stretches of sequence currently of unknown function. The minigene analysis methods optimised in this project and described above could be used in all of the investigations described below.

Recent work by our group has revealed the presence of a potential promoter element located downstream to both the main core promoter and the adjacent bi-directional promoter [R. Simone, personal communication]. This region has already been cloned and adapted to form a Gateway® entry clone and can therefore be incorporated into a *MAPT* minigene. It would be interesting to confirm whether this sequence is a promoter element, whether it can express full-length tau, and whether its transcripts are alternatively spliced. It may be that the *MAPT* gene has a second transcription start site in addition to the major core promoter, with this promoter perhaps limited to producing certain isoforms, for example 3R-tau. If so, the activity of this promoter would likely change during development and neuronal differentiation. The absence of this putative promoter – or similar elements elsewhere in intron -1 – in the current minigenes could potentially account for the over-expression of 4R-tau transcripts from these constructs.

Another valuable investigation would be to study the effect of genetic variation within the core promoter on minigene expression. In section 6.2.1 a TG repeat polymorphism located at the 3' end of the CP element was described (figure 6.2).

Work within our group has shown that TDP-43 binds to this region and that the strength of this binding is dependent upon the length of the dinucleotide repeat. Luciferase reporter gene assays have, in turn, associated repeat length with luciferase activity and there are early indications that this polymorphism is linked significantly to PSP risk [Roberto Simone, personal communication]. It would therefore be interesting to create a series of minigenes, each with expression driven by a CP element that differs only by the number of TG repeats it contains. If, as expected, repeat length is correlated with minigene expression as a result of differential TDP-43 binding, splicing of the alternate exons may also be affected. Indeed, TDP-43 is involved in multiple levels of gene expression and is a known splicing factor; however, there is no evidence to date that this protein is involved in the splicing of *MAPT* transcripts. Knockdown of endogenous TDP-43 should neutralise any differences in binding resulting from the repeat length of the minigene variants. Such analyses would form the most comprehensive study to date on the role of TDP-43 in *MAPT* expression and would greatly enhance our current understanding in this area of research.

6.4.2.5 *Trans*-acting factors

Once the stable minigene cell models have been established, they will provide a platform for the study of specific *trans*-acting factors on *MAPT* expression. As described above, one such factor of interest is the TDP-43 protein. Other factors that were identified in this project are hnRNP U and, potentially, β-actin. Common methods for assessing the role of *trans*-acting factors in gene expression comprise either knockdown or over-expression of the factor of interest, followed by quantification of the effect on expression of the target gene. These two methods are complementary and should elicit opposing effects; for example increased expression following knockdown should correlate with decreased expression following over-expression of the same factor, and vice versa.

The most widely used method for protein knockdown is RNA interference (RNAi), which prevents the translation of target mRNA transcripts into protein. This involves the transfection of synthetic small interfering RNA molecules

(siRNAs) that are approximately 20 nucleotides in length and demonstrate complementarity to the target transcript, in this case the mRNA transcripts of the *trans*-acting factor. The binding of these siRNAs to the target transcript produces short stretches of double-stranded RNA that are subsequently targeted for degradation, thus preventing translation of the target transcript into protein. Over-expression is simply achieved by transfecting the cell model with an appropriate plasmid vector containing the gene for the *trans*-acting factor under investigation.

A slight variation of this method could be used to investigate the role of micro RNAs (miRNAs) in *MAPT* minigene expression. MicroRNAs are expressed endogenously and behave similarly to siRNAs by binding to specific target transcripts and suppressing their translation, thus negatively regulating gene expression. MicroRNAs bind to the 3'UTRs of genes, as discussed in section 3.13, and therefore factors such as genetic variation and alternative polyA site usage could affect the extent of their regulation of target gene expression.

A recent study has linked a specific miRNA, denoted miR-132, to PSP by identifying a relative reduction in the abundance of this miRNA in the brains of PSP patients compared to healthy controls. Furthermore, knockdown of miR-132 in murine Neuro2a cells was shown to cause an increase in expression of the splicing factor polypyramidine tract-binding protein 2 (PTBP2), which plays a role in *MAPT* splicing. This, in turn, caused a significant reduction in the production of 4R-tau isoforms [157]. Thus, increased levels of miR-132 in the brains of PSP patients could potentially contribute to the over-expression of 4R-tau characteristic of this disease by modulating the expression of a *trans*-acting factor. This could be confirmed in human cells by knockdown and/or over-expression of miR-132 in the minigene cellular models followed by quantification of the effect on 4R/3R mRNA and protein ratios. If an effect is observed, correlations with PTPB2 expression levels may add weight to the proposed mechanism.

Such a study could provide the first replication of the miR-132/PTPB2/PSP findings in human cells and, potentially, investigate a role for the promoter in this type of regulation. If alterations to miR-132 expression differentially affect minigenes containing different promoter elements, this may add credence to the gene loop theory described in section 6.2.2, where the 3'UTR and promoter of a gene physically interact. Comparisons between the *MAPT* promoter- and CMV promoter-driven minigenes would particularly inform this investigation.

6.4.3 Investigation of the gene loop theory in the 3'UTR-mediated regulation *MAPT* expression

To further investigate the gene loop theory of 3'UTR-mediated gene expression, versions on the *MAPT* minigenes could be created in which the full-length 3'UTR is replaced by one of the deletion fragments – either the 5', middle or 3' end of the 3'UTR – as described in section 3.13. It would be interesting to see whether the pattern and/or rate of tau isoform expression differs when specific sections of the 3'UTR are absent, or indeed, whether such differences depend on the identity of the promoter element included in the minigenes. Again, comparisons between the *MAPT* core promoter and CMV promoters would be of significant value in this instance.

If specific promoter/3'UTR fragment combinations were shown to differentially affect minigene expression, it may be prudent to scan the relevant sequences for potential *trans*-acting protein binding sites that may be involved in gene loop formation (using tools such as those provided by the UCSC genome browser). Ssu72 binding would be of particular interest due to its reported involvement in gene loop-mediated regulation of the *FMP27* gene (section 6.2.2) [232]. An effect of Ssu72 knockdown/overexpression on expression of the minigenes would indicate a role for Ssu72 – and potentially gene loop formation – in *MAPT* expression. Immunoprecipitation of the Ssu72 protein from the cell chromatin extracts described in chapter 5 (used above in the experiment presented in figure 6.1), would confirm – or disprove – the binding of this protein to the predicted target regions of the *MAPT* promoter/3'UTR.

Moving away from the minigene model, chromosome conformation capture (3C) provides a high throughput method of analysing the natural organisation of chromosomes within a cell and has previously been used to identify a loop structure at the β-globin locus in erythroid cells [260]. It could therefore be used to confirm – or refute – the 'gene loop' hypothesis of *MAPT* regulation.

6.4.4 Natural antisense transcription and the bi-directional promoter

The luciferase reporter gene study of the *MAPT* bi-directional NAT promoter (the 'NP' element, chapter 3) showed this region to be capable of initiating transcription in both the sense and antisense orientations as well as altering transcription from the core promoter. It would be interesting to determine whether this region – in conjunction with the core promoter – could also affect alternative splicing at exons 2, 3 and 10. This, as before, could be done by altering the promoter element of the *MAPT* minigene. Added value may be gained from comparing expression of such minigenes in F1 and SH cell models. It was shown in section 3.12.1.2 that the addition of the NP element to the core promoter increases relative luciferase expression in F1 cells but reduces expression in SH cells, though the mechanism behind this could not be determined. As the cell line differences must result from differences in the expression of *trans*-acting factors, it is reasonable to suggest that this may also affect splicing, either directly (through interactions with the spliceosome) or indirectly (through the modifications to transcription rate).

Initial work by our group has shown that siRNA knockdown of the antisense transcript originating from the NP region increases the relative luciferase activity conferred by the CP and CP+NP constructs described in chapter 3 (data not shown). In complement, its overexpression reduces luciferase activity. These methods could also be applied to the minigene model to assess whether there is an effect on *MAPT* expression and, in particular, alternative splicing.

If so, would the two alleles of the rs3744457 polymorphism (section 3.12.3) differentially affect this function? The luciferase and genotyping analyses of this

polymorphism suggest that an effect, if any, would be very subtle and therefore minigene methods may not be sensitive enough in this instance. A more informative experiment may be to quantify the expression of the non-coding antisense transcript in two cell lines, each homozygous for one allele of the polymorphism (i.e. C/C or T/T). This could be done using the quantitative RT-PCR method described above on cellular RNA extracts.

If allelic differences in non-coding antisense transcript expression are observed endogenously, this may support and expand any findings relating to *MAPT* isoform expression gained from the NP minigene variants. Further insight may also be gained from looking at the differential binding of factors to the alleles of rs3744457. The EGR family of zinc finger transcription factors – namely EGR1 and EGR2 – have predicted binding sites in the region containing rs3744457, with the C-allele abolishing or weakening EGR binding. Thus, ChIP experiments using antibodies against EGR1 and EGR2 may identify a *trans*-acting factor that differentially binds to the alleles of rs3744457, implicating this polymorphism and the domain in which it sits in the expression of two *MAPT* non-coding transcripts.

6.4.5 Chromatin immunoprecipitation (ChIP)

The ChIP experiments described in this project must be repeated in two further biological replicates before conclusions can be reasonably drawn. In particular, the β-actin IP must be optimised before its contribution to the Pol II-hnRNP U association can be analysed. Additional IPs may also prove valuable, and these have been indicated in the above sections as appropriate. The main extension to the rs242557 ChIP experiments, however, must be to properly quantify the PCR products produced following each IP in each cell line. This will provide a better and more accurate comparison of factor binding to the alleles of rs242557. This could be done by quantifying the intensity of the bands produced following resolution by agarose gel electrophoresis, as achieved using the ImageJ software for the exon 10 and N-terminal exon minigene quantifications, or by real-time quantitative PCR, as described previously.

6.5 Final comments

The aetiology of PSP has yet to be fully determined, though small pieces of the puzzle have been and continue to be revealed. This project aimed to link three factors known to be altered in PSP, with the hope of describing a basic mechanism to which future studies can build upon. To a certain extent, this has been achieved. The A-allele of rs242557, known to significantly increase PSP risk, was shown to reduce the strength of a transcriptional repressor domain, thereby inducing a significantly higher level of core promoter activity than the G-allele variant; a phenomenon also reported in conjunction with PSP. This increase in transcriptional activity was shown to be accompanied by a significant increase in relative 4R-tau mRNA transcripts which, if translated into a similar increase at the protein level, would account for the increased 4R-tau expression observed in the PSP brain.

Perhaps the most valuable aspect of this project, however, was the creation of *MAPT* minigenes. Following a few tweaks to correct erroneous splicing events, the *MAPT* minigene cell models have the potential to inform a wide variety of studies at multiple levels of expression. As interest grows in the role of *MAPT* exons 2 and 3 in neurodegeneration, these corrected models could prove a valuable tool. Mutation screening, regulatory studies, microRNA analysis, non-coding RNA function, alternative poly(A) site usage, gene loop formation, differential *trans*-acting factor binding and, of course, genetic analyses are all possible using these models. Discussions have already taken place regarding their use for screening potential therapeutic agents aimed at reducing 4R transcript levels.

Other groups are taking different approaches to clarifying the role of rs242557 in PSP and these studies should complement the work described here. In particular, Richard Wade-Martins and colleagues have described a viral method of delivering the whole of the *MAPT* gene (the H1 variant) when cloned into a bacterial artificial chromosome (‘infectious’ BAC or iBAC) construct [261]. They are currently investigating the effect of the rs242557 alleles – replacing the wildtype

G-allele of the iBAC with the A-allele by site-directed mutagenesis – on *MAPT* alternative splicing, though the results of this project have yet to be published. Once completed, this study should hopefully support and further inform the results described here.

Another emerging method uses TALE-like effector nuclease (TALEN) technology to edit the genome of patient-derived induced pluripotent stem cells (iPS cells). This provides another method of producing isogenic cell models that are free of confounding from, among other things, different genetic and epigenetic backgrounds [262].

Gene expression analyses using minigene methods are often considered too artificial, with *in vivo* methods preferred due to their greater and more immediate biological relevance. The recent publication by Trabzuni and colleagues suggests, however, that reported H1/H2 differences in *MAPT* transcription rate detected by *in vivo* analyses are, in fact, likely to be artefactual [133]. This is due to the discovery of in-probe polymorphisms in the expression arrays that are commonly used in this type of study. It has been shown here that cell type and differentiation status have a significant influence on gene expression and current analysis methods using brain tissue appear to lack sufficient resolution to account for this. Indeed, recent *in vivo* expression analyses of 4R-tau expression in six brain regions counters the results of the Trabzuni study, not only by reporting significant differences in expression between specific brain regions, but by finding a general trend for increased 4R-tau expression for H1 chromosomes compared to H2 chromosomes [134]. Perhaps the only way to resolve these conflicting reports regarding haplotype-specific *MAPT* expression is to look in individual cell populations by methods such as laser cell capture.

This study has, however, demonstrated the value of *in vitro* methods and shown how functional analysis at the basic sequence level can be used to build a mechanism that can further inform *in vivo* experiments. This is particularly true when investigating the role of common variation on gene expression as the effect

is often too subtle for accurate quantification *in vivo*. This project has shown the significant influence common variation can have on gene expression and has provided a blueprint for a cellular model that can investigate such variation at multiple levels. The models described here have the potential to significantly impact upon our understanding of the role of *MAPT* expression in neurodegerative disease.

References

1. Pandya-Jones, A. and D.L. Black, *Co-transcriptional splicing of constitutive and alternative exons.* RNA, 2009. **15**(10): p. 1896-908.

2. Maniatis, D. and R. Reed, *An extensive network of coupling among gene expression machines.* Nature, 2002. **416**(6880): p. 499-506.

3. Cramer, P., et al., *Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer.* Mol Cell, 1999. **4**(2): p. 251-8.

4. Montes, M., et al., *Functional coupling of transcription and splicing.* Gene, 2012.

5. Montes, M., et al., *TCERG1 regulates alternative splicing of the Bcl-x gene by modulating the rate of RNA polymerase II transcription.* Mol Cell Biol, 2012. **32**(4): p. 751-62.

6. Kornblihtt, A.R., *Coupling transcription and alternative splicing.* Adv Exp Med Biol, 2007. **623**: p. 175-89.

7. Kornblihtt, A.R., et al., *Multiple links between transcription and splicing.* RNA, 2004. **10**(10): p. 1489-98.

8. Bentley, D., *The mRNA assembly line: transcription and processing machines in the same factory.* Curr Opin Cell Biol, 2002. **14**(3): p. 336-42.

9. Ameur, A., et al., *Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain.* Nat Struct Mol Biol, 2011. **18**(12): p. 1435-40.

10. Cramer, P., et al., *Functional association between promoter structure and transcript alternative splicing.* Proc Natl Acad Sci U S A, 1997. **94**(21): p. 11456-60.

11. Rademakers, R., et al., *High-density SNP haplotyping suggests altered regulation of tau gene expression in progressive supranuclear palsy.* Hum Mol Genet, 2005. **14**(21): p. 3281-92.

12. Myers, A.J., et al., *The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts.* Neurobiol Dis, 2007. **25**(3): p. 561-70.

13. Croft, L., et al., *ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome.* Nat Genet, 2000. **24**(4): p. 340-1.

14. Aranda-Abreu, G.E., et al., *Possible Cis-acting signal that could be involved in the localization of different mRNAs in neuronal axons.* Theor Biol Med Model, 2005. **2**: p. 33.

15. Archambault, J. and J.D. Friesen, *Genetics of eukaryotic RNA polymerases I, II, and III.* Microbiol Rev, 1993. **57**(3): p. 703-24.

16. McCracken, S., et al., *5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II.* Genes Dev, 1997. **11**(24): p. 3306-18.

17. McCracken, S., et al., *The C-terminal domain of RNA polymerase II couples mRNA processing to transcription.* Nature, 1997. **385**(6614): p. 357-61.

18. Phatnani, H.P. and A.L. Greenleaf, *Phosphorylation and functions of the RNA polymerase II CTD.* Genes Dev, 2006. **20**(21): p. 2922-36.

19. Cho, E.J., et al., *Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain.* Genes Dev, 2001. **15**(24): p. 3319-29.

20. Buratowski, S., *The CTD code.* Nat Struct Biol, 2003. **10**(9): p. 679-80.

21. Egloff, S. and S. Murphy, *Cracking the RNA polymerase II CTD code.* Trends Genet, 2008. **24**(6): p. 280-8.

22. Komarnitsky, P., E.J. Cho, and S. Buratowski, *Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription.* Genes Dev, 2000. **14**(19): p. 2452-60.

23. Nikolov, D.B. and S.K. Burley, *RNA polymerase II transcription initiation: a structural view.* Proc Natl Acad Sci U S A, 1997. **94**(1): p. 15-22.

24. Sims, R.J., 3rd, R. Belotserkovskaya, and D. Reinberg, *Elongation by RNA polymerase II: the short and long of it.* Genes Dev, 2004. **18**(20): p. 2437-68.

25. Ho, C.K. and S. Shuman, *Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme.* Mol Cell, 1999. **3**(3): p. 405-11.

26. Schroeder, S.C., et al., *Dynamic association of capping enzymes with transcribing RNA polymerase II.* Genes Dev, 2000. **14**(19): p. 2435-40.

27. Kameoka, S., P. Duque, and M.M. Konarska, *p54(nrb) associates with the 5' splice site within large transcription/splicing complexes.* EMBO J, 2004. **23**(8): p. 1782-91.

28. Calvo, O. and J.L. Manley, *Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination.* Mol Cell, 2001. **7**(5): p. 1013-23.

29. Custodio, N., et al., *Inefficient processing impairs release of RNA from the site of transcription.* EMBO J, 1999. **18**(10): p. 2855-66.

30. Kramer, A., *The structure and function of proteins involved in mammalian pre-mRNA splicing.* Annu Rev Biochem, 1996. **65**: p. 367-409.

31. Will, C.L. and R. Luhrmann, *Spliceosomal UsnRNP biogenesis, structure and function.* Curr Opin Cell Biol, 2001. **13**(3): p. 290-301.

32. Gornemann, J., et al., *Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex.* Mol Cell, 2005. **19**(1): p. 53-63.

33. Berglund, J.A., N. Abovich, and M. Rosbash, *A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition.* Genes Dev, 1998. **12**(6): p. 858-67.

34. Berglund, J.A., et al., *The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC.* Cell, 1997. **89**(5): p. 781-7.

35. MacMillan, A.M., et al., *Dynamic association of proteins with the pre-mRNA branch region.* Genes Dev, 1994. **8**(24): p. 3008-20.

36. Izquierdo, J.M. and J. Valcarcel, *A simple principle to explain the evolution of pre-mRNA splicing.* Genes Dev, 2006. **20**(13): p. 1679-84.

37. Mendes Soares, L.M. and J. Valcarcel, *The expanding transcriptome: the genome as the 'Book of Sand'.* EMBO J, 2006. **25**(5): p. 923-31.

38. Wu, J.Y. and T. Maniatis, *Specific interactions between proteins implicated in splice site selection and regulated alternative splicing.* Cell, 1993. **75**(6): p. 1061-70.

39. Lin, S. and X.D. Fu, *SR proteins and related factors in alternative splicing.* Adv Exp Med Biol, 2007. **623**: p. 107-22.

40. Zhang, C., et al., *Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls.* Science, 2010. **329**(5990): p. 439-43.

41. Barash, Y., et al., *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-9.

42. Beyer, A.L. and Y.N. Osheim, *Splice site selection, rate of splicing, and alternative splicing on nascent transcripts.* Genes Dev, 1988. **2**(6): p. 754-65.

43. Osheim, Y.N., O.L. Miller, Jr., and A.L. Beyer, *Visualization of Drosophila melanogaster chorion genes undergoing amplification.* Mol Cell Biol, 1988. **8**(7): p. 2811-21.

44. Custodio, N., et al., *In vivo recruitment of exon junction complex proteins to transcription sites in mammalian cell nuclei.* RNA, 2004. **10**(4): p. 622-33.

45. Johnson, C., et al., *Tracking COL1A1 RNA in osteogenesis imperfecta. splice-defective transcripts initiate transport from the gene but are retained within the SC35 domain.* J Cell Biol, 2000. **150**(3): p. 417-32.

46. Fong, N., et al., *A 10 residue motif at the C-terminus of the RNA pol II CTD is required for transcription, splicing and 3' end processing.* EMBO J, 2003. **22**(16): p. 4274-82.

47. Rosonina, E. and B.J. Blencowe, *Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3'-end cleavage.* RNA, 2004. **10**(4): p. 581-9.

48. Du, L. and S.L. Warren, *A functional interaction between the carboxy-terminal domain of RNA polymerase II and pre-mRNA splicing.* J Cell Biol, 1997. **136**(1): p. 5-18.

49. Misteli, T. and D.L. Spector, *RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo.* Mol Cell, 1999. **3**(6): p. 697-705.

50. de la Mata, M. and A.R. Kornblihtt, *RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20.* Nat Struct Mol Biol, 2006. **13**(11): p. 973-80.

51. Hirose, Y., R. Tacke, and J.L. Manley, *Phosphorylated RNA polymerase II stimulates pre-mRNA splicing.* Genes Dev, 1999. **13**(10): p. 1234-9.

52. Das, R., et al., *SR proteins function in coupling RNAP II transcription to pre-mRNA splicing.* Mol Cell, 2007. **26**(6): p. 867-81.

53. Kotovic, K.M., et al., *Cotranscriptional recruitment of the U1 snRNP to intron-containing genes in yeast.* Mol Cell Biol, 2003. **23**(16): p. 5768-79.

54. Bourquin, J.P., et al., *A serine/arginine-rich nuclear matrix cyclophilin interacts with the C-terminal domain of RNA polymerase II.* Nucleic Acids Res, 1997. **25**(11): p. 2055-61.

55. Yuryev, A., et al., *The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins.* Proc Natl Acad Sci U S A, 1996. **93**(14): p. 6975-80.

56. Das, R., et al., *Functional coupling of RNAP II transcription to spliceosome assembly.* Genes Dev, 2006. **20**(9): p. 1100-9.

57. Lin, S., et al., *The splicing factor SC35 has an active role in transcriptional elongation.* Nat Struct Mol Biol, 2008. **15**(8): p. 819-26.

58. O'Brien, T. and J.T. Lis, *RNA polymerase II pauses at the 5' end of the transcriptionally induced Drosophila hsp70 gene.* Mol Cell Biol, 1991. **11**(10): p. 5285-90.

59. Bentley, D.L. and M. Groudine, *A block to elongation is largely responsible for decreased transcription of c-myc in differentiated HL60 cells.* Nature, 1986. **321**(6071): p. 702-6.

60. Kao, S.Y., et al., *Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product.* Nature, 1987. **330**(6147): p. 489-93.

61. Krumm, A., et al., *The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region.* Genes Dev, 1992. **6**(11): p. 2201-13.

62. Kadener, S., et al., *Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing.* EMBO J, 2001. **20**(20): p. 5759-68.

63. Eperon, L.P., et al., *Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase?* Cell, 1988. **54**(3): p. 393-401.

64. Solnick, D., *Alternative splicing caused by RNA secondary structure.* Cell, 1985. **43**(3 Pt 2): p. 667-76.

65. Solnick, D. and S.I. Lee, *Amount of RNA secondary structure required to induce an alternative splice.* Mol Cell Biol, 1987. **7**(9): p. 3194-8.

66. Eperon, L.P., J.P. Estibeiro, and I.C. Eperon, *The role of nucleotide sequences in splice site selection in eukaryotic pre-messenger RNA.* Nature, 1986. **324**(6094): p. 280-2.

67. Roberts, G.C., et al., *Co-transcriptional commitment to alternative splice site selection.* Nucleic Acids Res, 1998. **26**(24): p. 5568-72.

68. de la Mata, M., et al., *A slow RNA polymerase II affects alternative splicing in vivo.* Mol Cell, 2003. **12**(2): p. 525-32.

69. Coulter, D.E. and A.L. Greenleaf, *Properties of mutationally altered RNA polymerases II of Drosophila.* J Biol Chem, 1982. **257**(4): p. 1945-52.

70. Glover-Cutter, K., et al., *RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes.* Nat Struct Mol Biol, 2008. **15**(1): p. 71-8.

71. Carrillo Oesterreich, F., S. Preibisch, and K.M. Neugebauer, *Global analysis of nascent RNA reveals transcriptional pausing in terminal exons.* Mol Cell, 2010. **40**(4): p. 571-81.

72. Alexander, R.D., et al., *Splicing-dependent RNA polymerase pausing in yeast.* Mol Cell. **40**(4): p. 582-93.

73. Izban, M.G. and D.S. Luse, *Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing.* Genes Dev, 1991. **5**(4): p. 683-96.

74. Kouzarides, T., *SnapShot: Histone-modifying enzymes.* Cell, 2007. **131**(4): p. 822.

75. Orphanides, G. and D. Reinberg, *RNA polymerase II elongation through chromatin.* Nature, 2000. **407**(6803): p. 471-5.

76. Gunderson, F.Q. and T.L. Johnson, *Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly.* PLoS Genet, 2009. **5**(10): p. e1000682.

77. Schwartz, S., E. Meshorer, and G. Ast, *Chromatin organization marks exon-intron structure.* Nat Struct Mol Biol, 2009. **16**(9): p. 990-5.

78. Tilgner, H., et al., *Nucleosome positioning as a determinant of exon recognition.* Nat Struct Mol Biol, 2009. **16**(9): p. 996-1001.

79. Listerman, I., A.K. Sapra, and K.M. Neugebauer, *Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells.* Nat Struct Mol Biol, 2006. **13**(9): p. 815-22.

80. Batsche, E., M. Yaniv, and C. Muchardt, *The human SWI/SNF subunit Brm is a regulator of alternative splicing.* Nat Struct Mol Biol, 2006. **13**(1): p. 22-9.

81. Bannister, A.J., R. Schneider, and T. Kouzarides, *Histone methylation: dynamic or static?* Cell, 2002. **109**(7): p. 801-6.

82. Goto, Y., et al., *Differential patterns of histone methylation and acetylation distinguish active and repressed alleles at X-linked genes.* Cytogenet Genome Res, 2002. **99**(1-4): p. 66-74.

83. Sugiyama, K., et al., *Aurora-B associated protein phosphatases as negative regulators of kinase activation.* Oncogene, 2002. **21**(20): p. 3103-11.

84. Hnilicova, J., et al., *Histone deacetylase activity modulates alternative splicing.* PLoS One. **6**(2): p. e16727.

85. Nogues, G., et al., *Transcriptional activators differ in their abilities to control alternative splicing.* J Biol Chem, 2002. **277**(45): p. 43110-4.

86. Loomis, R.J., et al., *Chromatin binding of SRp20 and ASF/SF2 and dissociation from mitotic chromosomes is modulated by histone H3 serine 10 phosphorylation.* Mol Cell, 2009. **33**(4): p. 450-61.

87. Sims, R.J., 3rd, et al., *Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing.* Mol Cell, 2007. **28**(4): p. 665-76.

88. Kolasinska-Zwierz, P., et al., *Differential chromatin marking of introns and expressed exons by H3K36me3.* Nat Genet, 2009. **41**(3): p. 376-81.

89. Hon, G., W. Wang, and B. Ren, *Discovery and annotation of functional chromatin signatures in the human genome.* PLoS Comput Biol, 2009. **5**(11): p. e1000566.

90. Carstens, R.P., E.J. Wagner, and M.A. Garcia-Blanco, *An intronic splicing silencer causes skipping of the IIIb exon of fibroblast growth factor receptor 2 through involvement of polypyrimidine tract binding protein.* Mol Cell Biol, 2000. **20**(19): p. 7388-400.

91. Luco, R.F., et al., *Regulation of alternative splicing by histone modifications.* Science. **327**(5968): p. 996-1000.

92. Shukla, S., et al., *CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.* Nature. **479**(7371): p. 74-9.

93. Lee, J.S., E. Smith, and A. Shilatifard, *The language of histone crosstalk.* Cell, 2010. **142**(5): p. 682-5.

94. Li, B., M. Carey, and J.L. Workman, *The role of chromatin during transcription.* Cell, 2007. **128**(4): p. 707-19.

95. Sisodia, S.S., B. Sollner-Webb, and D.W. Cleveland, *Specificity of RNA maturation pathways: RNAs transcribed by RNA polymerase III are not substrates for splicing or polyadenylation.* Mol Cell Biol, 1987. **7**(10): p. 3602-12.

96. Dickson, D.W., et al., *Neuropathology of variants of progressive supranuclear palsy.* Curr Opin Neurol, 2010. **23**(4): p. 394-400.

97. Vandrovcova, J., et al., *Disentangling the Role of the tau Gene Locus in Sporadic Tauopathies.* Curr Alzheimer Res, 2010.

98. Hardy, J., et al., *Tangle diseases and the tau haplotypes.* Alzheimer Dis Assoc Disord, 2006. **20**(1): p. 60-2.

99. Baker, M., et al., *Association of an extended haplotype in the tau gene with progressive supranuclear palsy.* Hum Mol Genet, 1999. **8**(4): p. 711-5.

100. Bennett, P., et al., *Direct genetic evidence for involvement of tau in progressive supranuclear palsy. European Study Group on Atypical Parkinsonism Consortium.* Neurology, 1998. **51**(4): p. 982-5.

101. Higgins, J.J., et al., *An extended 5'-tau susceptibility haplotype in progressive supranuclear palsy.* Neurology, 2000. **55**(9): p. 1364-7.

102. Morris, H.R., et al., *The tau gene A0 polymorphism in progressive supranuclear palsy and related neurodegenerative diseases.* J Neurol Neurosurg Psychiatry, 1999. **66**(5): p. 665-7.

103. Oliva, R., et al., *Significant changes in the tau A0 and A3 alleles in progressive supranuclear palsy and improved genotyping by silver detection.* Arch Neurol, 1998. **55**(8): p. 1122-4.

104. Hoglinger, G.U., et al., *Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy.* Nat Genet, 2011. **43**(7): p. 699-705.

105. Dickson, D.W., R. Rademakers, and M.L. Hutton, *Progressive supranuclear palsy: pathology and genetics.* Brain Pathol, 2007. **17**(1): p. 74-82.

106. Choumert, A., et al., *G303V tau mutation presenting with progressive supranuclear palsy-like features.* Mov Disord, 2011.

107. Boeve, B.F., *Progressive supranuclear palsy.* Parkinsonism Relat Disord, 2012. **18 Suppl 1**: p. S192-4.

108. Stamelou, M., et al., *Rational therapeutic approaches to progressive supranuclear palsy.* Brain, 2010.

109. Hoppitt, T., et al., *A systematic review of the incidence and prevalence of long-term neurological conditions in the UK.* Neuroepidemiology, 2011. **36**(1): p. 19-28.

110. Litvan, I., *Update on progressive supranuclear palsy.* Curr Neurol Neurosci Rep, 2004. **4**(4): p. 296-302.

111. Hayesmoore, J.B., et al., *The effect of age and the H1c MAPT haplotype on MAPT expression in human brain.* Neurobiol Aging, 2009. **30**(10): p. 1652-6.

112. Lang, A.E., *Treatment of progressive supranuclear palsy and corticobasal degeneration.* Mov Disord, 2005. **20 Suppl 12**: p. S83-91.

113. Rohrer, J.D., et al., *Novel L284R MAPT mutation in a family with an autosomal dominant progressive supranuclear palsy syndrome.* Neurodegener Dis, 2011. **8**(3): p. 149-52.

114. Ogaki, K., et al., *Analyses of the MAPT, PGRN, and C9orf72 mutations in Japanese patients with FTLD, PSP, and CBS.* Parkinsonism Relat Disord, 2012.

115. McKee, A.C., et al., *TDP-43 proteinopathy and motor neuron disease in chronic traumatic encephalopathy.* J Neuropathol Exp Neurol, 2010. **69**(9): p. 918-29.

116. Golbe, L.I., et al., *Follow-up study of risk factors in progressive supranuclear palsy.* Neurology, 1996. **47**(1): p. 148-54.

117. Flaherty, D.B., et al., *Phosphorylation of human tau protein by microtubule-associated kinases: GSK3beta and cdk5 are key participants.* J Neurosci Res, 2000. **62**(3): p. 463-72.

118. Grover, A., et al., *5' splice site mutations in tau associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10.* J Biol Chem, 1999. **274**(21): p. 15134-43.

119. Caceres, A. and K.S. Kosik, *Inhibition of neurite polarity by tau antisense oligonucleotides in primary cerebellar neurons.* Nature, 1990. **343**(6257): p. 461-3.

120. Shin, R.W., et al., *Hydrated autoclave pretreatment enhances tau immunoreactivity in formalin-fixed normal and Alzheimer's disease brain tissues.* Lab Invest, 1991. **64**(5): p. 693-702.

121. Binder, L.I., A. Frankfurter, and L.I. Rebhun, *The distribution of tau in the mammalian central nervous system.* J Cell Biol, 1985. **101**(4): p. 1371-8.

122. Gallo, J.M., W. Noble, and T.R. Martin, *RNA and protein-dependent mechanisms in tauopathies: consequences for therapeutic strategies.* Cell Mol Life Sci, 2007. **64**(13): p. 1701-14.

123. Neve, R.L., et al., *Identification of cDNA clones for the human microtubule-associated protein tau and chromosomal localization of the genes for tau and microtubule-associated protein 2.* Brain Res, 1986. **387**(3): p. 271-80.

124. D'Souza, I. and G.D. Schellenberg, *Regulation of tau isoform expression and dementia.* Biochim Biophys Acta, 2005. **1739**(2-3): p. 104-15.

125. Cleveland, D.W., S.Y. Hwo, and M.W. Kirschner, *Purification of tau, a microtubule-associated protein that induces assembly of microtubules from purified tubulin.* J Mol Biol, 1977. **116**(2): p. 207-25.

126. Cleveland, D.W., S.Y. Hwo, and M.W. Kirschner, *Physical and chemical properties of purified tau factor and the role of tau in microtubule assembly.* J Mol Biol, 1977. **116**(2): p. 227-47.

127. Takei, Y., et al., *Defects in axonal elongation and neuronal migration in mice with disrupted tau and map1b genes.* J Cell Biol, 2000. **150**(5): p. 989-1000.

128. Harada, A., et al., *Altered microtubule organization in small-calibre axons of mice lacking tau protein.* Nature, 1994. **369**(6480): p. 488-91.

129. Iqbal, K., et al., *Mechanisms of tau-induced neurodegeneration.* Acta Neuropathol, 2009. **118**(1): p. 53-69.

130. Panda, D., et al., *Differential regulation of microtubule dynamics by three- and four-repeat tau: implications for the onset of neurodegenerative disease.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9548-53.

131. Voss, K., et al., *Hsp70 alters tau function and aggregation in an isoform specific manner.* Biochemistry, 2012. **51**(4): p. 888-98.

132. Takuma, H., S. Arawaka, and H. Mori, *Isoforms changes of tau protein during development in various species.* Brain Res Dev Brain Res, 2003. **142**(2): p. 121-7.

133. Trabzuni, D., et al., *MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies.* Hum Mol Genet, 2012. **21**(18): p. 4094-103.

134. Majounie, E., et al., *Variation in tau isoform expression in different brain regions and disease states.* Neurobiol Aging.

135. Goode, B.L., et al., *Structural and functional differences between 3-repeat and 4-repeat tau isoforms. Implications for normal tau function and the onset of neurodegenetative disease.* J Biol Chem, 2000. **275**(49): p. 38182-9.

136. Kanemaru, K., et al., *Fetal-type phosphorylation of the tau in paired helical filaments.* J Neurochem, 1992. **58**(5): p. 1667-75.

137. Matsuo, E.S., et al., *Biopsy-derived adult human brain tau is phosphorylated at many of the same sites as Alzheimer's disease paired helical filament tau.* Neuron, 1994. **13**(4): p. 989-1002.

138. Mawal-Dewan, M., et al., *The phosphorylation state of tau in the developing rat brain is regulated by phosphoprotein phosphatases.* J Biol Chem, 1994. **269**(49): p. 30981-7.

139. Billingsley, M.L. and R.L. Kincaid, *Regulated phosphorylation and dephosphorylation of tau protein: effects on microtubule interaction, intracellular trafficking and neurodegeneration.* Biochem J, 1997. **323 (Pt 3)**: p. 577-91.

140. Lund, H., et al., *Tau-tubulin kinase 1 expression, phosphorylation and co-localization with phospho-Ser422 tau in the Alzheimer's disease brain.* Brain Pathol, 2012.

141. Hanger, D.P., et al., *New phosphorylation sites identified in hyperphosphorylated tau (paired helical filament-tau) from Alzheimer's disease brain using nanoelectrospray mass spectrometry.* J Neurochem, 1998. **71**(6): p. 2465-76.

142. Vega, I.E., et al., *Increase in tau tyrosine phosphorylation correlates with the formation of tau aggregates.* Brain Res Mol Brain Res, 2005. **138**(2): p. 135-44.

143. Taniguchi, T., et al., *Phosphorylation of tau is regulated by PKN.* J Biol Chem, 2001. **276**(13): p. 10025-31.

144. Derkinderen, P., et al., *Tyrosine 394 is phosphorylated in Alzheimer's paired helical filament tau and in fetal tau with c-Abl as the candidate tyrosine kinase.* J Neurosci, 2005. **25**(28): p. 6584-93.

145. Hanger, D.P., A. Seereeram, and W. Noble, *Mediators of tau phosphorylation in the pathogenesis of Alzheimer's disease.* Expert Rev Neurother, 2009. **9**(11): p. 1647-66.

146. Lebouvier, T., et al., *The microtubule-associated protein tau is also phosphorylated on tyrosine.* J Alzheimers Dis, 2009. **18**(1): p. 1-9.

147. Chen, X., et al., *Study of tauopathies by comparing Drosophila and human tau in Drosophila.* Cell Tissue Res, 2007. **329**(1): p. 169-78.

148. Dou, F., et al., *Chaperones increase association of tau protein with microtubules.* Proc Natl Acad Sci U S A, 2003. **100**(2): p. 721-6.

149. Andreadis, A., W.M. Brown, and K.S. Kosik, *Structure and novel exons of the human tau gene.* Biochemistry, 1992. **31**(43): p. 10626-33.

150. Leroy, O., et al., *Brain-specific change in alternative splicing of Tau exon 6 in myotonic dystrophy type 1.* Biochim Biophys Acta, 2006. **1762**(4): p. 460-7.

151. Sergeant, N., et al., *Dysregulation of human brain microtubule-associated tau mRNA maturation in myotonic dystrophy type 1.* Hum Mol Genet, 2001. **10**(19): p. 2143-55.

152. Jiang, H., et al., *Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons.* Hum Mol Genet, 2004. **13**(24): p. 3079-88.

153. Goedert, M. and R. Jakes, *Expression of separate isoforms of human tau protein: correlation with the tau pattern in brain and effects on tubulin polymerization.* EMBO J, 1990. **9**(13): p. 4225-30.

154. Mukhopadhyay, R. and J.H. Hoh, *AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force.* FEBS Lett, 2001. **505**(3): p. 374-8.

155. Andreadis, A., J.A. Broderick, and K.S. Kosik, *Relative exon affinities and suboptimal splice site signals lead to non-equivalence of two cassette exons.* Nucleic Acids Res, 1995. **23**(17): p. 3585-93.

156. Chen, S., et al., *MAPT isoforms: differential transcriptional profiles related to 3R and 4R splice variants.* J Alzheimers Dis, 2012. **22**(4): p. 1313-29.

157. Smith, P.Y., et al., *MicroRNA-132 loss is associated with tau exon 10 inclusion in progressive supranuclear palsy.* Hum Mol Genet, 2011. **20**(20): p. 4016-24.

158. Espinoza, M., et al., *Differential incorporation of tau isoforms in Alzheimer's disease.* J Alzheimers Dis, 2008. **14**(1): p. 1-16.

159. Iseki, E., et al., *Immunohistochemical investigation of neurofibrillary tangles and their tau isoforms in brains of limbic neurofibrillary tangle dementia.* Neurosci Lett, 2006. **405**(1-2): p. 29-33.

160. Jellinger, K.A. and J. Attems, *Neurofibrillary tangle-predominant dementia: comparison with classical Alzheimer disease.* Acta Neuropathol, 2007. **113**(2): p. 107-17.

161. Kitamura, T., et al., *Relationship between microtubule-binding repeats and morphology of neurofibrillary tangle in Alzheimer's disease.* Acta Neurol Scand, 2005. **112**(5): p. 327-34.

162. Lace, G., et al., *Hippocampal tau pathology is related to neuroanatomical connections: an ageing population-based study.* Brain, 2009. **132**(Pt 5): p. 1324-34.

163. Uchihara, T., et al., *Tangle evolution linked to differential 3- and 4-repeat tau isoform deposition: a double immunofluorolabeling study using two monoclonal antibodies.* Histochem Cell Biol, 2012.

164. Uchihara, T., et al., *Specific detection of pathological three-repeat tau after pretreatment with potassium permanganate and oxalic acid in PSP/CBD brains.* Brain Pathol, 2011. **21**(2): p. 180-8.

165. Yoshida, M., *Cellular tau pathology and immunohistochemical study of tau isoforms in sporadic tauopathies.* Neuropathology, 2006. **26**(5): p. 457-70.

166. Zody, M.C., et al., *Evolutionary toggling of the MAPT 17q21.31 inversion region.* Nat Genet, 2008. **40**(9): p. 1076-83.

167. Hardy, J., et al., *Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens.* Biochem Soc Trans, 2005. **33**(Pt 4): p. 582-5.

168. Stefansson, H., et al., *A common inversion under selection in Europeans.* Nat Genet, 2005. **37**(2): p. 129-37.

169. Pittman, A.M., et al., *The structure of the tau haplotype in controls and in progressive supranuclear palsy.* Hum Mol Genet, 2004. **13**(12): p. 1267-74.

170. Yoshiyama, Y., V.M. Lee, and J.Q. Trojanowski, *Frontotemporal dementia and tauopathy.* Curr Neurol Neurosci Rep, 2001. **1**(5): p. 413-21.

171. Hasegawa, M., et al., *FTDP-17 mutations N279K and S305N in tau produce increased splicing of exon 10.* FEBS Lett, 1999. **443**(2): p. 93-6.

172. Niblock, M. and J.M. Gallo, *Tau alternative splicing in familial and sporadic tauopathies.* Biochem Soc Trans, 2012. **40**(4): p. 677-80.

173. Anfossi, M., et al., *MAPT V363I variation in a sporadic case of frontotemporal dementia: variable penetrant mutation or rare polymorphism?* Alzheimer Dis Assoc Disord, 2012. **25**(1): p. 96-9.

174. Conrad, C., et al., *Genetic evidence for the involvement of tau in progressive supranuclear palsy.* Ann Neurol, 1997. **41**(2): p. 277-81.

175. Vandrovcova, J., et al., *Association of MAPT haplotype-tagging SNPs with sporadic Parkinson's disease.* Neurobiol Aging, 2009. **30**(9): p. 1477-82.

176. Houlden, H., et al., *Corticobasal degeneration and progressive supranuclear palsy share a common tau haplotype.* Neurology, 2001. **56**(12): p. 1702-6.

177. Myers, A.J., et al., *The H1c haplotype at the MAPT locus is associated with Alzheimer's disease.* Hum Mol Genet, 2005. **14**(16): p. 2399-404.

178. Zhang, J., et al., *The tau gene haplotype h1 confers a susceptibility to Parkinson's disease.* Eur Neurol, 2005. **53**(1): p. 15-21.

179. Healy, D.G., et al., *Tau gene and Parkinson's disease: a case-control study and meta-analysis.* J Neurol Neurosurg Psychiatry, 2004. **75**(7): p. 962-5.

180. Simon-Sanchez, J., et al., *Genome-wide association study reveals genetic risk underlying Parkinson's disease.* Nat Genet, 2009.

181. Pittman, A.M., et al., *Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration.* J Med Genet, 2005. **42**(11): p. 837-46.

182. Williams, D.R., et al., *Genetic variation at the tau locus and clinical syndromes associated with progressive supranuclear palsy.* Mov Disord, 2007. **22**(6): p. 895-7.

183. Zou, F., et al., *Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants.* PLoS Genet, 2012. **8**(6): p. e1002707.

184. Sundar, P.D., et al., *Two sites in the MAPT region confer genetic risk for Guam ALS/PDC and dementia.* Hum Mol Genet, 2007. **16**(3): p. 295-306.

185. Laws, S.M., et al., *Fine mapping of the MAPT locus using quantitative trait analysis identifies possible causal variants in Alzheimer's disease.* Mol Psychiatry, 2007. **12**(5): p. 510-7.

186. Liu, Q.Y., et al., *An exploratory study on STX6, MOBP, MAPT, and EIF2AK3 and late-onset Alzheimer's disease.* Neurobiol Aging, 2012.

187. Abraham, R., et al., *An association study of common variation at the MAPT locus with late-onset Alzheimer's disease.* Am J Med Genet B Neuropsychiatr Genet, 2009. **150B**(8): p. 1152-1155.

188. Mukherjee, O., et al., *Haplotype-based association analysis of the MAPT locus in late onset Alzheimer's disease.* BMC Genet, 2007. **8**: p. 3.

189. Seto-Salvia, N., et al., *Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes.* Arch Neurol. **68**(3): p. 359-64.

190. Mateo, I., et al., *Synergistic effect of heme oxygenase-1 and tau genetic variants on Alzheimer's disease risk.* Dement Geriatr Cogn Disord, 2008. **26**(4): p. 339-42.

191. Perry, G., et al., *Is oxidative damage the fundamental pathogenic mechanism of Alzheimer's and other neurodegenerative diseases?* Free Radic Biol Med, 2002. **33**(11): p. 1475-9.

192. Melov, S., et al., *Mitochondrial oxidative stress causes hyperphosphorylation of tau.* PLoS One, 2007. **2**(6): p. e536.

193. Wider, C., et al., *An evaluation of the impact of MAPT, SNCA and APOE on the burden of Alzheimer's and Lewy body pathology.* J Neurol Neurosurg Psychiatry, 2012. **83**(4): p. 424-9.

194. Wider, C., et al., *Association of the MAPT locus with Parkinson's disease.* Eur J Neurol, 2010.

195. Refenes, N., et al., *Role of the H1 haplotype of microtubule-associated protein tau (MAPT) gene in Greek patients with Parkinson's disease.* BMC Neurol, 2009. **9**: p. 26.

196. Fung, H.C., et al., *Association of tau haplotype-tagging polymorphisms with Parkinson's disease in diverse ethnic Parkinson's disease cohorts.* Neurodegener Dis, 2006. **3**(6): p. 327-33.

197. Elbaz, A., et al., *Independent and joint effects of the MAPT and SNCA genes in Parkinson disease.* Ann Neurol. **69**(5): p. 778-92.

198. Trotta, L., et al., *SNCA and MAPT genes: Independent and joint effects in Parkinson disease in the Italian population.* Parkinsonism Relat Disord. **18**(3): p. 257-62.

199. Lill, C.M., et al., *Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database.* PLoS Genet, 2012. **8**(3): p. e1002548.

200. Laws, S.M., et al., *Genetic analysis of MAPT haplotype diversity in frontotemporal dementia.* Neurobiol Aging, 2008. **29**(8): p. 1276-8.

201. Compta, Y., et al., *High cerebrospinal tau levels are associated with the rs242557 tau gene variant and low cerebrospinal beta-amyloid in Parkinson disease.* Neurosci Lett, 2011. **487**(2): p. 169-73.

202. Kauwe, J.S., et al., *Variation in MAPT is associated with cerebrospinal fluid tau levels in the presence of amyloid-beta deposition.* Proc Natl Acad Sci U S A, 2008. **105**(23): p. 8050-4.

203. Ezquerra, M., et al., *Different MAPT haplotypes are associated with Parkinson's disease and progressive supranuclear palsy.* Neurobiol Aging, 2011.

204. Caffrey, T.M., et al., *Haplotype-specific expression of exon 10 at the human MAPT locus.* Hum Mol Genet, 2006. **15**(24): p. 3529-37.

205. Caffrey, T.M., C. Joachim, and R. Wade-Martins, *Haplotype-specific expression of the N-terminal exons 2 and 3 at the human MAPT locus.* Neurobiol Aging, 2008. **29**(12): p. 1923-9.

206. Andreadis, A., et al., *The exon trapping assay partly discriminates against alternatively spliced exons.* Nucleic Acids Res, 1993. **21**(9): p. 2217-21.

207. Jiang, Z., et al., *Aberrant splicing of tau pre-mRNA caused by intronic mutations associated with the inherited dementia frontotemporal dementia with parkinsonism linked to chromosome 17.* Mol Cell Biol, 2000. **20**(11): p. 4036-48.

208. Wang, J., et al., *Tau exon 10, whose missplicing causes frontotemporal dementia, is regulated by an intricate interplay of cis elements and trans factors.* J Neurochem, 2004. **88**(5): p. 1078-90.

209. Anfossi, M., et al., *Compound heterozygosity of 2 novel MAPT mutations in frontotemporal dementia.* Neurobiol Aging, 2011. **32**(4): p. 757 e1-757 e11.

210. Yu, Q., J. Guo, and J. Zhou, *A minimal length between tau exon 10 and 11 is required for correct splicing of exon 10.* J Neurochem, 2004. **90**(1): p. 164-72.

211. Wang, Y., et al., *Tau exons 2 and 10, which are misregulated in neurodegenerative diseases, are partly regulated by silencers which bind a SRp30c.SRp55 complex that either recruits or antagonizes htra2beta1.* J Biol Chem, 2005. **280**(14): p. 14230-9.

212. Li, K., M.C. Arikan, and A. Andreadis, *Modulation of the membrane-binding domain of tau protein: splicing regulation of exon 2.* Brain Res Mol Brain Res, 2003. **116**(1-2): p. 94-105.

213. Arikan, M.C., et al., *Modulation of the membrane-binding projection domain of tau protein: splicing regulation of exon 3.* Brain Res Mol Brain Res, 2002. **101**(1-2): p. 109-21.

214. Wei, M.L. and A. Andreadis, *Splicing of a regulated exon reveals additional complexity in the axonal microtubule-associated protein tau.* J Neurochem, 1998. **70**(4): p. 1346-56.

215. Gao, Q.S., et al., *Complex regulation of tau exon 10, whose missplicing causes frontotemporal dementia.* J Neurochem, 2000. **74**(2): p. 490-500.

216. Kar, A., et al., *RBM4 interacts with an intronic element and stimulates tau exon 10 inclusion.* J Biol Chem, 2006. **281**(34): p. 24479-88.

217. Wang, Y., et al., *Heterogeneous nuclear ribonucleoprotein E3 modestly activates splicing of tau exon 10 via its proximal downstream intron, a hotspot for frontotemporal dementia mutations.* Gene, 2010. **451**(1-2): p. 23-31.

218. Wang, Y., et al., *An SRp75/hnRNPG complex interacting with hnRNPE2 regulates the 5' splice site of tau exon 10, whose misregulation causes frontotemporal dementia.* Gene, 2011. **485**(2): p. 130-8.

219. Donahue, C.P., et al., *Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing.* J Biol Chem, 2006. **281**(33): p. 23302-6.

220. Ko, L.W., et al., *Assembly of filamentous tau aggregates in human neuronal cells.* J Alzheimers Dis, 2004. **6**(6): p. 605-22; discussion 673-81.

221. Bhaskar, K., et al., *Tyrosine phosphorylation of tau accompanies disease progression in transgenic mouse models of tauopathy.* Neuropathol Appl Neurobiol, 2010.

222. Lapointe, N.E., et al., *Tau 6D and 6P isoforms inhibit polymerization of full-length tau in vitro.* Biochemistry, 2009. **48**(51): p. 12290-7.

223. Rodriguez-Martin, T., et al., *Correction of tau mis-splicing caused by FTDP-17 MAPT mutations by spliceosome-mediated RNA trans-splicing.* Hum Mol Genet, 2009. **18**(17): p. 3266-73.

224. Rodriguez-Martin, T., et al., *Reprogramming of tau alternative splicing by spliceosome-mediated RNA trans-splicing: implications for tauopathies.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15659-64.

225. Kalbfuss, B., S.A. Mabon, and T. Misteli, *Correction of alternative splicing of tau in frontotemporal dementia and parkinsonism linked to chromosome 17.* J Biol Chem, 2001. **276**(46): p. 42986-93.

226. Dawson, H.N., et al., *The tau N279K exon 10 splicing mutation recapitulates frontotemporal dementia and parkinsonism linked to chromosome 17 tauopathy in a mouse model.* J Neurosci, 2007. **27**(34): p. 9155-68.

227. Ezquerra, M., et al., *Different MAPT haplotypes are associated with Parkinson's disease and progressive supranuclear palsy.* Neurobiol Aging, 2009.

228. Lapidot, M. and Y. Pilpel, *Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms.* EMBO Rep, 2006. **7**(12): p. 1216-22.

229. Sadot, E., et al., *Short- and long-term mechanisms of tau regulation in PC12 cells.* J Cell Sci, 1995. **108 (Pt 8)**: p. 2857-64.

230. Buratti, E., et al., *Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping.* EMBO J, 2001. **20**(7): p. 1774-84.

231. Abhyankar, M.M., C. Urekar, and P.P. Reddi, *A novel CpG-free vertebrate insulator silences the testis-specific SP-10 gene in somatic tissues: role for TDP-43 in insulator function.* J Biol Chem, 2007. **282**(50): p. 36143-54.

232. Tan-Wong, S.M., et al., *Gene loops enhance transcriptional directionality.* Science, 2012. **338**(6107): p. 671-5.

233. Aranda-Abreu, G.E., et al., *Embryonic lethal abnormal vision-like RNA-binding proteins regulate neurite outgrowth and tau expression in PC12 cells.* J Neurosci, 1999. **19**(16): p. 6907-17.

234. Cooper, T.A., *Use of minigene systems to dissect alternative splicing elements.* Methods, 2005. **37**(4): p. 331-40.

235. Hartley, J.L., G.F. Temple, and M.A. Brasch, *DNA cloning using in vitro site-specific recombination.* Genome Res, 2000. **10**(11): p. 1788-95.

236. Sasaki, Y., et al., *Multi-gene gateway clone design for expression of multiple heterologous genes in living cells: eukaryotic clones containing two and three ORF multi-gene cassettes expressed from a single promoter.* J Biotechnol, 2008. **136**(3-4): p. 103-12.

237. Sasaki, Y., et al., *Evidence for high specificity and efficiency of multiple recombination signals in mixed DNA cloning by the Multisite Gateway system.* J Biotechnol, 2004. **107**(3): p. 233-43.

238. Bushman, W., et al., *Control of directionality in lambda site specific recombination.* Science, 1985. **230**(4728): p. 906-11.

239. Landy, A., *Dynamic, structural, and regulatory aspects of lambda site-specific recombination.* Annu Rev Biochem, 1989. **58**: p. 913-49.

240. Weisberg, R.A., et al., *Role for DNA homology in site-specific recombination. The isolation and characterization of a site affinity mutant of coliphage lambda.* J Mol Biol, 1983. **170**(2): p. 319-42.

241. Potter, C.J. and L. Luo, *Splinkerette PCR for mapping transposable elements in Drosophila.* PLoS One, 2010. **5**(4): p. e10168.

242. Anaya, J.F., R. De Silva, and A.J. Lees, *Tau gene promoter rs242557 and allele-specific protein binding.* Translational Neuroscience, 2011. **2**(2): p. 176-205.

243. Kukalev, A., et al., *Actin and hnRNP U cooperate for productive transcription by RNA polymerase II.* Nat Struct Mol Biol, 2005. **12**(3): p. 238-44.

244. Hager, G.L., et al., *Dynamics of nuclear receptor movement and transcription.* Biochim Biophys Acta, 2004. **1677**(1-3): p. 46-51.

245. Barber, R.D., et al., *GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues.* Physiol Genomics, 2005. **21**(3): p. 389-95.

246. Sun, M., et al., *Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity.* Genome Res, 2006. **16**(7): p. 922-33.

247. Hastings, M.L., et al., *Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels.* Nucleic Acids Res, 1997. **25**(21): p. 4296-300.

248. Ou, S.H., et al., *Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs.* J Virol, 1995. **69**(6): p. 3584-96.

249. Ayala, Y.M., T. Misteli, and F.E. Baralle, *TDP-43 regulates retinoblastoma protein phosphorylation through the repression of cyclin-dependent kinase 6 expression.* Proc Natl Acad Sci U S A, 2008. **105**(10): p. 3785-9.

250. Neumann, M., et al., *Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis.* Science, 2006. **314**(5796): p. 130-3.

251. 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, 2006. **351**(3): p. 602-11.

252. Yokota, O., et al., *Phosphorylated TDP-43 pathology and hippocampal sclerosis in progressive supranuclear palsy.* Acta Neuropathol, 2010. **120**(1): p. 55-66.

253. Mapendano, C.K., et al., *Crosstalk between mRNA 3' end processing and transcription initiation.* Mol Cell, 2010. **40**(3): p. 410-22.

254. Shell, S.A., et al., *Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection.* J Biol Chem, 2005. **280**(48): p. 39950-61.

255. Sandberg, R., et al., *Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.* Science, 2008. **320**(5883): p. 1643-7.

256. Dewey, C.N., I.B. Rogozin, and E.V. Koonin, *Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns.* BMC Genomics, 2006. **7**: p. 311.

257. Gibb, G.M., et al., *Differential involvement and heterogeneous phosphorylation of tau isoforms in progressive supranuclear palsy.* Brain Res Mol Brain Res, 2004. **121**(1-2): p. 95-101.

258. Jeganathan, S., et al., *Global hairpin folding of tau in solution.* Biochemistry, 2006. **45**(7): p. 2283-93.

259. Horowitz, P.M., et al., *N-terminal fragments of tau inhibit full-length tau polymerization in vitro.* Biochemistry, 2006. **45**(42): p. 12859-66.

260. Tolhuis, B., et al., *Looping and interaction between hypersensitive sites in the active beta-globin locus.* Mol Cell, 2002. **10**(6): p. 1453-65.

261. Wade-Martins, R., et al., *An infectious transfer and expression system for genomic DNA loci in human and mouse cells.* Nat Biotechnol, 2001. **19**(11): p. 1067-70.

262. Ding, Q., et al., *A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models.* Cell Stem Cell, 2012.

Appendix A

CP: the *MAPT* core promoter (chr17:43971166-43972505)

In the below multiple sequence alignment of the two CP elements, sequence matches are denoted *, with sequence differences highlighted in **red**. Exon 0 is highlighted in **green**.

```
CP_H1 CAAATGCTCTGCGATGTGTTAAGCACTGTTTGAAATTCGTCTAATTTAAGATTTTTTTTT 60
CP_H2 CAAATGCTCTGCGATGTGTTAAGCACTGTTTGAAATTCGTCTAA-----GATTTTTTTTT 55
      ******************************************     **********

CP_H1 CTGACGTAACGGTTAGAT----------TCACGTTTCTTTTTTTTTAAGTACAGTTCTAC 110
CP_H2 CTGACGTAACGGTTAGATACATCATAGATCACGTTTCTTTTTTTTTAAGTACAGTTCTAC 115
      ******************          ********************************

CP_H1 TGTATTGTAACTGAGTTAGCTTGCTTTAAGCCGATTTGTTAAGGAAAGGATTCACCTTGG 170
CP_H2 TGTATTGTAACTGAGTTAGCTTGCTTTAAGCCGATTTGTTAAGGAAAGGATTCACCTTGG 175
      ************************************************************

CP_H1 TCAGTAACAAAAAAGGTGGGAAAAAAGCAAGGAGAAAGGAAGCAGCCTGGGGGAAAGAGA 230
CP_H2 TCAGTAACAAAAAAGGTGGGAAAAAAGCAAGGAGAAAGGAAGCAGCCTGGGGGAAAGAGA 235
      ************************************************************

CP_H1 CCTTAGCCAGGGGGGCGGTTTCGGGACTACGAAGGGTCGGGGCGGACGGACTCGAGGGCC 290
CP_H2 CCTTAGCCAGGGGGGCGGTTTCGGGACTACGAAGGGTCGGGGCGGACGGACCCGAGGGCC 295
      ************************************************* ********

CP_H1 GGCCACGTGGAAGGCCGCTCAGGACTTCTGTAGGAGAGGACACCGCCCCAGGCTGACTGA 350
CP_H2 GCCCACGTGGAAGGCCGCTCAGGACCTCTGTAGGAGAGGACACCGCCCCAGGCTGACTGA 355
      * ********************** ***********************************

CP_H1 AAGTAAAGGGCAGCGGACCCAGCGGCGGAGCCACTGGCCTTGCCCCGACCCCGCATGGCC 410
CP_H2 AAGTAAAGGGCAGCGGACCCAGCGGCGGAGCCACTGGCCTTGCCCCGACCCCGCATGGCC 415
      ************************************************************

CP_H1 CGAAGGAGGACACCCACCCCCGCAACGACACAAAGACTCCAACTACAGGAGGTGGAGAAA 470
CP_H2 CGAAGGAGGACACCCACCCCCGCAACGAGACAAAGACTCCAACTACAGGAGGTGGAGAAA 475
      **************************** ******************************

CP_H1 GCGCGTGCGCCACGGAACGCGCGTGCGCGCTGCGGTCAGCGCCGCGGCCTGAGGCGTAGC 530
CP_H2 GCGCGTGCGCCACGGAACGCGCGTGCGCGCTGCGGTCAGCGCCGCGGCCTGAGGCGTAGC 535
      ************************************************************

CP_H1 GGGAGGGGGACCGCGAAAGGGCAGCGCCGAGAGGAACGAGCCGGGAGACGCCGGACGGCC 590
CP_H2 GGGAGGGGGACCGCGAAAGGGCAGCGCCGAGAGGAACGAGCCGGGAGACGCCGGACGGCC 595
      ************************************************************

CP_H1 GAGCGGCAGGGCGCTCGCGCGCGCCCACTAGTGGCCGGAGGAGAAGGCTCCCGCGGAGGC 650
CP_H2 GAGCGGCAGGGCGCTCGCGCGCGCCCACTGGTGGCCGGAGGAGAAGGCTCCCGCGGAGGC 655
      ****************************.*****************************

CP_H1 CGCGCTGCCCGCCCCCTCCCCTGGGGAGGCTCGCGTTCCCGCTGCTCGCGCCTGCGCCGC 710
CP_H2 CGCGCTGCCCGCCCCCTCCCCTGGGGAGGCTCGCGTTCCCGCTGCTCGCGCCTGCGCCGC 715
      ************************************************************

CP_H1 CCGCCGGCCTCAGGAACGCGCCCTCTTCGCCGGCGCGCGCCCTCGCAGTCACCGCCACCC 770
CP_H2 CCGCCGGCCTCAGGAACGCGCCCTCTTCGCCGGCGCGCGCCCTCGCAGTCACCGCCACCC 775
      ************************************************************

CP_H1 ACCAGCTCCGGCACCAACAGCAGCGCCGCTGCCACCGCCCACCTTCTGCCGCCGCCACCA 830
CP_H2 AACAGCTCCGGCACCAACAGCAGCGCCGCTGCCACCGCCCACCTTCTGCCGCCGCCACCA 835
      *.********************************************************
```

```
CP_H1 CAGCCACCTTCTCCTCCTCCGCTGTCCTCTCCCGTCCTCGCCTCTGTCGACTATCAGGTA 890
CP_H2 CAGCCACCTTCTCCTCCTCCGCTGTCCTCTCC-GTCCTCGCCTCTGTCGACTATCAGGTA 894
      ********************************* *************************

CP_H1 AGCGCCGCGGCTCCGAAATCTGCCTCGCCGTCCGCCTCTGTGCACCCCTGCGCCGCCGCC 950
CP_H2 AGCGCCGCGGCTCCGAAATCTGCCTCGCCGTCCGCCTCTGTGCACCCCTGCGCCGCCGCC 954
      ************************************************************

CP_H1 CCTCGCCCTCCCTCTCCGCAGACTGGGGCTTCGTGCGCCGGGCATCGGTCGGGGCCACCG 1010
CP_H2 CCTCGCCCTCCCTCTCCGCAGACTGGAGCTTCGTGCGCCGGGCATCGGTCGGGGCCACCG 1014
      ************************.***********************************

CP_H1 CAGGGCCCCTCCCTGCCTCCCCTGCTCGGGGGCTGGGGCCAGGGCGGCCTGGAAAGGGAC 1070
CP_H2 CAGGGCCCCTCCCTGCCTCCCCTGCTCGGGGGCTGGGGCCAGGGCGGCCTGGAAAGGGAC 1074
      ************************************************************

CP_H1 CTGAGCAAGGGATGCACGCACGCGTGAGTGCGCGCGTGTGTGTGTGCTGGAGGGTCTTCA 1130
CP_H2 CTGAGCAAGGGATGCACGCACGCGTGAGTGCGCGCGTGTGTGTGTGCTGGAGGGTCTTCA 1134
      ************************************************************

CP_H1 CCACCAGATTCGCGCAGACCCCAGGTGGAGGCTGTGCCGGCAGGGTGGGGCGCGGCGGCG 1190
CP_H2 CCACCAGATTCGCGCAGACCCCAGGTGGAGGCTGTGCCGGCAGGGTGGGGCGCGGCGGCG 1194
      ************************************************************

CP_H1 GTGACTTGGGGGAGGGGGCTGCCCTTCACTCTCGACTGCAGCCTTTTGCCGCAATGGGCG 1250
CP_H2 GTGACTTGGGGGAGGGGGCTGCCCTTCACTCTCGACTGCAGCCTTTTGCCGCAATGGGCG 1254
      ************************************************************

CP_H1 TGTGTGTGTGTGTGTGTGTGTGTGTGTGTG------GAGGGGTCCGATAACGACCC 1310
CP_H2 TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGGAGGGGTCCGATAACGACCC 1312
      ************************************************************

CP_H1 CCGAAACCGAATCTGAAATCCGCTGTCC 1338
CP_H2 CCGAAACCGAATCTGAAATCCGCTGTCC 1340
     ****************************
```

Appendix B

SD: the rs242557 SNP domain (chr17:44019339-44020150)

In the below multiple sequence alignment of the three SD elements, sequence matches are denoted *, with sequence differences highlighted in **red**. The rs242557 polymorphism is highlighted in **green**.

```
SD_H1B TGGGACAGATCCTCAGTGGAACATGACTCTGTAACGAGAGCATTTTGTTTTGTCAAAATG 60
SD_H1C TGGGACAGATCCTCAGTGGAACATGACTCTGTAACGAGAGCATTTTGTTTTGTCAAAATG 60
SD_H2  TGGGACAGATCCTCAGTGGAACATGACTCTGTAACGAGAGCATTTTGTTTTGTCAAAATG 60
       ************************************************************

SD_H1B AGAACATATTATTGCCTTTCATCTGATTGTAAACATAATACATGTTTATAAAACAGTATA 120
SD_H1C AGAACATATTATTGCCTTTCATCTGATTGTAAACATAATACATGTTTATAAAACAGTATA 120
SD_H2  AGAACATATTATTGCCTTTCATCTGATTGTAAACATAATACATGTTTATAAAACAGTATA 120
       ************************************************************

SD_H1B ATGAGACAAAAATGTAGACACTAATAAGGGAAAATCTCCCTAATTGTATTTCTCTTCACA 180
SD_H1C ATGAGACAAAAATGTAGACACTAATAAGGGAAAATCTCCCTAATTGTATTTCTCTTCACA 180
SD_H2  ATGAGACAAAAATGTAGACGCTAATAAGGGAAAATCTCCCTAATTGTATTTCTCTTCACA 180
       ******************.*****************************************

SD_H1B GAGAAAGCCCCTGTTGGGCATATATACTCTAGTTTGTTTATTTGTTTGACTACACATATA 240
SD_H1C GAGAAAGCCCCTGTTGGGCATATATACTCTAGTTTGTTTATTTGTTTGACTACACATATA 240
SD_H2  GAGAAAGCCCCTGTTGGGCATATATACTCTAGTTTGTTTATTTGTTTGACTACACATATA 240
       ************************************************************

SD_H1B TGTATTCTTTTCTTATGTATAAAAATTCTGAACATGCACATTTCTGCAACTACTGTTTTC 300
SD_H1C TGTATTCTTTTCTTATGTATAAAAATTCTGAACATGCACATTTCTGCAACTACTGTTTTC 300
SD_H2  TGTATTCTTTTCTTATGTATAAAAATTCTGAACATGCACATTTCTGCAACTACTGTTTTC 300
       ************************************************************

SD_H1B ACTTGATGATGCATGGACCTCTCTAGAGTGTACGTTTCTTCTTCCTTACAAAGCAGTTGG 360
SD_H1C ACTTGATGATGCATGGACCTCTCTAGAGTGTACGTTTCTTCTTCCTTACAAAGCAGTTGG 360
SD_H2  ACTTAATGATGCATGGACCTCTCTAGAGTGTACGTTTCTTCCCTTACAAAGCAGTTGG 360
       ****.*********************************** ******************

SD_H1B CTTCGCCCAGGGTGCACCAGGACACGGTTTTGGCTCTGTCCCCAGGGTGTCACGGGACCA 420
SD_H1C CTTCGCCCAGGGTACACCAGGACACGGTTTTGGCTCTGTCCCCAGGGTGTCACGGGACCA 420
SD_H2  CTTCGCCCAGGGTGCACCAGGACACGGTTTTGGCTCTGTCCCCAGGGTGTCACGGGACCA 420
       *************.**********************************************

SD_H1B GGGGATGATCTCACAGGGTCTGCCATCTGCCCTGCCTGGCCGGAGGCTGCATCGAGAGGG 480
SD_H1C GGGGATGATCTCACAGGGTCTGCCATCTGCCCTGCCTGGCCGGAGGCTGCATCGAGAGGG 480
SD_H2  GGGGATGATCTCACAGGGTCTGCCATCTGCCCTGCCTGGCCGGAGGCTGCATCGAGAGGG 480
       ************************************************************

SD_H1B CCAAGGGGCACCACGTGTCGTGGGTACTGTCAAACAAGAGCCTTCAGAGCCTTCCACAGT 540
SD_H1C CCAAGGGGCACCACGTGTCGTGGGTACTGTCAAACAAGAGCCTTCAGAGCCTTCCACAGT 540
SD_H2  CCAAGGGGCACCACGTGTCGTGGGTACTGTCAAACAAGAGCCTTCAGAGCCTTCCACAGT 540
       ************************************************************

SD_H1B CTTTCTTTTGCTTCCCAGCATTGCTTCCCCGCTGGTGGACTCTGAATCTAGAACTAGCTC 600
SD_H1C CTTTCTTTTGCTTCCCAGCATTGCTTCCCCGCTGGTGGACTCTGAATCTAGAACTAGCTC 600
SD_H2  CTTTCTTTTGCTTCCCAGCATTGCTTCCCCGCTGGTGGACTCTGAATCTAGAACTAGCTC 600
       ************************************************************

SD_H1B CAGGCGCCTCTCCAAATTCAGACGGGAGCTGGGGCACTATTATAATGCAAATCTAGGCAA 660
SD_H1C CAGGCGCCTCTCCAAATTCAGACGGGAGCTGGGGCACTATTATAATGCAAATCTAGGCAA 660
SD_H2  CAGGCGCCTCTCCAAATTCAGACGGGAGCTGGGGCACTATTATAATGCAAATCTAGGCAA 660
       ************************************************************
```

```
SD_H1B AGCCCTCCCAATACCAGGATCCAGAATGGGGTGGGGCCCTTTGCCCTGAAAAGCTGTTTA 720
SD_H1C AGCCCTCCCAATACCAGGATCCAGAATGGGGTGGGGCCCTTTGCCCTGAAAAGCTGTTTA 720
SD_H2  AGCCCTCCCAATACCAGGATCCAGAATGGGGTGGGGCCCTTTGCCCTGAAAAGCTGTTTA 720
       ************************************************************

SD_H1B GTTTGAAAATACAAACAGGAGACAGAAAAGTTTGGCTAAATTAATGGATAAAGTTTTAAC 780
SD_H1C GTTTGAAAATACAAACAGGAGACAGAAAAGTTTGGCTAAATTAATGGATAAAGTTTTAAC 780
SD_H2  GTTTGAAAATACAAACAGGAGACAGAAAAGTTTGGCTAAATTAATGGATAAAGTTTTAAC 780
       ************************************************************

SD_H1B GATGGTAACCATAGTAGGGTTCATCGACAGCC 812
SD_H1C GATGGTAACCATAGTAGGGTTCATCGACAGCC 812
SD_H2  GATGGTAACCATAGTAGGGTTCATCGACAGCC 812
       ********************************
```

Appendix C

NP: the NAT promoter region (chr17:43972506:43973404)

In the below multiple sequence alignment of the three NP elements, sequence matches are denoted *, with sequence differences highlighted in **red**. The predicted bi-directional promoter is highlighted in **green**.

```
NP_H1B CTGCCGCTGTTCGCCATCAGCTCTAAGAAAGACGTGGATCGGGTTCTAGAAAAGATGACT 60
NP_H1C CTGCCGCTGTTCGCCATCAGCTCTAAGAAAGACGTGGATCGGGTTCTAGAAAAGATGACT 60
NP_H2  CTGCCGCTGTTCGCCATCAGCTCTAAGAAAGACGTGGATCGGGTTCTAGAAAAGATGACT 60
       ************************************************************

NP_H1B CCCTGCACGCCCCTCCCTGCACCTCCCGAGCAGTGATTCCGACAGGGCCTTCACTGCCCC 120
NP_H1C CCCTGCACGCCCCTCCCTGCACCTCCCGAGCAGTGATTCCGACAGGGCCTTCACTGCCCC 120
NP_H2  CCCTGCACGCCCCTCTCTGCACCTCCCGAGCAGTGATTCCGACAGGGCCTTCACTGCCCC 120
       *************** ****************************************** 

NP_H1B TGATTTTAGGCGGGGGCCGGCCCCCTCCCCTTTTCCTCCTTCAGAAACCCGTAGGGGACA 180
NP_H1C TGATTTTAGGCGGGGGCCGGCCCCCTCCCCTTTTCCTCCTTCAGAAACCCGTAGGGGACA 180
NP_H2  TGATTTTAGGCGGGGGCCGGCCCCCTCCCCTTTTCCTCCTTCAGAAACCCGTAGGGGACA 180
       ************************************************************

NP_H1B TTTGGGGGCTGGGAGAAATCGAGGAGATGGGGAGGGGTCCACGCGCTGTCACTTTAGTTG 240
NP_H1C TTTGGGGGCTGGGAGAAATCGAGGAGATGGGGAGGGGTCCACGCGCTGTCACTTTAGTTG 240
NP_H2  TTTGGGGGCTGGGAGAAATCGAGGAGATGGGGAGGGGTCCACGCGCTGTCACTTTAGTTG 240
       ************************************************************

NP_H1B CCCTTCCCCCTGCGCACGCCTGGCACAGAGACGCGAGCAGCGCCGTGCCTGAGAACAGTG 300
NP_H1C CCCTTCCCCCTGCGCACGCCTGGCACAGAGACGCGAGCAGCGCCGTGCCTGAGAACAGTG 300
NP_H2  CCCTTCCCCCTGCGCACGCCTGGCACAGAGACGCGAGCAGCGCCGTGCCTGAGAACAGTG 300
       ************************************************************

NP_H1B CGCGGATCCCACTGTGCACGCTCGCAAAGGCAGGGTTCACCTGGCCTGGCGATGTGGACG 360
NP_H1C CGCGGATCCCACTGTGCACGCTCGCAAAGGCAGGGTTCACCTGGCCTGGCGATGTGGACG 360
NP_H2  CGCGGATCCCACTGTGCACGCTCGCAAAGGCAGGGTTCACCTGGCCTGGCGATGTGGACG 360
       ************************************************************

NP_H1B GACTCGGCGGCCGCTGGTCCCCGTTCGCGGGCACGCACAGCCGCAGCCACGCACGGATGG 420
NP_H1C GACTCGGCGGCCGCTGGTCCCCGTTCGCGGGCACGCACAGCCGCAGCCATGCACGGATGG 420
NP_H2  GACTCGGCGGCCGCTGGTCCCCGTTCGCGGGCACGCACAGCCGCAGCCATGCACGGATGG 420
       **************************************** ********** 

NP_H1B GCGCGGGGCTGCAGGTGCATCTCGGGGCGGATTTCTTTCTCAGCGCTCGGAGCGCAGGGC 480
NP_H1C GCGCGGGGCTGCAGGTGCATCTCGGGGCGGATTTCTTTCTCAGCGCTCGGAGCGCAGGGC 480
NP_H2  GCGCGGGGCTGCAGGTGCATCTCGGGGCGGATTTCTTTCTCAGCGCTCGGAGCGCAGGGC 480
       ************************************************************

NP_H1B GCCCGGCGTGTGCGCTCCCTGCCGGAGGCGCGGGGCTGGCGCGCAGGGCTCGCCCCTCAC 540
NP_H1C GCCCGGCGTGTGCGCTCCCTGCCGGAGGCGCGGGGCTGGCGCGCAGGGCTCGCCCCTCAC 540
NP_H2  GCCCGGCGTGTGCGCTCCCTGCCGGAGGCGCGGGGCTGGCGCGCAGGGCTCGCCCCTCAC 540
       ************************************************************

NP_H1B TGCGGCAGTGGGTGTGGACCCTGGTGGGCGAGGAAGGGGGAGGATAGGCTGTGCCTCCTC 600
NP_H1C TGCGGCAGTGGGTGTGGACCCTGGTGGGCGAGGAAGGGGGAGGATAGGCTGTGCCTCCTC 600
NP_H2  TGCGGCAGTGGGTGTGGACCCTGGTGGGCGAGGAGGGGGGAGGATAGGCTGTGCCTCCTC 600
       ********************************.***************************

NP_H1B CCACTCCCGCCCCCAGCCCCCCTTTTTTTCCCCCTCGGAACGCGAGGTGCCATCTTTTTT 660
NP_H1C CCACTCCCGCCCCCAGCCCCCCTTTTTTTCCCCCTCGGAACGCGAGGTGCCATCTTTTTT 660
NP_H2  CCACTCCCGCCCCCACCCCCCCTTTTTTTCCCCCTCGGAACGCGAGGTGCCATCTTTTTT 660
       ************** ************************************************
```

```
NP_H1B CGGCGTGTCACGTCTTTACGGTGCCATGCCAAACCGGGTGGCCGGGCTTCATAGGACAGG 720
NP_H1C CGGCGTGTCACGTCTTTACGGTGCCATGCCAAACCGGGTGGCCGGGCTTCATAGGACAGG 720
NP_H2  CGGCGTGTCACGTCTTTACGGTGCCATGCCAAACCGGGTGGCCGGGCTTCATAGGACAGG 720
       ************************************************************

NP_H1B GCGGGGCCTGGCATTAAAGGGAGGGGGACAATCAGCGCTGAAATCTTGGCGTTTTGCTGC 780
NP_H1C GCGGGGCCTGGCATTAAAGGGAGGGGGACAATCAGCGCTGAAATCTTGGCGTTTTGCTGC 780
NP_H2  GCGGGGCCTGGCATTAAAGGGAGGGGGACAATCAGCGCTGAAATCTTGGCGTTTTGCTGC 780
       ************************************************************

NP_H1B TGCGGGCGTGAGCACTGGGGGCGTTCGCCCAGCACCTTCTTCGGGGGCTCTTTGCTTTGT 840
NP_H1C TGCGGGCGTGAGCACTGGGGGCGTTCGCCCAGCACCTTCTTCGGGGGCTCTTTGCTTTGT 840
NP_H2  TGCGGGCGTGAGCACTGGGGGCGTTCGCCCAGCACCTTCTTCGGGGGCTCTTTGCTTTGT 840
       ************************************************************

NP_H1B CTGTAGAGGTTACGTGATCTGCGCTCCCAGCCCTGGTTTCTGGCTTTTATTCTGAGGGT 899
NP_H1C CTGTAGAGGTTACGTGATCTGCGCTCCCAGCCCTGGTTTCTGGCTTTTATTCTGAGGGT 899
NP_H2  CTGTAGAGGTTACGTGATCTGCGCTCCCAGCCCTGGTTTCTGGCTTTTATTCTGAGGGT 899
       ************************************************************
```

Appendix D

3'UTR Fragment 1 (Fr1): chr17:44101545-44102731

In the below multiple sequence alignment of the three Fr1 variants, sequence matches are denoted *, with sequence differences highlighted in **red**. The region of Fr1 that overlaps with Fr2 is highlighted in blue, with the ***AatII*** internal restriction site underlined and italicised.

```
Fr1_H1B CCTGGGGCGGTCAATAATTGTGGGGAGGAGAGAATGAGAGAGTGTGGAAAAAAAAAGAAT 60
Fr1_H1C CCTGGGGCGGTCAATAATTGTGGAGAGGAGAGAATGAGAGAGTGTGGAAAAAAAAAGAAT 60
Fr1_H2  CCTGGGGCGGTCAATAATCGTGGAGAGGAGAGAATGAGAGAGTGTGGAAAAAAAAAGAAT 60
        ****************** ****.************************************

Fr1_H1B AATGACCCGGCCCCCGCCCTCTGCCCCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAA 120
Fr1_H1C AATGACCCGGCCCCCGCCCTCTGCCCCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAA 120
Fr1_H2  AATGACCCGGCCCCCGCCCTCTGCCCCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAA 120
        ************************************************************

Fr1_H1B TCACTTAACCTGCTTTTGTCACTCGGCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAG 180
Fr1_H1C TCACTTAACCTGCTTTTGTCACTCGGCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAG 180
Fr1_H2  TCACTTAACCTGCTTTTGTCACTCGGCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAG 180
        ************************************************************

Fr1_H1B TAAGAGCAAATTTCATCTTTCCAAATTGATGGGTGGGCTAGTAATAAAATATTT-AAAAA 239
Fr1_H1C TAAGAGCAAATTTCATCTTTCCAAATTGATGGGTGGGCTAGTAATAAAATATTTAAAAAA 240
Fr1_H2  TAAGAGCAAATTTCATCTTTCCAAATTGATGGGTGGGCTAGTAATAAAATATTTTAAAAA 240
        ************************************************************ *****

Fr1_H1B AAACATTCAAAAACATGGCCACATCCAACATTTCCTCAGGCAATTCCTTTTGATTCTTTT 299
Fr1_H1C AAACATTCAAAAACATGGCCACATCCAACATTTCCTCAGGCAATTCCTTTTGATTCTTTT 300
Fr1_H2  AAACATTCAAAAACATGGCCACATCCAACATTTCCTCAGGCAATTCCTTTTGATTCTTTT 300
        ************************************************************

Fr1_H1B TTCTT-CCCCCTCCATGTAGAAGAGGGAGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGG 358
Fr1_H1C TTCTT-CCCCCTCCATGTAGAAGAGGGAGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGG 359
Fr1_H2  TTCTTCCCCCCTCCATGTAGAAGAGGGGGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGG 360
        ***** ********************.******************************

Fr1_H1B ATTTCAAGGGACTGGGGGTGCCAACCACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGC 418
Fr1_H1C ATTTCAAGGGACTGGGGGTGCCAACCACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGC 419
Fr1_H2  ATTTCAAGGGACTGGGGGTGCCAACCACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGC 420
        ************************************************************

Fr1_H1B AGTGGCAGCAACAAAGGATTTGAAACTTGGTGTGTTCGTGGAGCCACAGGCAGACGATGT 478
Fr1_H1C AGTGGCAGCAACAAAGGATTTGAAACTTGGTGTGTTCGTGGAGCCACAGGCAGACGATGT 479
Fr1_H2  AGTGGCAGCAACAAAGGATTTGAAACTTGGTGTGTTCGTGGAGCCACAGGCAGACGATGT 480
        ************************************************************

Fr1_H1B CAACCTTGTGTGAGTGTGACGGGGGTTGGGGTGGGGCGGGAGGCCACGGGGGAGGCCGAG 538
Fr1_H1C CAACCTTGTGTGAGTGTGACGGGGGTTGGGGTGGGACGGGAGGCCACGGGGGAGGCCGAG 539
Fr1_H2  CAACCTTGTGTGAGTGTGACGGGGGTTGGGGTGGGGCGGGAGGCCACGGGGGAGGCCGAG 540
        ************************************.************************

Fr1_H1B GCAGGGGCTGGGCAGAGGGGAGAGGAAGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACG 598
Fr1_H1C GCAGGGGCTGGGCAGAGGGGAGAGGAAGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACG 599
Fr1_H2  GCAGGGGCTGGGCAGAGGGGAGAGGAAGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACG 600
        ************************************************************
```

```
Fr1_H1B TGCTGGAGAGTAGACATCCCCCTCCTTGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGC 658
Fr1_H1C TGCTGGAGAGTAGACATCCCCCTCCTTGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGC 659
Fr1_H2  TGCTGGAGAGTAGACATCCCCCTCCTTGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGC 660
        ************************************************************

Fr1_H1B AGCGTCTGAGCGGCCGCCTGTCCTTGGTGGCCGGAGGTGGGGGCCTGCTGTGGGTCAGTG 718
Fr1_H1C AGCGTCTGAGCGGCCGCCTGTCCTTGGTGGCCGGGGGTGGGGGCCTGCTGTGGGTCAGTG 719
Fr1_H2  AGCGTCTGAGCGGCCGCCTGTCCTTGGTGGCCGGGGGTGGGGGCCTGCTGTGGGTCAGTG 720
        ***********************************.************************

Fr1_H1B TGCCACCCTCTGCAGGGCAGCCTGTGGGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAG 778
Fr1_H1C TGCCACCCTCTGCAGGGCAGCCTGTGGGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAG 779
Fr1_H2  TGCCACCCTCTGCAGGGCAGCCTGTGGGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAG 780
        ************************************************************

Fr1_H1B CTGGCAGGAGGGTGGCACTTCGTGGATGACCTCCTTAGAAAAGACTGACCTTGATGTCTT 838
Fr1_H1C CTGGCAGGAGGGTGGCACTTCGTGGATGACCTCCTTAGAAAAGACTGACCTTGATGTCTT 839
Fr1_H2  CTGGCAGGAGGGTGGCACTTCGTGGATGACCTCCTTAGAAAAGACTGACCTTGATGTCTT 840
        ************************************************************

Fr1_H1B GAGAGCGCTGGCCTCTTCCTCCCTCCCTGCAGGGTAGGGGGCCTGAGTTGAGGGGCTTCC 898
Fr1_H1C GAGAGCGCTGGCCTCTTCCTCCCTCCCTGCAGGGTAGGGGGCCTGAGTTGAGGGGCTTCC 899
Fr1_H2  GAGAGCGCTGGCCTCTTCCTCCCTCCCTGCAGGGTAGGGGACCTGAGTTGAGGGGCTTCC 900
        ****************************************.*******************

Fr1_H1B CTCT--GCTCCACAGAAACCCTGTTTTATTGAGTTCTGAAGGTTGGAACTGCTGCCATGA 956
Fr1_H1C CTCT--GCTCCACAGAAACCCTGTTTTATTGAGTTCTGAAGGTTGGAACTGCTGCCATGA 957
Fr1_H2  CTCTCTGCTCCACAGAAACCCTGTTTTATTGAGTTCTGAAGGTTGGAACTGCTGCCATGA 960
        ****  ************************************************************

Fr1_H1B TTTTGGCCACTTTGCAGACCTGGGACTTTAGGGCTAACCAGTTCTCTTTGTAAGGACTTG 1016
Fr1_H1C TTTTGGCCACTTTGCAGACCTGGGACTTTAGGGCTAACCAGTTCTCTTTGTAAGGACTTG 1017
Fr1_H2  TTTTGGCCACTTTGCAGACCTGGGACTTTAGGGCTAACCAGTTCTCTTTGTAAGGACTTG 1020
        ************************************************************

Fr1_H1B TGCCTCTTGGGAGACGTCCACCCGTTTCCAAGCCTGGGCCACTGGCATCTCTGGAGTGTG 1076
Fr1_H1C TGCCTCTTGGGAGACGTCCACCCGTTTCCAAGCCTGGGCCACTGGCATCTCTGGAGTGTG 1077
Fr1_H2  TGCCTCTTGGGAGACGTCCACCCGTTTCCAAGCCTGGGCCACCGGCATCTCTGGAGTGTG 1080
        ****************************************** ****************

Fr1_H1B TGGGGGTCTGGGAGGCAGGTCCCGAGCCCCCTGTCCTTCCCACGGCCACTGCAGTCACCC 1136
Fr1_H1C TGGGGGTCTGGGAGGCAGGTCCCGAGCCCCCTGTCCTTCCCACGGCCACTGCAGTCACCC 1137
Fr1_H2  TGGGGGTCTGGGAGGCGGGTCCCGAGCCCCCTGTCCTTCCCACGGCCACTGCAGTCACCC 1140
        ****************.*******************************************

Fr1_H1B C-GTCTGCGCCGCTGTGCTGTTGTCTGCCGTGAGAGCCCAATCACTGCCTA 1186
Fr1_H1C CTGTCTGCGCCGCTGTGCTGTTGTCTGCCGTGAGAGCCCAATCACTGCCTA 1188
Fr1_H2  CTGTCTGCCCCGCTGTGCTGTTGTCTGCCGTGAGAGCCCAATCACTGCCTA 1191
        * ****** ******************************************
```

Appendix E

Fragment 2 (Fr2): chr17:44102418-44104245

In the below multiple sequence alignment of the three Fr2 variants, sequence matches are denoted *, with sequence differences highlighted in **red**. The regions of Fr2 that overlaps with Fr1 (5' end) and Fr3 (3' end) are highlighted in blue, with the ***AatII*** and ***XbaI*** internal restriction sites underlined and italicised.

```
Fr2_H1B TAGGGGGCCTGAGTTGAGGGGCTTCCCTCT--GCTCCACAGAAACCCTGTTTTATTGAGT 57
Fr2_H1C TAGGGGGCCTGAGTTGAGGGGCTTCCCTCT--GCTCCACAGAAACCCTGTTTTATTGAGT 58
Fr2_H2  TAGGGGGCCTGAGTTGAGGGGCTTCCCTCTCTGCTCCACAGAAACCCTGTTTTATTGAGT 59
        ****************************  ****************************

Fr2_H1B TCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACTTTGCAGACCTGGGACTTTAGGGC 117
Fr2_H1C TCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACTTTGCAGACCTGGGACTTTAGGGC 118
Fr2_H2  TCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACCTTGCAGACCTGGGACTTTAGGGC 119
        ********************************** ************************

Fr2_H1B TAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGACGTCCACCCGTTTCCAAGCC 177
Fr2_H1C TAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGACGTCCACCCGTTTCCAAGCC 178
Fr2_H2  TAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGACGTCCACCCGTTTCCAAGCC 179
        ************************************************************

Fr2_H1B TGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAGGCAGGTCCCGAGCCCCCTGT 237
Fr2_H1C TGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAGGCAGGTCCCGAGCCCCCTGT 238
Fr2_H2  TGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAGGCGGGTCCCGAGCCCCCTGT 239
        ****************************************.*******************

Fr2_H1B CCTTCCCACGGCCACTGCAGTCACCCC-GTCTGCGCCGCTGTGCTGTTGTCTGCCGTGAG 296
Fr2_H1C CCTTCCCACGGCCACTGCAGTCACCCCTGTCTGCGCCGCTGTGCTGTTGTCTGCCGTGAG 298
Fr2_H2  CCTTCCCACGGCCACTGCAGTCACCCCTGTCTGCCCCGCTGTGCTGTTGTCTGCCGTGAG 299
        *************************** ****** ************************

Fr2_H1B AGCCCAATCACTGCCTATACCCCTCATCACACGTCACAATGTCCCGAATTCCCAGCCTCA 356
Fr2_H1C AGCCCAATCACTGCCTATACCCCTCATCACACGTCACAATGTCCCGAATTCCCAGCCTCA 358
Fr2_H2  AGCCCAATCACTGCCTATACCCCTCAT--CACGTCACAATGTCCCGAATTCCCAGCCTCA 357
        **************************  ********************************

Fr2_H1B CCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGTACCTACTCCATACTGAGGG 416
Fr2_H1C CCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGTACCTACTCCATACTGAGGG 418
Fr2_H2  CCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGTACCTACTCCATACTGAGGG 417
        ************************************************************

Fr2_H1B TGAAATTAAGGGAAGGCAAAGTCCAGGCACAAGAGTGGGACCCCAGCCTCTCACTCTCAG 476
Fr2_H1C TGAAATTAAGGGAAGGCAAAGTCCAGGCACAAGAGTGGGACCCCAGCCTCTCACTCTCAG 478
Fr2_H2  TGAAATTAAGGGAAGGCAAAGTCCAGGCACCAGAGTGGGACCCCAGCCTCTCACTCTCAG 477
        ******************************.*****************************

Fr2_H1B TTCCACTCATCCAACTGGGACCCTCACCACGAATCTCATGATCTGATTCGGTTCCCTGTC 536
Fr2_H1C TTCCACTCATCCAACTGGGACCCTCACCACGAATCTCATGATCTGATTCGGTTCCCTGTC 538
Fr2_H2  TTCCACTCATCCAACTGGGACCCTCACCACGAATCTCACGATCTGATTCGGTTCCCTGTC 537
        ********************************** ********************

Fr2_H1B TCCTCCTTCCGTCACAGATGTGAGCCAGGGCACTGCTCAGCTGTGACCCTAGGTGTTTCT 596
Fr2_H1C TCCTCCTCCCGTCACAGATGTGAGCCAGGGCACTGCTCAGCTGTGACCCTAGGTGTTTCT 598
Fr2_H2  TCCTCCTCCCGTCACAGATGTGAGCCAGGGCACTGCTCAGCTGTGACCCTAGGTGTTTCT 597
        ******* ************************************************
```

```
Fr2_H1B GCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCCTGGCCCCTTCCTGTGCTGA 656
Fr2_H1C GCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCCTGGCCCCTTCCTGTGCTGA 658
Fr2_H2  GCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCCTGGCCCCTTCCTGTGCTGA 657
        ************************************************************

Fr2_H1B GCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGGCACCAGGATGGAAGGGCAA 716
Fr2_H1C GCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGGCACCAGGATGGAAGGGCAA 718
Fr2_H2  GCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGGCACCAGGATGGAAGGGCAA 717
        ************************************************************

Fr2_H1B GGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTGCACCCTAGCTTGTAGCTGC 776
Fr2_H1C GGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTGCACCCTAGCTTGTAGCTGC 778
Fr2_H2  GGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTGCACCCTAGCTTGTAGCTGC 777
        ************************************************************

Fr2_H1B CAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATGCATAGTATCAGCCCTCCAC 836
Fr2_H1C CAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATGCATAGTATCAGCCCTCCAC 838
Fr2_H2  CAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATGCATAGTATCAGCCCTCCAC 837
        ************************************************************

Fr2_H1B ACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTTTCCCCCAGTCCCAGCTGGA 896
Fr2_H1C ACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTTTCCCCCAGTCCCAGCTGGA 898
Fr2_H2  ACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTTCCCCCCAGTCCCAGCTGGA 897
        ****************************************** ******************

Fr2_H1B AGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACATAGATGTTGCCCTGCCCTC 956
Fr2_H1C AGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACATAGATGTTGCCCTGCCCTC 958
Fr2_H2  AGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACATAGATGTTGCCCTGCCCTC 957
        ************************************************************

Fr2_H1B CCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTTTATGCTTGGATTCACCAGA 1016
Fr2_H1C CCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTTTATGCTTGGATTCACCAGA 1018
Fr2_H2  CCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTTTATGCTTGGATTCACCAGA 1017
        ************************************************************

Fr2_H1B GTGACTATGATAGTGAAAAGAAAAAAAAAAAAAAAAA-GGACGCATGTATCTTGAAATG 1075
Fr2_H1C GTGACTATGATAGTGAAAAGAAAAAAAAAAAAAAAAA GGACGCATGTATCTTGAAATG 1078
Fr2_H2  GTGACTATGATAGTGAAAA---AAAAAAAAAAAAAAAAA GGACGCATGTATCTTGAAATG 1074
        ******************   ****************.**********************

Fr2_H1B CTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTCACCCCCACACTGGGACTC 1135
Fr2_H1C CTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTCACCCCCACACTGGGACTC 1138
Fr2_H2  CTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTCACCCCCACACTGGGACTC 1134
        ************************************************************

Fr2_H1B GTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTTTTGAAAGGCTTTCCTCAG 1195
Fr2_H1C GTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTTTTGAAAGGCTTTCCTCAG 1198
Fr2_H2  GTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTTTTGAAAGGCTTTCCTCAG 1194
        ************************************************************

Fr2_H1B CACCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAGGCCGTTCAGCTGTGACGA 1255
Fr2_H1C CACCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAGGCCGTTCAGCTGTGACGA 1258
Fr2_H2  CATCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAGGCCGTTCAGCTGTGACGA 1254
        ** ********************************************************

Fr2_H1B AGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCCTTCCCTACTTCCCCTTGG 1315
Fr2_H1C AGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCCTTCCCTACTTCCCCTTGG 1318
Fr2_H2  AGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCCTTCCCTACTTCCCCTTGG 1314
        ************************************************************

Fr2_H1B GGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTCTGGCTTGCGGCGCGAGGA 1375
Fr2_H1C GGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTCTGGCTTGCGGCGCGAGGA 1378
Fr2_H2  GGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTCTGGCTTGCGGCGCGAGGA 1374
        ************************************************************
```

```
Fr2_H1B TGGTTCTCTCTGGTCATAGCCCGAAGTCTCATGGCAGTCCCAAAGGAGGCTTACAACTCC 1435
Fr2_H1C TGGTTCTCTCTGGTCATAGCCCGAAGTCTCATGGCAGTCCCAAAGGAGGCTTACAACTCC 1438
Fr2_H2  TGGTTCTCTCTGGTCATAGCCCGAAGTCTCACAGCAGTCCCAAAGGAGGCTTACAACTCC 1434
        ******************************* .**************************

Fr2_H1B TGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATCTGCAGCTCCCAGAAGCTC 1495
Fr2_H1C TGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATCTGCAGCTCCCAGAAGCTC 1498
Fr2_H2  TGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATCTGCAGCTCCCAGAAGCTC 1494
        ************************************************************

Fr2_H1B CGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGGGGCA 1555
Fr2_H1C CGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGGGGCA 1558
Fr2_H2  CGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGGGGCA 1554
        ********************************************************** 

Fr2_H1B GCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATTGGGATGAATTGCCTGTCC 1615
Fr2_H1C GCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATTGGGATGAATTGCCTGTCC 1618
Fr2_H2  GCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATTGGGATGAATTGCCTGTCC 1614
        ************************************************************

Fr2_H1B TGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTCAGGA 1675
Fr2_H1C TGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTCAGGA 1678
Fr2_H2  TGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTCAGGA 1674
        ************************************************************

Fr2_H1B GACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCTGACCTTGCACAAACTCCA 1735
Fr2_H1C GACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCTGACCTTGCACAAACTCCA 1738
Fr2_H2  GACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCTGACCTTGCACAAACTCCA 1734
        ************************************************************

Fr2_H1B TCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAACTCA 1795
Fr2_H1C TCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAACTCA 1798
Fr2_H2  TCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAACTCA 1794
        ************************************************************

Fr2_H1B GCAGCCTCAGGCCCAATTCTGCCACTTCTGGT 1827
Fr2_H1C GCAGCCTCAGGCCCAATTCTGCCACTTCTGGT 1830
Fr2_H2  GCAGCCTCAGGCCCAATTCTGCCACTTCTGGT 1826
        ********************************
```

Appendix F

Fragment 3 (Fr3): chr17:44103934-44105914

In the below multiple sequence alignment of the three Fr3 variants, sequence matches are denoted *, with sequence differences highlighted in **red**. The region of Fr3 that overlaps with Fr2 is highlighted in blue, with the ***XbaI*** internal restriction site underlined and italicised.

```
H1B CAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGGGGCAGCGCAGCCTCCCACCAAGGG 60
H1C CAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGGGGCAGCGCAGCCTCCCACCAAGGG 58
H2  CAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGGGGCAGCGCAGCCTCCCACCAAGGG 60
    ************************************************************

H1B CCCTGCGACCACAGCAGGGATTGGGATGAATTGCCTGTCCTGGATCTGCTCTAGAGGCCC 120
H1C CCCTGCGACCACAGCAGGGATTGGGATGAATTGCCTGTCCTGGATCTGCTCTAGAGGCCC 118
H2  CCCTGCGACCACAGCAGGGATTGGGATGAATTGCCTGTCCTGGATCTGCTCTAGAGGCCC 120
    ************************************************************

H1B AAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTCAGGAGACACTGTTCCCAAAGCCTT 180
H1C AAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTCAGGAGACACTGTTCCCAAAGCCTT 178
H2  AAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTCAGGAGACACTGTTCCCAAAGCCTT 180
    ************************************************************

H1B GACCAGAGCACCTCAGCCCGCTGACCTTGCACAAACTCCATCTGCTGCCATGAGAAAAGG 240
H1C GACCAGAGCACCTCAGCCCGCTGACCTTGCACAAACTCCATCTGCTGCCATGAGAAAAGG 238
H2  GACCAGAGCACCTCAGCCCGCTGACCTTGCACAAACTCCATCTGCTGCCATGAGAAAAGG 240
    ************************************************************

H1B GAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAACTCAGCAGCCTCAGGCCCAATTCT 300
H1C GAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAACTCAGCAGCCTCAGGCCCAATTCT 298
H2  GAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAACTCAGCAGCCTCAGGCCCAATTCT 300
    ************************************************************

H1B GCCACTTCTGGTTTGGGTACAGTTAAAGGCAACCCTGAGGGACTTGGCAGTAGAAATCCA 360
H1C GCCACTTCTGGTTTGGGTACAGTTAAAGGCAACCCTGAGGGACTTGGCAGTAGAAATCCA 358
H2  GCCACTTCTGGTTTGGGTACAGTTAAAGGCAACCCTGAGGGACTTG-CAGTAGAAATCCA 359
    ********************************************** ************

H1B GGGCCTCCCCTGGGGCTGGCAGCTTCGTGTGCAGCTAGAGCTTTACCTAAAAGGAAGTCT 420
H1C GGGCCTCCCCTGGGGCTGGCAGCTTCGTGTGCAGCTAGAGCTTTACCTGAAAGGAAGTCT 418
H2  GGGCCTCCCCTGGGGCTGGCAGCTTCGTGTGCAGCTAGAGCTTTACCTGCAAGGAAGTCT 419
    **************************************************..**********

H1B CTGGGCCCAGAACTCTCCACCAAGAGCCTCCCTGCCGTTCGCTGAGTCCCAGCAATTCTC 480
H1C CTGGGCCCAGAACTCTCCACCAAGAGCCTCCCTGCCGTTCGCTGAGTCCCAGCAATTCTC 478
H2  CTGGGCCCAGAACTCTCCACCAAGAGCCTCCCTGCCGTTCGCTGAGTCCCAGCAATTCT- 478
    ************************************************************

H1B CTAAGTTGAAGGGATCTGAGAAGGAGAAGGAAATGTGGGGTAGATTTGGTGGTGGTTAGA 540
H1C CTAAGTTGAAGGGATCTGAGAAGGAGAAGGAAATGTGGGGTAGATTTGGTGGTGGTTAGA 538
H2  --AAGTTGAAGGGATCTGAGAAGGAGAAGGAAATGTGGGGTAGATTTGGTGGTGGTTAGA 536
      ********************************************************

H1B GATATGCCCCCCTCATTACTGCCAACAGTTTCGGCTGCATTTCTTCACGCACCTCGGTTC 600
H1C GATATGCCCCCCTCATTACTGCCAACAGTTTCGGCTGCATTTCTTCACGCACCTCGGTTC 598
H2  GATATGCCCCCCTCATTACTGCCAACAGTTTCGGCCGCATTTCTTCACGCACCTCGGTTC 596
    ********************************** ************************
```

```
H1B CTCTTCCTGAAGTTCTTGTGCCCTGCTCTTCAGCACCATGGGCCTTCTTATACGGAAGGC 660
H1C CTCTTCCTGAAGTTCTTGTGCCCTGCTCTTCAGCACCATGGGCCTTCTTATACGGAAGGC 658
H2  CTCTTCCTGAAGTTCTTGTGCCCTGCTCTTCAGCACCATGGGCCT---TATACGGAAGGC 653
    ******************************************   ************

H1B TCTGGGATCTCCCCCTTGTGGGG-CAGGCTCTTGGGGCCAGCCTAAGATCATGGTTTAGG 719
H1C TCTGGGATCTCCCCCTTGTGGGGGCAGGCTCTTGGGGCCAGCCTAAGATCATGGTTTAGG 718
H2  TCTGGGATCTCCCCCTTGTGGGG-CAGGCTCTTGGGGCCAGCCTAAGATCATGGTTTAGG 712
    ********************** ***********************************

H1B GTGATCAGTGCTGGCAGATAAATTGAAAAGGCACGCTGGCTTGTGATCTTAAATGAGGAC 779
H1C GTGATCAGTGCTGGCAGATAAATTGAAAAGGCACGCTGGCTTGTGATCTTAAATGAGGAC 778
H2  GTGATCAGTGCTGGCAGATAAATTGAAAAGGCACGCTGGCTTGTGATCTTAAATGAGGAC 772
    ************************************************************

H1B AATCCCCCCAGGGCTGGGCACTCCTCCCCTCCCCTCACTTCTCCCACCTGCAGAGCCAGT 839
H1C AATCCCCCCAGGGCTGGGCACTCCTCCCCTCCCCTCACTTCTCCCACCTGCAGAGCCAGT 838
H2  AATCCCCCCAGGGCTGGGCACTCCTCCCCTCCCCTCACTTCTCCCACCTGCAGAGCCAGT 832
    ************************************************************

H1B GTCCTTGGGTGGGCTAGATAGGATATACTGTATGCCGGCTCCTTCAAGCTGCTGACTCAC 899
H1C GTCCTTGGGTGGGCTAGATAGGATATACTGTATGCCGGCTCCTTCAAGCTGCTGACTCAC 898
H2  GTCCTTGGGTGGGCTAGATAGGATATACTGTATGCCGGCTCCTTCAAGCTGCTGACTCAC 892
    ************************************************************

H1B TTTATCAATAGTTCCATTTAAATTGACTTCAGTGGTGAGACTGTATCCTGTTTGCTATTG 959
H1C TTTATCAATAGTTCCATTTAAATTGACTTCAGTGGTGAGACTGTATCCTGTTTGCTATTG 958
H2  TTTATCAATAGTTCCATTTAAATTGACTTCAGTGGTGAGACTGTATCCTGTTTGCTATTG 952
    ************************************************************

H1B CTTGTTGTGCTATGGGGGGAGGGGGGAGGAATGTGTAAGATAGTTAACATGGGCAAAGGG 1019
H1C CTTGTTGTGCTATGGGGGGAGGGGGGAGGAATGTGTAAGATAGTTAACATGGGCAAAGGG 1018
H2  CTTGTTGTGCTATGGGGGGAGGGGGGAGGAATGTGTAAGATAGTTAACATGGGCAAAGGG 1012
    ************************************************************

H1B AGATCTTGGGGTGCAGCACTTAAACTGCCTCGTAACCCTTTTCATGATTTCAACCACATT 1079
H1C AGATCTTGGGGTGCAGCACTTAAACTGCCTCGTAACCCTTTTCATGATTTCAACCACATT 1078
H2  AGATCTTGGGGTGCAGCACTTAAACTGCCTCGTAACCCTTTTCATGATTTCAACCACATT 1072
    ************************************************************

H1B TGCTAGAGGGAGGGAGCAGCCACGGAGTTAGAGGCCCTTGGGGTTTCTCTTTTCCACTGA 1139
H1C TGCTAGAGGGAGGGAGCAGCCACGGAGTTAGAGGCCCTTGGGGTTTCTCTTTTCCACTGA 1138
H2  TGCTAGAGGGAGGGAGCAGCCACGGAGTTAGAGGCCCTTGGGGTTTCTCTTTTCCACTGA 1132
    ************************************************************

H1B CAGGCTTTCCCAGGCAGCTGGCTAGTTCATTCCCTCCCCAGCCAGGTGCAGGCGTAGGAA 1199
H1C CAGGCTTTCCCAGGCAGCTGGCTAGTTCATTCCCTCCCCAGCCAGGTGCAGGCGTAGGAA 1198
H2  CAGGCTTTCCCAGGCAGCTGGCTAGTTCATTCCCTCCCCAGCCAGGTGCAGGCGTAGGAA 1192
    ************************************************************

H1B TATGGACATCTGGTTGCTTTGGCCCGCTGCCCTCTTTCAGGGGTCCTAAGCCCACAATCA 1259
H1C TATGGACATCTGGTTGCTTTGGCCTGCTGCCCTCTTTCAGGGGTCCTAAGCCCACAATCA 1258
H2  TATGGACATCTGGTTGCTTTGGCCTGCTGCCCTCTTTCAGGGGTCCTAAGCCCACAATCA 1252
    ************************ ***********************************

H1B TGCCTCCCTAAGACCTTGGCATCCTTCCCTCTAAGCCGTTGGCACCTCTGTGCCACCTCT 1319
H1C TGCCTCCCTAAGACCTTGGCATCCTTCCCTCTAAGCCGTTGGCACCTCTGTGCCACCTCT 1318
H2  TGCCTCCCTAAGACCTTGGCATCCTTCCCTCTAAGCCGTTGGCACCTCTGTGCCACCTCT 1312
    ************************************************************

H1B CACACTGGCTCCAGACACACAGCCTGTGCTTTTGGAGCTGAGATCACTCGCTTCACCCTC 1379
H1C CACACTGGCTCCAGACACACAGCCTGTGCTTTTGGAGCTGAGATCACTCGCTTCACCCTC 1378
H2  CACACTGGCTCCAGACACACAGCCTGTGCTTTTGGAGCTGAGATCACTCGCTTCACCCTC 1372
    ************************************************************
```

```
H1B CTCATCTTTGTTCTCCAAGTAAAGCCACGAGGTCGGGGCGAGGGCAGAGGTGATCACCTG 1439
H1C CTCATCTTTGTTCTCCAAGTAAAGCCACGAGGTCGGGGCGAGGGCAGAGGTGATCACCTG 1438
H2  CTCATCTTTGTTCTCCAAGTAAAGCCACGAGGTCGGGGCGAGGGCAGAGGTGATCACCTG 1432
    ************************************************************

H1B CGTGTCCCATCTACAGACCTGCAGCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAA 1499
H1C CGTGTCCCATCTACAGACCAGCGGCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAA 1498
H2  CGTGTCCCATCTACAGACCTGCGGCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAA 1492
    ********************:**.**************************************

H1B AGGGTTACCCTGGGCACTGGCCTAGAGCCTCACCTCCTAATAGACTTAGCCCCATGAGTT 1559
H1C AGGGTTACCCTGGGCACTGGCCTAGAGCCTCACCTCCTAATAGACTTAGCCCCATGAGTT 1558
H2  AGGGTTACCCTGGGCACTGGCCTAGAGCCTCACCTCCTAATAGACTTAGCCCCATGAGTT 1552
    ************************************************************

H1B TGCCATGTTGAGCAGGACTATTTCTGGCACTTGCAAGTCCCATGATTTCTTCGGTAATTC 1619
H1C TGCCATGTTGAGCAGGACTATTTCTGGCACTTGCAAGTCCCATGATTTCTTCGGTAATTC 1618
H2  TGCCATGTTGAGCAGGACTATTTCTGGCACTTGCAAGTCCCATGATTTCTTCGGTAATTC 1612
    ************************************************************

H1B TGAGGGTGGGGGGAGGGACATGAAATCATCTTAGCTTAGCTTTCTGTCTGTGAATGTCTA 1679
H1C TGAGGGTGGGGGGAGGGACATGAAATCATCTTAGCTTAGCTTTCTGTCTGTGAATGTCTA 1678
H2  TGAGGGTGGGGGGAGGGACATGAAATCATCTTAGCTTAGCTTTCTGTCTGTGAATGTCTA 1672
    ************************************************************

H1B TATAGTGTATTGTGTGTTTTAACAAATGATTTACACTGACTGTTGCTGTAAAAGTGAATT 1739
H1C TATAGTGTATTGTGTGTTTTAACAAATGATTTACACTGACTGTTGCTGTAAAAGTGAATT 1738
H2  TATAGTGTATTGTGTGTTTTAACAAATGATTTACACTGACTGTTGCTGTAAAAGTGAATT 1732
    ************************************************************

H1B TGGAAATAAAGTTATTACTCTGATTAAATAAGGTCTCCATTCATGGATTCCAAGGACAAG 1799
H1C TGGAAATAAAGTTATTACTCTGATTAAATAAGGTCTCCATTCATGGATTCCAAGGACAAG 1798
H2  TGGAAATAAAGTTATTACTCTGATTAAATAAGGTCTCCATTCATGGATTCCAAGGACAAG 1792
    ************************************************************

H1B AAAGTCATATAGAATGTCTATTTTTTAAGTTCTTTCCCACGCACCCTTAGATAATTTAGC 1859
H1C AAAGTCATATAGAATGTCTATTTTTTAAGTTCTTTCCCACGCACCCTTAGATAATTTAGC 1858
H2  AAAGTCATATAGAATGTCTATTTTTTAAGTTCTTTCCCACGCACCCTTAGATAATTTAGC 1852
    ************************************************************

H1B TCAGAACAGGAAATGATAGTATTAATAAAAGCTGGACATCAGGATTAACAGCTCTCTCTG 1919
H1C TCAGAACAGGAAATGATAGTATTAATAAAAGCTGGACATCAGGATTAACAGCTCTCTCTG 1918
H2  TCAGAACAGGAAATGATAGTATTAATAAAAGCTGGACATCAGGATTAACAGCTCTCTCTG 1912
    ************************************************************

H1B GGGCCCTGAAGGTGAGAGTTCTCAGACTTGCTCATTTGCAGTTGCTTCTTTGTGATGCTG 1979
H1C GGGCCCTGAAGGTGAGAGTTCTCAGACTTGCTCATTTGCAGTTGCTTCTTTGTGATGCTG 1978
H2  GGGCCCTGAAGGTGAGAGTTCTCAGACTTGCTCATTTGCAGTTGCTTCTTTGTGATGCTG 1972
    ************************************************************

H1B GC 1981
H1C GC 1980
H2  GC 1974
    **
```

Appendix G

CP H1X mutations

In the below multiple sequence alignment of the wildtype and mutated H1 CP variants, sequence matches are denoted * and the **A120G** and **G596T** mutations are highlighted in **red**. Exon 0 is highlighted in **green**.

```
CP_H1  CAAATGCTCTGCGATGTGTTAAGCACTGTTTGAAATTCGTCTAATTTAAGATTTTTTTTT 60
CP_H1X CAAATGCTCTGCGATGTGTTAAGCACTGTTTGAAATTCGTCTAATTTAAGATTTTTTTTT 60
       ************************************************************

CP_H1  CTGACGTAACGGTTAGATTCACGTTTCTTTTTTTTTAAGTACAGTTCTACTGTATTGTAA 120
CP_H1X CTGACGTAACGGTTAGATTCACGTTTCTTTTTTTTTAAGTACAGTTCTACTGTATTGTAG 120
       ************************************************************.

CP_H1  CTGAGTTAGCTTGCTTTAAGCCGATTTGTTAAGGAAAGGATTCACCTTGGTCAGTAACAA 180
CP_H1X CTGAGTTAGCTTGCTTTAAGCCGATTTGTTAAGGAAAGGATTCACCTTGGTCAGTAACAA 180
       ************************************************************

CP_H1  AAAAGGTGGGAAAAAAGCAAGGAGAAAGGAAGCAGCCTGGGGGAAAGAGACCTTAGCCAG 240
CP_H1X AAAAGGTGGGAAAAAAGCAAGGAGAAAGGAAGCAGCCTGGGGGAAAGAGACCTTAGCCAG 240
       ************************************************************

CP_H1  GGGGGCGGTTTCGGGACTACGAAGGGTCGGGGCGGACGGACTCGAGGGCCGGCCACGTGG 300
CP_H1X GGGGGCGGTTTCGGGACTACGAAGGGTCGGGGCGGACGGACTCGAGGGCCGGCCACGTGG 300
       ************************************************************

CP_H1  AAGGCCGCTCAGGACTTCTGTAGGAGAGGACACCGCCCCAGGCTGACTGAAAGTAAAGGG 360
CP_H1X AAGGCCGCTCAGGACTTCTGTAGGAGAGGACACCGCCCCAGGCTGACTGAAAGTAAAGGG 360
       ************************************************************

CP_H1  CAGCGGACCCAGCGGCGGAGCCACTGGCCTTGCCCCGACCCCGCATGGCCCGAAGGAGGA 420
CP_H1X CAGCGGACCCAGCGGCGGAGCCACTGGCCTTGCCCCGACCCCGCATGGCCCGAAGGAGGA 420
       ************************************************************

CP_H1  CACCCACCCCCGCAACGACACAAAGACTCCAACTACAGGAGGTGGAGAAAGCGCGTGCGC 480
CP_H1X CACCCACCCCCGCAACGACACAAAGACTCCAACTACAGGAGGTGGAGAAAGCGCGTGCGC 480
       ************************************************************

CP_H1  CACGGAACGCGCGTGCGCGCTGCGGTCAGCGCCGCGGCCTGAGGCGTAGCGGGAGGGGGA 540
CP_H1X CACGGAACGCGCGTGCGCGCTGCGGTCAGCGCCGCGGCCTGAGGCGTAGCGGGAGGGGGA 540
       ************************************************************

CP_H1  CCGCGAAAGGGCAGCGCCGAGAGGAACGAGCCGGGAGACGCCGGACGGCCGAGCGGCAGG 600
CP_H1X CCGCGAAAGGGCAGCGCCGAGAGGAACGAGCCGGGAGACGCCGGACGGCCGAGCGTCAGG 600
       ************************************************************ ****

CP_H1  GCGCTCGCGCGCGCCCACTAGTGGCCGGAGGAGAAGGCTCCCGCGGAGGCCGCGCTGCCC 660
CP_H1X GCGCTCGCGCGCGCCCACTAGTGGCCGGAGGAGAAGGCTCCCGCGGAGGCCGCGCTGCCC 660
       ************************************************************

CP_H1  GCCCCCTCCCCTGGGGAGGCTCGCGTTCCCGCTGCTCGCGCCTGCGCCGCCCGCCGGCCT 720
CP_H1X GCCCCCTCCCCTGGGGAGGCTCGCGTTCCCGCTGCTCGCGCCTGCGCCGCCCGCCGGCCT 720
       ************************************************************

CP_H1  CAGGAACGCGCCCTCTTCGCCGGCGCGCGCCCTCGCAGTCACCGCCACCCACCAGCTCCG 780
CP_H1X CAGGAACGCGCCCTCTTCGCCGGCGCGCGCCCTCGCAGTCACCGCCACCCACCAGCTCCG 780
       ************************************************************

CP_H1  GCACCAACAGCAGCGCCGCTGCCACCGCCCACCTTCTGCCGCCGCCACCACAGCCACCTT 840
CP_H1X GCACCAACAGCAGCGCCGCTGCCACCGCCCACCTTCTGCCGCCGCCACCACAGCCACCTT 840
       ************************************************************
```

```
CP_H1  CTCCTCCTCCGCTGTCCTCTCCCGTCCTCGCCTCTGTCGACTATCAGGTAAGCGCCGCGG 900
CP_H1X CTCCTCCTCCGCTGTCCTCTCCCGTCCTCGCCTCTGTCGACTATCAGGTAAGCGCCGCGG 900
       ************************************************************

CP_H1  CTCCGAAATCTGCCTCGCCGTCCGCCTCTGTGCACCCCTGCGCCGCCGCCCCTCGCCCTC 960
CP_H1X CTCCGAAATCTGCCTCGCCGTCCGCCTCTGTGCACCCCTGCGCCGCCGCCCCTCGCCCTC 960
       ************************************************************

CP_H1  CCTCTCCGCAGACTGGGGCTTCGTGCGCCGGGCATCGGTCGGGGCCACCGCAGGGCCCCT 1020
CP_H1X CCTCTCCGCAGACTGGGGCTTCGTGCGCCGGGCATCGGTCGGGGCCACCGCAGGGCCCCT 1020
       ************************************************************

CP_H1  CCCTGCCTCCCCTGCTCGGGGGCTGGGGCCAGGGCGGCCTGGAAAGGGACCTGAGCAAGG 1080
CP_H1X CCCTGCCTCCCCTGCTCGGGGGCTGGGGCCAGGGCGGCCTGGAAAGGGACCTGAGCAAGG 1080
       ************************************************************

CP_H1  GATGCACGCACGCGTGAGTGCGCGCGTGTGTGTGTGCTGGAGGGTCTTCACCACCAGATT 1140
CP_H1X GATGCACGCACGCGTGAGTGCGCGCGTGTGTGTGTGCTGGAGGGTCTTCACCACCAGATT 1140
       ************************************************************

CP_H1  CGCGCAGACCCCAGGTGGAGGCTGTGCCGGCAGGGTGGGGCGCGGCGGCGGTGACTTGGG 1200
CP_H1X CGCGCAGACCCCAGGTGGAGGCTGTGCCGGCAGGGTGGGGCGCGGCGGCGGTGACTTGGG 1200
       ************************************************************

CP_H1  GGAGGGGGCTGCCCTTCACTCTCGACTGCAGCCTTTTGCCGCAATGGGCGTGTGTGTGTG 1260
CP_H1X GGAGGGGGCTGCCCTTCACTCTCGACTGCAGCCTTTTGCCGCAATGGGCGTGTGTGTGTG 1260
       ************************************************************

CP_H1  TGTGTGTGTGTGTGTGTGTGTGTGGAGGGGTCCGATAACGACCCCCGAAACCGAATCTGA 1320
CP_H1X TGTGTGTGTGTGTGTGTGTGTGTGGAGGGGTCCGATAACGACCCCCGAAACCGAATCTGA 1320
       ************************************************************

CP_H1  AATCCGCTGTCC 1332
CP_H1X AATCCGCTGTCC 1332
       ************
```

Appendix H

The CMV immediate early promoter

```
  1 CTCTGCTTAT ATAGACCTCC CACCGTACAC GCCTACCGCC CATTTGCGTC AATGGGGCGG
 61 AGTTGTTACG ACATTTTGGA AAGTCCCGTT GATTTTGGTG CCAAAACAAA CTCCCATTGA
 21 CGTCAATGGG GTGGAGACTT GGAAATCCCC GTGAGTCAAA CCGCTATCCA CGCCCATTGA
181 TGTACTGCCA AAACCGCATC ACCATGGTAA TAGCGATGAC TAATACGTAG ATGTACTGCC
241 AAGTAGGAAA GTCCCATAAG GTCATGTACT GGGCATAATG CCAGGCGGGC CATTTACCGT
301 CATTGACGTC AATAGGGGGC GTACTTGGCA TATGATACAC TTGATGTACT GCCAAGTGGG
361 CAGTTTACCG TAAATACTCC ACCCATTGAC GTCAATGGAA AGTCCCTATT GGCGTTACTA
421 TGGGAACATA CGTCATTATT GACGTCAATG GGCGGGGGTC GTTGGGCGGT CAGCCAGGCG
481 GGCCATTTAC CGTAAGTTAT GTAACGCGGA ACTCCATATA TGGGCTATGA ACTAATGACC
541 CCGTAATTGA TTACTATTAA TAACT
```

Appendix I

CP H1B vs H1C Minigenes

In the below multiple sequence alignment of the two CP minigenes, sequence matches are denoted *, with sequence differences highlighted in **red**. Exon 0 is highlighted in **green**.

```
CP_H1B          GGGGACAAGTTTGTACAAAAAAGCAGGCTTCCAAATGCTCTGCGATGTGTTAAGCACTGT 60
CP_H1C          GGGGACAAGTTTGTACAAAAAAGCAGGCTTCCAAATGCTCTGCGATGTGTTAAGCACTGT 60
                ************************************************************

CP_H1B          TTGAAATTCGTCTAATTTAAGATTTTTTTTTCTGACGTAACGGTTAGATTCACGTTTCTT 120
CP_H1C          TTGAAATTCGTCTAATTTAAGATTTTTTTTTCTGACGTAACGGTTAGATTCACGTTTCTT 120
                ************************************************************

CP_H1B          TTTTTTTAAGTACAGTTCTACTGTATTGTAACTGAGTTAGCTTGCTTTAAGCCGATTTGT 180
CP_H1C          TTTTTTTAAGTACAGTTCTACTGTATTGTAACTGAGTTAGCTTGCTTTAAGCCGATTTGT 180
                ************************************************************

CP_H1B          TAAGGAAAGGATTCACCTTGGTCAGTAACAAAAAAGGTGGGAAAAAAGCAAGGAGAAAGG 240
CP_H1C          TAAGGAAAGGATTCACCTTGGTCAGTAACAAAAAAGGTGGGAAAAAAGCAAGGAGAAAGG 240
                ************************************************************

CP_H1B          AAGCAGCCTGGGGGAAAGAGACCTTAGCCAGGGGGGCGGTTTCGGGACTACGAAGGGTCG 300
CP_H1C          AAGCAGCCTGGGGGAAAGAGACCTTAGCCAGGGGGGCGGTTTCGGGACTACGAAGGGTCG 300
                ************************************************************

CP_H1B          GGGCGGACGGACTCGAGGGCCGGCCACGTGGAAGGCCGCTCAGGACTTCTGTAGGAGAGG 360
CP_H1C          GGGCGGACGGACTCGAGGGCCGGCCACGTGGAAGGCCGCTCAGGACTTCTGTAGGAGAGG 360
                ************************************************************

CP_H1B          ACACCGCCCCAGGCTGACTGAAAGTAAAGGGCAGCGGACCCAGCGGCGGAGCCACTGGCC 420
CP_H1C          ACACCGCCCCAGGCTGACTGAAAGTAAAGGGCAGCGGACCCAGCGGCGGAGCCACTGGCC 420
                ************************************************************

CP_H1B          TTGCCCCGACCCCGCATGGCCCGAAGGAGGACACCCACCCCCGCAACGACACAAAGACTC 480
CP_H1C          TTGCCCCGACCCCGCATGGCCCGAAGGAGGACACCCACCCCCGCAACGACACAAAGACTC 480
                ************************************************************

CP_H1B          CAACTACAGGAGGTGGAGAAAGCGCGTGCGCCACGGAACGCGCGTGCGCGCTGCGGTCAG 540
CP_H1C          CAACTACAGGAGGTGGAGAAAGCGCGTGCGCCACGGAACGCGCGTGCGCGCTGCGGTCAG 540
                ************************************************************

CP_H1B          CGCCGCGGCCTGAGGCGTAGCGGGAGGGGGACCGCGAAAGGGCAGCGCCGAGAGGAACGA 600
CP_H1C          CGCCGCGGCCTGAGGCGTAGCGGGAGGGGGACCGCGAAAGGGCAGCGCCGAGAGGAACGA 600
                ************************************************************

CP_H1B          GCCGGGAGACGCCGGACGGCCGAGCGGCAGGGCGCTCGCGCGCGCCCACTAGTGGCCGGA 660
CP_H1C          GCCGGGAGACGCCGGACGGCCGAGCGGCAGGGCGCTCGCGCGCGCCCACTAGTGGCCGGA 660
                ************************************************************

CP_H1B          GGAGAAGGCTCCCGCGGAGGCCGCGCTGCCCGCCCCCTCCCCTGGGGAGGCTCGCGTTCC 720
CP_H1C          GGAGAAGGCTCCCGCGGAGGCCGCGCTGCCCGCCCCCTCCCCTGGGGAGGCTCGCGTTCC 720
                ************************************************************

CP_H1B          CGCTGCTCGCGCCTGCGCCGCCCGCCGGCCTCAGGAACGCGCCCTCTTCGCCGGCGCGCG 780
CP_H1C          CGCTGCTCGCGCCTGCGCCGCCCGCCGGCCTCAGGAACGCGCCCTCTTCGCCGGCGCGCG 780
                ************************************************************

CP_H1B          CCCTCGCAGTCACCGCCACCCACCAGCTCCGGCACCAACAGCAGCGCCGCTGCCACCGCC 840
CP_H1C          CCCTCGCAGTCACCGCCACCCACCAGCTCCGGCACCAACAGCAGCGCCGCTGCCACCGCC 840
                ************************************************************

CP_H1B          CACCTTCTGCCGCCGCCACCACAGCCACCTTCTCCTCCTCCGCTGTCCTCTCCCGTCCTC 900
CP_H1C          CACCTTCTGCCGCCGCCACCACAGCCACCTTCTCCTCCTCCGCTGTCCTCTCCCGTCCTC 900
                ************************************************************
```

```
CP_H1B          GCCTCTGTCGACTATCAGGTAAGCGCCGCGGCTCCGAAATCTGCCTCGCCGTCCGCCTCT 960
CP_H1C          GCCTCTGTCGACTATCAGGTAAGCGCCGCGGCTCCGAAATCTGCCTCGCCGTCCGCCTCT 960
                ************************************************************

CP_H1B          GTGCACCCCTGCGCCGCCGCCCCTCGCCCTCCCTCTCCGCAGACTGGGGCTTCGTGCGCC 1020
CP_H1C          GTGCACCCCTGCGCCGCCGCCCCTCGCCCTCCCTCTCCGCAGACTGGGGCTTCGTGCGCC 1020
                ************************************************************

CP_H1B          GGGCATCGGTCGGGGCCACCGCAGGGCCCCTCCCTGCCTCCCCTGCTCGGGGGCTGGGGC 1080
CP_H1C          GGGCATCGGTCGGGGCCACCGCAGGGCCCCTCCCTGCCTCCCCTGCTCGGGGGCTGGGGC 1080
                ************************************************************

CP_H1B          CAGGGCGGCCTGGAAAGGGACCTGAGCAAGGGATGCACGCACGCGTGAGTGCGCGCGTGT 1140
CP_H1C          CAGGGCGGCCTGGAAAGGGACCTGAGCAAGGGATGCACGCACGCGTGAGTGCGCGCGTGT 1140
                ************************************************************

CP_H1B          GTGTGTGCTGGAGGGTCTTCACCACCAGATTCGCGCAGACCCCAGGTGGAGGCTGTGCCG 1200
CP_H1C          GTGTGTGCTGGAGGGTCTTCACCACCAGATTCGCGCAGACCCCAGGTGGAGGCTGTGCCG 1200
                ************************************************************

CP_H1B          GCAGGGTGGGGCGCGGCGGCGGTGACTTGGGGGAGGGGGCTGCCCTTCACTCTCGACTGC 1260
CP_H1C          GCAGGGTGGGGCGCGGCGGCGGTGACTTGGGGGAGGGGGCTGCCCTTCACTCTCGACTGC 1260
                ************************************************************

CP_H1B          AGCCTTTTGCCGCAATGGGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGGAGGGGT 1320
CP_H1C          AGCCTTTTGCCGCAATGGGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGGAGGGGT 1320
                ************************************************************

CP_H1B          CCGATAACGACCCCCGAAACCGAATCTGAAATCCGCTGTCCACAACTTTGTATACAAAAG 1380
CP_H1C          CCGATAACGACCCCCGAAACCGAATCTGAAATCCGCTGTCCACAACTTTGTATACAAAAG 1380
                ************************************************************

CP_H1B          TTGTCACAACTTTGTATACAAAAGTTGTCCCTGGTGGTGTTGAATATGAACTGCTGCGGT 1440
CP_H1C          TTGTCACAACTTTGTATACAAAAGTTGTCCCTGGTGGTGTTGAATATGAACTGCTGCGGT 1440
                ************************************************************

CP_H1B          GTTGGTAAATTAAGCAAGCAGATAGATGTAAATAACGCTTGGGCAGGAATATGGAGCACG 1500
CP_H1C          GTTGGTAAATTAAGCAAGCAGATAGATGTAAATAACGCTTGGGCAGGAATATGGAGCACG 1500
                ************************************************************

CP_H1B          GGATGAGGATGGGCGGCCAACTGTTAGAGAGGGTAGCAGGGAGGCTGAGATCTGCCTGCC 1560
CP_H1C          GGATGAGGATGGGCGGCCAACTGTTAGAGAGGGTAGCAGGGAGGCTGAGATCTGCCTGCC 1560
                ************************************************************

CP_H1B          ATGAACTGGGAGGAGAGGCTCCTCTCTCTCTTCACCCCCACTCTGCCCCCCAACACTCCT 1620
CP_H1C          ATGAACTGGGAGGAGAGGCTCCTCTCTCTCTTCACCCCCACTCTGCCCCCCAACACTCCT 1620
                ************************************************************

CP_H1B          CAGAACTTATCCTCTCCTCTTCTTTCCCCAGGTGAACTTTGAACCAGGATGGCTGAGCCC 1680
CP_H1C          CAGAACTTATCCTCTCCTCTTCTTTCCCCAGGTGAACTTTGAACCAGGATGGCTGAGCCC 1680
                ************************************************************

CP_H1B          CGCCAGGAGTTCGAAGTGATGGAAGATCACGCTGGGACGTACGGGTTGGGGGACAGGAAA 1740
CP_H1C          CGCCAGGAGTTCGAAGTGATGGAAGATCACGCTGGGACGTACGGGTTGGGGGACAGGAAA 1740
                ************************************************************

CP_H1B          GATCAGGGGGGCTACACCATGCACCAAGACCAAGAGGGTGACACGGACGCTGGCCTGAAA 1800
CP_H1C          GATCAGGGGGGCTACACCATGCACCAAGACCAAGAGGGTGACACGGACGCTGGCCTGAAA 1800
                ************************************************************

CP_H1B          GGTTAGTGGACAGCCATGCACAGCAGGCCCAGATCACTGCAAGCCAAGGGGTGGCGGGAA 1860
CP_H1C          GGTTAGTGGACAGCCATGCACAGCAGGCCCAGATCACTGCAAGCCAAGGGGTGGCGGGAA 1860
                ************************************************************

CP_H1B          CAGTTTGCATCCAGAATTGCAAAGAAATTTTAAATACATTATTGTCTTAGACTGTCAGTA 1920
CP_H1C          CAGTTTGCATCCAGAATTGCAAAGAAATTTTAAATACATTATTGTCTTAGACTGTCAGTA 1920
                ************************************************************

CP_H1B          AAGTAAAGCCTCATTAATTTGAGTGGGCCAAGATAACTCAAGCAGTGAGATAATGGCCAG 1980
CP_H1C          AAGTAAAGCCTCATTAATTTGAGTGGGCCAAGATAACTCAAGCAGTGAGATAATGGCCAG 1980
                ************************************************************
```

```
CP_H1B          ACTCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGAAGGCCCAGGCAGGAGGATCCCTT 2040
CP_H1C          ACACGGTGGCTCACGCCTGTAATCCCAGCACTTTGGAAGGCCCAGGCAGGAGGATCCCTT 2040
                ** ************************************************************

CP_H1B          GAGGCCAGGAATTTGAGACCGGCCTGGGCAACATAGCAAGACCCCGTCTCTAAAATAATT 2100
CP_H1C          GAGGCCAGGAATTTGAGACCGGCCTGGGCAACATAGCAAGACCCCGTCTCTAAAATAATT 2100
                ************************************************************

CP_H1B          TAAAAATTAGCCAGGTGTTGTGGTGCATGTCTATAGTCCTAGCTACTCAGGATGCTGAGG 2160
CP_H1C          TAAAAATTAGCCAGGTGTTGTGGTGCATGTCTATAGTCCTAGCTACTCAGGATGCTGAGG 2160
                ************************************************************

CP_H1B          CAGAAGGATCACTTGAGCCCAGGAGTTCAAGGTTGCAGTAAGCTGTGATTATAAAACTGC 2220
CP_H1C          CAGAAGGATCACTTGAGCCCAGGAGTTCAAGGTTGCAGTAAGCTGTGATTATAAAACTGC 2220
                ************************************************************

CP_H1B          ACTCCAGCCTGAGCAACAGAGCAAGACCCTGTCAAAAAAAAAAGAAAAGAAAAAAGAAAG 2280
CP_H1C          ACTCCAGCCTGAGCAACAGAGCAAGACCCTGTCAAAAAAAAAAGAAAAGAAAAAAGAAAG 2280
                ************************************************************

CP_H1B          AAAGAAATTTACCTTGAGTTACCCACATGAGTGAATGTAGGGACAGAGATTTTAGGGCCT 2340
CP_H1C          AAAGAAATTTACCTTGAGTTACCCACATGAGTGAATGTAGGGACAGAGATTTTAGGGCCT 2340
                ************************************************************

CP_H1B          TAACAATCTCTCAAATACAGGGTACTTTTTGAGGCATTAGCCACACCTGTTAGCTTATAA 2400
CP_H1C          TAACAATCTCTCAAATACAGGGTACTTTTTGAGGCATTAGCCACACCTGTTAGCTTATAA 2400
                ************************************************************

CP_H1B          ATCAGTGGTATTGATTAGCATGTAAAATATGTGACTTTAAACATTGCTTTTTATCTCTTA 2460
CP_H1C          ATCAGTGGTATTGATTAGCATGTAAAATATGTGACTTTAAACATTGCTTTTTATCTCTTA 2460
                ************************************************************

CP_H1B          CTTAGATCAGGCCTGAGTGGCCTCTCTTTAGCAAGAGTTGGTTAGCCCTGGGATTCTTAC 2520
CP_H1C          CTTAGATCAGGCCTGAGTGGCCTCTCTTTAGCAAGAGTTGGTTAGCCCTGGGATTCTTAC 2520
                ************************************************************

CP_H1B          TGTAGCCACATTAATAAACAACATCGACTTCTAAACATTCTATAATACCATCTTTTGGCC 2580
CP_H1C          TGTAGCCACATTAATAAACAACATCGACTTCTAAACATTCTATAATACTATCTTTTGGCC 2580
                ********************************************** ************

CP_H1B          AAATTGACTTCGCCTCTTCTCGAGCACAGGGAAGGGACAATTCAGCCCTTCTAGGAGGAG 2640
CP_H1C          AAATTTACTTCGCCTCTTCTCGAGCACAGGGAAGGGACAATTCAGCCCTTCTAGGAGGAG 2640
                ***** ********************************************************

CP_H1B          GAGGAGGTAGTTTTCTCATTTCTATTAAGGCAACAAAAGCTGCCTTACTAAGGACATTCT 2700
CP_H1C          GAGGAGGTAGTTTTCTCATTTCTATTAAGGCAACAAAAGCTGCCTTACTAAGGACATTCT 2700
                ************************************************************

CP_H1B          TGGTGGAGGGCGTGACTGTCAACCACTGTGATCATTTGGGCCTCTCTTGCCCAGGCTTCC 2760
CP_H1C          TGGTGGAGGGCGTGACTGTCAACCACTGTGATCATTTGGGCCTCTCTTGCCCAGGCTTCC 2760
                ************************************************************

CP_H1B          CATTCTGAAAGGACAGTTTTATTGTAGGTACACATGGCTGCCATTTCAAATGTAACTCAC 2820
CP_H1C          CATTCTGAAAGGACAGTTTTATTGTAGGTACACATGGCTGCCATTTCAAATGTAACTCAC 2820
                ************************************************************

CP_H1B          AGCTTGTCCATCAGTCCTTGGAGGTCTTTCTATGAAAGGAGCTTGGTGGCGTCCAAACAC 2880
CP_H1C          AGCTTGTCCATCAGTCCTTGGAGGTCTTTCTATGAAAGGAGCTTGGTGGCGTCCAAACAC 2880
                ************************************************************

CP_H1B          CACCCAATGTCCACTTAGAAGTAAGCACCGTGTCTGCCCTGAGCTGACTCCTTTTCCAAG 2940
CP_H1C          CACCCAATGTCCACTTAGAAGTAAGCACCGTGTCTGCCCTGAGCTGACTCCTTTTCCAAG 2940
                ************************************************************

CP_H1B          GAAGGGGTTGGATCGCTGAGTGTTTTTCCAGGTGTCTACTTGTTGTTAATTAATAGCAAT 3000
CP_H1C          GAAGGGGTTGGATCGCTGAGTGTTTTTCCAGGTGTCTACTTGTTGTTAATTAATAGCAAT 3000
                ************************************************************

CP_H1B          GACAAAGCAGAAGGTTCATGCGTAGCTCGGCTTTCTGGTATTTGCTGCCCGTTGACCAAT 3060
CP_H1C          GACAAAGCAGAAGGTTCATGCGTAGCTCGGCTTTCTGGTATTTGCTGCCCGTTGACCAAT 3060
                ************************************************************
```

```
CP_H1B          GGAAGATAAACCTTTGCCTCAGGTGGCACCACTAGCTGGTTAAGAGGCACTTTGTCCTTT 3120
CP_H1C          GGAAGATAAACCTTTGCCTCAGGTGGCACCACTAGCTGGTTAAGAGGCACTTTGTCCTCT 3120
                ************************************************************ *

CP_H1B          CACCCAGGAGCAAACGCACATCACCTGTGTCCTCATCTGATGGCCCTGGTGTGGGGCACA 3180
CP_H1C          CACCCAGGAGCAAACGCACATCACCTGTGTCCTCATCTGATGGCCCTGGTGTGGG-CACA 3179
                ************************************************************ ****

CP_H1B          GTCGTGTTGGCAGGGAGGGAGGTGGGGTTGGTCCCCTTTGTGGGTTTGTTGCGAGGCCGT 3240
CP_H1C          GTCGTGTTGGCAGGGAGGGAGGTGGGGTTGGTCCCCTTTGTGGGTTTGTTGCGAGGCCGT 3239
                ************************************************************

CP_H1B          GTTCCAGCTGTTTCCACAGGGAGCGATTTTCAGCTCCACAGGACACTGCTCCCCAGTTCC 3300
CP_H1C          GTTCCAGCTGTTTCCACAGGGAGCGATTTTAAGCTCCACAGGACACTGCTCCCCAGTTCC 3299
                ****************************** ******************************

CP_H1B          TCCTGAGAACAAAAGGGGGCGCTGGGGAGAGGCCACCGTTCTGAGGGCTCACTGTATGTG 3360
CP_H1C          TCCTGAGAACAAAAGGGGGCGCTGGGGAGAGGCCACCGTTCTGAGGGCTCACTGTATGTG 3359
                ************************************************************

CP_H1B          TTCCAGAATCTCCCCTGCAGACCCCCACTGAGGACGGATCTGAGGAACCGGGCTCTGAAA 3420
CP_H1C          TTCCAGAATCTCCCCTGCAGACCCCCACTGAGGACGGATCTGAGGAACCGGGCTCTGAAA 3419
                ************************************************************

CP_H1B          CCTCTGATGCTAAGAGCACTCCAACAGCGGAAGGTGGGCCCCCCTTCAGACGCCCCCTCC 3480
CP_H1C          CCTCTGATGCTAAGAGCACTCCAACAGCGGAAGGTGGGCCCCCCTTCAGACGCCCCCTCC 3479
                ************************************************************

CP_H1B          ATGCCTCCAGCCTGTGCTTAGCCGTGCTTTGAGCCTCCCTCCTGGCTGCATCTGCTGCTC 3540
CP_H1C          ATGCCTCCAGCCTGTGCTTAGCCGTGCTTTGAGCCTCCCTCCTGGCTGCATCTGCTGCTC 3539
                ************************************************************

CP_H1B          CCCCTGGCTGAGAGATGTGCTCACTCCTTCGGTGCTTTGCAGGACAGCGTGGTGGGAGCT 3600
CP_H1C          CCCCTGGCTGAGAGATGTGCTCACTCCTTCGGTGCTTTGCAGGACAGCGTGGTGGGAGCT 3599
                ************************************************************

CP_H1B          GAGCCTTGCGTCGATGCCTTGCTTGCTGGTGCTGAGTGTGGGCACCTTCATCCCGTGTGT 3660
CP_H1C          GAGCCTTGCGTCGATGCCTTGCTTGCTGGTGCTGAGTGTGGGCACCTTCATCCCGTGTGT 3659
                ************************************************************

CP_H1B          GCTCTGGAGGCAGCCACCCTTGGACAGTCCCGCGCACAGCTCCACAAAGCCCCGCTCCAT 3720
CP_H1C          GCTCTGGAGGCAGCCACCCTTGGACAGTCCCGCGCACAGCTCCACAAAGCCCCGCTCCAT 3719
                ************************************************************

CP_H1B          ACGATTGTCCTCCCACACCCCCTTCAAAGGCCCCCTCCTCTCTCTTTCTTCAGGGGCCAG 3780
CP_H1C          ACGATTGTCCTCCCACACCCCCTTCAAAAGCCCCCTCCTCTCTCTTTCTTCAGGGGCCAG 3779
                ************************** ******************************

CP_H1B          TAGGTCCCAGAGCAGCCATTTGGCTGAGGGAAGGGGCAGGTCAGTGGACATCTGATCTTG 3840
CP_H1C          TAGGTCCCAGAGCAGCCATTTGGCTGAGGGAAGGGGCAGGTCAGTGGACATCTGATCTTG 3839
                ************************************************************

CP_H1B          GTTTAGTATCATTCATTTTGGGGGCTCTGGGTGTGGCCTGGGCCTCTGGACTTTGGCCAC 3900
CP_H1C          GTTTAGTATCCTTCATTTTGGGGGCTCTGGGTGTGGCCTGGGCCTCTGGACTTTGGCCAC 3899
                ********** ***************************************************

CP_H1B          GGTGTTTGTTCCAGCCCTTCTCCTAACCTGTCCTTTCCAGACACTCGGCATCTAGGTTAT 3960
CP_H1C          GGTGTTTGTTCCAGCCCTTCTCCTAACCTGTCCTTTCCAGACACTCGGCATCTAGGTTAT 3959
                ************************************************************

CP_H1B          TAGCACCTCGCATACTTTCTGACATGCTCCTCAGTCCTGATTTTGACCATCTTCTCTTGC 4020
CP_H1C          TAGCACCTCGCATACTTTCCGACATGCTCCTCAGTCCTGATTTTGACCATCTTCTCTTGC 4019
                ****************** ****************************************

CP_H1B          TTCCCATCTGTGTCAGTCAAGCCGCGGAAAGCCTTCAAAGCTGACAACTCCTTATGTGTA 4080
CP_H1C          TTCCCATCTGTGTCAGTCAAGCCGCGGAAAGCCTTCAAAGCTGACAACTCCTTATGTGTA 4079
                ************************************************************

CP_H1B          CCCGGAAAGGCCTGGGAGTGTGCCAGGGCATTGCTCGGGAGGGACGCTGATTTGGAAGCA 4140
CP_H1C          CCCGGAAAGGCCTGGGAGTGTGCCAGGGCATTGCTCGGGAGGGACGCTGATTTGGAAGCA 4139
                ************************************************************
```

```
CP_H1B          TTTACCTGATGAGAGACTGACAGCAGCTCCTGGTAGCCGAGCTTTCCCTCCTGCCTCTGC 4200
CP_H1C          TTTACCTGATGAGAGACTGACAGCAGCTCCTGGTAGCCGAGCTTTCCCTCCTGCCTCTGC 4199
                ************************************************************

CP_H1B          TGTGAAGGTGGACCCATCCAACAGTCAAATGCCTGACTCTGGACAGGAGCGGACCTATTT 4260
CP_H1C          TGTGAAGGTGGACCCATCCAACAGTCAAACGCCTGACTCTGGACAGGAGCGGACCTATTT 4259
                *************************** ****************************

CP_H1B          ATTGCCATGCAAGGGACTCTGCACTTTTGAATTGTGGGTCATGGGCTTGGATTTAGGGGT 4320
CP_H1C          ATTGCCATGCAAGGGACTCTGCACTTTTGAATTGTGGGTCATGGGCTTGGATTTAGGGGT 4319
                ************************************************************

CP_H1B          TAGAGCTGGGAGAAGTCTTGGAAGTCACCTAGAGATGACACTGCCATTTTGCAGATGAGG 4380
CP_H1C          TAGAGCTGGGAGAAGTCTTGGAAGTCACCTAGAGATGACACTGCCATTTTGCAGATGAGG 4379
                ************************************************************

CP_H1B          AAACCGTCCAATAAAAATGGACCAAGGACTTGCCCAAAGCCTCACAGCAAAACCATAGGC 4440
CP_H1C          AAACCGTCCAATCAAAATGGACCAAGGACTTGCCCAAAGCCTCACAGCAAAACCATAGGC 4439
                ************ ************************************************

CP_H1B          CCCCGCACTAACCCCAGAGTCCCTGTGCTGTCTTAAGGATCATATAGTTGTAAGCAATCA 4500
CP_H1C          CCCCGCACTAACCCCAGAGTCCCTGTGCTGTCTTAAGGATCATATAGTTGTAAGCAATCA 4499
                ************************************************************

CP_H1B          TCTGGTTTTCAGTATTTCTTCTTTTAAAATGCCTGGGGCCATGCCCAGCAGTCTGTTTCA 4560
CP_H1C          TCTGGTTTTCAGTATTTCTTCTTTTAAAATGCCTGGGGCCATGCCCAGCAGTCTGTTTCA 4559
                ************************************************************

CP_H1B          CTGCAGCGTTTACACAGGGCTGCCGGGCTTTCCTGGTGGATGAGCTGGGCGGTTCATGAG 4620
CP_H1C          CTGCAGCGTTTACACAGGGCTGCCGGGCTTTCCTGGTGGATGAGCTGGGCGGTTCATGAG 4619
                ************************************************************

CP_H1B          CCAGAACCACTCAGCAGCATGTCAGTGTGCTTCCTGGGGAGCTGGTAGCAGGGGCTCCGG 4680
CP_H1C          CCAGAACCACTCAGCAGCATGTCAGTGTGCTTCCTGGGGAGCTGGTAGCAGGGGCTCCGG 4679
                ************************************************************

CP_H1B          GCCCTACTTCAGGGCTGCTTTCTGGCATATGGCTGATCCCCTCCTCACTCCTCCTCCCTG 4740
CP_H1C          GCCCTACTTCAGGGCTGCTTTCTGGCATATGGCTGATCCCCTCCTCACTCCTCCTCCCTG 4739
                ************************************************************

CP_H1B          CATTGCTCCTGCGCAAGAAGCAAAGGTGAGGGGCTGGGTATGGCTCGTCCTGGCCCCTCT 4800
CP_H1C          CATTGCTCCTGCGCAAGAAGCAAAGGTGAGGGGCTGGGTATGGCTCGTCCTGGCCCCTCT 4799
                ************************************************************

CP_H1B          AAGGTGGATCTCGGTGGTTTCTAGATGTGACAGCACCCTTAGTGGATGAGGGAGCTCCCG 4860
CP_H1C          AAGGTGGATCTCGGTGGTTTCTAGATGTGACAGCACCCTTAGTGGATGAGGGAGCTCCCG 4859
                ************************************************************

CP_H1B          GCAAGCAGGCTGCCGCGCAGCCCCACACGGAGATCCCAGAAGGAACCACAGGTGAGGGTA 4920
CP_H1C          GCAAGCAGGCTGCCGCGCAGCCCCACACGGAGATCCCAGAAGGAACCACAGGTGAGGGTA 4919
                ************************************************************

CP_H1B          AGCCCCAGAGACCCCCAGGCAGTCAAGGCCCTGCTGGGTGCCCCAGCTGACCTGTGACAG 4980
CP_H1C          AGCCCCAGAGACCCCCAGGCAGTCAAGGCCCTGCTGGGTGCCCCAGCTGACCTGTGACAG 4979
                ************************************************************

CP_H1B          AAGTGAGGGAGCTTTGCGTGTTTATCCTCCTGTGGGGCAGGAACATGGGTGGATTCTGGC 5040
CP_H1C          AAGTGAGGGAGCTTTGCGTGTTTATCCTCCTGTGGGGCAGGAACATGGGTGGATTCTGGC 5039
                ************************************************************

CP_H1B          TCCTGGGAATCTTGGGTTGTGAGTAGCTCGATGCCTTGGTGCTCAGTTACCTCCCTGGCT 5100
CP_H1C          TCCTGGGAATCTTGGGTTGTGAGTAGCTCGATGCCTTGGTGCTCAGTTACCTCCCTGGCT 5099
                ************************************************************

CP_H1B          GCCTGCCAGCCTCTCAGAGCATTTAGGGCCTTCTGGACTTCTAGATGCTCCTCATCTTGC 5160
CP_H1C          GCCTGCCAGCCTCTCAGAGCATTTAGGGCCTTCTGGACTTCTAGATGCTCCTCATCTTGC 5159
                ************************************************************

CP_H1B          CTCAGTCAGCGCGTCAGTTCCAGAGACTTCTCTGCAGGGTTTTCTGGGGCAGGTGGTGGC 5220
CP_H1C          CTCAGTCAGCGCGTCAGTTCCAGAGACTTCTCTGCAGGGTTTTCTGGGGCAGGTGGTGGC 5219
                ************************************************************
```

```
CP_H1B          AGACCCGTGCCTTCTTGACACCTGAGGTCAGTCCACCCTCCTGCTCAGACTGCCCAGCAC 5280
CP_H1C          AGACCCGTGCCTTCTTGACACCTGAGGTCAGTCCACCCTCCTGCTCAGACTGCCCAGCAC 5279
                ************************************************************

CP_H1B          AGGGTCACCTCCCAAGGGGTGGACCCCAAGATCACCTGAGCGCACAGAGGGTGCAGATGA 5340
CP_H1C          AGGGTCACCTCCCAAGGGGTGGACCCCAAGATCACCTGAGCGCACAGAGGGTGCAGATGA 5339
                ************************************************************

CP_H1B          CTGGACCACACCTTTTGGTGATCTTAATGAGGTGGTCCCAGAGGAGCTCAGACATGCAAT 5400
CP_H1C          CTGGACCACACCTTTTGGTGATCTTAATGAGGTGGTCCCAGAGGAGCTCAGACATGCAAT 5399
                ************************************************************

CP_H1B          CTAGCATCCAGTTCTGGGACTCTGTCTCCTTTTCAAACGTATTCATGTAGAACAGGCATG 5460
CP_H1C          CTAGCATCCAGTTCTGGGACTCTGTCTCCTTTTCAAACGTATTCATGTAGAACAGGCATG 5459
                ************************************************************

CP_H1B          ACGAGAATGCCTTGTCAACATGGGTGATGGGGAATCAATCAGACAGGGCGCATGCCCCGT 5520
CP_H1C          ACGAGAATGCCTTGTCAACATGGGTGATGGGGAATCAATCAGACAGGGCGCATGCCCCGT 5519
                ************************************************************

CP_H1B          GAGCCCATTGCCCGCCCTCCCATGCCCTCAGCAGCTGCCTGGGGACAGCCAATGGCCTGG 5580
CP_H1C          GAGCCCATTGCCCGCCCTCCCATGCCCTCAGCAGCTGCCTGGGGACAGCCAATGGCCTGG 5579
                ************************************************************

CP_H1B          GTGTTTCTGAGGCTACCACATGGCTTCCAGGAAACTCGAGAACCTTTCTCTCCCTTGCCT 5640
CP_H1C          GTGTTTCTGAGGCTACCACATGGCTTCCAGGAAACTCGAGAACCTTTCTCTCCCTTGCCT 5639
                ************************************************************

CP_H1B          ACACTCTTCACACAGGCCTGTGCTGGCCAGCGGTGGGGATCCGGCATTCCTATCTTAGGT 5700
CP_H1C          ACACTCTTCACACAGGCCTGTGCTGGCCAGCGGTGGGGATCCGGCATTCCTATCTTAGGT 5699
                ************************************************************

CP_H1B          GCAGAAAGTGACTGACTCATTGCAGGCCTGGGAGATAAGACTGATGGCCCAGCCAGCAAG 5760
CP_H1C          GCAGAGAGTGACTGACTCATTGCAGGCCTGGGAGATAAGACTGATGGCCCAGCCAGCAAG 5759
                ***** ********************************************************

CP_H1B          ATGTATGGATTTCTCAGAGGCAGTGGCCTCTGTCATTGTCCTCAGGAAATGCTGGTGATT 5820
CP_H1C          ATGTATGGATTTCTCAGAGGCAGTGGCCTCTGTCATTGTCCTCAGGAAATGCTGGTGATT 5819
                ************************************************************

CP_H1B          CTGGTGGCCTGAGGTCAATGCATGTCAACGTGGCCAACTTGCCTTATAAACTTTTTTTCT 5880
CP_H1C          CTGGTGGCCTGAGGTCAATGCATGTCAACGTGGCCAACTTGCCTTATAAACTTTTTTTCT 5879
                ************************************************************

CP_H1B          GGACAATTGCGTGCACTGTCCTGTAACAGTGTCCTGTTGTTTATGATGCAGAAATAGGTG 5940
CP_H1C          GGACAATTGCGTGCACTGTCCTGTAACAGTGTCCTGTTGTTTATGATGCAGAAATAGGTG 5939
                ************************************************************

CP_H1B          TTTTTAAAGCCTATTGATTTTGGTACTATTAATGTGGTCAGGAACTTTCTCAGTCTTTCT 6000
CP_H1C          TTTTTAAAGCCTATTGATTTTGGTACTATTAATGTGGTCAGGAACTTTCTCAGTCTTTCT 5999
                ************************************************************

CP_H1B          TGTTTGGGGTGAGCTGTGGCTTCCTAAACAGGAACCCAAGACACCCCCAAAAGCTGCTCA 6060
CP_H1C          TGTTTGGGGTGAGCTGTGGCTTCCTAAACAGGAACCCAAGACACCCCCAAAAGCTGCTCA 6059
                ************************************************************

CP_H1B          CCAGCACTGCCAGCCTCCCTCTTACCAAGTAGCACCCGTTCAGGACATTCTGCGAAAGGC 6120
CP_H1C          CCAGCACTGCCAGCCTCCCTCTTACCAAGTAGCACCCGTTCAGGACATTCTGCGAAAGGC 6119
                ************************************************************

CP_H1B          ATTTGCCCAGAAGTTGGGAGGAAGGAAATGTAACATTTTGGGGCACCTACCATATGCCAG 6180
CP_H1C          ATTTGCCCAGAAGTTGGGAGGAAGGAAATGTAACATTTTGGGGCACCTACCATATGCCAG 6179
                ************************************************************

CP_H1B          GCACCAGGCTAAACGTGTTCACACAAATTCTCTTACTAACCCTCACCATCCTTCTACAAG 6240
CP_H1C          GCACCAGGCTAAACGTGTTCACACAAATTCTCTTACTAACCCTCACCATCCTTCTACAAG 6239
                ************************************************************

CP_H1B          ACAAACTAGTATCTTCATCTTGGGGTTCAAGATGAGGAAATGGAGGCTCAGAGAGGTTGA 6300
CP_H1C          ACAAACTAGTATCTTCATCTTGGGGTTCAAGATGAGGAAATGGAGGCTCAGAGAGGTTGA 6299
                ************************************************************
```

```
CP_H1B          ATGAATGCCGGTGCCTGGATATGAACCCCATCTGCCTGACTCCGCAACCCAGGCAAAGTC 6360
CP_H1C          ATGAATGCCGGTGCCTGGATATGAACCCCATCTGCCTGACTCCGCAACCCAGGCAAAGTC 6359
                ************************************************************

CP_H1B          TTTCCTTGAACTTCCCAGCAGCCACTGCTTAGACACAGCCTCCACAACCATGGCTCAGCA 6420
CP_H1C          TTTCCTTGAACTTCCCAGCAGCCACTGCTTAGACACAGCCTCCACAACCATGGCTCAGCA 6419
                ************************************************************

CP_H1B          GCAAATTGCTTCTCTGACCTCACTCAGCCTGTGTGTCCTTGTTGAGTGAGGCATTCAGGA 6480
CP_H1C          GCAAATTGCTTCTCTGACCTCACTCAGCCTGTGTGTCCTTGTTGAGTGAGGCATTCAGGA 6479
                ************************************************************

CP_H1B          CCCTGGTCCCAAAGTGGAGAAAGTCTTTCCTACTAGGTCATAGCTACACCTGCATGTGGG 6540
CP_H1C          CCCTGGTCCCAAAGTGGAGAAAGTCTTTCCTACTAGGTCATAGCTACACCTGCATGTGGG 6539
                ************************************************************

CP_H1B          TGCTGTGCCTTTTGTTTAGTGAACTTTTATCACCAGCATCCTCAGCAATGACATTTGCAG 6600
CP_H1C          TGCTGTGCCTTTTGTTTAGTGAACTTTTATCACCAGCATCCTCAGCAATGACATTTGCAG 6599
                ************************************************************

CP_H1B          AGAAGCCAGAGCTGAGGCACCTTGGTATTCTTGGGATGTGACTTTCCTGAATGTTTAAGG 6660
CP_H1C          AGAAGCCAGAGCTGAGGCACCTTGGTATTCTTGGGATGTGACTTTCCTGAATGTTTAAGG 6659
                ************************************************************

CP_H1B          GAAAATGCCCGAAGGTACAGAGAGCTTGGTTTCTAGTAAACAATAACTGTCTTGCTTTTA 6720
CP_H1C          GAAAATGCCCGAAGGTACAGAGAGCTTGGTTTCTAGTAAACAATAACTGTCTTGCTTTTA 6719
                ************************************************************

CP_H1B          CCCCCCTTCATTTGCTGACACATACACCAGCACCCAACTTTTCTATACAAAGTTGTCCAG 6780
CP_H1C          CCCCCCTTCATTTGCTGACACATACACCAGCACC-AACTTTTCTATACAAAGTTGTCCAG 6778
                ********************************* ************************

CP_H1B          CTGAAGAAGCAGGCATTGGAGACACCCCCAGCCTGGAAGACGAAGCTGCTGGTCACGTGA 6840
CP_H1C          CTGAAGAAGCAGGCATTGGAGACACCCCCAGCCTGGAAGACGAAGCTGCTGGTCACGTGA 6838
                ************************************************************

CP_H1B          CCCAAGCTCGCATGGTCAGTAAAAGCAAAGACGGGACTGGAAGCGATGACAAAAAAGCCA 6900
CP_H1C          CCCAAGCTCGCATGGTCAGTAAAAGCAAAGACGGGACTGGAAGCGATGACAAAAAAGCCA 6898
                ************************************************************

CP_H1B          AGGGGGCTGATGGTAAAACGAAGATCGCCACACCGCGGGGAGCAGCCCCTCCAGGCCAGA 6960
CP_H1C          AGGGGGCTGATGGTAAAACGAAGATCGCCACACCGCGGGGAGCAGCCCCTCCAGGCCAGA 6958
                ************************************************************

CP_H1B          AGGGCCAGGCCAACGCCACCAGGATTCCAGCAAAAACCCCGCCCGCTCCAAAGACACCAC 7020
CP_H1C          AGGGCCAGGCCAACGCCACCAGGATTCCAGCAAAAACCCCGCCCGCTCCAAAGACACCAC 7018
                ************************************************************

CP_H1B          CCAGCTCTGGTGAACCTCCAAAATCAGGGGATCGCAGCGGCTACAGCAGCCCCGGCTCCC 7080
CP_H1C          CCAGCTCTGGTGAACCTCCAAAATCAGGGGATCGCAGCGGCTACAGCAGCCCCGGCTCCC 7078
                ************************************************************

CP_H1B          CAGGCACTCCCGGCAGCCGCTCCCGCACCCCGTCCCTTCCAACCCCACCCACCCGGGAGC 7140
CP_H1C          CAGGCACTCCCGGCAGCCGCTCCCGCACCCCGTCCCTTCCAACCCCACCCACCCGGGAGC 7138
                ************************************************************

CP_H1B          CCAAGAAGGTGGCAGTGGTCCGTACTCCACCCAAGTCGCCGTCTTCCGCCAAGAGCCGCC 7200
CP_H1C          CCAAGAAGGTGGCAGTGGTCCGTACTCCACCCAAGTCGCCGTCTTCCGCCAAGAGCCGCC 7198
                ************************************************************

CP_H1B          TGCAGACAGCCCCCGTGCCCATGCCAGACCTGAAGAATGTCAAGTCCAAGATCGGCTCCA 7260
CP_H1C          TGCAGACAGCCCCCGTGCCCATGCCAGACCTGAAGAATGTCAAGTCCAAGATCGGCTCCA 7258
                ************************************************************

CP_H1B          CTGAGAACCTGAAGCACCAGCCGGGAGGCGGGAAGTCTAGAGTGAGAGTGGCTGGCTGCG 7320
CP_H1C          CTGAGAACCTGAAGCACCAGCCGGGAGGCGGGAAGTCTAGAGTGAGAGTGGCTGGCTGCG 7318
                ************************************************************

CP_H1B          CGTGGAGGTGTGGGGGGCTGCGCCTGGAGGGGTAGGGCTGTGCCTGGAAGGGTAGGGCTG 7380
CP_H1C          CGTGGAGGTGTGGGGGGCTGCGCCTGGAGGGGTAGGGCTGTGCCTGGAAGGGTAGGGCTG 7378
                ************************************************************
```

```
CP_H1B          CGCCTGGAGGTGCGCGGTTGAGCGTGGAGTCGTGGGACTGTGCATGGAGGTGTGGGGCTC 7440
CP_H1C          CGCCTGGAGGTGCGCGGTTGAGCGTGGAGTCGTGGGACTGTGCATGGAGGTGTGGGGCTC 7438
                ************************************************************

CP_H1B          CCCGCACCTGAGCACCCCCGCATAACACCCCAGTCCCCTCTGGACCCTCTTCAAGGAAGT 7500
CP_H1C          CCCGCACCTGAGCACCCCCGCATAACACCCCAGTCCCCTCTGGACCCTCTTCAAGGAAGT 7498
                ************************************************************

CP_H1B          TCAGTTCTTTATTGGGCTCTCCACTACACTGTGAGTGCCCTCCTCAGGCGAGAGAACGTT 7560
CP_H1C          TCAGTTCTTTATTGGGCTCTCCACTACACTGTGAGTGCCCTCCTCAGGCGAGAGAACGTT 7558
                ************************************************************

CP_H1B          CTGGCTCTTCTCTTGCCCCTTCAGCCCCTGTTAATCGGACAGAGATGGCAGGGCTGTGTC 7620
CP_H1C          CTGGCTCTTCTCTTGCCCCTTCAGCCCCTGTTAATCGGACAGAGATGGCAGGGCTGTGTC 7618
                ************************************************************

CP_H1B          TCCACGGCCGGAGGCTCTCATAGTCAGGGCACCCACAGCGGTTCCCCACCTGCCTTCTGG 7680
CP_H1C          TCCACGGCCGGAGGCTCTCATAGTCAGGGCACCCACAGCGGTTCCCCACCTGCCTTCTGG 7678
                ************************************************************

CP_H1B          GCAGAATACACTGCCACCCATAGGTCAGCATCTCCACTCGTGGGCCATCTGCTTAGGTTG 7740
CP_H1C          GCAGAATACACTGCCACCCATAGGTCAGCATCTCCACTCGTGGGCCATCTGCTTAGGTTG 7738
                ************************************************************

CP_H1B          GGTTCCTCTGGATTCTGGGGAGATTGGGGGTTCTGTTTTGATCAGCTGATTCTTCTGGGA 7800
CP_H1C          GGTTCCTCTGGATTCTGGGGAGATTGGGGGTTCTGTTTTGATCAGCTGATTCTTCTGGGA 7798
                ************************************************************

CP_H1B          GCAAGTGGGTGCTCGCGAGCTCTCCAGCTTCCTAAAGGTGGAGAAGCACAGACTTCGGGG 7860
CP_H1C          GCAAGTGGGTGCTCGCGAGCTCTCCAGCTTCCTAAAGGTGGAGAAGCACAGACTTCGGGG 7858
                ************************************************************

CP_H1B          GCCTGGCCTGGATCCCTTTCCCCATTCCTGTCCCTGTGCCCCTCGTCTGGGTGCGTTACC 7920
CP_H1C          GCCTGGCCTGGATCCCTTTCCCCATTCCTGTCCCTGTGCCCCTCGTCTGGGTGCGTTACC 7918
                ************************************************************

CP_H1B          ATGGTTTTCTATTTCATAGTTCTTAGGCAAATTGGTAAAAATCATTTCTCATCAAAACGC 7980
CP_H1C          ATGGTTTTCTATTTCATAGTTCTTAGGCAAATTGGTAAAAATCATTTCTCATCAAAACGC 7978
                ************************************************************

CP_H1B          TGATATTTTCACACCTCCCTGGTGTCTGCAGAAAGAACCTTCCAGAAATGCAGTCGTGGG 8040
CP_H1C          TGATATTTTCACACCTCCCTGGTGTCTGCAGAAAGAACCTTCCAGAAATGCAGTCGTGGG 8038
                ************************************************************

CP_H1B          AGACCCATCCAGGCCACCCCTGCTTATGGAAGAGCTGAGAAAAAGCCCCACGGGGGCATT 8100
CP_H1C          AGACCCATCCAGGCCACCCCTGCTTATGGAAGAGCTGAGAAAAAGCCCCACGGGAGCATT 8098
                ********************************************************** *****

CP_H1B          TGCTCAGCTTCCGTTACGCACCTAGTGGCATTGTGGGTGGGAGAGGGCTGGTGGGTGGAT 8160
CP_H1C          TGCTCAGCTTCCGTTACGCACCTAGTGGCATTGTGGGTGGGAGAGGGCTGGTGGGTGGAT 8158
                ************************************************************

CP_H1B          GGAAGGAGAAGGCACAGCCCCCCCTTGCAGGGACAGAGCCCTCGTACAGAAGGGACACCC 8220
CP_H1C          GGAAGGAGAAGGCACAGCCCCCCCTTGCAGGGACAGAGCCCTCGTACAGAAGGGACACCC 8218
                ************************************************************

CP_H1B          CACATTTGTCTTCCCCACAAAGCGGCCTGTGTCCTGCCTACGGGGTCAGGGCTTCTCAAA 8280
CP_H1C          CACATTTGTCTTCCCCACAAAGCGGCCTGTGTCCTGCCTACGGGGTCAGGGCTTCTCAAA 8278
                ************************************************************

CP_H1B          CCTGGCTGTGTGTCAGAATCACCAGGGGAACTTTTCAAAACTAGAGAGACTGAAGCCAGA 8340
CP_H1C          CCTGGCTGTGTGTCAGAATCACCAGGGGAACTTTTCAAAACTAGAGAGACTGAAGCCAGA 8338
                ************************************************************

CP_H1B          CTCCTAGATTCTAATTCTAGGTCAGGGCTAGGGGCTGAGATTGTAAAAATCCACAGGTGA 8400
CP_H1C          CTCCTAGATTCTAATTCTAGGTCAGGGCTAGGGGCTGAGATTGTAAAAATCCACAGGTGA 8398
                ************************************************************

CP_H1B          TTCTGATGCCCGGCAGGCTTGAGAACAGCCGCAGGGAGTTCTCTGGGAATGTGCCGGTGG 8460
CP_H1C          TTCTGATGCCCGGCAGGCTTGAGAACAGCCGCAGGGAGTTCTCTGGGAATGTGCCGGTGG 8458
                ************************************************************
```

```
CP_H1B          GTCTAGCCAGGTGTGAGTGGAGATGCCGGGGAACTTCCTATTACTCACTCGTCAGTGTGG 8520
CP_H1C          GTCTAGCCAGGTGTGAGTGGAGATGCCGGGGAACTTCCTATTACTCACTCGTCAGTGTGG 8518
                ************************************************************

CP_H1B          CCGAACTCATTTTTCACTTGACCTCAGGCTGGTGAACGCTCCCCTCTGGGGTTCAGGCCT 8580
CP_H1C          CCGAACACATTTTTCACTTGACCTCAGGCTGGTGAACGCTCCCCTCTGGGGTTCAGGCCT 8578
                ****** ***********************************************************

CP_H1B          CACGATGCCATCCTTTTGTGAAGTGAGGACCTGCAATCCCAGCTTCGTAAAGCCCGCTGG 8640
CP_H1C          CACGATGCCATCCTTTTGTGAAGTGAGGACCTGCAATCCCAGCTTCGTAAAGCCCGCTGG 8638
                ************************************************************

CP_H1B          AAATCACTCACACTTCTGGGATGCCTTCAGAGCAGCCCTCTATCCCTTCAGCTCCCCTGG 8700
CP_H1C          AAATCACTCACACTTCTGGGATGCCTTCAGAGCAGCCCTCTATCCCTTCAGCTCCCCTGG 8698
                ************************************************************

CP_H1B          GATGTGACTCAACCTCCCGTCACTCCCCAGACTGCCTCTGCCAAGTCCGAAAGTGGAGGC 8760
CP_H1C          GATGTGACTCAACCTCCCGTCACTCCCCAGACTGCCTCTGCCAAGTCCGAAAGTGGAGGC 8758
                ************************************************************

CP_H1B          ATCCTTGCGAGCAAGTAGGCGGGTCCAGGGTGGCGCATGTCACTCATCGAAAGTGGAGGC 8820
CP_H1C          ATCCTTGCGAGCAAGTAGGCGGGTCCAGGGTGGCGCATGTCACTCATCGAAAGTGGAGGC 8818
                ************************************************************

CP_H1B          GTCCTTGCGAGCAAGCAGGCGGGTCCAGGGTGGCGTGTCACTCATCCTTTTTTCTGGCTA 8880
CP_H1C          GTCCTTGCGAGCAAGCAGGCGGGTCCAGGGTGGCGTGTCACTCATCCTTTTTTCTGGCTA 8878
                ************************************************************

CP_H1B          CCAAAGGTGCAGATAATTAATAAGAAGCTGGATCTTAGCAACGTCCAGTCCAAGTGTGGC 8940
CP_H1C          CCAAAGGTGCAGATAATTAATAAGAAGCTGGATCTTAGCAACGTCCAGTCCAAGTGTGGC 8938
                ************************************************************

CP_H1B          TCAAAGGATAATATCAAACACGTCCCGGGAGGCGGCAGTGTGAGTACCTTCACACGTCCC 9000
CP_H1C          TCAAAGGATAATATCAAACACGTCCCGGGAGGCGGCAGTGTGAGTACCTTCACACGTCCC 8998
                ************************************************************

CP_H1B          ATGCGCCGTGCTGTGGCTTGAATTATTAGGAAGTGGTGTGAGTGCGTACACTTGCGAGAC 9060
CP_H1C          ATGCGCCGTGCTGTGGCTTGAATTATTAGGAAGTGGTGTGAGTGCGTACACTTGCGAGAC 9058
                ************************************************************

CP_H1B          ACTGCATAGAATAAATCCTTCTTGGGCTCTCAGGATCTGGCTGCGACCTCTGGGTGAATG 9120
CP_H1C          ACTGCATAGAATAAATCCTTCTTGGGCTCTCAGGATCTGGCTGCGACCTCTGGGTGAATG 9118
                ************************************************************

CP_H1B          TAGCCCGGCTCCCCACATTCCCCCACACGGTCCACTGTTCCCAGAAGCCCCTTCCTCATA 9180
CP_H1C          TAGCCCGGCTCCCCACATTCCCCCACACGGTCCACTGTTCCCAGAAGCCCCTTCCTCATA 9178
                ************************************************************

CP_H1B          TTCTAGGAGGGGGTGTCCCAGCATTTCTGGGTCCCCCAGCCTGCGCAGGCTGTGTGGACA 9240
CP_H1C          TTCTAGGAGGGGGTGTCCCAGCATTTCTGGGTCCCCCAGCCTGCGCAGGCTGTGTGGACA 9238
                ************************************************************

CP_H1B          GAATAGGGCAGATGACGGACCCTCTCTCCGGACCCTGCCTGGGAAGCTGAGAATACCCAT 9300
CP_H1C          GAATAGGGCAGATGACGGACCCTCTCTCCGGACCCTGCCTGGGAAGCTGAGAATACCCAT 9298
                ************************************************************

CP_H1B          CAAAGTCTCCTTCCACTCATGCCCAGCCCTGTCCCCAGGAGCCCCATAGCCCATTGGAAG 9360
CP_H1C          CAAAGTCTCCTTCCACTCATGCCCAGCCCTGTCCCCAGGAGCCCCATAGCCCATTGGAAG 9358
                ************************************************************

CP_H1B          TTGGGCTGAAGGTGGTGGCACCTGAGACTGGGCTGCCGCCTCCTCCCCCGACACCTGGGC 9420
CP_H1C          TTGGGCTGAAGGTGGTGGCACCTGAGACTGGGCTGCCGCCTCCTCCCCCGACACCTGGGC 9418
                ************************************************************

CP_H1B          AGGTTGACGTTGAGTGGCTCCACTGTGGACAGGTGACCCGTTTGTTCTGATGAGCGGACA 9480
CP_H1C          AGGTTGACGTTGAGTGGCTCCACTGTGGACAGGTGACCCGTTTGTTCTGATGAGCGGACA 9478
                ************************************************************

CP_H1B          CCAAGGTCTTACTGTCCTGCTCAGCTGCTGCTCCTACACGTTCAAGGCAGGAGCCGATTC 9540
CP_H1C          CCAAGGTCTTACTGTCCTGCTCAGCTGCTGCTCCTACACGTTCAAGGCAGGAGCCGATTC 9538
                ************************************************************
```

```
CP_H1B          CTAAGCCTCCAGCTTATGCTTAGCCTGCGCCACCCTCTGGCAGAGACTCCAGATGCAAAG 9600
CP_H1C          CTAAGCCTCCAGCTTATGCTTAGCCTGCGCCACCCTCTGGCAGAGACTCCAGATGCAAAG 9598
                ************************************************************

CP_H1B          AGCCAAACCAAAGTGGCATGCCTCGAGCTTACTGAGACACTAAATCTGTTGGTTTCTGCT 9660
CP_H1C          AGCCAAACCAAAATGGCATGCCTCGAGCTTACTGAGACACTAAATCTGTTGGTTTCTGCT 9658
                *********** ************************************************

CP_H1B          GTGCCACCTACCCACCCTGTTGGTGTTGCTTTGTTCCTATTGCTAAAGACAGGAATGTCC 9720
CP_H1C          GTGCCACCTACCCACCCTGTTGGTGTTGCTTTGTTCCTATTGCTAAAGACAGGAATGTCC 9718
                ************************************************************

CP_H1B          AGGACACTGAGTGTGCAGGTGCCTGCTGGTTCTCACGTCCGAGCTGCTGAACTCCGCTGG 9780
CP_H1C          AGGACACTGAGTGTGCAGGTGCCTGCTGGTTCTCACGTCCGAGCTGCTGAACTCCGCTGG 9778
                ************************************************************

CP_H1B          GTCCTGCTTACTGATGGTCTTTGCTCTAGTGCTTTCCAGGGTCCGTGGAAGCTTTTCCTG 9840
CP_H1C          GTCCTGCTTACTGATGGTCTTTGCTCTAGTGCTTTCCAGGGTCCGTGGAAGCTTTTCCTG 9838
                ************************************************************

CP_H1B          GAATAAAGCCCACGCATCGACCCTCACAGCGCCTCCCCTCTTTGAGGCCCAGCAGATACC 9900
CP_H1C          GAATAAAGCCCACGCATCGACCCTCACAGCGCCTCCCCTCTTTGAGGCCCAGCAGATACC 9898
                ************************************************************

CP_H1B          CCACTCCTGCCTTTCCAGCAAGATTTTTCAGATGCTGTGCATACTCATCATATTGATCAC 9960
CP_H1C          CCACTCCTGCCTTTCCAGCAAGATTTTTCAGATGCTGTGCATACTCATCATATTGATCAC 9958
                ************************************************************

CP_H1B          TTTTTTCTTCATGCCTGATTGTGATCTGTCAATTTCATGTCAGGAAAGGGAGTGACATTT 10020
CP_H1C          TTTTTTCTTCATGCCTGATTGTGATCTGTCAATTTCATGTCAGGAAAGGGAGTGACATTT 10018
                ************************************************************

CP_H1B          TTACACTTAAGCGTTTGCTGAGCAAATGTCTGGGTCTTGCACAATGACAATGGGTCTCTG 10080
CP_H1C          TTACACTTAAGCGTTTGCTGAGCAAATGTCTGGGTCTTGCACAATGACAATGGGTCCCTG 10078
                ************************************************************ ***

CP_H1B          TTTTTCCCAGAGGCTCTTTTGTTCTGCAGGGATTGAAGACACTCCAGTCCCACAGTCCCC 10140
CP_H1C          TTTTTCCCAGAGGCTCTTTTGTTCTGCAGGGATTGAAGACACTCCAGTCCCACAGTCCCC 10138
                ************************************************************

CP_H1B          AGCTCCCCTGGGGCAGGGTTGGCAGAATTTCGACAACACATTTTTCCACCCTGACTAGGA 10200
CP_H1C          AGCTCCCCTGGGGCAGGGTTGGCAGAATTTCGACAACACATTTTTCCACCCTGACTAGGA 10198
                ************************************************************

CP_H1B          TGTGCTCCTCATGGCAGCTGGGAACCACTGTCCAATAAGGGCCTGGGCTTACACAGCTGC 10260
CP_H1C          TGTGCTCCTCATGGCAGCTGGGAACCACTGTCCAATAAGGGCCTGGGCTTACACAGCTGC 10258
                ************************************************************

CP_H1B          TTCTCATTGAGTTACACCCTTAATAAAATAATCCCATTTTATCCTTTTTGTCTCTCTGTC 10320
CP_H1C          TTCTCATTGAGTTACACCCTTAATAAAATAATCCCATTTTATCCTTTTTGTCTCTCTGTC 10318
                ************************************************************

CP_H1B          TTCCTCTCTCTCTGCCTTTCCTCTTCTCTCTCCTCCTCTCTCATCTCCAGGTGCAAATAG 10380
CP_H1C          TTCCTCTCTCTCTGCCTTTCCTCTTCTCTCTCCTCCTCTCTCATCTCCAGGTGCAAATAG 10378
                ************************************************************

CP_H1B          TCTACAAACCAGTTGACCTGAGCAAGGTGACCTCCAAGTGTGGCTCATTAGGCAACATCC 10440
CP_H1C          TCTACAAACCAGTTGACCTGAGCAAGGTGACCTCCAAGTGTGGCTCATTAGGCAACATCC 10438
                ************************************************************

CP_H1B          ATCATAAACCAGGACGTCGAGGTGGCCAGGTGGAAGTAAAATCTGAGAAGCTTGACTTCA 10500
CP_H1C          ATCATAAACCAGGACGTCGAGGTGGCCAGGTGGAAGTAAAATCTGAGAAGCTTGACTTCA 10498
                ************************************************************

CP_H1B          AGGACAGAGTCCAGTCGAAGATTGGGTCCCTGGACAATATCACCCACGTCCCTGGCGGAG 10560
CP_H1C          AGGACAGAGTCCAGTCGAAGATTGGGTCCCTGGACAATATCACCCACGTCCCTGGCGGAG 10558
                ************************************************************

CP_H1B          GAAATAAAAAGATTGAAACCCACAAGCTGACCTTCCGCGAGAACGCCAAAGCCAAGACAG 10620
CP_H1C          GAAATAAAAAGATTGAAACCCACAAGCTGACCTTCCGCGAGAACGCCAAAGCCAAGACAG 10618
                ************************************************************
```

```
CP_H1B          ACCACGGGGCGGAGATCGTGTACAAGTCGCCAGTGGTGTCTGGGGACACGTCTCCACGGC 10680
CP_H1C          ACCACGGGGCGGAGATCGTGTACAAGTCGCCAGTGGTGTCTGGGGACACGTCTCCACGGC 10678
                ************************************************************

CP_H1B          ATCTCAGCAATGTCTCCTCCACCGGCAGCATCGACATGGTAGACTCGCCCCAGCTCGCCA 10740
CP_H1C          ATCTCAGCAATGTCTCCTCCACCGGCAGCATCGACATGGTAGACTCGCCCCAGCTCGCCA 10738
                ************************************************************

CP_H1B          CGCTAGCTGACGAGGTGTCTGCCTCCCTGGCCAAGCAGGGTTTGGATTACAAGGATGACG 10800
CP_H1C          CGCTAGCTGACGAGGTGTCTGCCTCCCTGGCCAAGCAGGGTTTGGATTACAAGGATGACG 10798
                ************************************************************

CP_H1B          ACGATAAGTAAACAACTTTGTATAATAAAGTTGTCCCTGGGGCGGTCAATAATTGTGGAG 10860
CP_H1C          ACGATAAGTAA----CTTTGTATAATAAAGTTGTCCCTGGGGCGGTCAATAATTGTGGAG 10854
                ***********    *******************************************

CP_H1B          AGGAGAGAATGAGAGAGTGTGGAAAAAAAAAAGAATAATGACCCGGCCCCCGCCCTCTGCC 10920
CP_H1C          AGGAGAGAATGAGAGAGTGTGGAAAAAAAAAAGAATAATGACCCGGCCCCCGCCCTCTGCC 10914
                ************************************************************

CP_H1B          CCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAATCACTTAACCTGCTTTTGTCACTCG 10980
CP_H1C          CCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAATCACTTAACCTGCTTTTGTCACTCG 10974
                ************************************************************

CP_H1B          GCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAGTAAGAGCAAATTTCATCTTTCCAAA 11040
CP_H1C          GCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAGTAAGAGCAAATTTCATCTTTCCAAA 11034
                ************************************************************

CP_H1B          TTGATGGGTGGGCTAGTAATAAAATATTTAAAAAAAAACATTCAAAAACATGGCCACATC 11100
CP_H1C          TTGATGGGTGGGCTAGTAATAAAATATTTAAAAAAAAACATTCAAAAACATGGCCACATC 11094
                ************************************************************

CP_H1B          CAACATTTCCTCAGGCAATTCCTTTTGATTCTTTTTTCTTCCCCCTCCATGTAGAAGAGG 11160
CP_H1C          CAACATTTCCTCAGGCAATTCCTTTTGATTCTTTTTTCTTCCCCCTCCATGTAGAAGAGG 11154
                ************************************************************

CP_H1B          GAGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGGATTTCAAGGGACTGGGGGTGCCAACC 11220
CP_H1C          GAGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGGATTTCAAGGGACTGGGGGTGCCAACC 11214
                ************************************************************

CP_H1B          ACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGCAGTGGCAGCAACAAAGGATTTGAAAC 11280
CP_H1C          ACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGCAGTGGCAGCAACAAAGGATTTGAAAC 11274
                ************************************************************

CP_H1B          TTGGTGTGTTCGTGGAGCCACAGGCAGACGATGTCAACCTTGTGTGAGTGTGACGGGGGT 11340
CP_H1C          TTGGTGTGTTCGTGGAGCCACAGGCAGACGATGTCAACCTTGTGTGAGTGTGACGGGGGT 11334
                ************************************************************

CP_H1B          TGGGGTGGGGCGGGAGGCCACGGGGGAGGCCGAGGCAGGGGCTGGGCAGAGGGGAGAGGA 11400
CP_H1C          TGGGGTGGGGCGGGAGGCCACGGGGGAGGCCGAGGCAGGGGCTGGGCAGAGGGGAGAGGA 11394
                ************************************************************

CP_H1B          AGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACGTGCTGGAGAGTAGACATCCCCCTCCT 11460
CP_H1C          AGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACGTGCTGGAGAGTAGACATCCCCCTCCT 11454
                ************************************************************

CP_H1B          TGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGCAGCGTCTGAGCGGCCGCCTGTCCTTG 11520
CP_H1C          TGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGCAGCGTCTGAGCGGCCGCCTGTCCTTG 11514
                ************************************************************

CP_H1B          GTGGCCGGGGGTGGGGGCCTGCTGTGGGTCAGTGTGCCACCCTCTGCAGGGCAGCCTGTG 11580
CP_H1C          GTGGCCGGGGGTGGGGGCCTGCTGTGGGTCAGTGTGCCACCCTCTGCAGGGCAGCCTGTG 11574
                ************************************************************

CP_H1B          GGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAGCTGGCAGGAGGGTGGCACTTCGTGGA 11640
CP_H1C          GGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAGCTGGCAGGAGGGTGGCACTTCGTGGA 11634
                ************************************************************

CP_H1B          TGACCTCCTTAGAAAAGACTGACCTTGATGTCTTGAGAGCGCTGGCCTCTTCCTCCCTCC 11700
CP_H1C          TGACCTCCTTAGAAAAGACTGACCTTGATGTCTTGAGAGCGCTGGCCTCTTCCTCCCTCC 11694
                ************************************************************
```

```
CP_H1B          CTGCAGGGTAGGGGGCCTGAGTTGAGGGGCTTCCCTCTGCTCCACAGAAACCCTGTTTTA 11760
CP_H1C          CTGCAGGGTAGGGGGCCTGAGTTGAGGGGCTTCCCTCTGCTCCACAGAAACCCTGTTTTA 11754
                ************************************************************

CP_H1B          TTGAGTTCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACTTTGCAGACCTGGGACTT 11820
CP_H1C          TTGAGTTCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACTTTGCAGACCTGGGACTT 11814
                ************************************************************

CP_H1B          TAGGGCTAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGACGTCCACCCGTTTC 11880
CP_H1C          TAGGGCTAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGACGTCCACCCGTTTC 11874
                ************************************************************

CP_H1B          CAAGCCTGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAGGCAGGTCCCGAGCC 11940
CP_H1C          CAAGCCTGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAGGCAGGTCCCGAGCC 11934
                ************************************************************

CP_H1B          CCCTGTCCTTCCCACGGCCACTGCAGTCACCCCGTCTGCGCCGCTGTGCTGTTGTCTGCC 12000
CP_H1C          CCCTGTCCTTCCCACGGCCACTGCAGTCACCCCGTCTGCGCCGCTGTGCTGTTGTCTGCC 11994
                ************************************************************

CP_H1B          GTGAGAGCCCAATCACTGCCTATACCCCTCATCACACGTCACAATGTCCCGAATTCCCAG 12060
CP_H1C          GTGAGAGCCCAATCACTGCCTATACCCCTCATCACACGTCACAATGTCCCGAATTCCCAG 12054
                ************************************************************

CP_H1B          CCTCACCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGTACCTACTCCATACT 12120
CP_H1C          CCTCACCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGTACCTACTCCATACT 12114
                ************************************************************

CP_H1B          GAGGGTGAAATTAAGGGAAGGCAAAGTCCAGGCACAAGAGTGGGACCCCAGCCTCTCACT 12180
CP_H1C          GAGGGTGAAATTAAGGGAAGGCAAAGTCCAGGCACAAGAGTGGGACCCCAGCCTCTCACT 12174
                ************************************************************

CP_H1B          CTCAGTTCCACTCATCCAACTGGGACCCTCACCACGAATCTCATGATCTGATTCGGTTCC 12240
CP_H1C          CTCAGTTCCACTCATCCAACTGGGACCCTCACCACGAATCTCATGATCTGATTCGGTTCC 12234
                ************************************************************

CP_H1B          CTGTCTCCTTCTCCCGTCACAGATGTGAGCCAGGGCACTGCTCAGCTGTGACCCTAGGTG 12300
CP_H1C          CTGTCTCCTCCTCCCGTCACAGATGTGAGCCAGGGCACTGCTCAGCTGTGACCCTAGGTG 12294
                ********* ********************************************************

CP_H1B          TTTCTGCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCCTGGCCCCTTCCTGT 12360
CP_H1C          TTTCTGCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCCTGGCCCCTTCCTGT 12354
                ************************************************************

CP_H1B          GCTGAGCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGGCACCAGGATGGAAG 12420
CP_H1C          GCTGAGCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGGCACCAGGATGGAAG 12414
                ************************************************************

CP_H1B          GGCAAGGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTGCACCCTAGCTTGTA 12480
CP_H1C          GGCAAGGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTGCACCCTAGCTTGTA 12474
                ************************************************************

CP_H1B          GCTGCCAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATGCATAGTATCAGCCC 12540
CP_H1C          GCTGCCAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATGCATAGTATCAGCCC 12534
                ************************************************************

CP_H1B          TCCACACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTTTCCCCCAGTCCCAG 12600
CP_H1C          TCCACACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTTTCCCCCAGTCCCAG 12594
                ************************************************************

CP_H1B          CTGGAAGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACATAGATGTTGCCCTG 12660
CP_H1C          CTGGAAGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACATAGATGTTGCCCTG 12654
                ************************************************************

CP_H1B          CCCTCCCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTTTATGCTTGGATTCA 12720
CP_H1C          CCCTCCCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTTTATGCTTGGATTCA 12714
                ************************************************************

CP_H1B          CCAGAGTGACTATGATAGTGAAAAGAAAAAAAAAAAAAAAA---GGACGCATGTATCTTGA 12777
CP_H1C          CCAGAGTGACTATGATAGTGAAAAGAAAAAAAAAAAAAAAAAAGGACGCATGTATCTTGA 12774
                ****************************************   *****************
```

```
CP_H1B          AATGCTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTCACCCCCACACTGGG 12837
CP_H1C          AATGCTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTCACCCCCACACTGGG 12834
                ************************************************************

CP_H1B          ACTCGTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTTTTGAAAGGCTTTCC 12897
CP_H1C          ACTCGTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTTTTGAAAGGCTTTCC 12894
                ************************************************************

CP_H1B          TCAGCACCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAGGCCGTTCAGCTGTG 12957
CP_H1C          TCAGCACCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAGGCCGTTCAGCTGTG 12954
                ************************************************************

CP_H1B          ACGAAGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCCTTCCCTACTTCCCC 13017
CP_H1C          ACGAAGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCCTTCCCTACTTCCCC 13014
                ************************************************************

CP_H1B          TTGGGGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTCTGGCTTGCGGCGCG 13077
CP_H1C          TTGGGGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTCTGGCTTGCGGCGCG 13074
                ************************************************************

CP_H1B          AGGATGGTTCTCTCTGGTCATAGCCCGAAGTCTCATGGCAGTCCCAAAGGAGGCTTACAA 13137
CP_H1C          AGGATGGTTCTCTCTGGTCATAGCCCGAAGTCTCATGGCAGTCCCAAAGGAGGCTTACAA 13134
                ************************************************************

CP_H1B          CTCCTGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATCTGCAGCTCCCAGAA 13197
CP_H1C          CTCCTGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATCTGCAGCTCCCAGAA 13194
                ************************************************************

CP_H1B          GCTCCGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGG 13257
CP_H1C          GCTCCGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGCTCGCCCTCTGGAGG 13254
                ************************************************************

CP_H1B          GGCAGCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATTGGGATGAATTGCCT 13317
CP_H1C          GGCAGCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATTGGGATGAATTGCCT 13314
                ************************************************************

CP_H1B          GTCCTGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTC 13377
CP_H1C          GTCCTGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGATGACTTGACAAGTC 13374
                ************************************************************

CP_H1B          AGGAGACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCTGACCTTGCACAAAC 13437
CP_H1C          AGGAGACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCTGACCTTGCACAAAC 13434
                ************************************************************

CP_H1B          TCCATCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAA 13497
CP_H1C          TCCATCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATTGCTGCCTAAAGAAA 13494
                ************************************************************

CP_H1B          CTCAGCAGCCTCAGGCCCAATTCTGCCACTTCTGGTTTGGGTACAGTTAAAGGCAACCCT 13557
CP_H1C          CTCAGCAGCCTCAGGCCCAATTCTGCCACTTCTGGTTTGGGTACAGTTAAAGGCAACCCT 13554
                ************************************************************

CP_H1B          GAGGGACTTGGCAGTAGAAATCCAGGGCCTCCCCTGGGGCTGGCAGCTTCGTGTGCAGCT 13617
CP_H1C          GAGGGACTTGGCAGTAGAAATCCAGGGCCTCCCCTGGGGCTGGCAGCTTCGTGTGCAGCT 13614
                ************************************************************

CP_H1B          AGAGCTTTACCTGAAAGGAAGTCTCTGGGCCCAGAACTCTCCACCAAGAGCCTCCCTGCC 13677
CP_H1C          AGAGCTTTACCTGAAAGGAAGTCTCTGGGCCCAGAACTCTCCACCAAGAGCCTCCCTGCC 13674
                ************************************************************

CP_H1B          GTTCGCTGAGTCCCAGCAATTCTCCTAAGTTGAAGGGATCTGAGAAGGAGAAGGAAATGT 13737
CP_H1C          GTTCGCTGAGTCCCAGCAATTCTCCTAAGTTGAAGGGACCTGAGAAGGAGAAGGAAGTGT 13734
                ************************************ ********** ********** ***

CP_H1B          GGGGTAGATTTGGTGGTGGTTAGAGATATGCCCCCCTCATTACTGCCAACAGTTTCGGCT 13797
CP_H1C          GGGGTAGATTTGGTGGTGGTTAGAGATATGCCCCCCTCATTACTGCCAACAGTTTCGGCT 13794
                ************************************************************

CP_H1B          GCATTTCTTCACGCACCTCGGTTCCTCTTCCTGAAGTTCTTGTGCCCTGCTCTTCAGCAC 13857
CP_H1C          GCATTTCTTCACGCACCTCGGTTCCTCTTCCTGAAGTTCTTGTGCCCTGCTCTTCAGCAC 13854
                ************************************************************
```

```
CP_H1B          CATGGGCCTTCTTATACGGAAGGCTCTGGGATCTCCCCCTTGTGGGGCAGGCTCTTGGGG 13917
CP_H1C          CATGGGCCTTCTTATACGGAAGGCTCTGGGATCTCCCCCTTGTGGGGCAGGCTCTTGGGG 13914
                ************************************************************

CP_H1B          CCAGCCTAAGATCATGGTTTAGGGTGATCAGTGCTGGCAGATAAATTGAAAAGGCACGCT 13977
CP_H1C          CCAGCCTAAGATCATGGTTTAGGGTGATCAGTGCTGGCAGATAAATTGAAAAGGCACGCT 13974
                ************************************************************

CP_H1B          GGCTTGTGATCTTAAATGAGGACAATCCCCCCAGGGCTGGGCACTCCTCCCCTCCCCTCA 14037
CP_H1C          GGCTTGTGATCTTAAATGAGGACAATCCCCCCAGGGCTGGGCACTCCTCCCCTCCCCTCA 14034
                ************************************************************

CP_H1B          CTTCTCCCACCTGCAGAGCCAGTGTCCTTGGGTGGGCTAGATAGGATATACTGTATGCCG 14097
CP_H1C          CTTCTCCCACCTGCAGAGCCAGTGTCCTTGGGTGGGCTAGATAGGATATACTGTATGCCG 14094
                ************************************************************

CP_H1B          GCTCCTTCAAGCTGCTGACTCACTTTATCAATAGTTCCATTTAAATTGACTTCAGTGGTG 14157
CP_H1C          GCTCCTTCAAGCTGCTGACTCACTTTATCAATAGTTCCATTTAAATTGACTTCAGTGGTG 14154
                ************************************************************

CP_H1B          AGACTGTATCCTGTTTGCTATTGCTTGTTGTGCTATGGGGGGAGGGGGGAGGAATGTGTA 14217
CP_H1C          AGACTGTATCCTGTTTGCTATTGCTTGTTGTGCTATGGGGGGAGGGGGGAGGAATGTGTA 14214
                ************************************************************

CP_H1B          AGATAGTTAACATGGGCAAAGGGAGATCTTGGGGTGCAGCACTTAAACTGCCTCGTAACC 14277
CP_H1C          AGATAGTTAACATGGGCAAAGGGAGATCTTGGGGTGCAGCACTTAAACTGCCTCGTAACC 14274
                ************************************************************

CP_H1B          CTTTTCATGATTTCAACCACATTTGCTAGAGGGAGGGAGCAGCCACGGAGTTAGAGGCCC 14337
CP_H1C          CTTTTCATGATTTCAACCACATTTGCTAGAGGGAGGGAGCAGCCACGGAGTTAGAGGCCC 14334
                ************************************************************

CP_H1B          TTGGGGTTTCTCTTTTCCACTGACAGGCTTTCCCAGGCAGCTGGCTAGTTCATTCCCTCC 14397
CP_H1C          TTGGGGTTTCTCTTTTCCACTGACAGGCTTTCCCAGGCAGCTGGCTAGTTCATTCCCTCC 14394
                ************************************************************

CP_H1B          CCAGCCAGGTGCAGGCGTAGGAATATGGACATCTGGTTGCTTTGGCCTGCTGCCCTCTTT 14457
CP_H1C          CCAGCCAGGTGCAGGCGTAGGAATATGGACATCTGGTTGCTTTGGCCTGCTGCCCTCTTT 14454
                ************************************************************

CP_H1B          CAGGGGTCCTAAGCCCACAATCATGCCTCCCTAAGACCTTGGCATCCTTCCCTCTAAGCC 14517
CP_H1C          CAGGGGTCCTAAGCCCACAATCATGCCTCCCTAAGACCTTGGCATCCTTCCCTCTAAGCC 14514
                ************************************************************

CP_H1B          GTTGGCACCTCTGTGCCACCTCTCACACTGGCTCCAGACACACAGCCTGTGCTTTTGGAG 14577
CP_H1C          GTTGGCACCTCTGTGCCACCTCTCACACTGGCTCCAGACACACAGCCTGTGCTTTTGGAG 14574
                ************************************************************

CP_H1B          CTGAGATCACTCGCTTCACCCTCCTCATCTTTGTTCTCCAAGTAAAGCCACGAGGTCGGG 14637
CP_H1C          CTGAGATCACTCGCTTCACCCTCCTCATCTTTGTTCTCCAAGTAAAGCCACGAGGTCGGG 14634
                ************************************************************

CP_H1B          GCGAGGGCAGAGGTGATCACCTGCGTGTCCCATCTACAGACCTGCAGCTTCATAAAACTT 14697
CP_H1C          GCGAGGGCAGAGGTGATCACCTGCGTGTCCCATCTACAGACCTGCAGCTTCATAAAACTT 14694
                ************************************************************

CP_H1B          CTGATTTCTCTTCAGCTTTGAAAAGGGTTACCCTGGGCACTGGCCTAGAGCCTCACCTCC 14757
CP_H1C          CTGATTTCTCTTCAGCTTTGAAAAGGGTTACCCTGGGCACTGGCCTAGAGCCTCACCTCC 14754
                ************************************************************

CP_H1B          TAATAGACTTAGCCCCATGAGTTTGCCATGTTGAGCAGGACTATTTCTGGCACTTGCAAG 14817
CP_H1C          TAATAGACTTAGCCCCATGAGTTTGCCATGTTGAGCAGGACTATTTCTGGCACTTGCAAG 14814
                ************************************************************

CP_H1B          TCCCATGATTTCTTCGGTAATTCTGAGGGTGGGGGGAGGGACATGAAATCATCTTAGCTT 14877
CP_H1C          TCCCATGATTTCTTCGGTAATTCTGAGGGTGGGGGGAGGGACATGAAATCATCTTAGCTT 14874
                ************************************************************

CP_H1B          AGCTTTCTGTCTGTGAATGTCTATATAGTGTATTGTGTGTTTTAACAAATGATTTACACT 14937
CP_H1C          AGCTTTCTGTCTGTGAATGTCTATATAGTGTATTGTGTGTTTTAACAAATGATTTACACT 14934
                ************************************************************
```

```
CP_H1B          GACTGTTGCTGTAAAAGTGAATTTGGAAATAAAGTTATTACTCTGATTAAATAAGGTCTC 14997
CP_H1C          GACTGTTGCTGTAAAAGTGAATTTGGAAATAAAGTTATTACTCTGATTAAATAAGGTCTC 14994
                ************************************************************

CP_H1B          CATTCATGGATTCCAAGGACAAGAAAGTCATATAGAATGTCTATTTTTTAAGTTCTTTCC 15057
CP_H1C          CATTCATGGATTCCAAGGACAAGAAAGTCATATAGAATGTCTATTTTTTAAGTTCTTTCC 15054
                ************************************************************

CP_H1B          CACGCACCCTTAGATAATTTAGCTCAGAACAGGAAATGATAGTATTAATAAAAGCTGGAC 15117
CP_H1C          CACGCACCCTTAGATAATTTAGCTCAGAACAGGAAATGATAGTATTAATAAAAGCTGGAC 15114
                ************************************************************

CP_H1B          ATCAGGATTAACAGCTCTCTCTGGGGCCCTGAAGGTGAGAGTTCTCAGACTTGCTCATTT 15177
CP_H1C          ATCAGGATTAACAGCTCTCTCTGGGGCCCTGAAGGTGAGAGTTCTCAGACTTGCTCATTT 15174
                ************************************************************

CP_H1B          GCAGTTGCTTCTTTGTGATGCTGGCTACCCAGCTTTCTTGTACAAAGTGGT 15228
CP_H1C          GCAGTTGCTTCTTTGTGATGCTGGCTACCCAGCTTTCTTGTACAAAGTGGT 15225
                ***************************************************
```

Appendix J

CP+rs242557 H1B vs H1C Minigenes

In the below multiple sequence alignment of the two CP+rs242557 minigenes, sequence matches are denoted *, with sequence differences highlighted in **red**. Exon 0 is highlighted in **green**.

```
CP+H1B-G        GGGGACAAGTTTGTACAAAAAAGCAGGCTTCCAAATGCTCTGCGATGTGTTAAGCACTGT 60
CP+H1C-A        GGGGACAAGTTTGTACAAAAAAGCAGGCTTCCAAATGCTCTGCGATGTGTTAAGCACTGT 60
                ************************************************************

CP+H1B-G        TTGAAATTCGTCTAATTTAAGATTTTTTTTTCTGACGTAACGGTTAGATTCACGTTTCTT 120
CP+H1C-A        TTGAAATTCGTCTAATTTAAGATTTTTTTTTCTGACGTAACGGTTAGATTCACGTTTCTT 120
                ************************************************************

CP+H1B-G        TTTTTTTAAGTACAGTTCTACTGTATTGTAACTGAGTTAGCTTGCTTTAAGCCGATTTGT 180
CP+H1C-A        TTTTTTTAAGTACAGTTCTACTGTATTGTAACTGAGTTAGCTTGCTTTAAGCCGATTTGT 180
                ************************************************************

CP+H1B-G        TAAGGAAAGGATTCACCTTGGTCAGTAACAAAAAAGGTGGGAAAAAAGCAAGGAGAAAGG 240
CP+H1C-A        TAAGGAAAGGATTCACCTTGGTCAGTAACAAAAAAGGTGGGAAAAAAGCAAGGAGAAAGG 240
                ************************************************************

CP+H1B-G        AAGCAGCCTGGGGGAAAGAGACCTTAGCCAGGGGGGCGGTTTCGGGACTACGAAGGGTCG 300
CP+H1C-A        AAGCAGCCTGGGGGAAAGAGACCTTAGCCAGGGGGGCGGTTTCGGGACTACGAAGGGTCG 300
                ************************************************************

CP+H1B-G        GGGCGGACGGACTCGAGGGCCGGCCACGTGGAAGGCCGCTCAGGACTTCTGTAGGAGAGG 360
CP+H1C-A        GGGCGGACGGACTCGAGGGCCGGCCACGTGGAAGGCCGCTCAGGACTTCTGTAGGAGAGG 360
                ************************************************************

CP+H1B-G        ACACCGCCCCAGGCTGACTGAAAGTAAAGGGCAGCGGACCCAGCGGCGGAGCCACTGGCC 420
CP+H1C-A        ACACCGCCCCAGGCTGACTGAAAGTAAAGGGCAGCGGACCCAGCGGCGGAGCCACTGGCC 420
                ************************************************************

CP+H1B-G        TTGCCCCGACCCCGCGTGGCCCGAAGGAGGACACCCACCCCCGCAACGACACAAAGACTC 480
CP+H1C-A        TTGCCCCGACCCCGCATGGCCCGAAGGAGGACACCCACCCCCGCAACGACACAAAGACTC 480
                *************** ****************************************

CP+H1B-G        CAACTACAGGAGGTGGAGAAAGCGCGTGCGCCACGGAACGCGCGTGCGCGCTGCGGTCAG 540
CP+H1C-A        CAACTACAGGAGGTGGAGAAAGCGCGTGCGCCACGGAACGCGCGTGCGCGCTGCGGTCAG 540
                ************************************************************

CP+H1B-G        CGCCGCGGCCTGAGGCGTAGCGGGAGGGGGACCGCGAAAGGGCAGCGCCGAGAGGAACGA 600
CP+H1C-A        CGCCGCGGCCTGAGGCGTAGCGGGAGGGGGACCGCGAAAGGGCAGCGCCGAGAGGAACGA 600
                ************************************************************

CP+H1B-G        GCCGGGAGACGCCGGACGGCCGAGCGGCAGGGCGCTCGCGCGCGCCCACTAGTGGCCGGA 660
CP+H1C-A        GCCGGGAGACGCCGGACGGCCGAGCGGCAGGGCGCTCGCGCGCGCCCACTAGTGGCCGGA 660
                ************************************************************

CP+H1B-G        GGAGAAGGCTCCCGCGGAGGCCGCGCTGCCCGCCCCCTCCCCTGGGGAGGCTCGCGTTCC 720
CP+H1C-A        GGAGAAGGCTCCCGCGGAGGCCGCGCTGCCCGCCCCCTCCCCTGGGGAGGCTCGCGTTCC 720
                ************************************************************

CP+H1B-G        CGCTGCTCGCGCCTGCGCCGCCCGCCGGCCTCAGGAACGCGCCCTCTTCGCCGGCGCGCG 780
CP+H1C-A        CGCTGCTCGCGCCTGCGCCGCCCGCCGGCCTCAGGAACGCGCCCTCTTCGCCGGCGCGCG 780
                ************************************************************

CP+H1B-G        CCCTCGCAGTCACCGCCACCCACCAGCTCCGGCACCAACAGCAGCGCCGCTGCCACCGCC 840
CP+H1C-A        CCCTCGCAGTCACCGCCACCCACCAGCTCCGGCACCAACAGCAGCGCCGCTGCCACCGCC 840
                ************************************************************

CP+H1B-G        CACCTTCTGCCGCCGCCACCACAGCCACCTTCTCCTCCTCCGCTGTCCTCTCCCGTCCTC 900
CP+H1C-A        CACCTTCTGCCGCCGCCACCACAGCCACCTTCTCCTCCTCCGCTGTCCTCTCCCGTCCTC 900
                ************************************************************
```

```
CP+H1B-G        GCCTCTGTCGACTATCAGGTAAGCGCCGCGGCTCCGAAATCTGCCTCGCCGTCCGCCTCT 960
CP+H1C-A        GCCTCTGTCGACTATCAGGTAAGCGCCGCGGCTCCGAAATCTGCCTCGCCGTCCGCCTCT 960
                ************************************************************

CP+H1B-G        GTGCACCCCTGCGCCGCCGCCCCTCGCCCTCCCTCTCCGCAGACTGGGGCTTCGTGCGCC 1020
CP+H1C-A        GTGCACCCCTGCGCCGCCGCCCCTCGCCCTCCCTCTCCGCAGACTGGGGCTTCGTGCGCC 1020
                ************************************************************

CP+H1B-G        GGGCATCGGTCGGGGCCACCGCAGGGCCCCTCCCTGCCTCCCCTGCTCGGGGGCTGGGGC 1080
CP+H1C-A        GGGCATCGGTCGGGGCCACCGCAGGGCCCCTCCCTGCCTCCCCTGCTCGGGGGCTGGGGC 1080
                ************************************************************

CP+H1B-G        CAGGGCGGCCTGGAAAGGGACCTGAGCAAGGGATGCACGCACGCGTGAGTGCGCGCGTGT 1140
CP+H1C-A        CAGGGCGGCCTGGAAAGGGACCTGAGCAAGGGATGCACGCACGCGTGAGTGCGCGCGTGT 1140
                ************************************************************

CP+H1B-G        GTGTGTGCTGGAGGGTCTTCACCACCAGATTCGCGCAGACCCCAGGTGGAGGCTGTGCCG 1200
CP+H1C-A        GTGTGTGCTGGAGGGTCTTCACCACCAGATTCGCGCAGACCCCAGGTGGAGGCTGTGCCG 1200
                ************************************************************

CP+H1B-G        GCAGGGTGGGGCGCGACGGCGGTGACTTGGGGGAGGGGGCTGCCCTTCACTCTCGACTGC 1260
CP+H1C-A        GCAGGGTGGGGCGCGGCGGCGGTGACTTGGGGGAGGGGGCTGCCCTTCACTCTCGACTGC 1260
                *************** ********************************************

CP+H1B-G        AGCCTTTTGCCGCAATGGGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTG----GAGGGGT 1316
CP+H1C-A        AGCCTTTTGCCGCAATGGGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAGGGGT 1320
                ************************************************    *******

CP+H1B-G        CCGATAACGACCCCCGAAACCGAATCTGAAATCCGCTGTCCGCTAGCTGGGACAGATCCT 1376
CP+H1C-A        CCGATAACGACCCCCGAAACCGAATCTGAAATCCGCTGTCCGCTAGCTGGGACAGATCCT 1380
                ************************************************************

CP+H1B-G        CAGTGGAACATGACTCTGTAACGAGAGCATTTTGTTTTGTCAAAATGAGAACATATTATT 1436
CP+H1C-A        CAGTGGAACATGACTCTGTAACGAGAGCATTTTGTTTTGTCAAAATGAGAACATATTATT 1440
                ************************************************************

CP+H1B-G        GCCTTTCATCTGATTGTAAACATAATACATGTTTATAAAACAGTATAATGAGACAAAAAT 1496
CP+H1C-A        GCCTTTCATCTGATTGTAAACATAATACATGTTTATAAAACAGTATAATGAGACAAAAAT 1500
                ************************************************************

CP+H1B-G        GTAGACACTAATAAGGGAAAATCTCCCTAATTGTATTTCTCTTCACAGAGAAAGCCCCTG 1556
CP+H1C-A        GTAGACACTAATAAGGGAAAATCTCCCTAATTGTATTTCTCTTCACAGAGAAAGCCCCTG 1560
                ************************************************************

CP+H1B-G        TTGGGCATATATACTCTAGTTTGCTTATTTGTTTGACTACACATATATGTATTCTTTTCT 1616
CP+H1C-A        TTGGGCATATATACTCTAGTTTGTTTATTTGTTTGACTACACATATATGTATTCTTTTCT 1620
                ********************** ***********************************

CP+H1B-G        TATGTATAAAAATTCTGAACATGCACATTTCTGCAACTACTGTTTTCACTTGATGATGCA 1676
CP+H1C-A        TATGTATAAAAATTCTGAACATGCACATTTCTGCAACTACTGTTTTCACTTGATGATGCA 1680
                ************************************************************

CP+H1B-G        TGGACCTCTCTAGAGTGTACGTTTCTTCTTCCTTACAAAGCAGTTGGCTTCGCCCAGGGT 1736
CP+H1C-A        TGGACCTCTCTAGAGTGTACGTTTCTTCTTCCTTACAAAGCAGTTGGCTTCGCCCAGGGT 1740
                ************************************************************

CP+H1B-G        GCACCAGGACACGGTTTTGGCTCTGTCCCCAGGGTGTCACGGGACCAGGGGATGATCTCA 1796
CP+H1C-A        ACACCAGGACACGGTTTTGGCTCTGTCCCCAGGGTGTCACGGGACCAGGGGATGATCTCA 1800
                 ***********************************************************

CP+H1B-G        CAGGGTCTGCCATCTGCCCTGCCTGGCCGGAGGCTGCATCGAGAGGGCCAAGGGGCACCA 1856
CP+H1C-A        CAGGGTCTGCCATCTGCCCTGCCTGGCCGGAGGCTGCATCGAGAGGGCCAAGGGGCACCA 1860
                ************************************************************

CP+H1B-G        CGTGTCGTGGGTACTGTCAAACAAGAGCCTTCAGAGCCTTCCACAGTCTTTCTTTTGCTT 1916
CP+H1C-A        CGTGTCGTGGGTACTGTCAAACAAGAGCCTTCAGAGCCTTCCACAGTCTTTCTTTTGCTT 1920
                ************************************************************

CP+H1B-G        CCCAGCATTGCTTCCCCGCTGGTGGACTCTGAATCTAGAACTAGCTCCAGGCGCCTCTCC 1976
CP+H1C-A        CCCAGCATTGCTTCCCCGCTGGTGGACTCTGAATCTAGAACTAGCTCCAGGCGCCTCTCC 1980
                ************************************************************
```

```
CP+H1B-G        AAATTCAGACGGGAGCTGGGGCACTATTATAATGCAAATCTAGGCAAAGCCCTCCCAATA 2036
CP+H1C-A        AAATTCAGACGGGAGCTGGGGCACTATTATAATGCAAATCTAGGCAAAGCCCTCCCAATA 2040
                ************************************************************

CP+H1B-G        CCAGGATCCAGAATGGGGTGGGGCCCTTTGCCCTGAAAAGCTGTTTAGTTTGAAAATACA 2096
CP+H1C-A        CCAGGATCCAGAATGGGGTGGGGCCCTTTGCCCTGAAAAGCTGTTTAGTTTGAAAATACA 2100
                ************************************************************

CP+H1B-G        AACAGGAGACAGAAAAGTTTGGCTAAATTAATGGATAAAGTTTTAACGATGGTAACCATA 2156
CP+H1C-A        AACAGGAGACAGAAAAGTTTGGCTAAATTAATGGATAAAGTTTTAACGATGGTAACCATA 2160
                ************************************************************

CP+H1B-G        GTAGGGTTCATCGACAGCCACAACTTTGTATACAAAAGTTGTCCCTGGTGGTGTTGAATA 2216
CP+H1C-A        GTAGGGTTCATCGACAGCCACAACTTTGTATACAAAAGTTGTCCCTGGTGGTGTTGAATA 2220
                ************************************************************

CP+H1B-G        TGAACTGCTGCGGTGTTGGTAAATTAAGCAAGCAGATAGATGTAAATAACGCTTGGGCAG 2276
CP+H1C-A        TGAACTGCTGCGGTGTTGGTAAATTAAGCAAGCAGATAGATGTAAATAACGCTTGGGCAG 2280
                ************************************************************

CP+H1B-G        GAATATGGAGCACGGGATGAGGATGGGCGGCCAACTGTTAGAGAGGGTAGCAGGGAGGCT 2336
CP+H1C-A        GAATATGGAGCACGGGATGAGGATGGGCGGCCAACTGTTAGAGAGGGTAGCAGGGAGGCT 2340
                ************************************************************

CP+H1B-G        GAGATCTGCCTGCCATGAACTGGGAGGAGAGGCTCCTCTCTCTCTTCACCCCCACTCTGC 2396
CP+H1C-A        GAGATCTGCCTGCCATGAACTGGGAGGAGAGGCTCCTCTCTCTCTTCACCCCCACTCTGC 2400
                ************************************************************

CP+H1B-G        CCCCCAACACTCCTCAGAACTTATCCTCTCCTCTTCTTTCCCCAGGTGAACTTTGAACCA 2456
CP+H1C-A        CCCCCAACACTCCTCAGAACTTATCCTCTCCTCTTCTTTCCCCAGGTGAACTTTGAACCA 2460
                ************************************************************

CP+H1B-G        GGATGGCTGAGCCCCGCCAGGAGTTCGAAGTGATGGAAGATCACGCTGGGACGTACGGGT 2516
CP+H1C-A        GGATGGCTGAGCCCCGCCAGGAGTTCGAAGTGATGGAAGATCACGCTGGGACGTACGGGT 2520
                ************************************************************

CP+H1B-G        TGGGGGACAGGAAAGATCAGGGGGGCTACACCATGCACCAAGACCAAGAGGGTGACACGG 2576
CP+H1C-A        TGGGGGACAGGAAAGATCAGGGGGGCTACACCATGCACCAAGACCAAGAGGGTGACACGG 2580
                ************************************************************

CP+H1B-G        ACGCTGGCCTGAAAGGTTAGTGGACAGCCATGCACAGCAGGCCCAGATCACTGCAAGCCA 2636
CP+H1C-A        ACGCTGGCCTGAAAGGTTAGTGGACAGCCATGCACAGCAGGCCCAGATCACTGCAAGCCA 2640
                ************************************************************

CP+H1B-G        AGGGGTGGCGGGAACAGTTTGCATCCAGAATTGCAAAGAAATTTTAAATACATTATTGTC 2696
CP+H1C-A        AGGGGTGGCGGGAACAGTTTGCATCCAGAATTGCAAAGAAATTTTAAATACATTATTGTC 2700
                ************************************************************

CP+H1B-G        TTAGACTGTCAGTAAAGTAAAGCCTCATTAATTTGAGTGGGCCAAGATAACTCAAGCAGT 2756
CP+H1C-A        TTAGACTGTCAGTAAAGTAAAGCCTCATTAATTTGAGTGGGCCAAGATAACTCAAGCAGT 2760
                ************************************************************

CP+H1B-G        GAGATAATGGCCAGACTCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGAAGGCCCAGG 2816
CP+H1C-A        GAGATAATGGCCAGACACGGTGGCTCACGCCTGTAATCCCAGCACTTTGGAAGGCCCAGG 2820
                *************** ********************************************

CP+H1B-G        CAGGAGGATCCCTTGAGGCCAGGAATTTGAGACCGGCCTGGGCAACATAGCAAGACCCCG 2876
CP+H1C-A        CAGGAGGATCCCTTGAGGCCAGGAATTTGAGACCGGCCTGGGCAACATAGCAAGACCCCG 2880
                ************************************************************

CP+H1B-G        TCTCTAAAATAATTTAAAAATTAGCCAGGTGTTGTGGTGCATGTCTATAGTCCTAGCTAC 2936
CP+H1C-A        TCTCTAAAATAATTTAAAAATTAGCCAGGTGTTGTGGTGCATGTCTATAGTCCTAGCTAC 2940
                ************************************************************

CP+H1B-G        TCAGGATGCTGAGGCAGAAGGATCACTTGAGCCCAGGAGTTCAAGGTTGCAGTAAGCTGT 2996
CP+H1C-A        TCAGGATGCTGAGGCAGAAGGATCACTTGAGCCCAGGAGTTCAAGGTTGCAGTAAGCTGT 3000
                ************************************************************

CP+H1B-G        GATTATAAAACTGCACTCCAGCCTGAGCAACAGAGCAAGACCCTGTCAAAAAAAAAAGAA 3056
CP+H1C-A        GATTATAAAACTGCACTCCAGCCTGAGCAACAGAGCAAGACCCTGTCAAAAAAAAAAGAA 3060
                ************************************************************
```

```
CP+H1B-G        AAGAAAAAAGAAAGAAAGAAATTTACCTTGAGTTACCCACATGAGTGAATGTAGGGACAG 3116
CP+H1C-A        AAGAAAAAAGAAAGAAAGAAATTTACCTTGAGTTACCCACATGAGTGAATGTAGGGACAG 3120
                ************************************************************

CP+H1B-G        AGATTTTAGGGCCTTAACAATCTCTCAAATACAGGGTACTTTTTGAGGCATTAGCCACAC 3176
CP+H1C-A        AGATTTTAGGGCCTTAACAATCTCTCAAATACAGGGTACTTTTTGAGGCATTAGCCACAC 3180
                ************************************************************

CP+H1B-G        CTGTTAGCTTATAAATCAGTGGTATTGATTAGCATGTAAAATATGTGACTTTAAACATTG 3236
CP+H1C-A        CTGTTAGCTTATAAATCAGTGGTATTGATTAGCATGTAAAATATGTGACTTTAAACATTG 3240
                ************************************************************

CP+H1B-G        CTTTTTATCTCTTACTTAGATCAGGCCTGAGTGGCCTCTCTTTAGCAAGAGTTGGTTAGC 3296
CP+H1C-A        CTTTTTATCTCTTACTTAGATCAGGCCTGAGTGGCCTCTCTTTAGCAAGAGTTGGTTAGC 3300
                ************************************************************

CP+H1B-G        CCTGGGATTCTTACTGTAGCCACATTAATAAACAACATCGACTTCTAAACATTCTATAAT 3356
CP+H1C-A        CCTGGGATTCTTACTGTAGCCACATTAATAAACAACATCGACTTCTAAACATTCTATAAT 3360
                ************************************************************

CP+H1B-G        ACCATCTTTTGGCCAAATTGACTTCGCCTCTTCTCGAGCACAGGGAAGGGACAATTCAGC 3416
CP+H1C-A        ACTATCTTTTGGCCAAATTTACTTCGCCTCTTCTCGAGCACAGGGAAGGGACAATTCAGC 3420
                ** ***************** ****************************************

CP+H1B-G        CCTTCTAGGAGGAGGAGGAGGTAGTTTTCTCATTTCTATTAAGGCAACAAAAGCTGCCTT 3476
CP+H1C-A        CCTTCTAGGAGGAGGAGGAGGTAGTTTTCTCATTTCTATTAAGGCAACAAAAGCTGCCTT 3480
                ************************************************************

CP+H1B-G        ACTAAGGACATTCTTGGTGGAGGGCGTGACTGTCAACCACTGTGATCATTTGGGCCTCTC 3536
CP+H1C-A        ACTAAGGACATTCTTGGTGGAGGGCGTGACTGTCAACCACTGTGATCATTTGGGCCTCTC 3540
                ************************************************************

CP+H1B-G        TTGCCCAGGCTTCCCATTCTGAAAGGACAGTTTTATTGTAGGTACACATGGCTGCCATTT 3596
CP+H1C-A        TTGCCCAGGCTTCCCATTCTGAAAGGACAGTTTTATTGTAGGTACACATGGCTGCCATTT 3600
                ************************************************************

CP+H1B-G        CAAATGTAACTCACAGCTTGTCCATCAGTCCTTGGAGGTCTTTCTATGAAAGGAGCTTGG 3656
CP+H1C-A        CAAATGTAACTCACAGCTTGTCCATCAGTCCTTGGAGGTCTTTCTATGAAAGGAGCTTGG 3660
                ************************************************************

CP+H1B-G        TGGCGTCCAAACACCACCCAATGTCCACTTAGAAGTAAGCACCGTGTCTGCCCTGAGCTG 3716
CP+H1C-A        TGGCGTCCAAACACCACCCAATGTCCACTTAGAAGTAAGCACCGTGTCTGCCCTGAGCTG 3720
                ************************************************************

CP+H1B-G        ACTCCTTTTCCAAGGAAGGGGTTGGATCGCTGAGTGTTTTTCCAGGTGTCTACTTGTTGT 3776
CP+H1C-A        ACTCCTTTTCCAAGGAAGGGGTTGGATCGCTGAGTGTTTTTCCAGGTGTCTACTTGTTGT 3780
                ************************************************************

CP+H1B-G        TAATTAATAGCAATGACAAAGCAGAAGGTTCATGCGTAGCTCGGCTTTCTGGTATTTGCT 3836
CP+H1C-A        TAATTAATAGCAATGACAAAGCAGAAGGTTCATGCGTAGCTCGGCTTTCTGGTATTTGCT 3840
                ************************************************************

CP+H1B-G        GCCCGTTGACCAATGGAAGATAAACCTTTGCCTCAGGTGGCACCACTAGCTGGTTAAGAG 3896
CP+H1C-A        GCCCGTTGACCAATGGAAGATAAACCTTTGCCTCAGGTGGCACCACTAGCTGGTTAAGAG 3900
                ************************************************************

CP+H1B-G        GCACTTTGTCCTTTCACCCAGGAGCAAACGCACATCACCTGTGTCCTCATCTGATGGCCC 3956
CP+H1C-A        GCACTTTGTCCTCTCACCCAGGAGCAAACGCACATCACCTGTGTCCTCATCTGATGGCCC 3960
                ************ ************************************************

CP+H1B-G        TGGTGTGGGGCACAGTCGTGTTGGCAGGGAGGGAGGTGGGGTTGGTCCCCTTTGTGGGTT 4016
CP+H1C-A        TGGTGTGGG-CACAGTCGTGTTGGCAGGGAGGGAGGTGGGGTTGGTCCCCTTTGTGGGTT 4019
                ********* ************************************************

CP+H1B-G        TGTTGCGAGGCCGTGTTCCAGCTGTTTCCACAGGGAGCGATTTTCAGCTCCACAGGACAC 4076
CP+H1C-A        TGTTGCGAGGCCGTGTTCCAGCTGTTTCCACAGGGAGCGATTTTAAGCTCCACAGGACAC 4079
                ******************************************* ***************

CP+H1B-G        TGCTCCCCAGTTCCTCCTGAGAACAAAAGGGGGCGCTGGGGAGAGGCCACCGTTCTGAGG 4136
CP+H1C-A        TGCTCCCCAGTTCCTCCTGAGAACAAAAGGGGGCGCTGGGGAGAGGCCACCGTTCTGAGG 4139
                ************************************************************
```

```
CP+H1B-G        GCTCACTGTATGTGTTCCAGAATCTCCCCTGCAGACCCCCACTGAGGACGGATCTGAGGA 4196
CP+H1C-A        GCTCACTGTATGTGTTCCAGAATCTCCCCTGCAGACCCCCACTGAGGACGGATCTGAGGA 4199
                ************************************************************

CP+H1B-G        ACCGGGCTCTGAAACCTCTGATGCTAAGAGCACTCCAACAGCGGAAGGTGGGCCCCCCTT 4256
CP+H1C-A        ACCGGGCTCTGAAACCTCTGATGCTAAGAGCACTCCAACAGCGGAAGGTGGGCCCCCCTT 4259
                ************************************************************

CP+H1B-G        CAGACGCCCCCTCCATGCCTCCAGCCTGTGCTTAGCCGTGCTTTGAGCCTCCCTCCTGGC 4316
CP+H1C-A        CAGACGCCCCCTCCATGCCTCCAGCCTGTGCTTAGCCGTGCTTTGAGCCTCCCTCCTGGC 4319
                ************************************************************

CP+H1B-G        TGCATCTGCTGCTCCCCCTGGCTGAGAGATGTGCTCACTCCTTCGGTGCTTTGCAGGACA 4376
CP+H1C-A        TGCATCTGCTGCTCCCCCTGGCTGAGAGATGTGCTCACTCCTTCGGTGCTTTGCAGGACA 4379
                ************************************************************

CP+H1B-G        GCGTGGTGGGAGCTGAGCCTTGCGTCGATGCCTTGCTTGCTGGTGCTGAGTGTGGGCACC 4436
CP+H1C-A        GCGTGGTGGGAGCTGAGCCTTGCGTCGATGCCTTGCTTGCTGGTGCTGAGTGTGGGCACC 4439
                ************************************************************

CP+H1B-G        TTCATCCCGTGTGTGCTCTGGAGGCAGCCACCCTTGGACAGTCCCGCGCACAGCTCCACA 4496
CP+H1C-A        TTCATCCCGTGTGTGCTCTGGAGGCAGCCACCCTTGGACAGTCCCGCGCACAGCTCCACA 4499
                ************************************************************

CP+H1B-G        AAGCCCCGCTCCATACGATTGTCCTCCCACACCCCCTTCAAAAGCCCCCTCCTCTCTCTT 4556
CP+H1C-A        AAGCCCCGCTCCATACGATTGTCCTCCCACACCCCCTTCAAAAGCCCCCTCCTCTCTCTT 4559
                ************************************************************

CP+H1B-G        TCTTCAGGGGCCAGTAGGTCCCAGAGCAGCCATTTGGCTGAGGGAAGGGGCAGGTCAGTG 4616
CP+H1C-A        TCTTCAGGGGCCAGTAGGTCCCAGAGCAGCCATTTGGCTGAGGGAAGGGGCAGGTCAGTG 4619
                ************************************************************

CP+H1B-G        GACATCTGATCTTGGTTTAGTATCCTTCATTTTGGGGGCTCTGGGTGTGGCCTGGGCCTC 4676
CP+H1C-A        GACATCTGATCTTGGTTTAGTATCCTTCATTTTGGGGGCTCTGGGTGTGGCCTGGGCCTC 4679
                ************************************************************

CP+H1B-G        TGGACTTTGGCCACGGTGTTTGTTCCAGCCCTTCTCCTAACCTGTCCTTTCCAGACACTC 4736
CP+H1C-A        TGGACTTTGGCCACGGTGTTTGTTCCAGCCCTTCTCCTAACCTGTCCTTTCCAGACACTC 4739
                ************************************************************

CP+H1B-G        GGCATCTAGGTTATTAGCACCTCGCATACTTTCTGACATGCTCCTCAGTCCTGATTTTGA 4796
CP+H1C-A        GGCATCTAGGTTATTAGCACCTCGCATACTTTCCGACATGCTCCTCAGTCCTGATTTTGA 4799
                ********************************* ************************

CP+H1B-G        CCATCTTCTCTTGCTTCCCATCTGTGTCAGTCAAGCCGCGGAAAGCCTTCAAAGCTGACA 4856
CP+H1C-A        CCATCTTCTCTTGCTTCCCATCTGTGTCAGTCAAGCCGCGGAAAGCCTTCAAAGCTGACA 4859
                ************************************************************

CP+H1B-G        ACTCCTTATGTGTACCCGGAAAGGCCTGGGAGTGTGCCAGGGCATTGCTCGGGAGGGACG 4916
CP+H1C-A        ACTCCTTATGTGTACCCGGAAAGGCCTGGGAGTGTGCCAGGGCATTGCTCGGGAGGGACG 4919
                ************************************************************

CP+H1B-G        CTGATTTGGAAGCATTTACCTGATGAGAGACTGACAGCAGCTCCTGGTAGCCGAGCTTTC 4976
CP+H1C-A        CTGATTTGGAAGCATTTACCTGATGAGAGACTGACAGCAGCTCCTGGTAGCCGAGCTTTC 4979
                ************************************************************

CP+H1B-G        CCTCCTGCCTCTGCTGTGAAGGTGGACCCATCCAACAGTCAAATGCCTGACTCTGGACAG 5036
CP+H1C-A        CCTCCTGCCTCTGCTGTGAAGGTGGACCCATCCAACAGTCAAACGCCTGACTCTGGACAG 5039
                ******************************************* ****************

CP+H1B-G        GAGCGGACCTATTTATTGCCATGCAAGGGACTCTGCACTTTTGAATTGTGGGTCATGGGC 5096
CP+H1C-A        GAGCGGACCTATTTATTGCCATGCAAGGGACTCTGCACTTTTGAATTGTGGGTCATGGGC 5099
                ************************************************************

CP+H1B-G        TTGGATTTAGGGGTTAGAGCTGGGAGAAGTCTTGGAAGTCACCTAGAGATGACACTGCCA 5156
CP+H1C-A        TTGGATTTAGGGGTTAGAGCTGGGAGAAGTCTTGGAAGTCACCTAGAGATGACACTGCCA 5159
                ************************************************************

CP+H1B-G        TTTTGCAGATGAGGAAACCGTCCAATAAAAATGGACCAAGGACTTGCCCAAAGCCTCACA 5216
CP+H1C-A        TTTTGCAGATGAGGAAACCGTCCAATCAAAATGGACCAAGGACTTGCCCAAAGCCTCACA 5219
                ************************** *********************************
```

```
CP+H1B-G        GCAAAACCATAGGCCCCCGCACTAACCCCAGAGTCCCTGTGCTGTCTTAAGAATCAAATA 5276
CP+H1C-A        GCAAAACCATAGGCCCCCGCACTAACCCCAGAGTCCCTGTGCTGTCTTAAGGATCATATA 5279
                ************************************************ **** ***

CP+H1B-G        GTTGTAAGCAATCATCTGGTTTTCAGTATTTCTTCTTTTAAAATGCCTGGGGCCATGCCC 5336
CP+H1C-A        GTTGTAAGCAATCATCTGGTTTTCAGTATTTCTTCTTTTAAAATGCCTGGGGCCATGCCC 5339
                ************************************************************

CP+H1B-G        AGCAGTCTGTTTCACTGCAGCGTTTACACAGGGCTGCCGGGCTTTCCTGGTGGATGAGCT 5396
CP+H1C-A        AGCAGTCTGTTTCACTGCAGCGTTTACACAGGGCTGCCGGGCTTTCCTGGTGGATGAGCT 5399
                ************************************************************

CP+H1B-G        GGGCGGTTCATGAGCCAGAACCACTCAGCAGCATGTCAGTGTGCTTCCTGGGGAGCTGGT 5456
CP+H1C-A        GGGCGGTTCATGAGCCAGAACCACTCAGCAGCATGTCAGTGTGCTTCCTGGGGAGCTGGT 5459
                ************************************************************

CP+H1B-G        AGCAGGGGCTCCGGGCCCTACTTCAGGGCTGCTTTCTGGCATATGGCTGATCCCCTCCTC 5516
CP+H1C-A        AGCAGGGGCTCCGGGCCCTACTTCAGGGCTGCTTTCTGGCATATGGCTGATCCCCTCCTC 5519
                ************************************************************

CP+H1B-G        ACTCCTCCTCCCTGCATTGCTCCTGCGCAAGAAGCAAAGGTGAGGGGCTGGGTATGGCTC 5576
CP+H1C-A        ACTCCTCCTCCCTGCATTGCTCCTGCGCAAGAAGCAAAGGTGAGGGGCTGGGTATGGCTC 5579
                ************************************************************

CP+H1B-G        GTCCTGGCCCCTCTAAGGTGGATCTCGGTGGTTTCTAGATGTGACAGCACCCTTAGTGGA 5636
CP+H1C-A        GTCCTGGCCCCTCTAAGGTGGATCTCGGTGGTTTCTAGATGTGACAGCACCCTTAGTGGA 5639
                ************************************************************

CP+H1B-G        TGAGGGAGCTCCCGGCAAGCAGGCTGCCGCGCAGCCCCACACGGAGATCCCAGAAGGAAC 5696
CP+H1C-A        TGAGGGAGCTCCCGGCAAGCAGGCTGCCGCGCAGCCCCACACGGAGATCCCAGAAGGAAC 5699
                ************************************************************

CP+H1B-G        CACAGGTGAGGGTAAGCCCCAGAGACCCCCAGGCAGTCAAGGCCCTGCTGGGTGCCCCAG 5756
CP+H1C-A        CACAGGTGAGGGTAAGCCCCAGAGACCCCCAGGCAGTCAAGGCCCTGCTGGGTGCCCCAG 5759
                ************************************************************

CP+H1B-G        CTGACCTGTGACAGAAGTGAGGGAGCTTTGCGTGTTTATCCTCCTGTGGGGCAGGAACAT 5816
CP+H1C-A        CTGACCTGTGACAGAAGTGAGGGAGCTTTGCGTGTTTATCCTCCTGTGGGGCAGGAACAT 5819
                ************************************************************

CP+H1B-G        GGGTGGATTCTGGCTCCTGGGAATCTTGGGTTGTGAGTAGCTCGATGCCTTGGTGCTCAG 5876
CP+H1C-A        GGGTGGATTCTGGCTCCTGGGAATCTTGGGTTGTGAGTAGCTCGATGCCTTGGTGCTCAG 5879
                ************************************************************

CP+H1B-G        TTACCTCCCTGGCTGCCTGCCAGCCTCTCAGAGCATTTAGGGCCTTCTGGACTTCTAGAT 5936
CP+H1C-A        TTACCTCCCTGGCTGCCTGCCAGCCTCTCAGAGCATTTAGGGCCTTCTGGACTTCTAGAT 5939
                ************************************************************

CP+H1B-G        GCTCCTCATCTTGCCTCAGTCAGCGCGTCAGTTCCAGAGACTTCTCTGCAGGGTTTTCTG 5996
CP+H1C-A        GCTCCTCATCTTGCCTCAGTCAGCGCGTCAGTTCCAGAGACTTCTCTGCAGGGTTTTCTG 5999
                ************************************************************

CP+H1B-G        GGGCAGGTGGTGGCAGACCCGTGCCTTCTTGACACCTGAGGTCAGTCCACCCTCCTGCTC 6056
CP+H1C-A        GGGCAGGTGGTGGCAGACCCGTGCCTTCTTGACACCTGAGGTCAGTCCACCCTCCTGCTC 6059
                ************************************************************

CP+H1B-G        AGACTGCCCAGCACAGGGTCACCTCCCAAGGGGTGGACCCCAAGATCACCTGAGCGCACA 6116
CP+H1C-A        AGACTGCCCAGCACAGGGTCACCTCCCAAGGGGTGGACCCCAAGATCACCTGAGCGCACA 6119
                ************************************************************

CP+H1B-G        GAGGGTGCAGATGACTGGACCACACCTTTTGGTGATCTTAATGAGGTGGTCCCAGAGGAG 6176
CP+H1C-A        GAGGGTGCAGATGACTGGACCACACCTTTTGGTGATCTTAATGAGGTGGTCCCAGAGGAG 6179
                ************************************************************

CP+H1B-G        CTCAGACATGCAATCTAGCATCCAGTTCTGGGACTCTGTCTCCTTTTCAAACGTATTCAT 6236
CP+H1C-A        CTCAGACATGCAATCTAGCATCCAGTTCTGGGACTCTGTCTCCTTTTCAAACGTATTCAT 6239
                ************************************************************

CP+H1B-G        GTAGAACAGGCATGACGAGAATGCCTTGTCAACATGGGTGATGGGGAATCAATCAGACAG 6296
CP+H1C-A        GTAGAACAGGCATGACGAGAATGCCTTGTCAACATGGGTGATGGGGAATCAATCAGACAG 6299
                ************************************************************
```

```
CP+H1B-G        GGCGCATGCCCCGTGAGCCCATTGCCCGCCCTCCCATGCCCTCAGCAGCTGCCTGGGGAC 6356
CP+H1C-A        GGCGCATGCCCCGTGAGCCCATTGCCCGCCCTCCCATGCCCTCAGCAGCTGCCTGGGGAC 6359
                ************************************************************

CP+H1B-G        AGCCAATGGCCTGGGTGTTTCTGAGGCTACCACATGGCTTCCAGGAAACTCGAGAACCTT 6416
CP+H1C-A        AGCCAATGGCCTGGGTGTTTCTGAGGCTACCACATGGCTTCCAGGAAACTCGAGAACCTT 6419
                ************************************************************

CP+H1B-G        TCTCTCCCTTGCCTACACTCTTCACACAGGCCTGTGCTGGCCAGCGGTGGGGATCCGGCA 6476
CP+H1C-A        TCTCTCCCTTGCCTACACTCTTCACACAGGCCTGTGCTGGCCAGCGGTGGGGATCCGGCA 6479
                ************************************************************

CP+H1B-G        TTCCTATCTTAGGTGCAGAAAGTGACTGACTCATTGCAGGCCTGGGAGATAAGACTGATG 6536
CP+H1C-A        TTCCTATCTTAGGTGCAGAGAGTGACTGACTCATTGCAGGCCTGGGAGATAAGACTGATG 6539
                ****************** ****************************************

CP+H1B-G        GCCCAGCCAGCAAGATGTATGGATTTCTCAGAGGCAGTGGCCTCTGTCATTGTCCTCAGG 6596
CP+H1C-A        GCCCAGCCAGCAAGATGTATGGATTTCTCAGAGGCAGTGGCCTCTGTCATTGTCCTCAGG 6599
                ************************************************************

CP+H1B-G        AAATGCTGGTGATTCTGGTGGCCTGAGGTCAATGCATGTCAACGTGGCCAACTTGCCTTA 6656
CP+H1C-A        AAATGCTGGTGATTCTGGTGGCCTGAGGTCAATGCATGTCAACGTGGCCAACTTGCCTTA 6659
                ************************************************************

CP+H1B-G        TAAACTTTTTTTCTGGACAATTGCGTGCACTGTCCTGTAACAGTGTCCTGTTGTTTATGA 6716
CP+H1C-A        TAAACTTTTTTTCTGGACAATTGCGTGCACTGTCCTGTAACAGTGTCCTGTTGTTTATGA 6719
                ************************************************************

CP+H1B-G        TGCAGAAATAGGTGTTTTTAAAGCCTATTGATTTTGGTACTATTAATGTGGTCAGGAACT 6776
CP+H1C-A        TGCAGAAATAGGTGTTTTTAAAGCCTATTGATTTTGGTACTATTAATGTGGTCAGGAACT 6779
                ************************************************************

CP+H1B-G        TTCTCAGTCTTTCTTGTTTGGGGTGAGCTGTGGCTTCCTAAACAGGAACCCAAGACACCC 6836
CP+H1C-A        TTCTCAGTCTTTCTTGTTTGGGGTGAGCTGTGGCTTCCTAAACAGGAACCCAAGACACCC 6839
                ************************************************************

CP+H1B-G        CCAAAAGCTGCTCACCAGCACTGCCAGCCTCCCTCTTACCAAGTAGCACCCGTTCAGGAC 6896
CP+H1C-A        CCAAAAGCTGCTCACCAGCACTGCCAGCCTCCCTCTTACCAAGTAGCACCCGTTCAGGAC 6899
                ************************************************************

CP+H1B-G        ATTCTGCGAAAGGCATTTGCCCAGAAGTTGGGAGGAAGGAAATGTAACATTTTGGGGCAC 6956
CP+H1C-A        ATTCTGCGAAAGGCATTTGCCCAGAAGTTGGGAGGAAGGAAATGTAACATTTTGGGGCAC 6959
                ************************************************************

CP+H1B-G        CTACCATATGCCAGGCACCAGGCTAAACGTGTTCACACAAATTCTCTTACTAACCCTCAC 7016
CP+H1C-A        CTACCATATGCCAGGCACCAGGCTAAACGTGTTCACACAAATTCTCTTACTAACCCTCAC 7019
                ************************************************************

CP+H1B-G        CATCCTTCTACAAGACAAACTAGTATCTTCATCTTGGGGTTCAAGATGAGGAAATGGAGG 7076
CP+H1C-A        CATCCTTCTACAAGACAAACTAGTATCTTCATCTTGGGGTTCAAGATGAGGAAATGGAGG 7079
                ************************************************************

CP+H1B-G        CTCAGAGAGGTTGAATGAATGCCGGTGCCTGGATATGAACCCCATCTGCCTGACTCCGCA 7136
CP+H1C-A        CTCAGAGAGGTTGAATGAATGCCGGTGCCTGGATATGAACCCCATCTGCCTGACTCCGCA 7139
                ************************************************************

CP+H1B-G        ACCCAGGCAAAGTCTTTCCTTGAACTTCCCAGCAGCCACTGCTTAGACACAGCCTCCACA 7196
CP+H1C-A        ACCCAGGCAAAGTCTTTCCTTGAACTTCCCAGCAGCCACTGCTTAGACACAGCCTCCACA 7199
                ************************************************************

CP+H1B-G        ACCATGGCTCAGCAGCAAATTGCTTCTCTGACCTCACTCAGCCTGTGTGTCCTTGTTGAG 7256
CP+H1C-A        ACCATGGCTCAGCAGCAAATTGCTTCTCTGACCTCACTCAGCCTGTGTGTCCTTGTTGAG 7259
                ************************************************************

CP+H1B-G        TGAGGCATTCAGGACCCTGGTCCCAAAGTGGAGAAAGTCTTTCCTACTAGGTCATAGCTA 7316
CP+H1C-A        TGAGGCATTCAGGACCCTGGTCCCAAAGTGGAGAAAGTCTTTCCTACTAGGTCATAGCTA 7319
                ************************************************************

CP+H1B-G        CACCTGCATGTGGGTGCTGTGCCTTTTGTTTAGTGAACTTTTATCACCAGCATCCTCAGC 7376
CP+H1C-A        CACCTGCATGTGGGTGCTGTGCCTTTTGTTTAGTGAACTTTTATCACCAGCATCCTCAGC 7379
                ************************************************************
```

```
CP+H1B-G        AATGACATTTGCAGAGAAGCCAGAGCTGAGGCACCTTGGTATTCTTGGGATGTGACTTTC 7436
CP+H1C-A        AATGACATTTGCAGAGAAGCCAGAGCTGAGGCACCTTGGTATTCTTGGGATGTGACTTTC 7439
                ************************************************************

CP+H1B-G        CTGAATGTTTAAGGGAAAATGCCCGAAGGTACAGAGAGCTTGGTTTCTAGTAAACAATAA 7496
CP+H1C-A        CTGAATGTTTAAGGGAAAATGCCCGAAGGTACAGAGAGCTTGGTTTCTAGTAAACAATAA 7499
                ************************************************************

CP+H1B-G        CTGTCTTGCTTTTACCCCCCTTCATTTGCTGACACATACACCAGCACCCAACTTTTCTAT 7556
CP+H1C-A        CTGTCTTGCTTTTACCCCCCTTCATTTGCTGACACATACACCAGCACCCAACTTTTCTAT 7559
                ************************************************************

CP+H1B-G        ACAAAGTTGTCCAGCTGAAGAAGCAGGCATTGGAGACACCCCCAGCCTGGAAGACGAAGC 7616
CP+H1C-A        ACAAAGAAGTCCAGCTGAAGAAGCAGGCATTGGAGACACCCCCAGCCTGGAAGACGAAGC 7619
                ******  ***************************************************

CP+H1B-G        TGCTGGTCACGTGACCCAAGCTCGCATGGTCAGTAAAAGCAAAGACGGGACTGGAAGCGA 7676
CP+H1C-A        TGCTGGTCACGTGACCCAAGCTCGCATGGTCAGTAAAAGCAAAGACGGGACTGGAAGCGA 7679
                ************************************************************

CP+H1B-G        TGACAAAAAAGCCAAGGGGGCTGATGGTAAAACGAAGATCGCCACACCGCGGGGAGCAGC 7736
CP+H1C-A        TGACAAAAAAGCCAAGGGGGCTGATGGTAAAACGAAGATCGCCACACCGCGGGGAGCAGC 7739
                ************************************************************

CP+H1B-G        CCCTCCAGGCCAGAAGGGCCAGGCCAACGCCACCAGGATTCCAGCAAAAACCCCGCCCGC 7796
CP+H1C-A        CCCTCCAGGCCAGAAGGGCCAGGCCAACGCCACCAGGATTCCAGCAAAAACCCCGCCCGC 7799
                ************************************************************

CP+H1B-G        TCCAAAGACACCACCCAGCTCTGGTGAACCTCCAAAATCAGGGGATCGCAGCGGCTACAG 7856
CP+H1C-A        TCCAAAGACACCACCCAGCTCTGGTGAACCTCCAAAATCAGGGGATCGCAGCGGCTACAG 7859
                ************************************************************

CP+H1B-G        CAGCCCCGGCTCCCCAGGCACTCCCGGCAGCCGCTCCCGCACCCCGTCCCTTCCAACCCC 7916
CP+H1C-A        CAGCCCCGGCTCCCCAGGCACTCCCGGCAGCCGCTCCCGCACCCCGTCCCTTCCAACCCC 7919
                ************************************************************

CP+H1B-G        ACCCACCCGGGAGCCCAAGAAGGTGGCAGTGGTCCGTACTCCACCCAAGTCGCCGTCTTC 7976
CP+H1C-A        ACCCACCCGGGAGCCCAAGAAGGTGGCAGTGGTCCGTACTCCACCCAAGTCGCCGTCTTC 7979
                ************************************************************

CP+H1B-G        CGCCAAGAGCCGCCTGCAGACAGCCCCCGTGCCCATGCCAGACCTGAAGAATGTCAAGTC 8036
CP+H1C-A        CGCCAAGAGCCGCCTGCAGACAGCCCCCGTGCCCATGCCAGACCTGAAGAATGTCAAGTC 8039
                ************************************************************

CP+H1B-G        CAAGATCGGCTCCACTGAGAACCTGAAGCACCAGCCGGGAGGCGGGAAGTCTAGAGTGAG 8096
CP+H1C-A        CAAGATCGGCTCCACTGAGAACCTGAAGCACCAGCCGGGAGGCGGGAAGTCTAGAGTGAG 8099
                ************************************************************

CP+H1B-G        AGTGGCTGGCTGCGCGTGGAGGTGTGGGGGGCTGCGCCTGGAGGGGTAGGGCTGTGCCTG 8156
CP+H1C-A        AGTGGCTGGCTGCGCGTGGAGGTGTGGGGGGCTGCGCCTGGAGGGGTAGGGCTGTGCCTG 8159
                ************************************************************

CP+H1B-G        GAAGGGTAGGGCTGCGCCTGGAGGTGCGCGGTTGAGCGTGGAGTCGTGGGACTGTGCATG 8216
CP+H1C-A        GAAGGGTAGGGCTGCGCCTGGAGGTGCGCGGTTGAGCGTGGAGTCGTGGGACTGTGCATG 8219
                ************************************************************

CP+H1B-G        GAGGTGTGGGGCTCCCCGCACCTGAGCACCCCCGCATAACACCCCAGTCCCCTCTGGACC 8276
CP+H1C-A        GAGGTGTGGGGCTCCCCGCACCTGAGCACCCCCGCATAACACCCCAGTCCCCTCTGGACC 8279
                ************************************************************

CP+H1B-G        CTCTTCAAGGAAGTTCAGTTCTTTATTGGGCTCTCCACTACACTGTGAGTGCCCTCCTCA 8336
CP+H1C-A        CTCTTCAAGGAAGTTCAGTTCTTTATTGGGCTCTCCACTACACTGTGAGTGCCCTCCTCA 8339
                ************************************************************

CP+H1B-G        GGCGAGAGAACGTTCTGGCTCTTCTCTTGCCCCTTCAGCCCCTGTTAATCGGACAGAGAT 8396
CP+H1C-A        GGCGAGAGAACGTTCTGGCTCTTCTCTTGCCCCTTCAGCCCCTGTTAATCGGACAGAGAT 8399
                ************************************************************

CP+H1B-G        GGCAGGGCTGTGTCTCCACGGCCGGAGGCTCTCATAGTCAGGGCACCCACAGCGGTTCCC 8456
CP+H1C-A        GGCAGGGCTGTGTCTCCACGGCCGGAGGCTCTCATAGTCAGGGCACCCACAGCGGTTCCC 8459
                ************************************************************
```

```
CP+H1B-G        CACCTGCCTTCTGGGCAGAATACACTGCCACCCATAGGTCAGCATCTCCACTCGTGGGCC 8516
CP+H1C-A        CACCTGCCTTCTGGGCAGAATACACTGCCACCCATAGGTCAGCATCTCCACTCGTGGGCC 8519
                ************************************************************

CP+H1B-G        ATCTGCTTAGGTTGGGTTCCTCTGGATTCTGGGGAGATTGGGGGTTCTGTTTTGATCAGC 8576
CP+H1C-A        ATCTGCTTAGGTTGGGTTCCTCTGGATTCTGGGGAGATTGGGGGTTCTGTTTTGATCAGC 8579
                ************************************************************

CP+H1B-G        TGATTCTTCTGGGAGCAAGTGGGTGCTCGCGAGCTCTCCAGCTTCCTAAAGGTGGAGAAG 8636
CP+H1C-A        TGATTCTTCTGGGAGCAAGTGGGTGCTCGCGAGCTCTCCAGCTTCCTAAAGGTGGAGAAG 8639
                ************************************************************

CP+H1B-G        CACAGACTTCAGGGGCCTGGCCTGGATCCCTTTCCCCATTCCTGTCCCTGTGCCCCTCGT 8696
CP+H1C-A        CACAGACTTCGGGGGCCTGGCCTGGATCCCTTTCCCCATTCCTGTCCCTGTGCCCCTCGT 8699
                ********** ***************************************************

CP+H1B-G        CTGGGTGCGTTACCATGGTTTTCTATTTCATAGTTCTTAGGCAAATTGGTAAAAATCATT 8756
CP+H1C-A        CTGGGTGCGTTACCATGGTTTTCTATTTCATAGTTCTTAGGCAAATTGGTAAAAATCATT 8759
                ************************************************************

CP+H1B-G        TCTCATCAAAACGCTGATATTTTCACACCTCCCTGGTGTCTGCAGAAAGAACCTTCCAGA 8816
CP+H1C-A        TCTCATCAAAACGCTGATATTTTCACACCTCCCTGGTGTCTGCAGAAAGAACCTTCCAGA 8819
                ************************************************************

CP+H1B-G        AATGCAGTCGTGGGAGACCCATCCAGGCCACCCCTGCTTATGGAAGAGCTGAGAAAAAGC 8876
CP+H1C-A        AATGCAGTCGTGGGAGACCCATCCAGGCCACCCCTGCTTATGGAAGAGCTGAGAAAAAGC 8879
                ************************************************************

CP+H1B-G        CCCACGGGGGCATTTGCTCAGCTTCCGTTACGCACCTAGTGGCATTGTGGGTGGGAGAGG 8936
CP+H1C-A        CCCACGGGAGCATTTGCTCAGCTTCCGTTACGCACCTAGTGGCATTGTGGGTGGGAGAGG 8939
                ******** ***************************************************

CP+H1B-G        GCTGGTGGGTGGATGGAAGGAGAAGGCACAGCCCCCCCTTGCAGGGACAGAGCCCTCGTA 8996
CP+H1C-A        GCTGGTGGGTGGATGGAAGGAGAAGGCACAGCCCCCCCTTGCAGGGACAGAGCCCTCGTA 8999
                ************************************************************

CP+H1B-G        CAGAAGGGACACCCCACATTTGTCTTCCCCACAAAGCGGCCTGTGTCCTGCCTACGGGGT 9056
CP+H1C-A        CAGAAGGGACACCCCACATTTGTCTTCCCCACAAAGCGGCCTGTGTCCTGCCTACGGGGT 9059
                ************************************************************

CP+H1B-G        CAGGGCTTCTCAAACCTGGCTGTGTGTCAGAATCACCAGGGGAACTTTTCAAAACTAGAG 9116
CP+H1C-A        CAGGGCTTCTCAAACCTGGCTGTGTGTCAGAATCACCAGGGGAACTTTTCAAAACTAGAG 9119
                ************************************************************

CP+H1B-G        AGACTGAAGCCAGACTCCTAGATTCTAATTCTAGGTCAGGGCTAGGGGCTGAGATTGTAA 9176
CP+H1C-A        AGACTGAAGCCAGACTCCTAGATTCTAATTCTAGGTCAGGGCTAGGGGCTGAGATTGTAA 9179
                ************************************************************

CP+H1B-G        AAATCCACAGGTGATTCTGATGCCCGGCAGGCTTGAGAACAGCCGCAGGGAGTTCTCTGG 9236
CP+H1C-A        AAATCCACAGGTGATTCTGATGCCCGGCAGGCTTGAGAACAGCCGCAGGGAGTTCTCTGG 9239
                ************************************************************

CP+H1B-G        GAATGTGCCGGTGGGTCTAGCCAGGTGTGAGTGGAGATGCCGGGGAACTTCCTATTACTC 9296
CP+H1C-A        GAATGTGCCGGTGGGTCTAGCCAGGTGTGAGTGGAGATGCCGGGGAACTTCCTATTACTC 9299
                ************************************************************

CP+H1B-G        ACTCGTCAGTGTGGCCGAACATATTTTTCACTTGACCTCAGGCTGGTGAACGCTCCCCTC 9356
CP+H1C-A        ACTCGTCAGTGTGGCCGAACACATTTTTCACTTGACCTCAGGCTGGTGAACGCTCCCCTC 9359
                ******************** ***************************************

CP+H1B-G        TGGGGTTCAGGCCTCACGATGCCATCCTTTTGTGAAGTGAGGACCTGCAATCCCAGCTTC 9416
CP+H1C-A        TGGGGTTCAGGCCTCACGATGCCATCCTTTTGTGAAGTGAGGACCTGCAATCCCAGCTTC 9419
                ************************************************************

CP+H1B-G        GTAAAGCCCGCTGGAAATCACTCACACTTCTGGGATGCCTTCAGAGCAGCCCTCTATCCC 9476
CP+H1C-A        GTAAAGCCCGCTGGAAATCACTCACACTTCTGGGATGCCTTCAGAGCAGCCCTCTATCCC 9479
                ************************************************************

CP+H1B-G        TTCAGCTCCCCTGGGATGTGACTCAACCTCCCGTCACTCCCCAGACTGCCTCTGCCAAGT 9536
CP+H1C-A        TTCAGCTCCCCTGGGATGTGACTCAACCTCCCGTCACTCCCCAGACTGCCTCTGCCAAGT 9539
                ************************************************************
```

```
CP+H1B-G        CCGAAAGTGGAGGCATCCTTGCGAGCAAGTAGGCGGGTCCAGGGTGGCGCATGTCACTCA 9596
CP+H1C-A        CCGAAAGTGGAGGCATCCTTGCGAGCAAGTAGGCGGGTCCAGGGTGGCGCATGTCACTCA 9599
                ************************************************************

CP+H1B-G        TCGAAAGTGGAGGCGTCCTTGCGAGCAAGCAGGCGGGTCCAGGGTGGCGTGTCACTCATC 9656
CP+H1C-A        TCGAAAGTGGAGGCGTCCTTGCGAGCAAGCAGGCGGGTCCAGGGTGGCGTGTCACTCATC 9659
                ************************************************************

CP+H1B-G        CTTTTTTCTGGCTACCAAAGGTGCAGATAATTAATAAGAAGCTGGATCTTAGCAACGTCC 9716
CP+H1C-A        CTTTTTTCTGGCTACCAAAGGTGCAGATAATTAATAAGAAGCTGGATCTTAGCAACGTCC 9719
                ************************************************************

CP+H1B-G        AGTCCAAGTGTGGCTCAAAGGATAATATCAAACACGTCCCGGGAGGCGGCAGTGTGAGTA 9776
CP+H1C-A        AGTCCAAGTGTGGCTCAAAGGATAATATCAAACACGTCCCGGGAGGCGGCAGTGTGAGTA 9779
                ************************************************************

CP+H1B-G        CCTTCACACGTCCCATGCGCCGTGCTGTGGCTTGAATTATTAGGAAGTGGTGTGAGTGCG 9836
CP+H1C-A        CCTTCACACGTCCCATGCGCCGTGCTGTGGCTTGAATTATTAGGAAGTGGTGTGAGTGCG 9839
                ************************************************************

CP+H1B-G        TACACTTGCGAGACACTGCATAGAATAAATCCTTCTTGGGCTCTCAGGATCTGGCTGCGA 9896
CP+H1C-A        TACACTTGCGAGACACTGCATAGAATAAATCCTTCTTGGGCTCTCAGGATCTGGCTGCGA 9899
                ************************************************************

CP+H1B-G        CCTCTGGGTGAATGTAGCCCGGCTCCCCACATTCCCCCACACGGTCCACTGTTCCCAGAA 9956
CP+H1C-A        CCTCTGGGTGAATGTAGCCCGGCTCCCCACATTCCCCCACACGGTCCACTGTTCCCAGAA 9959
                ************************************************************

CP+H1B-G        GCCCCTTCCTCATATTCTAGGAGGGGGTGTCCCAGCATTTCTGGGTCCCCCAGCCTGCGC 10016
CP+H1C-A        GCCCCTTCCTCATATTCTAGGAGGGGGTGTCCCAGCATTTCTGGGTCCCCCAGCCTGCGC 10019
                ************************************************************

CP+H1B-G        AGGCTGTGTGGACAGAATAGGGCAGATGACGGACCCTCTCTCCGGACCCTGCCTGGGAAG 10076
CP+H1C-A        AGGCTGTGTGGACAGAATAGGGCAGATGACGGACCCTCTCTCCGGACCCTGCCTGGGAAG 10079
                ************************************************************

CP+H1B-G        CTGAGAATACCCATCAAAGTCTCCTTCCACTCATGCCCAGCCCTGTCCCCAGGAGCCCCA 10136
CP+H1C-A        CTGAGAATACCCATCAAAGTCTCCTTCCACTCATGCCCAGCCCTGTCCCCAGGAGCCCCA 10139
                ************************************************************

CP+H1B-G        TAGCCCATTGGAAGTTGGGCTGAAGGTGGTGGCACCTGAGACTGGGCTGCCGCCTCCTCC 10196
CP+H1C-A        TAGCCCATTGGAAGTTGGGCTGAAGGTGGTGGCACCTGAGACTGGGCTGCCGCCTCCTCC 10199
                ************************************************************

CP+H1B-G        CCCGACACCTGGGCAGGTTGACGTTGAGTGGCTCCACTGTGGACAGGTGACCCGTTTGTT 10256
CP+H1C-A        CCCGACACCTGGGCAGGTTGACGTTGAGTGGCTCCACTGTGGACAGGTGACCCGTTTGTT 10259
                ************************************************************

CP+H1B-G        CTGATGAGCGGACACCAAGGTCTTACTGTCCTGCTCAGCTGCTGCTCCTACACGTTCAAG 10316
CP+H1C-A        CTGATGAGCGGACACCAAGGTCTTACTGTCCTGCTCAGCTGCTGCTCCTACACGTTCAAG 10319
                ************************************************************

CP+H1B-G        GCAGGAGCCGATTCCTAAGCCTCCAGCTTATGCTTAGCCTGCGCCACCCTCTGGCAGAGA 10376
CP+H1C-A        GCAGGAGCCGATTCCTAAGCCTCCAGCTTATGCTTAGCCTGCGCCACCCTCTGGCAGAGA 10379
                ************************************************************

CP+H1B-G        CTCCAGATGCAAAGAGCCAAACCAAAGTGGCATGCCTCGAGCTTACTGAGACACTAAATC 10436
CP+H1C-A        CTCCAGATGCAAAGAGCCAAACCAAAATGGCATGCCTCGAGCTTACTGAGACACTAAATC 10439
                ************************* ********************************

CP+H1B-G        TGTTGGTTTCTGCTGTGCCACCTACCCACCCTGTTGGTGTTGCTTTGTTCCTATTGCTAA 10496
CP+H1C-A        TGTTGGTTTCTGCTGTGCCACCTACCCACCCTGTTGGTGTTGCTTTGTTCCTATTGCTAA 10499
                ************************************************************

CP+H1B-G        AGACAGGAATGTCCAGGACACTGAGTGTGCAGGTGCCTGCTGGTTCTCACGTCCGAGCTG 10556
CP+H1C-A        AGACAGGAATGTCCAGGACACTGAGTGTGCAGGTGCCTGCTGGTTCTCACGTCCGAGCTG 10559
                ************************************************************

CP+H1B-G        CTGAACTCCGCTGGGTCCTGCTTACTGATGGTCTTTGCTCTAGTGCTTTCCAGGGTCCGT 10616
CP+H1C-A        CTGAACTCCGCTGGGTCCTGCTTACTGATGGTCTTTGCTCTAGTGCTTTCCAGGGTCCGT 10619
                ************************************************************
```

```
CP+H1B-G        GGAAGCTTTTCCTGGAATAAAGCCCACGCATCGACCCTCACAGCGCCTCCCCTCTTTGAG 10676
CP+H1C-A        GGAAGCTTTTCCTGGAATAAAGCCCACGCATCGACCCTCACAGCGCCTCCCCTCTTTGAG 10679
                ************************************************************

CP+H1B-G        GCCCAGCAGATACCCCACTCCTGCCTTTCCAGCAAGATTTTTCAGATGCTGTGCATACTC 10736
CP+H1C-A        GCCCAGCAGATACCCCACTCCTGCCTTTCCAGCAAGATTTTTCAGATGCTGTGCATACTC 10739
                ************************************************************

CP+H1B-G        ATCATATTGATCACTTTTTTCTTCATGCCTGATTGTGATCTGTCAATTTCATGTCAGGAA 10796
CP+H1C-A        ATCATATTGATCACTTTTTTCTTCATGCCTGATTGTGATCTGTCAATTTCATGTCAGGAA 10799
                ************************************************************

CP+H1B-G        AGGGAGTGACATTTTTACACTTAAGCGTTTGCTGAGCAAATGTCTGGGTCTTGCACAATG 10856
CP+H1C-A        AGGGAGTGACATTTTTACACTTAAGCGTTTGCTGAGCAAATGTCTGGGTCTTGCACAATG 10859
                ************************************************************

CP+H1B-G        ACAATGGGTCCCTGTTTTTCCCAGAGGCTCTTTTGTTCTGCAGGGATTGAAGACACTCCA 10916
CP+H1C-A        ACAATGGGTCCCTGTTTTTCCCAGAGGCTCTTTTGTTCTGCAGGGATTGAAGACACTCCA 10919
                ************************************************************

CP+H1B-G        GTCCCACAGTCCCCAGCTCCCCTGGGGCAGGGTTGGCAGAATTTCGACAACACATTTTTC 10976
CP+H1C-A        GTCCCACAGTCCCCAGCTCCCCTGGGGCAGGGTTGGCAGAATTTCGACAACACATTTTTC 10979
                ************************************************************

CP+H1B-G        CACCCTGACTAGGATGTGCTCCTCATGGCAGCTGGGAACCACTGTCCAATAAGGGCCTGG 11036
CP+H1C-A        CACCCTGACTAGGATGTGCTCCTCATGGCAGCTGGGAACCACTGTCCAATAAGGGCCTGG 11039
                ************************************************************

CP+H1B-G        GCTTACACAGCTGCTTCTCATTGAGTTACACCCTTAATAAAATAATCCCATTTTATCCTT 11096
CP+H1C-A        GCTTACACAGCTGCTTCTCATTGAGTTACACCCTTAATAAAATAATCCCATTTTATCCTT 11099
                ************************************************************

CP+H1B-G        TTTGTCTCTCTGTCTTCCTCTCTCTCTGCCTTTCCTCTTCTCTCTCCTCCTCTCTCATCT 11156
CP+H1C-A        TTTGTCTCTCTGTCTTCCTCTCTCTCTGCCTTTCCTCTTCTCTCTCCTCCTCTCTCATCT 11159
                ************************************************************

CP+H1B-G        CCAGGTGCAAATAGTCTACAAACCAGTTGACCTGAGCAAGGTGACCTCCAAGTGTGGCTC 11216
CP+H1C-A        CCAGGTGCAAATAGTCTACAAACCAGTTGACCTGAGCAAGGTGACCTCCAAGTGTGGCTC 11219
                ************************************************************

CP+H1B-G        ATTAGGCAACATCCATCATAAACCAGGACGTCGAGGTGGCCAGGTGGAAGTAAAATCTGA 11276
CP+H1C-A        ATTAGGCAACATCCATCATAAACCAGGACGTCGAGGTGGCCAGGTGGAAGTAAAATCTGA 11279
                ************************************************************

CP+H1B-G        GAAGCTTGACTTCAAGGACAGAGTCCAGTCGAAGATTGGGTCCCTGGACAATATCACCCA 11336
CP+H1C-A        GAAGCTTGACTTCAAGGACAGAGTCCAGTCGAAGATTGGGTCCCTGGACAATATCACCCA 11339
                ************************************************************

CP+H1B-G        CGTCCCTGGCGGAGGAAATAAAAAGATTGAAACCCACAAGCTGACCTTCCGCGAGAACGC 11396
CP+H1C-A        CGTCCCTGGCGGAGGAAATAAAAAGATTGAAACCCACAAGCTGACCTTCCGCGAGAACGC 11399
                ************************************************************

CP+H1B-G        CAAAGCCAAGACAGACCACGGGGCGGAGATCGTGTACAAGTCGCCAGTGGTGTCTGGGGA 11456
CP+H1C-A        CAAAGCCAAGACAGACCACGGGGCGGAGATCGTGTACAAGTCGCCAGTGGTGTCTGGGGA 11459
                ************************************************************

CP+H1B-G        CACGTCTCCACGGCATCTCAGCAATGTCTCCTCCACCGGCAGCATCGACATGGTAGACTC 11516
CP+H1C-A        CACGTCTCCACGGCATCTCAGCAATGTCTCCTCCACCGGCAGCATCGACATGGTAGACTC 11519
                ************************************************************

CP+H1B-G        GCCCCAGCTCGCCACGCTAGCTGACGAGGTGTCTGCCTCCCTGGCCAAGCAGGGTTTGGA 11576
CP+H1C-A        GCCCCAGCTCGCCACGCTAGCTGACGAGGTGTCTGCCTCCCTGGCCAAGCAGGGTTTGGA 11579
                ************************************************************

CP+H1B-G        TTACAAGGATGACGACGATAAGTGAACAACTTTGTATAATAAAGTTGTCCCTGGGGCGGT 11636
CP+H1C-A        TTACAAGGATGACGACGATAAGTGAACAACTTTGTATAATAAAGTTGTCCCTGGGGCGGT 11639
                ************************************************************

CP+H1B-G        CAATAATTGTGGGGAGGAGAGAATGAGAGAGTGTGGAAAAAAAAAAGAATAATGACCCGGC 11696
CP+H1C-A        CAATAATTGTGGAGAGGAGAGAATGAGAGAGTGTGGAAAAAAAAAGAATAATGACCCGGC 11699
                ************ ********************************************
```

```
CP+H1B-G        CCCCGCCCTCTGCCCCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAATCACTTAACCT 11756
CP+H1C-A        CCCCGCCCTCTGCCCCCAGCTGCTCCTCGCAGTTCGGTTAATTGGTTAATCACTTAACCT 11759
                ************************************************************

CP+H1B-G        GCTTTTGTCACTCGGCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAGTAAGAGCAAAT 11816
CP+H1C-A        GCTTTTGTCACTCGGCTTTGGCTCGGGACTTCAAAATCAGTGATGGGAGTAAGAGCAAAT 11819
                ************************************************************

CP+H1B-G        TTCATCTTTCCAAATTGATGGGTGGGCTAGTAATAAAATATTTAAAAAAAAACATTCAAA 11876
CP+H1C-A        TTCATCTTTCCAAATTGATGGGTGGGCTAGTAATAAAATATTTAAAAAAAAACATTCAAA 11879
                ************************************************************

CP+H1B-G        AACATGGCCACATCCAACATTTCCTCAGGCAATTCCTTTTGATTCTTTTTTCTTCCCCCT 11936
CP+H1C-A        AACATGGCCACATCCAACATTTCCTCAGGCAATTCCTTTTGATTCTTTTTTCTTCCCCCT 11939
                ************************************************************

CP+H1B-G        CCATGTAGAAGAGGGAGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGGATTTCAAGGGAC 11996
CP+H1C-A        CCATGTAGAAGAGGGAGAAGGAGAGGCTCTGAAAGCTGCTTCTGGGGGATTTCAAGGGAC 11999
                ************************************************************

CP+H1B-G        TGGGGGTGCCAACCACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGCAGTGGCAGCAAC 12056
CP+H1C-A        TGGGGGTGCCAACCACCTCTGGCCCTGTTGTGGGGGTGTCACAGAGGCAGTGGCAGCAAC 12059
                ************************************************************

CP+H1B-G        AAAGGATTTGAAACTTGGTGTGTTCGTGGAGCCACAGGCAGACGATGTCAACCTTGTGTG 12116
CP+H1C-A        AAAGGATTTGAAACTTGGTGTGTTCGTGGAGCCACAGGCAGACGATGTCAACCTTGTGTG 12119
                ************************************************************

CP+H1B-G        AGTGTGACGGGGGTTGGGGTGGGGCGGGAGGCCACGGGGGAGGCCGAGGCAGGGGCTGGG 12176
CP+H1C-A        AGTGTGACGGGGGTTGGGGTGGGGCGGGAGGCCACGGGGGAGGCCGAGGCAGGGGCTGGG 12179
                ************************************************************

CP+H1B-G        CAGAGGGGAGAGGAAGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACGTGCTGGAGAGTA 12236
CP+H1C-A        CAGAGGGGAGAGGAAGCACAAGAAGTGGGAGTGGGAGAGGAAGCCACGTGCTGGAGAGTA 12239
                ************************************************************

CP+H1B-G        GACATCCCCCTCCTTGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGCAGCGTCTGAGCG 12296
CP+H1C-A        GACATCCCCCTCCTTGCCGCTGGGAGAGCCAAGGCCTATGCCACCTGCAGCGTCTGAGCG 12299
                ************************************************************

CP+H1B-G        GCCGCCTGTCCTTGGTGGCCGGGGGTGGGGGCCTGCTGTGGGTCAGTGTGCCACCCTCTG 12356
CP+H1C-A        GCCGCCTGTCCTTGGTGGCCGGGGGTGGGGGCCTGCTGTGGGTCAGTGTGCCACCCTCTG 12359
                ************************************************************

CP+H1B-G        CAGGGCAGCCTGTGGGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAGCTGGCAGGAGGG 12416
CP+H1C-A        CAGGGCAGCCTGTGGGAGAAGGGACAGCGGGTAAAAAGAGAAGGCAAGCTGGCAGGAGGG 12419
                ************************************************************

CP+H1B-G        TGGCACTTCGTGGATGACCTCCTTAGAAAAGACTGACCTTGATGTCTTGAGAGCGCTGGC 12476
CP+H1C-A        TGGCACTTCGTGGATGACCTCCTTAGAAAAGACTGACCTTGATGTCTTGAGAGCGCTGGC 12479
                ************************************************************

CP+H1B-G        CTCTTCCTCCCTCCCTGCAGGGTAGGGGGCCTGAGTTGAGGGGCTTCCCTCTGCTCCACA 12536
CP+H1C-A        CTCTTCCTCCCTCCCTGCAGGGTAGGGGGCCTGAGTTGAGGGGCTTCCCTCTGCTCCACA 12539
                ************************************************************

CP+H1B-G        GAAACCCTGTTTTATTGAGTTCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACTTTG 12596
CP+H1C-A        GAAACCCTGTTTTATTGAGTTCTGAAGGTTGGAACTGCTGCCATGATTTTGGCCACTTTG 12599
                ************************************************************

CP+H1B-G        CAGACCTGGGACTTTAGGGCTAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGA 12656
CP+H1C-A        CAGACCTGGGACTTTAGGGCTAACCAGTTCTCTTTGTAAGGACTTGTGCCTCTTGGGAGA 12659
                ************************************************************

CP+H1B-G        CGTCCACCCGTTTCCAAGCCTGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAG 12716
CP+H1C-A        CGTCCACCCGTTTCCAAGCCTGGGCCACTGGCATCTCTGGAGTGTGTGGGGGTCTGGGAG 12719
                ************************************************************

CP+H1B-G        GCAGGTCCCGAGCCCCCTGTCCTTCCCACGGCCACTGCAGTCACCCCGTCTGCGCCGCTG 12776
CP+H1C-A        GCAGGTCCCGAGCCCCCTGTCCTTCCCACGGCCACTGCAGTCACCCCGTCTGCGCCGCTG 12779
                ************************************************************
```

```
CP+H1B-G        TGCTGTTGTCTGCCGTGAGAGCCCAATCACTGCCTATACCCCTCATCACACGTCACAATG 12836
CP+H1C-A        TGCTGTTGTCTGCCGTGAGAGCCCAATCACTGCCTATACCCCTCATCACACGTCACAATG 12839
                ************************************************************

CP+H1B-G        TCCCGAATTCCCAGCCTCACCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGT 12896
CP+H1C-A        TCCCGAATTCCCAGCCTCACCACCCCTTCTCAGTAATGACCCTGGTTGGTTGCAGGAGGT 12899
                ************************************************************

CP+H1B-G        ACCTACTCCATACTGAGGGTGAAATTAAGGGAAGGCAAAGTCCAGGCACAAGAGTGGGAC 12956
CP+H1C-A        ACCTACTCCATACTGAGGGTGAAATTAAGGGAAGGCAAAGTCCAGGCACAAGAGTGGGAC 12959
                ************************************************************

CP+H1B-G        CCCAGCCTCTCACTCTCAGTTCCACTCATCCAACTGGGACCCTCACCACGAATCTCATGA 13016
CP+H1C-A        CCCAGCCTCTCACTCTCAGTTCCACTCATCCAACTGGGACCCTCACCACGAATCTCATGA 13019
                ************************************************************

CP+H1B-G        TCTGATTCGGTTCCCTGTCTCCTTCTCCCGTCACAGATGTGAGCCAGGGCACTGCTCAGC 13076
CP+H1C-A        TCTGATTCGGTTCCCTGTCTCCTCCTCCCGTCACAGATGTGAGCCAGGGCACTGCTCAGC 13079
                ********************** ***********************************

CP+H1B-G        TGTGACCCTAGGTGTTTCTGCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCC 13136
CP+H1C-A        TGTGACCCTAGGTGTTTCTGCCTTGTTGACATGGAGAGAGCCCTTTCCCCTGAGAAGGCC 13139
                ************************************************************

CP+H1B-G        TGGCCCCTTCCTGTGCTGAGCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGG 13196
CP+H1C-A        TGGCCCCTTCCTGTGCTGAGCCCACAGCAGCAGGCTGGGTGTCTTGGTTGTCAGTGGTGG 13199
                ************************************************************

CP+H1B-G        CACCAGGATGGAAGGGCAAGGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTG 13256
CP+H1C-A        CACCAGGATGGAAGGGCAAGGCACCCAGGGCAGGCCCACAGTCCCGCTGTCCCCCACTTG 13259
                ************************************************************

CP+H1B-G        CACCCTAGCTTGTAGCTGCCAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATG 13316
CP+H1C-A        CACCCTAGCTTGTAGCTGCCAACCTCCCAGACAGCCCAGCCCGCTGCTCAGCTCCACATG 13319
                ************************************************************

CP+H1B-G        CATAGTATCAGCCCTCCACACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTT 13376
CP+H1C-A        CATAGTATCAGCCCTCCACACCCGACAAAGGGGAACACACCCCCTTGGAAATGGTTCTTT 13379
                ************************************************************

CP+H1B-G        TCCCCCAGTCCCAGCTGGAAGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACA 13436
CP+H1C-A        TCCCCCAGTCCCAGCTGGAAGCCATGCTGTCTGTTCTGCTGGAGCAGCTGAACATATACA 13439
                ************************************************************

CP+H1B-G        TAGATGTTGCCCTGCCCTCCCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTT 13496
CP+H1C-A        TAGATGTTGCCCTGCCCTCCCCATCTGCACCCTGTTGAGTTGTAGTTGGATTTGTCTGTT 13499
                ************************************************************

CP+H1B-G        TATGCTTGGATTCACCAGAGTGACTATGATAGTGAAAAGAAAAAAAAAAAAAAA---GGA 13553
CP+H1C-A        TATGCTTGGATTCACCAGAGTGACTATGATAGTGAAAAGAAAAAAAAAAAAAAAAAGGA 13559
                ********************************************************   ***

CP+H1B-G        CGCATGTATCTTGAAATGCTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTC 13613
CP+H1C-A        CGCATGTATCTTGAAATGCTTGTAAAGAGGTTTCTAACCCACCCTCACGAGGTGTCTCTC 13619
                ************************************************************

CP+H1B-G        ACCCCCACACTGGGACTCGTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTT 13673
CP+H1C-A        ACCCCCACACTGGGACTCGTGTGGCCTGTGTGGTGCCACCCTGCTGGGGCCTCCCAAGTT 13679
                ************************************************************

CP+H1B-G        TTGAAAGGCTTTCCTCAGCACCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAG 13733
CP+H1C-A        TTGAAAGGCTTTCCTCAGCACCTGGGACCCAACAGAGACCAGCTTCTAGCAGCTAAGGAG 13739
                ************************************************************

CP+H1B-G        GCCGTTCAGCTGTGACGAAGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCC 13793
CP+H1C-A        GCCGTTCAGCTGTGACGAAGGCCTGAAGCACAGGATTAGGACTGAAGCGATGATGTCCCC 13799
                ************************************************************

CP+H1B-G        TTCCCTACTTCCCCTTGGGGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTC 13853
CP+H1C-A        TTCCCTACTTCCCCTTGGGGCTCCCTGTGTCAGGGCACAGACTAGGTCTTGTGGCTGGTC 13859
                ************************************************************
```

```
CP+H1B-G        TGGCTTGCGGCGCGAGGATGGTTCTCTCTGGTCATAGCCCGAAGTCTCATGGCAGTCCCA 13913
CP+H1C-A        TGGCTTGCGGCGCGAGGATGGTTCTCTCTGGTCATAGCCCGAAGTCTCATGGCAGTCCCA 13919
                ************************************************************

CP+H1B-G        AAGGAGGCTTACAACTCCTGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATC 13973
CP+H1C-A        AAGGAGGCTTACAACTCCTGCATCACAAGAAAAAGGAAGCCACTGCCAGCTGGGGGGATC 13979
                ************************************************************

CP+H1B-G        TGCAGCTCCCAGAAGCTCCGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGC 14033
CP+H1C-A        TGCAGCTCCCAGAAGCTCCGTGAGCCTCAGCCACCCCTCAGACTGGGTTCCTCTCCAAGC 14039
                ************************************************************

CP+H1B-G        TCGCCCTCTGGAGGGGCAGCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATT 14093
CP+H1C-A        TCGCCCTCTGGAGGGGCAGCGCAGCCTCCCACCAAGGGCCCTGCGACCACAGCAGGGATT 14099
                ************************************************************

CP+H1B-G        GGGATGAATTGCCTGTCCTGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGA 14153
CP+H1C-A        GGGATGAATTGCCTGTCCTGGATCTGCTCTAGAGGCCCAAGCTGCCTGCCTGAGGAAGGA 14159
                ************************************************************

CP+H1B-G        TGACTTGACAAGTCAGGAGACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCT 14213
CP+H1C-A        TGACTTGACAAGTCAGGAGACACTGTTCCCAAAGCCTTGACCAGAGCACCTCAGCCCGCT 14219
                ************************************************************

CP+H1B-G        GACCTTGCACAAACTCCATCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATT 14273
CP+H1C-A        GACCTTGCACAAACTCCATCTGCTGCCATGAGAAAAGGGAAGCCGCCTTTGCAAAACATT 14279
                ************************************************************

CP+H1B-G        GCTGCCTAAAGAAACTCAGCAGCCTCAGGCCCAATTCTGCCACTTCTGGTTTGGGTACAG 14333
CP+H1C-A        GCTGCCTAAAGAAACTCAGCAGCCTCAGGCCCAATTCTGCCACTTCTGGTTTGGGTACAG 14339
                ************************************************************

CP+H1B-G        TTAAAGGCAACCCTGAGGGACTTGGCAGTAGAAATCCAGGGCCTCCCCTGGGGCTGGCAG 14393
CP+H1C-A        TTAAAGGCAACCCTGAGGGACTTGGCAGTAGAAATCCAGGGCCTCCCCTGGGGCTGGCAG 14399
                ************************************************************

CP+H1B-G        CTTCGTGTGCAGCTAGAGCTTTACCTAAAAGGAAGTCTCTGGGCCCAGAACTCTCCACCA 14453
CP+H1C-A        CTTCGTGTGCAGCTAGAGCTTTACCTGAAAGGAAGTCTCTGGGCCCAGAACTCTCCACCA 14459
                ************************* ********************************

CP+H1B-G        AGAGCCTCCCTGCCGTTCGCTGAGTCCCAGCAATTCTCCTAAGTTGAAGGGATCTGAGAA 14513
CP+H1C-A        AGAGCCTCCCTGCCGTTCGCTGAGTCCCAGCAATTCTCCTAAGTTGAAGGGACCTGAGAA 14519
                ************************************************** *******

CP+H1B-G        GGAGAAGGAAATGTGGGGTAGATTTGGTGGTGGTTAGAGATATGCCCCCCTCATTACTGC 14573
CP+H1C-A        GGAGAAGGAAGTGTGGGGTAGATTTGGTGGTGGTTAGAGATATGCCCCCCTCATTACTGC 14579
                ********** ************************************************

CP+H1B-G        CAACAGTTTCGGCTGCATTTCTTCACGCACCTCGGTTCCTCTTCCTGAAGTTCTTGTGCC 14633
CP+H1C-A        CAACAGTTTCGGCTGCATTTCTTCACGCACCTCGGTTCCTCTTCCTGAAGTTCTTGTGCC 14639
                ************************************************************

CP+H1B-G        CTGCTCTTCAGCACCATGGGCCTTCTTATACGGAAGGCTCTGGGATCTCCCCCTTGTGGG 14693
CP+H1C-A        CTGCTCTTCAGCACCATGGGCCTTCTTATACGGAAGGCTCTGGGATCTCCCCCTTGTGGG 14699
                ************************************************************

CP+H1B-G        GCAGGCTCTTGGGGCCAGCCTAAGATCATGGTTTAGGGTGATCAGTGCTGGCAGATAAAT 14753
CP+H1C-A        GCAGGCTCTTGGGGCCAGCCTAAGATCATGGTTTAGGGTGATCAGTGCTGGCAGATAAAT 14759
                ************************************************************

CP+H1B-G        TGAAAAGGCACGCTGGCTTGTGATCTTAAATGAGGACAATCCCCCCAGGGCTGGGCACTC 14813
CP+H1C-A        TGAAAAGGCACGCTGGCTTGTGATCTTAAATGAGGACAATCCCCCCAGGGCTGGGCACTC 14819
                ************************************************************

CP+H1B-G        CTCCCCTCCCCTCACTTCTCCCACCTGCAGAGCCAGTGTCCTTGGGTGGGCTAGATAGGA 14873
CP+H1C-A        CTCCCCTCCCCTCACTTCTCCCACCTGCAGAGCCAGTGTCCTTGGGTGGGCTAGATAGGA 14879
                ************************************************************

CP+H1B-G        TATACTGTATGCCGGCTCCTTCAAGCTGCTGACTCACTTTATCAATAGTTCCATTTAAAT 14933
CP+H1C-A        TATACTGTATGCCGGCTCCTTCAAGCTGCTGACTCACTTTATCAATAGTTCCATTTAAAT 14939
                ************************************************************
```

```
CP+H1B-G        TGACTTCAGTGGTGAGACTGTATCCTGTTTGCTATTGCTTGTTGTGCTATGGGGGGAGGG 14993
CP+H1C-A        TGACTTCAGTGGTGAGACTGTATCCTGTTTGCTATTGCTTGTTGTGCTATGGGGGGAGGG 14999
                ************************************************************

CP+H1B-G        GGGAGGAATGTGTAAGATAGTTAACATGGGCAAAGGGAGATCTTGGGGTGCAGCACTTAA 15053
CP+H1C-A        GGGAGGAATGTGTAAGATAGTTAACATGGGCAAAGGGAGATCTTGGGGTGCAGCACTTAA 15059
                ************************************************************

CP+H1B-G        ACTGCCTCGTAACCCTTTTCATGATTTCAACCACATTTGCTAGAGGGAGGGAGCAGCCAC 15113
CP+H1C-A        ACTGCCTCGTAACCCTTTTCATGATTTCAACCACATTTGCTAGAGGGAGGGAGCAGCCAC 15119
                ************************************************************

CP+H1B-G        GGAGTTAGAGGCCCTTGGGGTTTCTCTTTTCCACTGACAGGCTTTCCCAGGCAGCTGGCT 15173
CP+H1C-A        GGAGTTAGAGGCCCTTGGGGTTTCTCTTTTCCACTGACAGGCTTTCCCAGGCAGCTGGCT 15179
                ************************************************************

CP+H1B-G        AGTTCATTCCCTCCCCAGCCAGGTGCAGGCGTAGGAATATGGACATCTGGTTGCTTTGGC 15233
CP+H1C-A        AGTTCATTCCCTCCCCAGCCAGGTGCAGGCGTAGGAATATGGACATCTGGTTGCTTTGGC 15239
                ************************************************************

CP+H1B-G        CTGCTGCCCTCTTTCAGGGGTCCTAAGCCCACAATCATGCCTCCCTAAGACCTTGGCATC 15293
CP+H1C-A        CTGCTGCCCTCTTTCAGGGGTCCTAAGCCCACAATCATGCCTCCCTAAGACCTTGGCATC 15299
                ************************************************************

CP+H1B-G        CTTCCCTCTAAGCCGTTGGCACCTCTGTGCCACCTCTCACACTGGCTCCAGACACACAGC 15353
CP+H1C-A        CTTCCCTCTAAGCCGTTGGCACCTCTGTGCCACCTCTCACACTGGCTCCAGACACACAGC 15359
                ************************************************************

CP+H1B-G        CTGTGCTTTTGGAGCTGAGATCACTCGCTTCACCCTCCTCATCTTTGTTCTCCAAGTAAA 15413
CP+H1C-A        CTGTGCTTTTGGAGCTGAGATCACTCGCTTCACCCTCCTCATCTTTGTTCTCCAAGTAAA 15419
                ************************************************************

CP+H1B-G        GCCACGAGGTCGGGGCGAGGGCAGAGGTGATCACCTGCGTGTCCCATCTACAGACCTGCA 15473
CP+H1C-A        GCCACGAGGTCGGGGCGAGGGCAGAGGTGATCACCTGCGTGTCCCATCTACAGACCTGCA 15479
                ************************************************************

CP+H1B-G        GCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAAAGGGTTACCCTGGGCACTGGCCT 15533
CP+H1C-A        GCTTCATAAAACTTCTGATTTCTCTTCAGCTTTGAAAAGGGTTACCCTGGGCACTGGCCT 15539
                ************************************************************

CP+H1B-G        AGAGCCTCACCTCCTAATAGACTTAGCCCCATGAGTTTGCCATGTTGAGCAGGACTATTT 15593
CP+H1C-A        AGAGCCTCACCTCCTAATAGACTTAGCCCCATGAGTTTGCCATGTTGAGCAGGACTATTT 15599
                ************************************************************

CP+H1B-G        CTGGCACTTGCAAGTCCCATGATTTCTTCGGTAATTCTGAGGGTGGGGGGAGGGACATGA 15653
CP+H1C-A        CTGGCACTTGCAAGTCCCATGATTTCTTCGGTAATTCTGAGGGTGGGGGGAGGGACATGA 15659
                ************************************************************

CP+H1B-G        AATCATCTTAGCTTAGCTTTCTGTCTGTGAATGTCTATATAGTGTATTGTGTGTTTTAAC 15713
CP+H1C-A        AATCATCTTAGCTTAGCTTTCTGTCTGTGAATGTCTATATAGTGTATTGTGTGTTTTAAC 15719
                ************************************************************

CP+H1B-G        AAATGATTTACACTGACTGTTGCTGTAAAAGTGAATTTGGAAATAAAGTTATTACTCTGA 15773
CP+H1C-A        AAATGATTTACACTGACTGTTGCTGTAAAAGTGAATTTGGAAATAAAGTTATTACTCTGA 15779
                ************************************************************

CP+H1B-G        TTAAATAAGGTCTCCATTCATGGATTCCAAGGACAAGAAAGTCATATAGAATGTCTATTT 15833
CP+H1C-A        TTAAATAAGGTCTCCATTCATGGATTCCAAGGACAAGAAAGTCATATAGAATGTCTATTT 15839
                ************************************************************

CP+H1B-G        TTTAAGTTCTTTCCCACGCACCCTTAGATAATTTAGCTCAGAACAGGAAATGATAGTATT 15893
CP+H1C-A        TTTAAGTTCTTTCCCACGCACCCTTAGATAATTTAGCTCAGAACAGGAAATGATAGTATT 15899
                ************************************************************

CP+H1B-G        AATAAAAGCTGGACATCAGGATTAACAGCTCTCTCTGGGGCCCTGAAGGTGAGAGTTCTC 15953
CP+H1C-A        AATAAAAGCTGGACATCAGGATTAACAGCTCTCTCTGGGGCCCTGAAGGTGAGAGTTCTC 15959
                ************************************************************

CP+H1B-G        AGACTTGCTCATTTGCAGTTGCTTCTTTGTGATGCTGGCTACCCAGCTTTCTTGTACAAA 16013
CP+H1C-A        AGACTTGCTCATTTGCAGTTGCTTCTTTGTGATGCTGGCTACCCAGCTTTCTTGTACAAA 16019
                ************************************************************
```

```
CP+H1B-G        GTGGT 16018
CP+H1C-A        GTGGT 16024
                *****
```

Appendix K

UCSC annotations of the CP and SD regions

Both regions are enriched for transcription factor binding and exhibit open chromatin structures and DNase I hypersensitivity.