Analysis of the relationship between gene structure, coding ability and nonsense-mediated decay in mammals

David Anderson de Lima Morais

Department of Biology, McGill University
Montreal, Quebec, Canada
March 2010

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

ABSTRACT

Non-coding mRNAs have been, until recently, regarded as functionless products of junk DNA. However, large-scale genomic studies have enabled us to unveil complex pathways that depend on non-coding mRNAs. In this thesis, I developed a pipeline to perform large scale analysis of non-coding sequences in mammals. In Chapter II, we gathered evidences of a non-random population of *pseudogenic duplicated exons* (ΨEs, i.e., exons disabled by frameshifts and premature stop codons) in four mammalian genomes: human, mouse, rat and cow. I observed a consistent population of ΨEs, associated with 0.4–1.0% of genes. These ΨE populations exhibit codon substitution patterns that are typical of an endemic population of decaying sequences. Also, ΨEs are more often associated with functional categories such as 'ion binding' and 'nucleic-acid binding' than duplicated exons in general. We also found that ΨEs can participate in alternative splicing events and are not randomly distributed within the gene structure. *Pseudogenic exons* may function in gene regulation through generation of transcribed pseudogenes, or regulatory alternative transcripts. To further investigate the role of non-coding mRNA, we mapped more than 16 millions EST/mRNAs to genomic sequences in order to identify alternative splice forms (AS) that can be target for mRNA nonsense-mediated decay (NMD) in the same four mammalian species (Chapter IV). We found that at least 10% of the mammalian genes have an alternative splice form targeted for NMD (AS-NMD candidate). More than 25% of the genes with an AS-NMD candidate in mouse, rat and cow also have an ortholog in human that is target for NMD. This highly significant trend

clearly suggests that these AS-NMD candidates have a regulatory conserved function across these species. The AS-NMD candidates also showed a similar pattern of gene ontology enrichment in all four species. Furthermore, we mapped the AS-NMD candidates to mass spectrometry-derived proteomics data. We found evidence of translation in at least 10% of the AS-NMD candidates. This AS-NMD candidate analysis has provided large-scale evidence for a conserved mechanism of autoregulation dependent on alternative splicing coupled with NMD, across four mammals.

RÉSUMÉ

Les ARNms non-codants ont été, depuis récemment, considéré comme des produits non-fonctionnels de l'ADN génomique sans fonction codante (DNA junk). Cependant, des études génomiques à grande échelle nous ont permis de dévoiler des sentiers (chemins) complexes qui dépendent de séquences d'ARNm non-codant. Dans cette thèse, nous développons une méthodologie afin de produire des analyses à grande échelle de séquences non codantes chez les mammifères. Dans le Chapitre II, nous avons ramassé des preuves d'une population non aléatoire d'exons pseudogéniques dupliqués (ΨEs, i.e., exons invalidés par des décalages de trame (frameshifts) et des codons d'arrêt prématurés) dans quatre génomes mammaliens: humain, souris, rat et vache. Nous avons observés une population consistante de ΨEs associée avec 0.4-1.0% des gènes. Ces populations ΨE présentent des modèles de substitution de codons qui sont typiques d'une population endémique de séquences en dégénérescence. De plus, les ΨEs sont plus souvent associés avec des catégories fonctionnelles telles que des *liaisons ioniques* et des *liaisons d'acides nucléiques* que des exons dupliqués en général. Nous avons également constaté que les ΨEs peuvent participer à des événements alternatifs d'épissage et ne sont pas distribués aléatoirement dans la structure du gène. Les exons pseudogéniques peuvent fonctionner dans la régulation des gènes à travers la génération de pseudogènes transcrits, ou de transcrits alternatifs régulateurs. Afin d'investiguer davantage le rôle d'ARNm non-codant, nous avons cartographié plus de 16 millions de EST/mRNAs a des séquences

génomiques afin d'identifier des formes alternatives d'épissure ou alternative splice forms (AS) qui peuvent être la cible pour l'ARNm non-sens dégradé ou mRNA nonsense-mediated decay (NMD) dans les mêmes quatre espèces mammaliennes (Chapitre IV). Nous avons découvert qu'au moins 10% des gènes mammaliens ont une forme alternative d'épissure ciblé pour NMD (candidat AS-NMD). Plus de 25 % des gènes avec le candidat AS-NMD dans la souris, le rat et la vache ont aussi un orthologue dans l'humain qui est ciblé pour NMD. Cette tendance hautement significative suggère clairement que ces candidats AS-NMD ont une fonction régulatrice conservée à travers ces espèces. Ces candidats AS-NMD démontrent aussi un modèle similaire d'enrichissement de gène ontologique dans toutes les quatre espèces. Par ailleurs, nous avons cartographié les candidats AS-NMD à des données protéomiques de spectrométrique dérivé en série. Nous avons trouvé une preuve de translation dans au moins 10% des candidats AS-NMD. Cette analyse de candidat NMD a fournit une preuve à grande échelle pour un mçanisme conservé d'autorégulation dépendante sur l'épissage coulé alternative avec NMD, à travers quatre mammifères.

ACKNOWLEDGMENTS

Firstly, I would like to convey my deepest thanks to my supervisor Dr. Paul M. Harrison for having provided the necessary funding for my studies at McGill despite the 'financial load' associated with an international student. I also thank him for his support and encouragement throughout these four years. His passion for pseudogenes and genome evolution was really inspirational. I have always admired the way he conducted the laboratory, giving lots of freedom to students and providing the ideal environment to become independent. Thanks Paul, I will always follow your scientific standards throughout my career.

I would like to thank NSERC (Natural Sciences and Engineering Research Council of Canada) and McGill University for financially support my research.

I want to thank my supervisory committee and qualify members, Dr. Thomas Bureau, Dr. Daniel Schoen, Dr. Jacek Majewski, Dr. Brian McGill for their support and guidance throughout my studies.

I also would like to thank my lab mates Benoit V., Zhang Y., Sergey M., Deena G., Amit K, Djamel H, Manish K, for their valuable suggestions that helped my research much better.

I want to thank the McGill staff for their help and friendship, in special to Mrs. Anne-Marie Sdicu, Mrs. Becky Dolan, Mrs. Susan Bocti and Mrs. Ancil Gittens. Many thanks for Scott Bunnel for his help with the McGill cluster and his patience to answer my many question and requests.

Thanks to Michel Barrette and his effort to make me have an account in mammoth cluster. Without you the second part of my work would have been nearly impossible.

A special thank to Mr. Aristóteles José de Souza Silva (Tote) who helped me in Brazil and made possible my travel to Canada.

Thanks to my Canadians and Brazilians friend, Tong, Andrej, Josh, Anna P., Foca, Patriota, Hugo M, Auguinho, Amaro C., Henrique C., Bruno A., Poliana, Barraca, Lilica, Célia C.

I would like thank my Canadian family Jeanine, Pierre, Cathy, Patricia, Jean-Pierre D., Claude, my beloved 2 years old goddaughter Nora, my nieces and nephew, Naomi, Melina and Nathaniel. Thanks to all Lemire and St-Yve family for you love and support.

I want to thank my Brazilian family, my grandma Dida for her passion for life, my sister Danielle (and now her baby girl!) and my aunts and uncles for their support and love.

I am infinitely grateful to my wonderful wife Danielle Lemire, who accepted to live in this big city of Montreal despite her preference for the Quebec county side. You

have been my strength and my hope during the difficult times, and my joy throughout the years. I would not have been able to accomplish all of this without you. Thank you so much.

Thank to my father, a man who taught me that nothing is really impossible. Thanks for you love, support and encouragement. Thank to my mother Rosa, who was my biggest supporter to come to Canada to pursue my degree, despite her broken heart to know that I would be distant. Unfortunately it is you, now, who is distant, but I will always feel you presence with me.

This thesis is dedicated to you, mom, with all my love.

Rosa Maria de Lima Morais

In memoriam

TABLE OF CONTENT

LIST OF FIGURES AND TABLES

PREFACE

Thesis format and Style

This thesis is in manuscript-based format and consists of a set of two papers presented in accordance with the "Guidelines for thesis preparation" from the Faculty of Graduate Studies and Research. It contains an Introduction (Chapter I), which provides a relevant literature review as well as the rationale and objective of the study performed in the Chapter II. The Chapter III includes a background and objectives of the research in Chapter IV, and a statement linking the Chapter II and IV. The last chapter (V) presents the general conclusion. The research chapters (II and IV) are formatted as research manuscripts, containing the following sections: Abstract, Introduction, Materials and Methods, Results and Discussion, References, Figures and Tables.

This thesis is partially based on two manuscripts. Chapter II consists of reformatted version of the published manuscript, while chapter IV contains unpublished results.

Contribution of Co-Authors

Chapter II

This Chapter is a reformatted version of:

Morais DD, Harrison PM: Genomic evidence for non-random endemic populations of decaying exons from mammalian genes. *BMC genomics* 2009, 10:309.

Dr. Harrison and I designed and I performed all simulations presented in this paper. I wrote the manuscript and Dr. Harrison edited the late draft.

Chapter IV

This Chapter is a reformatted version of:

Morais DAL, Harrison PM: Large-scale evidence for a conserved functional role for NMD candidates across mammals. *PLoS ONE* 2010 5(7): e11695. doi:10.1371.

I designed and executed all the simulation presented in this paper. I wrote the manuscript and Dr. Harrison edited the late draft.

Original Contributions to Knowledge

Each manuscript was prepared for publication in peer-reviewed journals focusing on genomic analysis and bioinformatics. The research was conducted under the supervision of Dr. Paul M. Harrison. I collected the data, conducted the experiments, analyzed the results, executed the statistical testes and wrote all the manuscripts.

Chapter II

Exon duplication fate has been vastly discussed. However, the focus of these discussions has always been put on whether this phenomenon leads to neofunctionalization or to subfunctionalization. In our study we focus on exon duplication followed by pseudogenization (i.e. loss of protein coding ability by acquisition of disruptive mutations). We found a population of decaying duplicated exons arising for up ~1% of the genes in four mammalian genomes. We verified that this population of *pseudogenic duplicated* exons is distributed non-randomly in mammals. We also found evidence of transcription of those *pseudogenic exons* indicating that they might participate in the regulation of homologous genes.

Chapter IV

We analyzed alternative splice forms that target the mRNA for nonsense mediated decay in four mammalian genomes. We showed that at least 10% of all known genes possess an alternatively spliced mRNA target for NMD. We also found that intron retention plays an important role in NMD-targeting and that these retained introns possess features that distinguish them from other non-retained introns. We verified that NMD is associated with some classes of functional categories and that they are very often associated with cancer and other genetics disorders. Additionally we found evidences of a conserved functional mechanism of gene regulation.

CHAPTER I

LITERATURE REVIEW

Genome duplication and diversification

The importance of gene duplication as a major player in genome diversification was outlined a long time ago in a study made by Susumu Ohno (OHNO *et al.* 1968). Ever since then, several other studies have pointed out that gene duplication has indeed been rampant (LYNCH and CONERY 2000; LYNCH and CONERY 2003). Many eukaryotic organisms had their whole genome duplicated, sometimes more than once (SEOIGHE and GEHRING 2004; WOLFE and SHIELDS 1997).

Four primary mechanisms of gene duplication have been described: whole genome duplication (WGD) (VAN DE PEER *et al.* 2003), tandem duplication (SHOJA and ZHANG 2006), segmental transposition (BAILEY *et al.* 2004) and retrotransposition (YU *et al.* 2007). These four mechanisms can be divided into two classes: the large scale duplication, and the small scale duplication. This division has a major impact on the fate of the duplicated genes and how they must be analyzed (HAKES *et al.* 2007).

The longevity of duplicated genes seems to correlate positively with the scale of the duplication. For example, duplicated copies from WGD persist for a longer period and evade complete degradation at a higher frequency than those generated by segmental duplication (HAHN *et al.* 2007; NADEAU and SANKOFF 1997). Additionally, certain types of genes are more likely than others to be retained within the genome following duplication, such as: genes that are present

in many evolutionarily divergent lineages, those that are functionally constrained, genes involved in environmental responses and highly expressed genes (DAVIS and PETROV 2004).

Another reason for the persistence of duplicates is that genetic modification can occur in one or both paralogs after duplication. For instance, duplication can permit each copy of a multifunctional protein to specialize on a subset of ancestral activities, thereby reducing pleiotropy (LYNCH and CONERY 2000). Duplicates might also be preserved if each paralog degrades in a complementary fashion, or if one or both paralogs acquire novel function (HE and ZHANG 2005). The neofunctionalization model implies that one of the gene copies carries out the ancestral function, while the other copy evolves neutrally until, by chance, it acquires beneficial mutation(s) during early stages of its evolution. Once a new function is achieved, purifying selection is expected to take over the later stages of evolution. Neofunctionalization could occur with complete loss, partial degradation or retention of ancestral function (HE and ZHANG 2005). Subfunctionalization, on the other hand, implies that each paralog evolves in a complementary fashion, so that the act of both is necessary to accomplish the full suite of the ancestral activities. This model is also known as the duplication-degeneration-complementation model (LYNCH and CONERY 2000). Subfunctionalization could occur at the expression level through divergence of paralogous expression profiles in spatial, temporal or quantitative dimensions. It

also could occur through divergence of functional protein domains, or as a consequence of differential cellular localizations that could facilitate or catalyze the evolution of unique function in paralogs (LYNCH and CONERY 2000; LYNCH and FORCE 2000).

If a gene is to be fixated within the genome, the modification that enables the fixation must occur shortly after the duplication, or else one copy would likely become disabled (LYNCH and CONERY 2000). The majority of the duplicated genes appear to be transient, contributing little to long-term phenotypic evolution, and are left behind in the genome as 'genetic fossils', known as pseudogenes (HARRISON *et al.* 2002; LYNCH and CONERY 2003).

A problem of bias

Mutational variation is not generally biased toward any adaptation. However, this does not imply that mutational variation is random in other respects. The possible existence of a bias in spontaneous mutation, and the importance of assessing these biases quantitatively, have been recognized for a long time (LI *et al.* 1984). However, empirical studies are difficult, since spontaneous mutations are too rare to be investigated and studies are almost inevitably subject to experimental bias according to the methods by which mutation are detected (PETROV and HARTL 2000).

One solution found to the bias problem was to investigate patterns of spontaneous mutation from analyses of the observed pattern of nucleotide substitution in functional genes. The main problem with this approach is that functional genes are affected not only by the relative rates at which different mutations occur spontaneously, but also – and predominantly – by natural selection. Furthermore, some mutations can be more deleterious than others and therefore they will be underrepresented in any observed sample. This problem is more accentuated in exons of protein-coding genes, where insertions and deletions (indels) are less frequent than point mutations (PETROV and HARTL 2000).

An alternative to dealing with the biases inherent in studying functional genes is to study their nonfunctional counterpart, i.e., pseudogenes (ΨGs). Since ΨGs are expected to evolve without functional constraints (ZHANG *et al.* 2002), it can be said that all types of mutations should have equal chance of fixation in them, and that the pattern of substitution in ΨGs should be congruent with the pattern of mutations (LI *et al.* 1984).

Pseudogenes ΨG

In the late 1970s, researchers were mapping the chromosomal location of several genes when they stumbled on DNA sequences that were similar to functional genes but contained 'defects' such as truncations and deleterious

mutations (JACQ *et al.* 1977). These lookalike genes were dubbed pseudogenes (ΨGs). For many years the term pseudogene was used to define 'sequences structurally similar to functional genes but containing important defects, which make them unable to produce functional proteins'. However, with the annotation of several ΨGs from non-protein coding genes, the definition had to evolve. Pseudogenes are now defined as defunct copies of genes that have lost their potential as DNA templates for functional products (ZHENG *et al.* 2005).

Pseudogenes are formed by two independent mechanisms that are believed to have different implications on genome evolution. The nonprocessed ΨGs arise usually after partial or complete segmental duplication of genes and subsequent loss of functions by mutation (ECHOLS *et al.* 2002). Nonprocessed ΨGs often retain the same intron-exon structure of the paralog from which they were duplicated (usually termed 'parental genes'). Processed ΨGs are decayed or disabled genes that show symptoms of retrotransposition, namely, lack of introns of the parental gene, and also 3' polyadenine tails, if formed more recently; short direct repeats flanking the sequence (for young retrotranspositions); frequent 5' truncation, and genomic location different from that of the parent gene (BETRAN *et al.* 2002; HARRISON and GERSTEIN 2002). It has been demonstrated experimentally and computationally that processed ΨGs can be formed through the action of LINE-1 reverse transcriptases (ESNAULT *et al.* 2000; PAVLICEK *et al.* 2002).

More than 75% of all annotated ΨGs are processed (ZHENG and GERSTEIN 2006). Nonprocessed ΨGs often accumulate mutations that change their sequences beyond the threshold of detection through sequence alignment. As listed above, processed ΨGs possess features that are more easily detectable (poly(A)-tail, lack of introns, presence of flanking direct repeats).

Processed ΨGs are more common in mammalian species, but are less abundant in other animal species. It is reported that the human genome has 8000-12000 pseudogenes, while the mouse genome has 5000 (ZHANG *et al.* 2004; ZHANG *et al.* 2003), the *Caenorhabditis elegans* genome contains only 208 (HARRISON *et al.* 2001), and Drosophila only 34 processed ΨGs (HARRISON *et al.* 2003). These differences in processed pseudogene abundance might be the result of differences in gametogenesis with species having longer lampbrush stages also having more processed pseudogenes (WEINER *et al.* 1986). In the *Arabidopsis thaliana* genome, 411 processed ΨGs were found. ΨGs have also been found in large numbers in prokaryotes such as *Mycobacterium leprae, Shigella flexneri* and *Salmonella typhi* (COLE *et al.* 2001; PARKHILL *et al.* 2001; WEI *et al.* 2003).

Are ΨGs Junk DNA?

The term 'junk DNA' was firstly used to define DNA tracts that are functionless at a given moment in evolutionary time (OHNO 1972). As much as

95% of haploid genomes in multicellular eukaryotes have been widely considered as junk. Although it seems to be easy to define junk DNA, it is only as easy as defining function, which is not easy (ZUCKERKANDL 2002).

It is generally thought that if the sequences are not recruited for what must obviously be a function, they will be rapidly changed or lost. However, rapid sequence changes do not imply an absence of function. Indeed, sequences can be lost whether functionless or not (ZUCKERKANDL 2002). The first ΨGs found in the late 1970s, had clear defects in their sequences, such as, lack of promoter, the presence of frameshifts and nonsense mutations or loss of splice sites (JACQ *et al.* 1977). Since function, in general, was regarded as linked to conserved nucleotide and amino acid residues, it seems easy to understand why these sequences were called functionless.

The ambiguity of nonfunctionality of ΨGs first appeared in a work of a group investigating the gene encoding nitric oxide synthase (NOS) in snails (KORNEEV *et al.* 1999). They showed that the Ψ-NOS was transcribed in the central nervous system of *Lymnaea stagnalis* and that it could form a RNA-RNA duplex with the mRNA from the NOS gene *in vivo*. The transcript of the Ψ-NOS is a natural antisense RNA that seems to have evolved through gene duplication and inversion. The consequence of the RNA duplex is the suppression of the expression of the NOS, which has a role in memory formation (KORNEEV *et al.* 2005). This seems to contradict the idea of ΨGs' nonfunctionality. Yet, Ψ-NOS

has premature stop codons in all three frames and unlike its parental gene cannot encode a protein.

Several recent studies have shown that most of the mammalian genome is transcribed and more than half of the transcribed regions are mapped outside of known genes (CARNINCI 2006; WILLINGHAM and GINGERAS 2006). In the light of these findings, it is not unanticipated that ΨGs can contribute to the complexity of the transcriptome. Several studies have found evidence of ΨG transcripts in different species (CHANG and SLIGHTOM 1984; KHACHANE and HARRISON 2009; ZHENG *et al.* 2005).

Is the ΨG transcription an evidence of its function?

With the amount of evidence for ΨG transcription growing, the problem of nonfunctionality prompts another question. Is the transcription of ΨGs an indication of a stochastic cellular process or is it an intrinsic biological function? The literature indicates that both occur (ZHENG and GERSTEIN 2007). Duplicated ΨGs can have some transcriptional activity, and thus potentially influence the expression of paralogous genes (ZHENG and GERSTEIN 2007). Such functionality could be eventually lost with the accumulation of further mutations. Processed ΨGs can also be transcribed if they land next to an active promoter (MARQUES *et*

al. 2005). Some of these retroposed gene copies presumably have functions different from those of their parental genes.

Apparently there are many ways in which a ΨG can express functionality. Some microRNAs seem to have evolved in a similar fashion to the Nitric Oxide synthase ΨGs (KORNEEV *et al.* 1999). For instance, the *Arabidopsis thaliana* MIR161 and MIR163 are duplicated inverted ΨGs that target their paralogous protein-coding genes (ALLEN *et al.* 2004). The gene host of some nucleolar RNAs (snoRNAs), such as human U19 and U22, has been proposed to be derived from protein-coding genes that lost their coding ability (BORTOLIN and KISS 1998; MOORE 1996). In some cases, pseudogenes may have acted as a DNA reservoir for increasing antibody diversity through gene conversion in vertebrates (BALAKIREV and AYALA 2003). Conversely, ΨGs can swap a piece of their DNA into a functional paralog and, therefore, cause diseases (BISCHOF *et al.* 2006).

Recently a pseudogene classification was proposed based on the functionality of the ΨGs: 1- Living genes; 2- Ghost pseudogenes (some intermediate functionality); 3- Dead pseudogenes. The ghost and dead ΨGs have subdivisions according to their functions and type of disablements (ZHENG and GERSTEIN 2007).

Pseudogenic exons (ΨEs)

The correspondence between exons and functional protein domains has suggested that many proteins evolved by modular assembly. Exon duplication has been widely observed within genes, and it is thought to be one of the main processes by which proteins acquire new domains. Tandemly duplicated exons are often retained if they acquire functionality, in a similar manner to whole gene duplications (LETUNIC *et al.* 2002).

It is also expected that the evolutionary constraints that apply to duplicated genes, are applied to duplicated exons. In other words, if a recently duplicated exon does not acquire functionality or bring advantages to its protein shortly after the duplication event, it will accumulate nonsense mutations and therefore become pseudogenized. The term *pseudogenic exon* (ΨE) is used to define exons that lost their functionality by acquisition of nonsense mutations, frame-shifts or loss of splicing signal (MORAIS and HARRISON 2009).

For example, the human gene of αA-crystallin is a single-copy gene that posseses a ΨE in early stages of decay. The protein of the αA-crystallin is the major structural component of the ocular lenses of all vertebrates and is highly conserved throughout evolution (GHAHGHAEI *et al.* 2009). The first intron of the human αA-crystallin gene is highly similar to an alternatively spliced exon in rodents. The ΨE is perhaps the result of a failed experiment in its evolution, or its

loss is related to adaptations for new requirements of the human eye (JAWORSKI and PIATIGORSKY 1989).

Often mutation in introns can trigger the inclusion of ΨEs in genes, causing these genes to produce truncated proteins that frequently lead to diseases. For instance, inherited growth-hormone insensitivity (GHI) is a heterogeneous disorder that is often caused by mutations in the coding exons or flanking intronic sequences of the growth-hormone receptor gene (GHR). The most severe form, Laron syndrome (MIM 262500), is characterized by severe growth retardation and dysmorphic facial features, associated with elevated circulating growth hormone (GH). Metherell *et al.* (2001) found a point mutation in children with GHI leading to activation of an intronic *pseudogenic exon* resulting in inclusion of an additional 108 nucleotides between exons 6 and 7 in the majority of GHR transcripts (METHERELL *et al.* 2001). A transversion within intron 6 of the FGG human gene is responsible for the inclusion of a ΨE 75 nucleotides long associated with the mechanisms leading to an autosomal recessive coaugulopathy called afibrinogenaemia (SPENA *et al.* 2007).

Treatments to correct point mutations and prevent ΨE inclusion are being developed. One of the most promising treatments uses an antisense oligonucleotide. Rodríguez-Pascau *et al* (2009), restored the normal splicing pattern of the gene NPC1 (without a ΨE located in the intron 9) using a

morpholino oligonucleotide targeted to the cryptic splice site and transfected into fibroblasts. These results support the efficacy of antisense therapeutics for the treatment of Niemann-Pick type C disease (RODRIGUEZ-PASCAU *et al.* 2009). This strategy has also been employed to restore the correct splicing in several diseases models such as cystic fibrosis (FRIEDMAN *et al.* 1999), ocular albinism type I (VETRINI *et al.* 2006), afibrinogenemia (DAVIS *et al.* 2009) and congenital disorder of glycosylation type IA (VEGA *et al.* 2009).

Linking statement I

Given the frequency of exon duplication events in mammalian genomes, pseudogenization is expected to happen for those exons which fail to bring evolutive advantages to their gene hosts. These duplicated *pseudogenic exons* (ΨEs) must remain in the genome as 'fossils' until they are degraded beyond the threshold of detection by alignment tools. For those recently disabled or under selective constraint, inclusion in alternative transcripts is an option, although, seldom advantageous. I developed a pipeline to annotate duplicated ΨEs in four mammalian genomes and compared the ΨEs with annotated duplicated coding exons. Furthermore, I gathered evidence of a non-random distribution of these ΨEs, such as in their positioning in the intron-exon structure of their host genes, association with specific protein families and domains, over-representation for certain functional categories, and non-random conservation across the four mammalian species. Additionally, I analyzed the codon substitution pattern of duplicated exons and duplicated ΨEs. Finally, I verified the participation of these ΨEs in alternative splice events, and the predicted consequences of this for nonsense-mediated decay targeting.

References

ALLEN, E., Z. XIE, A. M. GUSTAFSON, G. H. SUNG, J. W. SPATAFORA *et al.*, 2004 Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet **36:** 1282-1290.

BAILEY, J. A., D. M. CHURCH, M. VENTURA, M. ROCCHI and E. E. EICHLER, 2004 Analysis of segmental duplications and genome assembly in the mouse. Genome Res **14:** 789-801.

BALAKIREV, E. S., and F. J. AYALA, 2003 Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet **37:** 123-151.

BARBERAN-SOLER, S., and A. M. ZAHLER, 2008 Alternative splicing regulation during C. elegans development: splicing factors as regulated targets. PLoS Genet **4:** e1000001.

BASERGA, S. J., and E. J. BENZ, JR., 1988 Nonsense mutations in the human beta-globin gene affect mRNA metabolism. Proc Natl Acad Sci U S A **85:** 2056-2060.

BEN-DOV, C., B. HARTMANN, J. LUNDGREN and J. VALCARCEL, 2008 Genome-wide analysis of alternative pre-mRNA splicing. J Biol Chem **283:** 1229-1233.

BERGET, S. M., 1995 Exon recognition in vertebrate splicing. J Biol Chem **270:** 2411-2414.

BETRAN, E., K. THORNTON and M. LONG, 2002 Retroposed new genes out of the X in Drosophila. Genome Res **12:** 1854-1859.

BISCHOF, J. M., A. P. CHIANG, T. E. SCHEETZ, E. M. STONE, T. L. CASAVANT *et al.*, 2006 Genome-wide identification of pseudogenes capable of disease-causing gene conversion. Hum Mutat **27:** 545-552.

BLENCOWE, B. J., 2006 Alternative splicing: new insights from global analyses. Cell **126:** 37-47.

BOON, K. L., R. J. GRAINGER, P. EHSANI, J. D. BARRASS, T. AUCHYNNIKAVA *et al.*, 2007 prp8 mutations that cause human retinitis pigmentosa lead to a U5 snRNP maturation defect in yeast. Nat Struct Mol Biol **14:** 1077-1083.

BORTOLIN, M. L., and T. KISS, 1998 Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. Rna **4:** 445-454.

CARMEL, I., S. TAL, I. VIG and G. AST, 2004 Comparative analysis detects dependencies among the 5' splice-site positions. Rna **10:** 828-840.

CARNINCI, P., 2006 Tagging mammalian transcription complexity. Trends Genet **22:** 501-510.

CHANG, L. Y., and J. L. SLIGHTOM, 1984 Isolation and nucleotide sequence analysis of the beta-type globin pseudogene from human, gorilla and chimpanzee. J Mol Biol **180:** 767-784.

COLE, S. T., K. EIGLMEIER, J. PARKHILL, K. D. JAMES, N. R. THOMSON *et al.*, 2001 Massive gene decay in the leprosy bacillus. Nature **409:** 1007-1011.

DAVIS, J. C., and D. A. PETROV, 2004 Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol **2:** E55.

DAVIS, R. L., V. M. HOMER, P. M. GEORGE and S. O. BRENNAN, 2009 A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. Hum Mutat **30:** 221-227.

ECHOLS, N., P. HARRISON, S. BALASUBRAMANIAN, N. M. LUSCOMBE, P. BERTONE *et al.*, 2002 Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res **30:** 2515-2523.

ESNAULT, C., J. MAESTRE and T. HEIDMANN, 2000 Human LINE retrotransposons generate processed pseudogenes. Nat Genet **24:** 363-367.

FISCHER, D. C., K. NOACK, I. B. RUNNEBAUM, D. O. WATERMANN, D. G. KIEBACK *et al.*, 2004 Expression of splicing factors in human ovarian cancer. Oncol Rep **11:** 1085-1090.

FRIEDMAN, K. J., J. KOLE, J. A. COHN, M. R. KNOWLES, L. M. SILVERMAN *et al.*, 1999 Correction of aberrant splicing of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by antisense oligonucleotides. J Biol Chem **274:** 36193-36199.

GALLO, J. M., W. NOBLE and T. R. MARTIN, 2007 RNA and protein-dependent mechanisms in tauopathies: consequences for therapeutic strategies. Cell Mol Life Sci **64:** 1701-1714.

GHAHGHAEI, A., A. REKAS, J. A. CARVER and R. C. AUGUSTEYN, 2009 Structure/function studies of dogfish alpha-crystallin, comparison with bovine alpha-crystallin. Mol Vis **15:** 2411-2420.

GRAVELEY, B. R., 2000 Sorting out the complexity of SR protein functions. Rna **6:** 1197-1211.

HAHN, M. W., J. P. DEMUTH and S. G. HAN, 2007 Accelerated rate of gene gain and loss in primates. Genetics **177:** 1941-1949.

HAKES, L., J. W. PINNEY, S. C. LOVELL, S. G. OLIVER and D. L. ROBERTSON, 2007 All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol **8:** R209.

HANSEN, K. D., L. F. LAREAU, M. BLANCHETTE, R. E. GREEN, Q. MENG *et al.*, 2009 Genome-wide identification of alternative splice forms down-

regulated by nonsense-mediated mRNA decay in Drosophila. PLoS Genet **5:** e1000525.

HARRISON, P. M., N. ECHOLS and M. B. GERSTEIN, 2001 Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. Nucleic Acids Res **29:** 818-830.

HARRISON, P. M., and M. GERSTEIN, 2002 Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol **318:** 1155-1174.

HARRISON, P. M., H. HEGYI, S. BALASUBRAMANIAN, N. M. LUSCOMBE, P. BERTONE *et al.*, 2002 Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res **12:** 272-280.

HARRISON, P. M., D. MILBURN, Z. ZHANG, P. BERTONE and M. GERSTEIN, 2003 Identification of pseudogenes in the Drosophila melanogaster genome. Nucleic Acids Res **31:** 1033-1037.

HE, X., and J. ZHANG, 2005 Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169:** 1157-1164.

HE, Y., and R. SMITH, 2009 Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B. Cell Mol Life Sci **66:** 1239-1256.

HOLBROOK, J. A., G. NEU-YILIK, M. W. HENTZE and A. E. KULOZIK, 2004 Nonsense-mediated decay approaches the clinic. Nat Genet **36:** 801-808.

JACQ, C., J. R. MILLER and G. G. BROWNLEE, 1977 A pseudogene structure in 5S DNA of Xenopus laevis. Cell **12:** 109-120.

JAWORSKI, C. J., and J. PIATIGORSKY, 1989 A pseudo-exon in the functional human alpha A-crystallin gene. Nature **337:** 752-754.

JENSEN, C. J., B. J. OLDFIELD and J. P. RUBIO, 2009 Splicing, cis genetic variation and disease. Biochem Soc Trans **37:** 1311-1315.

JOHNSON, J. M., J. CASTLE, P. GARRETT-ENGELE, Z. KAN, P. M. LOERCH *et al.*, 2003 Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302:** 2141-2144.

KASHIMA, I., A. YAMASHITA, N. IZUMI, N. KATAOKA, R. MORISHITA *et al.*, 2006 Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. Genes Dev **20:** 355-367.

KERR, T. P., C. A. SEWRY, S. A. ROBB and R. G. ROBERTS, 2001 Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? Hum Genet **109:** 402-407.

KHACHANE, A. N., and P. M. HARRISON, 2009 Assessing the genomic evidence for conserved transcribed pseudogenes under selection. BMC Genomics **10:** 435.

KORKKO, J., L. ALA-KOKKO, A. DE PAEPE, L. NUYTINCK, J. EARLEY *et al.*, 1998 Analysis of the COL1A1 and COL1A2 genes by PCR amplification and scanning by conformation-sensitive gel electrophoresis identifies only COL1A1 mutations in 15 patients with osteogenesis imperfecta type I: identification of common sequences of null-allele mutations. Am J Hum Genet **62:** 98-110.

KORNEEV, S. A., J. H. PARK and M. O'SHEA, 1999 Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. J Neurosci **19:** 7711-7720.

KORNEEV, S. A., V. STRAUB, I. KEMENES, E. I. KORNEEVA, S. R. OTT *et al.*, 2005 Timed and targeted differential regulation of nitric oxide synthase (NOS) and anti-NOS genes by reward conditioning leading to long-term memory formation. J Neurosci **25:** 1188-1192.

LAREAU, L. F., M. INADA, R. E. GREEN, J. C. WENGROD and S. E. BRENNER, 2007 Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature **446:** 926-929.

LE HIR, H., E. IZAURRALDE, L. E. MAQUAT and M. J. MOORE, 2000 The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. Embo J **19:** 6860-6869.

LETUNIC, I., R. R. COPLEY and P. BORK, 2002 Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet **11:** 1561-1567.

LEW, J. E., S. ENOMOTO and J. BERMAN, 1998 Telomere length regulation and telomeric chromatin require the nonsense-mediated mRNA decay pathway. Mol Cell Biol **18:** 6121-6130.

LEWIS, B. P., R. E. GREEN and S. E. BRENNER, 2003 Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A **100:** 189-192.

LI, W. H., C. I. WU and C. C. LUO, 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol **21:** 58-71.

LIPSCOMBE, D., 2005 Neuronal proteins custom designed by alternative splicing. Curr Opin Neurobiol **15:** 358-363.

LONG, J. C., and J. F. CACERES, 2009 The SR protein family of splicing factors: master regulators of gene expression. Biochem J **417:** 15-27.

LYNCH, M., and J. S. CONERY, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155.

LYNCH, M., and J. S. CONERY, 2003 The evolutionary demography of duplicate genes. J Struct Funct Genomics **3:** 35-44.

LYNCH, M., and A. FORCE, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459-473.

MANIATIS, T., and B. TASIC, 2002 Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature **418:** 236-243.

MAQUAT, L. E., 2004 Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol **5:** 89-99.

MARQUES, A. C., I. DUPANLOUP, N. VINCKENBOSCH, A. REYMOND and H. KAESSMANN, 2005 Emergence of young human genes after a burst of retroposition in primates. PLoS Biol **3:** e357.

MEDGHALCHI, S. M., P. A. FRISCHMEYER, J. T. MENDELL, A. G. KELLY, A. M. LAWLER *et al.*, 2001 Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. Hum Mol Genet **10:** 99-105.

MENDELL, J. T., N. A. SHARIFI, J. L. MEYERS, F. MARTINEZ-MURILLO and H. C. DIETZ, 2004 Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat Genet **36:** 1073-1078.

METHERELL, L. A., S. A. AKKER, P. B. MUNROE, S. J. ROSE, M. CAULFIELD *et al.*, 2001 Pseudoexon activation as a novel mechanism for disease resulting in atypical growth-hormone insensitivity. Am J Hum Genet **69:** 641-646.

MITROVICH, Q. M., and P. ANDERSON, 2000 Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes Dev **14:** 2173-2184.

MODREK, B., and C. J. LEE, 2003 Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet **34:** 177-180.

MODREK, B., A. RESCH, C. GRASSO and C. LEE, 2001 Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res **29:** 2850-2859.

MOORE, M. J., 1996 Gene expression. When the junk isn't junk. Nature **379:** 402-403.

MORAIS, D. D., and P. M. HARRISON, 2009 Genomic evidence for non-random endemic populations of decaying exons from mammalian genes. BMC Genomics **10:** 309.

MORRISON, M., K. S. HARRIS and M. B. ROTH, 1997 smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. Proc Natl Acad Sci U S A **94:** 9782-9785.

NADEAU, J. H., and D. SANKOFF, 1997 Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147:** 1259-1266.

NEKLASON, D. W., C. H. SOLOMON, A. L. DALTON, S. K. KUWADA and R. W. BURT, 2004 Intron 4 mutation in APC gene results in splice defect and attenuated FAP phenotype. Fam Cancer **3:** 35-40.

NI, J. Z., L. GRATE, J. P. DONOHUE, C. PRESTON, N. NOBIDA *et al.*, 2007 Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev **21:** 708-718.

OHNO, S., 1972 So much "junk" DNA in our genome. Brookhaven Symp Biol **23:** 366-370.

OHNO, S., U. WOLF and N. B. ATKIN, 1968 Evolution from fish to mammals by gene duplication. Hereditas **59:** 169-187.

PAN, Q., M. A. BAKOWSKI, Q. MORRIS, W. ZHANG, B. J. FREY *et al.*, 2005 Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet **21:** 73-77.

PARKE, D. W., 2002 Stickler syndrome: clinical care and molecular genetics. Am J Ophthalmol **134:** 746-748.

PARKHILL, J., G. DOUGAN, K. D. JAMES, N. R. THOMSON, D. PICKARD *et al.*, 2001 Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **413:** 848-852.

PATTON, M. A., and A. R. AFZAL, 2002 Robinow syndrome. J Med Genet **39:** 305-310.

PAVLICEK, A., J. PACES, R. ZIKA and J. HEJNAR, 2002 Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene **300:** 189-194.

PETROV, D. A., and D. L. HARTL, 2000 Pseudogene evolution and natural selection for a compact genome. J Hered **91:** 221-227.

PUSCH, M., 2002 Myotonia caused by mutations in the muscle chloride channel gene CLCN1. Hum Mutat **19:** 423-434.

RESCH, A., Y. XING, A. ALEKSEYENKO, B. MODREK and C. LEE, 2004 Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic Acids Res **32:** 1261-1269.

RODRIGUEZ-PASCAU, L., M. J. COLL, L. VILAGELIU and D. GRINBERG, 2009 Antisense oligonucleotide treatment for a pseudoexon-generating mutation in the NPC1 gene causing Niemann-Pick type C diseaseb. Hum Mutat **30:** E993-E1001.

ROSENFELD, P. J., G. S. COWLEY, T. L. MCGEE, M. A. SANDBERG, E. L. BERSON *et al.*, 1992 A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet **1:** 209-213.

SALTZMAN, A. L., Y. K. KIM, Q. PAN, M. M. FAGNANI, L. E. MAQUAT *et al.*, 2008 Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol Cell Biol **28:** 4320-4330.

SCHAAL, T. D., and T. MANIATIS, 1999 Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol Cell Biol **19:** 1705-1719.

SCHWABE, G. C., S. TINSCHERT, C. BUSCHOW, P. MEINECKE, G. WOLFF *et al.*, 2000 Distinct mutations in the receptor tyrosine kinase gene ROR2 cause brachydactyly type B. Am J Hum Genet **67:** 822-831.

SEOIGHE, C., and C. GEHRING, 2004 Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet **20:** 461-464.

SHOJA, V., and L. ZHANG, 2006 A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Mol Biol Evol **23:** 2134-2141.

SKOTHEIM, R. I., and M. NEES, 2007 Alternative splicing in cancer: noise, functional, or systematic? Int J Biochem Cell Biol **39:** 1432-1449.

SOREK, R., R. SHAMIR and G. AST, 2004 How prevalent is functional alternative splicing in the human genome? Trends Genet **20:** 68-71.

SPENA, S., R. ASSELTA, M. PLATE, G. CASTAMAN, S. DUGA *et al.*, 2007 Pseudo-exon activation caused by a deep-intronic mutation in the fibrinogen gamma-chain gene as a novel mechanism for congenital afibrinogenaemia. Br J Haematol **139:** 128-132.

STOILOV, P., R. DAOUD, O. NAYLER and S. STAMM, 2004 Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. Hum Mol Genet **13:** 509-524.

SU, Z., J. WANG, J. YU, X. HUANG and X. GU, 2006 Evolution of alternative splicing after gene duplication. Genome Res **16:** 182-189.

TANKO, Q., B. FRANKLIN, H. LYNCH and J. KNEZETIC, 2002 A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. Mutat Res **503:** 37-42.

TAZI, J., N. BAKKOUR and S. STAMM, 2009 Alternative splicing and disease. Biochim Biophys Acta **1792:** 14-26.

ULE, J., and R. B. DARNELL, 2006 RNA binding proteins and the regulation of neuronal synaptic plasticity. Curr Opin Neurobiol **16:** 102-110.

VAN DE PEER, Y., J. S. TAYLOR and A. MEYER, 2003 Are all fishes ancient polyploids? J Struct Funct Genomics **3:** 65-73.

VEGA, A. I., C. PEREZ-CERDA, L. R. DESVIAT, G. MATTHIJS, M. UGARTE *et al.*, 2009 Functional analysis of three splicing mutations identified in the PMM2 gene: toward a new therapy for congenital disorder of glycosylation type Ia. Hum Mutat **30:** 795-803.

VETRINI, F., R. TAMMARO, S. BONDANZA, E. M. SURACE, A. AURICCHIO *et al.*, 2006 Aberrant splicing in the ocular albinism type 1 gene (OA1/GPR143) is corrected in vitro by morpholino antisense oligonucleotides. Hum Mutat **27:** 420-426.

VITHANA, E. N., L. ABU-SAFIEH, M. J. ALLEN, A. CAREY, M. PAPAIOANNOU *et al.*, 2001 A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). Mol Cell **8:** 375-381.

WANG, Z., M. E. ROLISH, G. YEO, V. TUNG, M. MAWSON *et al.*, 2004 Systematic identification and analysis of exonic splicing silencers. Cell **119:** 831-845.

WANG, Z., X. XIAO, E. VAN NOSTRAND and C. B. BURGE, 2006 General and specific functions of exonic splicing silencers in splicing control. Mol Cell **23:** 61-70.

WEI, J., M. B. GOLDBERG, V. BURLAND, M. M. VENKATESAN, W. DENG *et al.*, 2003 Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect Immun **71:** 2775-2786.

WEINER, A. M., P. L. DEININGER and A. EFSTRATIADIS, 1986 Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem **55:** 631-661.

WILLINGHAM, A. T., and T. R. GINGERAS, 2006 TUF love for "junk" DNA. Cell **125:** 1215-1220.

WIRTH, B., L. BRICHTA and E. HAHNEN, 2006 Spinal muscular atrophy and therapeutic prospects. Prog Mol Subcell Biol **44:** 109-132.

WOLFE, K. H., and D. C. SHIELDS, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708-713.

XING, Y., and C. J. LEE, 2005 Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genet **1:** e34.

YEO, G. W., E. VAN NOSTRAND, D. HOLSTE, T. POGGIO and C. B. BURGE, 2005 Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A **102:** 2850-2855.

YU, Z., D. MORAIS, M. IVANGA and P. M. HARRISON, 2007 Analysis of the role of retrotransposition in gene evolution in vertebrates. BMC Bioinformatics **8:** 308.

ZHANG, Z., N. CARRIERO and M. GERSTEIN, 2004 Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet **20:** 62-67.

ZHANG, Z., P. HARRISON and M. GERSTEIN, 2002 Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res **12:** 1466-1482.

ZHANG, Z., P. M. HARRISON, Y. LIU and M. GERSTEIN, 2003 Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res **13:** 2541-2558.

ZHENG, D., and M. B. GERSTEIN, 2006 A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol **7 Suppl 1:** S13 11-10.

ZHENG, D., and M. B. GERSTEIN, 2007 The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet **23:** 219-224.

ZHENG, D., Z. ZHANG, P. M. HARRISON, J. KARRO, N. CARRIERO *et al.*, 2005 Integrated pseudogene annotation for human chromosome 22: evidence for transcription. J Mol Biol **349:** 27-45.

ZUCKERKANDL, E., 2002 Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. Genetica **115:** 105-129.

CHAPTER II

Genomic evidence for non-random endemic populations of decaying exons from mammalian genes

A version of this chapter appears as Morais, D. D., and P. M. Harrison, 2009 **Genomic evidence for non-random endemic populations of decaying exons from mammalian genes**. *BMC Genomics* 10: 309.

Abstract

Functional diversification of genes in mammalian genomes is engendered by a number of processes, *e.g.*, gene duplication and alternative splicing. Gene duplication is classically discussed as leading to *neofunctionalization* (generation of new functions), *subfunctionalization* (generation of a varied function), or *pseudogenization* (loss of the gene and its function). Here, we focus on the process of pseudogenization, but specifically for individual exons from genes. It is at present unclear to what extent pseudogenization of individual exon duplications affects gene evolution, *i.e.*, is it a random phenomenon, or is it associated with specific types of genes and encoded proteins, and positions in gene structures? We gathered genomic evidence for duplicated *pseudogenic exons* (ΨEs, *i.e.*, exons disabled by frameshifts and premature stop codons), to examine for significant trends in their distribution across four mammalian genomes (specifically human, cow, mouse and rat). Across these four genomes, we observed a consistent population of ΨEs, associated with 0.4–1.0% of genes. These ΨE populations exhibit codon substitution patterns that are typical of an endemic population of decaying sequences. In human, ΨEs have significant overrepresentation for functional categories related to 'ion binding' and 'nucleic-acid binding', compared to duplicated exons in general. Also, ΨEs tend to be associated with some protein domains that are abundant generally, *e.g.*, Zinc-finger and immunoglobulin

protein domains, but not others, *e.g.*, EGF-like domains. Positionally, ΨEs are also significantly associated with the 5' end of genes, but despite this, individual stop codons are positioned so that there is significant avoidance of potential targeting to nonsense-mediated decay. In human, ΨEs are often associated with alternative splicing (in 22 out of 284 genes with ΨEs in their milieu), and can have different parts of their sequence differentially spliced in alternative transcripts. Some unusual cases of ΨEs embedded within 5' and 3' non-coding exons are observed. Our results indicate the types of genes that harbour ΨEs, and demonstrate that ΨEs have non-random distribution within gene structures. These ΨEs may function in gene regulation through generation of transcribed pseudogenes, or regulatory alternate transcripts.

Background

Natural selection acts on phenotypes arising from a vast range of genomic variations: chromosomal and segmental duplications, local duplications, and smaller insertions, deletions and nucleotide substitutions. Local duplication arises not only for whole genes or multiples of genes, but also for pieces of genes and for individual exons. A pseudogene (ΨG), in the case of protein-coding genes, is a copy of a gene that has symptoms of protein-coding deficiency [1-6]. Symptoms of protein-coding deficiency include: *(i)* coding-sequence disablements (frame-shifts and premature stop codons); *(ii)* neutral codon substitution patterns (that yield values of Ka/Ks, the ratio of nonsynonymous to synonymous codon substitutions of ~1.0); *(iii)* protein domain truncations [2]; *(iv)* mutation of deeply-conserved residue positions essential for protein function or structural integrity [1]. Processed pseudogenes are made by reverse transcription and re-integration into the genome, and have been extensively studied elsewhere [1-6]. Non-processed pseudogenes can arise after local or segmental gene duplication, and subsequent loss of protein- coding ability through mutation. A similar situation can arise within an individual gene structure: one or more exons can become duplicated within the vicinity of a gene. Such partial gene duplications may then lose coding ability, becoming *pseudogenic exons* (ΨEs), in a similar way.

Here, we have gathered genomic evidence for the distribution of *pseudogenic exons* (ΨEs) in the chromosomal milieu of annotated genes of four mammals with high coverage genome assemblies and extensive transcriptional validation (human, cow, mouse and rat). Such ΨEs can have a functional role. For example, recently it has been described that ΨEs with stop codons that are alternatively spliced can target messenger RNAs to nonsense-mediated decay (NMD), in a way that causes changes in expression levels for other transcripts from the gene [7]. In our analysis, we define ΨEs specifically using coding-sequence disruptions (*i.e.*, frameshifts and premature stop codons). We find a non-random distribution of ΨEs in each mammalian genome, associated with certain subtypes of genes and positions within genes.

Results and discussion

A pipeline was derived to detect *pseudogenic exons* (ΨEs) in the immediate chromosomal milieu of genes (Figure 2.1; see *Methods* for details). A ΨE is defined as an exon copy whose coding ability is compromised by a frameshift or a premature stop codon. Such frameshifts and stop codons are the most obvious indicators of coding-sequence decay. The designated *parent exon* for a ΨE is the most similar exon in the surrounding annotated gene structure. In addition, we annotated duplicated exons (DEs) in the transcripts from each gene, as described in *Methods*

We focused on four mammalian genome assemblies with high (>7X) coverage (human, cow, mouse and rat), to analyze the extent of the occurrence of ΨEs. We examined for significant trends in the distribution of ΨEs for a variety of properties. In particular, we focused on assessing the peculiarities of the ΨEs in comparison to the general population of duplicated exons. We analyzed the following: *(i)* divergence from designated parent exons; *(ii)* association with protein families; *(iii)* association with Gene Ontology functional categories; *(iv)* position of ΨEs with respect to the intron-exon structure of the gene; *(v)* participation in alternative splicing, and *(vi)* coding sequence selection pressures, as judged by Ka/Ks values.

Table 2.1 summarizes the distribution of ΨEs. Strikingly, ΨEs occur at a consistent level across all of the mammalian genomes studied. The annotation pipeline identified between ~300 to ~600 cases of ΨEs per genome. These ΨEs occur for 0.4–1.0% of genes, with a frequency of 1.3– 2.0 ΨEs per gene. In addition, we determined ~4000– 7000 duplicated exons (DEs) within the annotated genes of each of the four studied mammals (Table 2.1). A substantial fraction (~12–22%) of the ΨEs are located on the strand opposite to the putative parent gene (Table 2.1), indicating some sort of inversion process in their generation.

(i) Divergence from designated parent exons

We analyzed the distribution of percentage sequence identity between the ΨEs and their respective designated parent exons. These distributions were compared to an equivalent distribution for DEs (Figure 2.2). This equivalent distribution is from comparison of the DEs to their most homologous exons within the same gene. The distributions generally have a mode for both DEs and ΨEs at 40– 50% (Figure 2.2). Therefore, ΨEs are not unusually divergent in terms of protein sequence identity with respect to DEs in general.

In addition, we examined distributions of Ks values for those exons which align to their designated parent exons with > = 70% amino-acid sequence identity (to avoid consideration of sequences with codon saturation) (Figure 2.3). Although recently, evidence has been uncovered indicating that Ks values are under selection in mammals [8], they can still be used in a comparative sense to compare the age trends in populations of sequences. In general, there is a notable tendency for very young exon duplications, with a peak appearing in the Ks distributions for all species at the interval 0.00–0.10, for ΨEs and for duplicated exons in general. Interestingly, also, a sizeable fraction of ΨEs appear to be derived from anciently duplicated exons (i.e., 30–60% having Ks > 1.4); such exons were likely duplicated earlier in vertebrate evolution, and became disabled later during mammalian speciation.

The distribution of exon sizes of DEs has medians in the range ~40–50 amino acid residues (Figure 2.4, Supplementary Table S2.1). However, ΨEs are substantially longer than DEs in general (median values in the range 70–110 amino acid residues, and broader distributions) (Figure 2.4). This larger size trend for ΨEs arises chiefly from the exon size trends for the specific gene families that tend to make large numbers of ΨEs, such as the Zinc-finger-containing (ZFC) genes (see Supplementary Table S2.2 and protein family section below). In aggregate, the majority of the ΨEs (> ~75%) have at least half of their designated parents' length, and ~55% have between 0.9–1.1 of their parents' length (Figure 2.5). A small percentage (6–13%) of the ΨEs is marginally longer than their parent exons (Figure 2.5); this is potentially because of neutrally-occurring insertions arising after duplication [9].

(ii) Association with protein families

Some gene families spawn large numbers of pseudogenes. Examples include olfactory receptors [10], ribosomal-protein genes [11], ABC transporters [12], and heat shock proteins [13]. Harrison and Gerstein noted previously that the gene families with the most non-processed pseudogenes tend to be involved in some form of interaction with the environment [1], *e.g*. through roles in immunity [14], chemosensation [1,15], or small-molecule transport [12]. Such gene families can also be linked to recent segmental duplications in mammals [16]. Here, we

examined which are the most common protein domain families in the ΨE and DE data sets (Supplementary Table S2.2). These numbers indicate the number of exons with at least one copy of each protein domain considered. Exons containing zinc-finger domains and immunoglobulin-like domains are consistently in the top five most abundant for both ΨEs and DEs. Genes for zinc-finger-containing (ZFC) proteins have undergone lineage-specific expansions over the course of mammalian evolution, so decaying ZFC exons are an expected consequence of this, and could perform regulatory roles as part of transcribed pseudogenes [17]. Transcribed pseudogenes have recently been shown to regulate the expression of homologous genes through the formation of small, interfering RNAs [18,19]. Immunoglobulin like domains are used in many proteins that are involved in various aspects of immunity, and have been previously noted to generate large numbers of pseudogenes [14]. The most notable difference between ΨEs and DEs in general, is that ΨEs rarely arise that contain EGF-like (epidermal growth factor-like) domains, whereas these exons are consistently abundant, generally (significant difference, $P < 0.05$, binomial statistics; Supplementary Table S2.2). EGF-like domains have expanded greatly in number over the course of mammalian evolution, and are found (with a small number of exceptions) either in the extracellular part of transmembrane proteins or in secreted proteins [20,21].

(iii) Association with Gene Ontology functional categories

We used Gene Ontology (GO) functional classification to assess which functional associations are the most common for ΨEs (Table 2.2). A pairwise comparison between lists of genes was performed to check over-represented terms according to various criteria, for ΨEs, and for DEs generally. In this analysis, we only studied the human, mouse and rat genomes, since these are the genomes with extensive GO functional annotation. Specifically of interest are the GO terms that are over-represented in ΨEs compared to DEs (Table 2.2). Significant over-representation is calculated using a Fisher's exact test with $P' < 0.05$, and a correction to P' for multiple hypothesis testing [22].

The top ten human DEs and ΨEs GO terms do not differ greatly from each other, in each of the species studied. However, each organism has distinct significant over-representations of GO terms. In the human genome, '*Ionbinding'* and '*Nucleic acid binding'* are significantly overrepresented in ΨEs, compared to DEs (Table 2.2). This overrepresentation appears to be chiefly due to ZFC transcription factors, which are obviously candidates for regulation through unproductive splicing and translation, or through the formation of regulatory transcribed pseudogenes. In mouse, *'receptor activity*' is significantly overrepresented in ΨEs compared to DEs, and '*transferase activity*' in rat. These

indicate that different types of gene have undergone *pseudogenic exon* formation in recent evolutionary time in each of these three organisms.

(iv) Position of ΨEs with respect to the intron-exon structure of the annotated gene

In general, the majority of ΨEs are located within the 5' half of the genes in every studied genome ($P < 0.01$, using $\chi 2$ tests; Table 2.1). This scenario suggests that proteins tend to become more complex through addition of exons to the 5' termini of their encoding genes. These exons could be inefficiently spliced and therefore will appear in only a few transcripts, while they may be selected against if they disrupt the normal gene function [23,24]. Interestingly, the ΨEs are significantly 5' of their parents in rat (Table 2.3). We suggest that this is due to lineage-specific activity related to specific gene families (Supplementary Table S2.2).

A key issue in examining the distribution of stop codons in ΨEs, is whether they would produce transcripts that are susceptible to nonsense-mediated decay (NMD). We examined for individual stop codons in the ΨEs that would lead to NMD targeting (Table 2.1). The number of such stop codons in ΨEs that would lead to NMD is significantly smaller than what is expected by chance ($P < 0.01$, using $\chi 2$ test), in human and cow, but not in the two rodent genomes. The expected distribution in this case, is calculated from the total size of the gene

introns divided appropriately, given the position of the stop codons in each ΨE. This indicates a selection pressure in these species, against the positioning of individual stop codons in ΨEs in places that would cause NMD. It has been shown that alternative splicing can be coupled to NMD to regulate the expression of other transcripts from a gene [25]. This mechanism has been dubbed *regulated unproductive splicing and translation* [25]. There may therefore be a selection pressure against placement of stop-codon-bearing exons in some genes, so that they are not affected by this mechanism.

We curated on the human ΨE data, to search for unexpected positional distributions in genes. In human, forty-five ΨEs were found embedded in an untranslated region (UTR). These UTR-embedded ΨEs are not highly conserved. Only eight of them are also found in chimp and rhesus (four in each species), and none of them are shared by the three primate species simultaneously. None of the embedded ΨEs are conserved in a non-primate species (cow, dog, mouse or rat). This is despite syntenic conservation of 28 out of the 45 genes in a non-primate species involved in the embedding, when manually compared in the UCSC Genome browser [26]. It is possible that these UTR-embedded ΨEs are remnants of overlapping gene arrangements. The manner of overlap for overlapping gene pairs changes very dynamically over evolution; for example, only 95 out of 255 human overlapping gene pairs were reported to be conserved as overlapping pairs in the mouse genome [27].

(v) Participation in alternative splicing

Alternative splice products containing premature stop codons can be degraded through nonsense-mediated decay (NMD), and consequently cause altered expression of protein-coding transcripts through changes in abundance of splicing factors [7]. We examined whether any ΨEs have been annotated as part of alternative splicings. To do this, we cross-referenced the ASD (alternative splicing database) [28] 'splicing event' annotation, with our ΨE list from the human genome. Of the human 284 genes that harbour a ΨE in their genomic milieu, 101 are present in the ASD alternative splicing database. Out of these, we found 22 genes (entailing 59 transcripts) with evidence of transcription of a ΨE as an alternative exon. Analyzing the alternatively-spliced forms in detail, we found four cases of an unusual topology of splicing (Supplementary Table S2.3). These four human ΨEs can be differentially spliced in a topologically novel manner, in which one portion of a ΨE is recruited in one splice form, while a different portion of it can take part in another splice form (Supplementary Table S2.3).

(vi) Ka/Ks analysis

Ka/Ks (*i.e.*, the normalized ratio of non-synonymous and synonymous codon site substitution rates) is a measure of selection on coding sequences; values < 1.0 can indicate purifying selection, whereas values ~1.0 are theoretically expected for neutral selection pressures. Values significantly > 1.0 indicate

positive selection over the whole of a sequence. We examined Ka/Ks values for the different populations of ΨEs and DEs. Ka/Ks values were calculated for all exon alignments with amino-acid sequence identity > 70%, to avoid consideration of saturated nucleotide sequences [2,3]. In general, the DEs exhibit a mode in the range 0.00–0.25 for Ka/Ks, indicating a tendency to purifying selection (Figure 2.6). In contrast, the ΨE populations do not exhibit such a mode, instead peaking in the range 0.25–0.75 (Figure 2.6). Zhang *et al.* have previously observed such a Ka/Ks peak for *pseudogenic* transcripts captured by transposons [29], and for processed pseudogenes [3]. Thus is to be expected for endemic populations of neutrally evolving sequences, from comparisons with their putative parent sequences. The reasons for such Ka/Ks values < 1.0 may include: *(i)* continued purifying selection on the putative parent sequence; *(ii)* an original protein-coding phase for the present-day ΨE. Interestingly, ~30% of ΨE cases, have Ka/Ks values > 1.5, which indicates that they may have undergone positive selection before becoming disabled.

Conclusion

We gathered genomic evidence to assess for non-random distribution of *pseudogenic exons* (ΨEs) in four mammalian genomes. We observed endemic populations of decaying exons consistently across genomes, arising for up to ~1% of genes. These ΨEs were defined using coding sequence disablements

(frameshifts and premature stop codons). Of course, other *pseudogenic exons* may exist (such as those arising from initial disablement of splicing signals); however, such *pseudogenic exons* would be likely to acquire coding-sequence disablements rapidly, soon after their initial disablement.

The *pseudogenic exons* (ΨEs) are longer than duplicated exons in general, are associated with genes encoding specific protein domain families, such as zinc-finger-containing proteins, and are noticeably lacking for genes containing domains that are otherwise abundant, such as EGF-like domains. The ΨEs also demonstrate species-specific over-representation of GO functional categories relative to duplicated exons in general; for example, in human, GO functional categories for *'ion-binding'* and *'nucleic acid binding'* are significantly over-represented, compared to duplicated exons generally. The ΨE populations indicate the sorts of genes that have undergone exon decay in recent mammalian evolution (recent enough, and in large enough amounts, for them not to be deleted from the genomic DNA). We find statistical evidence for selection pressure on avoidance of stop codon placements in ΨEs that would lead to nonsense-mediate decay. In addition, we find some interesting positioning of ΨEs in gene structures, such as embedding in UTRs, or partial alternative splicing. The ΨE populations are a potential resource for the formation of transcribed pseudogenes, which can function in the regulation of homologous genes through formation of small, interfering RNAs [18,19,30]. They may also be involved in alternative transcripts

that have a regulatory function [7]. The annotated ΨEs that we have analyzed will be a fertile source for study using large-scale micro-array expression techniques for these two potential regulatory functions. Also, the ΨE data sets will be useful for further gene evolution study in mammals. The data are available from the authors at http://biology.mcgill.ca/faculty/harrison/.

Methods

Genome data

The genome sequences and annotation files of four mammals analyzed in this paper (human, mouse, rat, and cow) were downloaded from the Ensemble Web site http://www.ensembl.org, in February 2007. The genome assemblies are: human= Homo_sapiens.NCBI36.43; mouse= Mus_musculus.NCBIM36.43; rat= Rattus_norvegicus.RGSC3.4.43; cow= Bos_taurus.Btau_3.1.43. These genomes were chosen, because: *(i)* the genome assemblies are high (>7X) coverage, and *(ii)* >85% of the gene annotations in these genomes have complete transcription validation. To identify the duplicated exons we compared each exon of each gene against the whole protein sequence of the same gene using BLASTp (e-value ≤ 10-4) [31]. Exon definitions were taken directly from the genome annotations. To detect ΨEs, each exon was compared with the whole genomic DNA of the same gene plus a 5000-nucleotides (nt) buffer, 5' and 3' of the gene (Figure 2.1). The vast majority (>85%) of the introns of mammalian protein-coding genes are

<5000 nt in length. As is illustrated with the data from the cow genome in Supplementary Table S2.4, the number of ΨEs that are detected, has only a small dependence on the size of this buffer. We used protein-level sequence alignment to detect ΨEs throughout the paper; this is so that we can exploit the signal of protein coding sequence that is still in these sequences to detect them in the genomic DNA

Identification of Duplicated Exons (DEs) and *Pseudogenic Exons* (ΨEs)

(1) Exon boundaries

The positioning of exon boundaries in encoded protein sequence was deduced and extracted from Ensembl Genbank- style annotation files, downloaded from http:// www.ensembl.org. The positioning was then used to map the exact location of an exon BLAST match [31].

(2) Homology detection

Each exon (amino acid sequence) was compared against its whole protein sequence using BLASTp to find duplicated exons with similarities with e-value ≤ 10-4. For ΨEs, each exon was compared (using tBLASTn [31], with e-value as above) against the genomic milieu of the encoding gene, which is defined as the genomic DNA of the gene (including introns), plus 5000 nucleotides, 5' and 3' of the gene.

(3) Fastx/y and Genewise realignments

After filtering for overlapping each tBLASTn match was realigned using FASTX/Y, as previously described [2,15,32]. The FASTX/Y [33] program allows longer alignments and also allows the identification of stop codon and frame-shifts in ΨEs. To confirm that the disablements of FASTX/Y were not an artifact we also aligned the ΨEs with GeneWise [34]. Only ΨEs confirmed by both methods were kept in our analyses.

(4) Filtering

ΨEs were filtered to remove olfactory receptors (ORs) and other single exon genes, since it is difficult to classify them as processed or duplicated *pseudogenic exons* [35]. Each match was compared with the Interpro http:// www.ebi.ac.uk/interpro/, Gene Ontology (GO [36]) descriptions and Ensembl http://www.ensembl.org protein family descriptions. If a ΨE was annotated in at least one of those databanks as an OR or other single-exon gene, it was removed from the analysis. To confirm the presence of stop-codons each ΨEs was realigned against its translated parent using bl2seq [37]; the output was parsed so that stop codons outside of a margin of 10 amino acids at the ends of the aligned subsequences were adjudged to be verified.

(5) Orthologs

The information about orthologs was extracted from the Biomart query system in the Ensembl database. As a further filter for the ortholog assignments, we performed a 'local gene order' test [38]. We compared the chromosomal milieu of genes bearing ΨEs, with the milieus of their orthologs, as follows. After identifying the ortholog of the gene containing the ΨE (step (5) above), we took a window (Wgenes) of 9 genes in either direction (the gene bearing the ΨE plus 4 genes 5' and 4 genes 3' of it) and 'BLASTed' against the equivalent 9 genes for the ortholog. We focused on the human genome; therefore, this local gene order test was performed for the human data vs. cow, mouse and rat genomes. The number of significant matches (BLASTp e-value $\leq$ 10-4, sequence identity >45%, and match $\geq$ 0.6 length of both orthologs) between the milieu of the two considered species was investigated. We allowed up to three gaps in total within the Wgenes windows.

(6) NMD targeting

To analyze for potential NMD targeting, we disregarded any ΨE located beyond the 5' and 3' UTRs, and also ΨEs located after the real stop codon, since they would not lead to nonsense-mediated decay (NMD). Then, we mapped the position of each stop codon in every ΨE to see if they are in a NMD area. If a stop codon would be located more than 55 nucleotides 5' to the last exon-exon junction

in a transcript containing the ΨE, then the ΨE was labeled as within a putative NMD target.

Functional categories

Gene Ontology (GO [36]) functional categories, Ensembl protein family and Pfam protein family descriptions, where retrieved using the Biomart tool [39]. GO functional category enrichment analyses of DEs and ΨEs were performed using FatiGo database [22].

Alternative splicing (AS)

We checked whether exons are alternatively spliced, by counting up the number of splice forms that a gene produces, and labeling the exons as *constitutive* if they appear in all splice forms, and *alternative* otherwise. We cross-referenced the coordinates of each ΨE with every event of alternative splicing annotation in the ASD database [28].

Analysis of Ka/Ks values

The program codeml of the PAML package [40] was used to calculate the maximum-likelihood Ka/Ks values of designated parent exons compared to ΨEs and DEs. The input codon alignments were generated using the PAL2NAL program [41]. Only pairs of sequence with ≥ 70% of identity and ≥ 40 amino

acids long were used in this analysis as the reliability of Ka/Ks analysis falls rapidly below this threshold [42].

Authors' contributions

DM performed the data analysis and wrote the initial draft of the manuscript. PH designed and directed the project, and wrote later drafts of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by grants from the National Science and Engineering Research Council of Canada, and from McGill University.

References

1. Harrison PM, Gerstein M: **Studying genomes through the aeons: protein families, pseudogenes and proteome evolution.** *J Mol Biol* 2002, **318(5):**1155-1174.

2. Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M: **Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein- coding ability.** *Nucleic acids research* 2005, **33(8):**2374-2383.

3. Zhang Z, Harrison PM, Liu Y, Gerstein M: **Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome.** *Genome research* 2003, **13(12):**2541-2558.

4. Zhang ZL, Harrison PM, Gerstein M: **Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes.** *J Mol Biol* 2002, **323(5):**811-822.

5. Balasubramanian S, Harrison P, Hegyi H, Bertone P, Luscombe N, *et al.*: **SNPs on human chromosomes 21 and 22 – analysis in terms of protein features and pseudogenes.** *Pharmacogenomics* 2002, **3(3):**393-402.

6. Harrison PM, Carriero N, Liu Y, Gerstein M: **A "polyORFomic" analysis of prokaryote genomes using disabled-homology filtering reveals conserved but undiscovered short ORFs.** *J Mol Biol* 2003, **333(5):**885-892.

7. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements.** *Nature* 2007, **446(7138):**926-929.

8. Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, *et al.*: **Widespread positive selection in synonymous sites of mammalian genes.** *Molecular biology and evolution* 2007, **24(8):**1821-1831.

9. Zhang Z, Gerstein M: **Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes.** *Nucleic acids research* 2003, **31(18):**5338-5348.

10. Gilad Y, Man O, Paabo S, Lancet D: **Human specific loss of olfactory receptor genes.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(6):**3324-3327.

11. Zhang Z, Harrison P, Gerstein M: **Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome.** *Genome research* 2002, **12(10):**1466-1482.

12. Liu Y, Harrison PM, Kunin V, Gerstein M: **Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes.** *Genome biology* 2004, **5(9):**R64.

13. Brocchieri L, Conway de Macario E, Macario AJ: **hsp70 genes in the human genome: Conservation and differentiation patterns predict a wide array of overlapping and specialized functions.** *BMC evolutionary biology* 2008, **8:**19.

14. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, *et al.*: **Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22.** *Genome research* 2002, **12(2):**272-280.

15. Harrison P, Kumar A, Lan N, Echols N, Snyder M, *et al.*: **A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution.** *J Mol Biol* 2002, **316(3):**409-419.

16. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, *et al.*: **Recent segmental duplications in the human genome.** *Science* 2002, **297(5583):**1003-1007.

17. Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, *et al.*: **Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes.** *Genome Res* 2006, **16:**584-594.

18. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, *et al.*: **Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes.** *Nature* 2008, **453(7194):**534-538.

19. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, *et al.*: **Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes.** *Nature* 2008, **453(7194):**539-543.

20. Lohmeyer M, Harrison PM, Kannan S, DeSantis M, O'Reilly NJ, *et al.*: **Chemical synthesis, structural modeling, and biological activity of the epidermal growth factor-like domain of human cripto.** *Biochemistry* 1997, **36(13):**3837-3845.

21. Doolittle RF: **Stein and Moore Award address. Reconstructing history with amino acid sequences.** *Protein Sci* 1992, **1(2):**191-200.

22. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J: **BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.** *Nucleic acids research* 2005:W460-464.

23. Zhang XH, Chasin LA: **Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons.** ***Proceedings of the National Academy of Sciences of the United States of America*** **2006, 103(36):13427-13432.**

24. Modrek B, Lee CJ: **Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss.** *Nat Genet* 2003, **34(2):**177-180.

25. Lareau LF, Brooks AN, Soergel DA, Meng Q, Brenner SE: **The coupling of alternative splicing and nonsense-mediated mRNA decay.** *Adv Exp Med Biol* 2007, **623:**190-211.

26. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, *et al.*: **The UCSC Genome Browser Database: 2008 update.** *Nucleic acids research* 2008:D773-779.

27. Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I: **Mammalian overlapping genes: the comparative perspective.** *Genome research* 2004, **14(2):**280-286.

28. Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, *et al.*: **ASD: a bioinformatics resource on alternative splicing.** *Nucleic acids research* 2006:D46-55.

29. Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE: **The evolutionary fate of MULE-mediated duplications of host gene fragments in rice.** *Genome research* 2005, **15(9):**1292-1297.

30. Korneev SA, Park JH, O'Shea M: **Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene.** *J Neurosci* 1999, **19(18):**7711-7720.

31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25(17):**3389-3402.

32. Harrison P, Yu Z: **Frame disruptions in human mRNA transcripts, and their relationship with splicing and protein structures.** *BMC genomics* 2007, **8:**371.

33. Pearson WR: **Using the FASTA program to search protein and DNA sequence databases.** *Methods Mol Biol* 1994, **24:**307-331.

34. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome research* 2004, **14(5):**988-995.

35. Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N, *et al.*: **Integrated pseudogene annotation for human chromosome 22: evidence for transcription.** *J Mol Biol* 2005, **349(1):**27-45.

36. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, *et al.*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic acids research* 2004:D258-261.

37. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174(2):**247-250.

38. Yu Z, Morais D, Ivanga M, Harrison PM: **Analysis of the role of retrotransposition in gene evolution in vertebrates.** *BMC Bioinformatics* 2007, **8:**308.

39. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, *et al.*: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21(16):**3439-3440.

40. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.

41. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic acids research* 2006:W609-612.

42. Letunic I, Copley RR, Bork P: **Common exon duplication in animals and its role in alternative splicing.** *Hum Mol Genet* 2002, **11(13):**1561-1567.

Table 2.1: Summary of the annotations

Feature	*Homo sapiens†*	*Mus musculus†*	*Rattus norvegicus†*	*Bos taurus†*
DE	6717 (1341)	4645 (1079)	4052 (993)	4389 (982)
ΨE	377 (284)	270 (209)	364 (218)	581 (298)
- 5' half	263	178	88	431
- 3' half	114	92	276	155
- Opposite strand	13%	12.2%	21.5%	14.31%
-Number of **ΨEs** that would lead to NMD targeting	55	48	138	194
- Orthologs and the Gene Order test	-------	36/67* (53.7%)	39/62* (62.9%)	45/75* (60%)

* Number that can be assigned orthologs, as determined in Ensembl annotation. We can see that most assigned orthologs are in syntenic positions.
† The number of genes bearing the exons is in brackets.

Table 2.2: Most common Gene Ontology functional categories †

Homo sapiens

All genes (Total = 31524)	ΨEs (Total = 284)	DEs (Total = 1341)
GO:0005515, protein binding (5864)	GO:0043167, ion binding (92)*§	GO:0005515, protein binding (372)
GO:0043167, ion binding (3861)	GO:0003676, nucleic acid binding (74)*§	GO:0043167, ion binding (349)*
GO:0003676, nucleic acid binding (3251)	GO:0005515, protein binding (34)	GO:0003676, nucleic acid binding (176)
GO:0016787, hydrolase activity (2053)	GO:0016740, transferase activity (18)	GO:0016787, hydrolase activity (105)
GO:0000166, nucleotide binding (1992)	GO:0004872, receptor activity (13)	GO:0004872, receptor activity (84)
GO:0004872, receptor activity (1765)	GO:0000166, nucleotide binding (12)	GO:0000166, nucleotide binding (57)
GO:0016740, transferase activity (1631)	GO:0016491, oxidoreductase activity (11)§	GO:0016740, transferase activity (36)
GO:0016491, oxidoreductase activity (723)	GO:0016787, hydrolase activity (10)	GO:0030246, carbohydrate binding (35)*
GO:0015075, ion transporter activity (541)	GO:0030246, carbohydrate binding (6)	GO:0005201, extracellular matrix structural constituent (22)*
GO:0008289, lipid binding (420)	GO:0046906, tetrapyrrole binding (3)	GO:0004857, enzyme inhibitor activity (21)

Mus musculus

All genes (Total = 28390)	ΨEs (Total = 209)	DEs (Total = 1079)
GO:0005515, protein binding (5553)	GO:0043167, ion binding (55)*	GO:0005515, protein binding (374)*
GO:0043167, ion binding (3672)	GO:0003676, nucleic acid GO:0004872, receptor activity (27)§ binding (43)*	GO:0043167, ion binding (321)*
GO:0003676, nucleic acid binding (3382)	GO:0016787, hydrolase activity (16)	GO:0003676, nucleic acid binding (173)
GO:0004872, receptor activity (2779)	GO:0005515, protein binding (14)	GO:0016787, hydrolase activity (114)
GO:0016787, hydrolase activity (2260)	GO:0016491, oxidoreductase activity (9)	GO:0004872, extracellular matrix (109)
GO:0000166, nucleotide binding (2061)	GO:0004857, enzyme inhibitor activity (7)	GO:0000166, nucleotide binding receptor activity (91)
GO:0016740, transferase activity (1805)	GO:0000166, nucleotide binding (7)	GO:0016740, transferase activity (51)
GO:0016491, oxidoreductase activity (911)	GO:0046906, tetrapyrrole binding (6)	GO:0030246, carbohydrate binding (41)*
GO:0015075, ion transporter		GO:0005201, structural

activity (598) GO:0008289, lipid binding (401)	GO:0016740, transferase activity (5)	constituent (40)* GO:0016491, oxidoreductase activity (25)
Rattus norvegicus		
All genes (Total = 27302)	ΨEs (Total = 218)	DEs (Total = 993)
GO:0005515, protein binding (2732) GO:0043167, ion binding (2238) GO:0004872, receptor activity (2063) GO:0003676, nucleic acid binding (1720) GO:0000166, nucleotide binding (1406) GO:0016787, hydrolase activity (1331) GO:0016740, transferase activity (1179) GO:0016491, oxidoreductase activity (594) GO:0015075, ion transporter activity (392) GO:0003735, structural constituent of ribosome (284)	GO:0043167, ion binding (23) GO:0016740, transferase activity (16)§ GO:0003676, nucleic acid binding (15) GO:0004872, receptor activity (14) GO:0005515, protein binding (12) GO:0016787, hydrolase activity (12) GO:0000166, nucleotide binding (10) GO:0016491, oxidoreductase activity (6) GO:0046906, tetrapyrrole binding (5) GO:0030246, carbohydrate binding (4)	GO:0043167, ion binding (158)* GO:0005515, protein binding (155)* GO:0003676, nucleic acid binding (74) GO:0016787, hydrolase activity (62) GO:0000166, nucleotide binding (46) GO:0004872, receptor activity (37) GO:0016740, transferase activity (29) GO:0016491, oxidoreductase activity (16) GO:0030246, carbohydrate binding (16) GO:0005201, extracellular matrix structural constituent (14)*

* Over-represented term when compared with all genes.
§ Over-represented term when compared with DEs.
† GO term counts are listed for human, mouse and rat.

Table 2.3: Position of ΨEs in related with their parents

	Number of ΨE 5' to parent	Number of ΨE beyond the 5' end of the gene	Number of ΨE 3' to parent	Number of ΨE beyond the 3' end of the gene
Homo sapiens	179	78	198	118
Mus musculus	131	54	139	84
Rattus norvegicus	213 †	42	151 †	51
Bos taurus	298	34	283	41

† Significantly non-random, $P < 0.05$, chi-squared test.

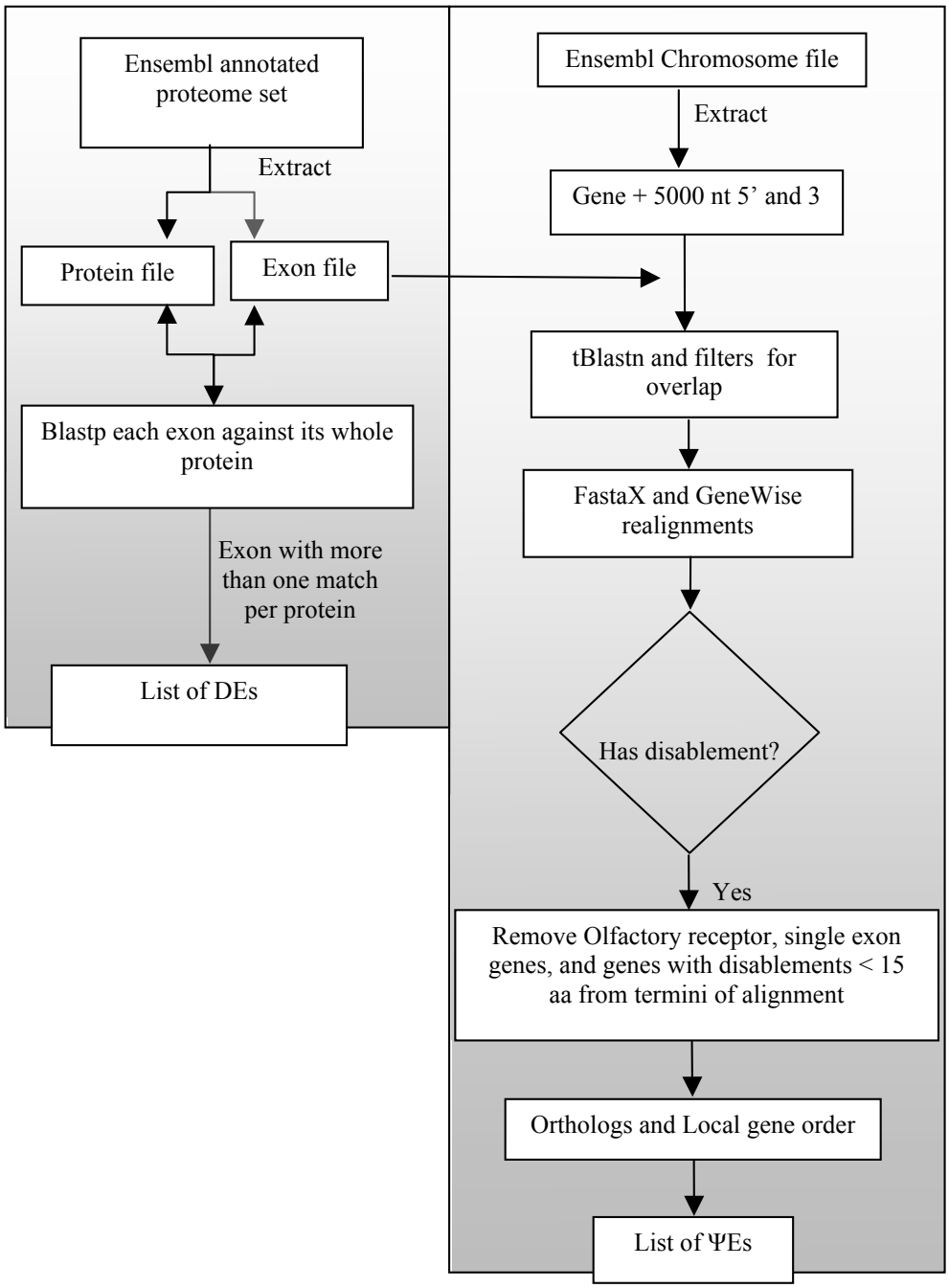
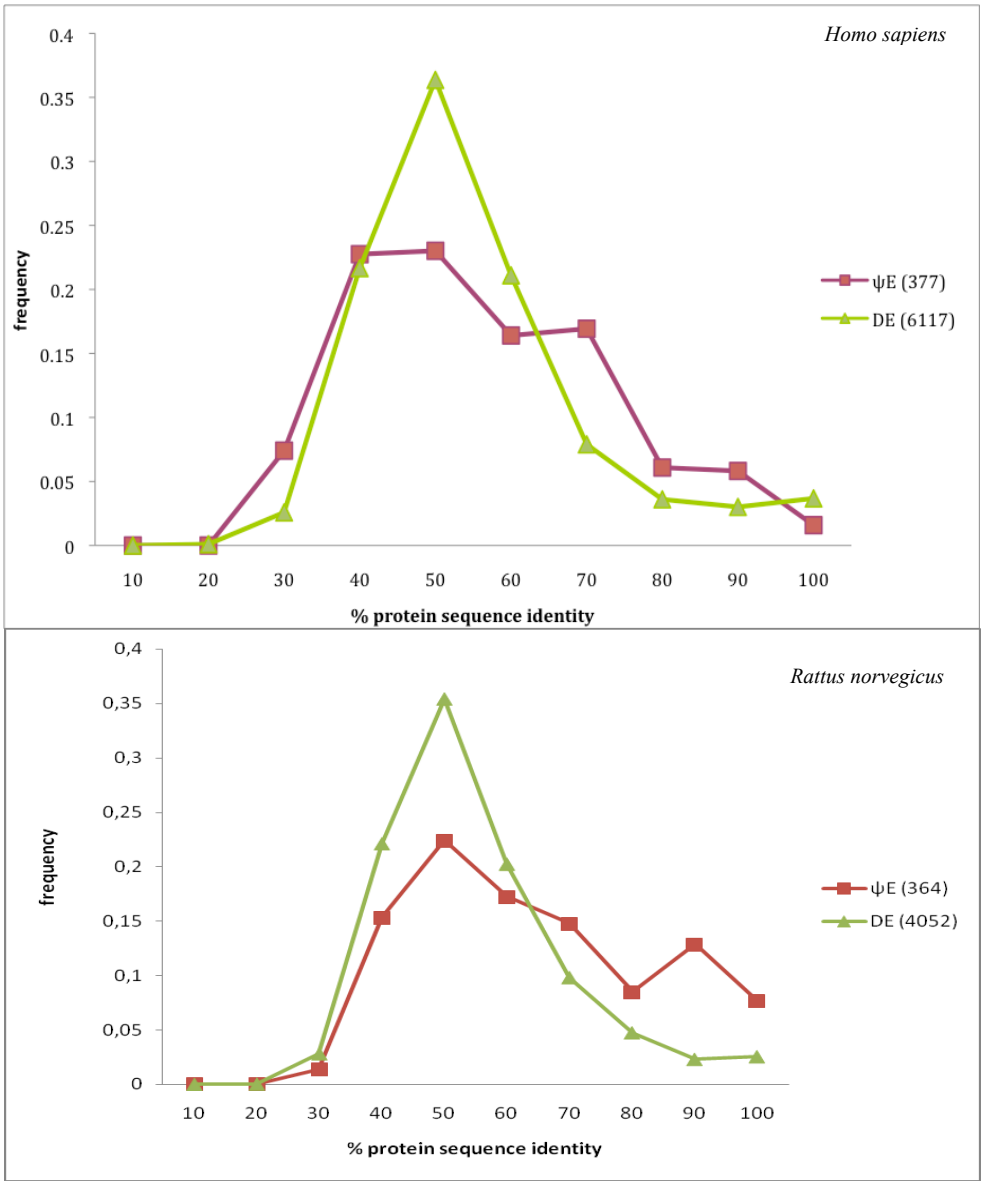


Figure 2.1: Pipeline annotation of DEs, ΨEs. The pipeline annotation is summarized.


Homo sapiens
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
frequency
10
20
30
40
50
60
70
80
90
100
% protein sequence identity
ψE (377)
DE (6117)
Rattus norvegicus
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
frequency
10
20
30
40
50
60
70
80
90
100
% protein sequence identity
ψE (364)
DE (4052)

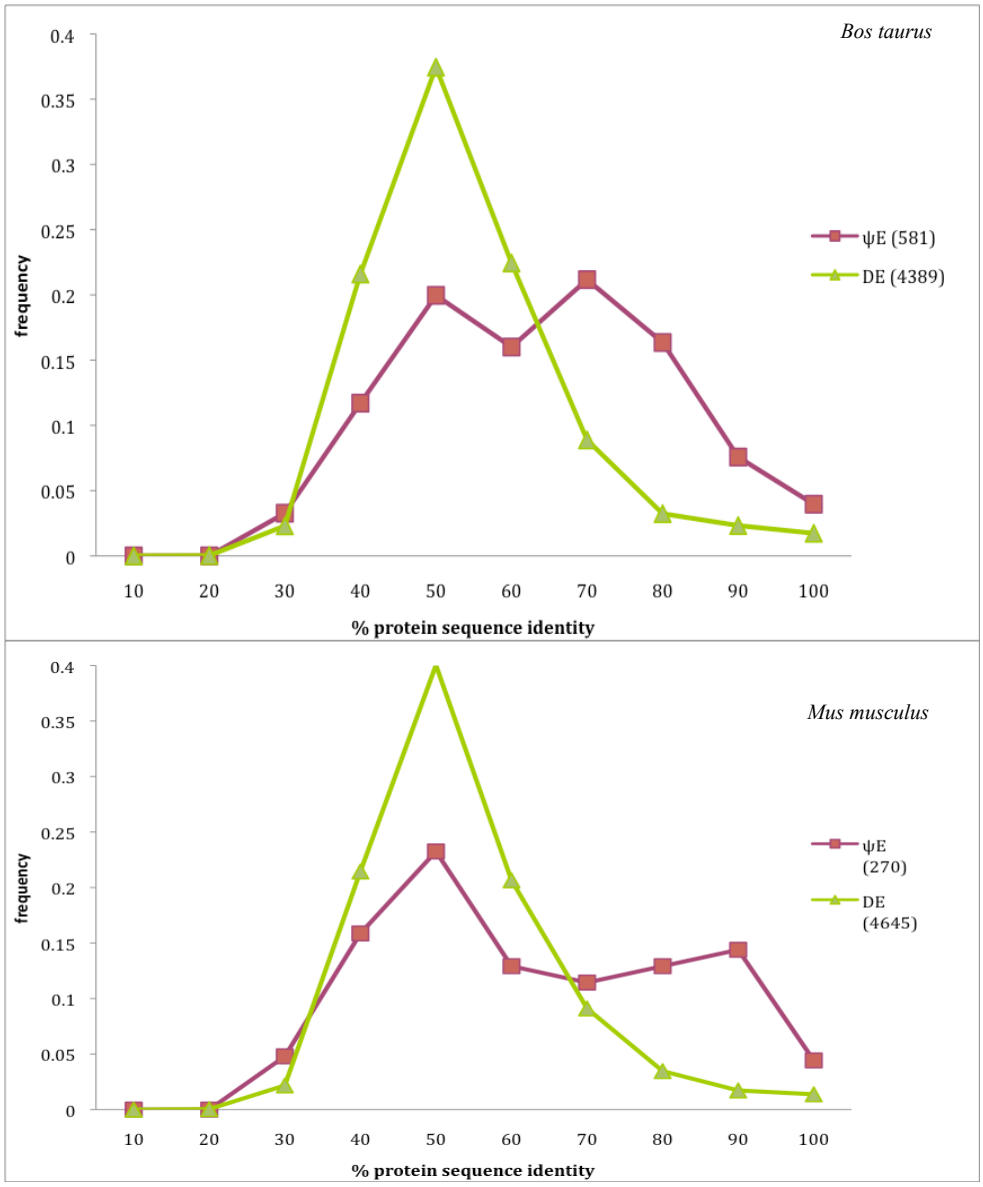
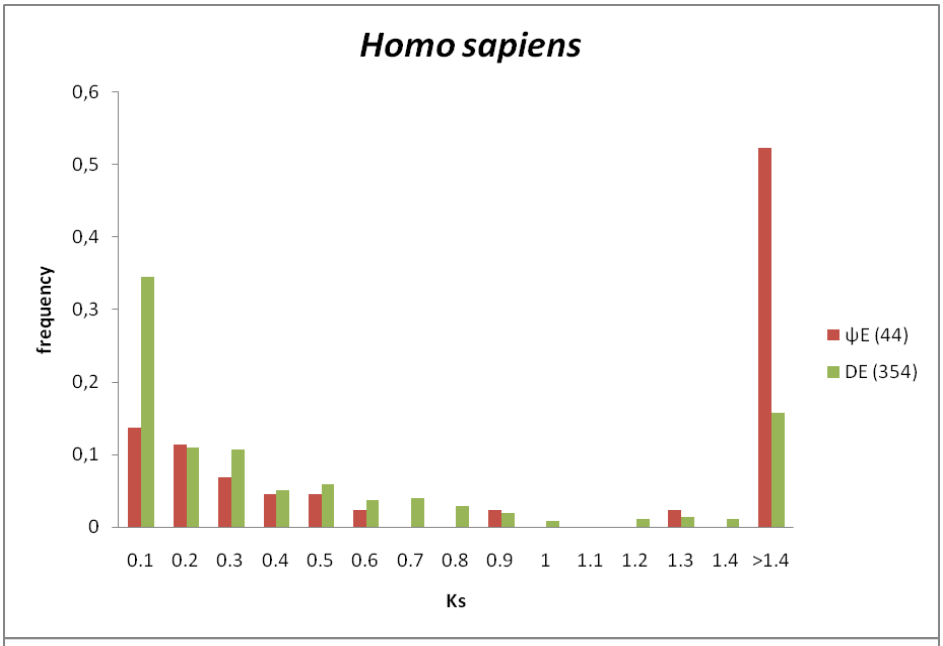
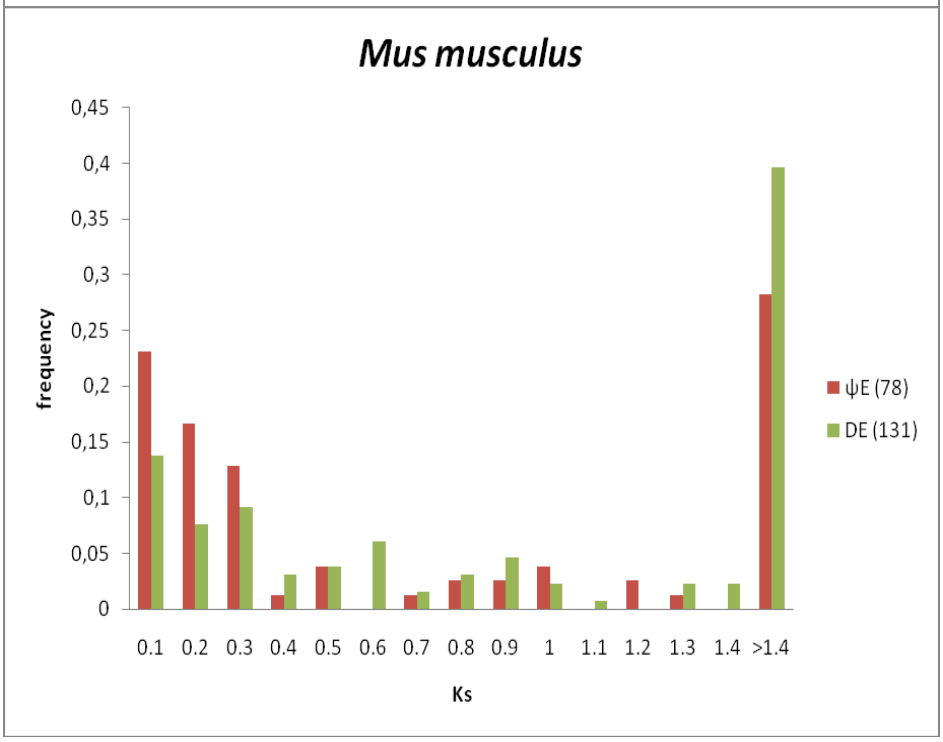


Figure 2.2: Distributions of protein sequence identity for DEs and ΨEs. These curves are for the data sets listed in Table 2.1. There are four panels for each of the four mammals analysed, labelled with the binomial species name. For each panel, the DE curve is green, and the ΨE curve is red. The bin label x is for all values such that, x-10 < value $\leq x$.


Homo sapiens
0,6
0,5
0,4
0,3
0,2
0,1
0
frequency
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 >1.4
Ks
ψE (44)
DE (354)


Mus musculus
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
frequency
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 >1.4
Ks
ψE (78)
DE (131)

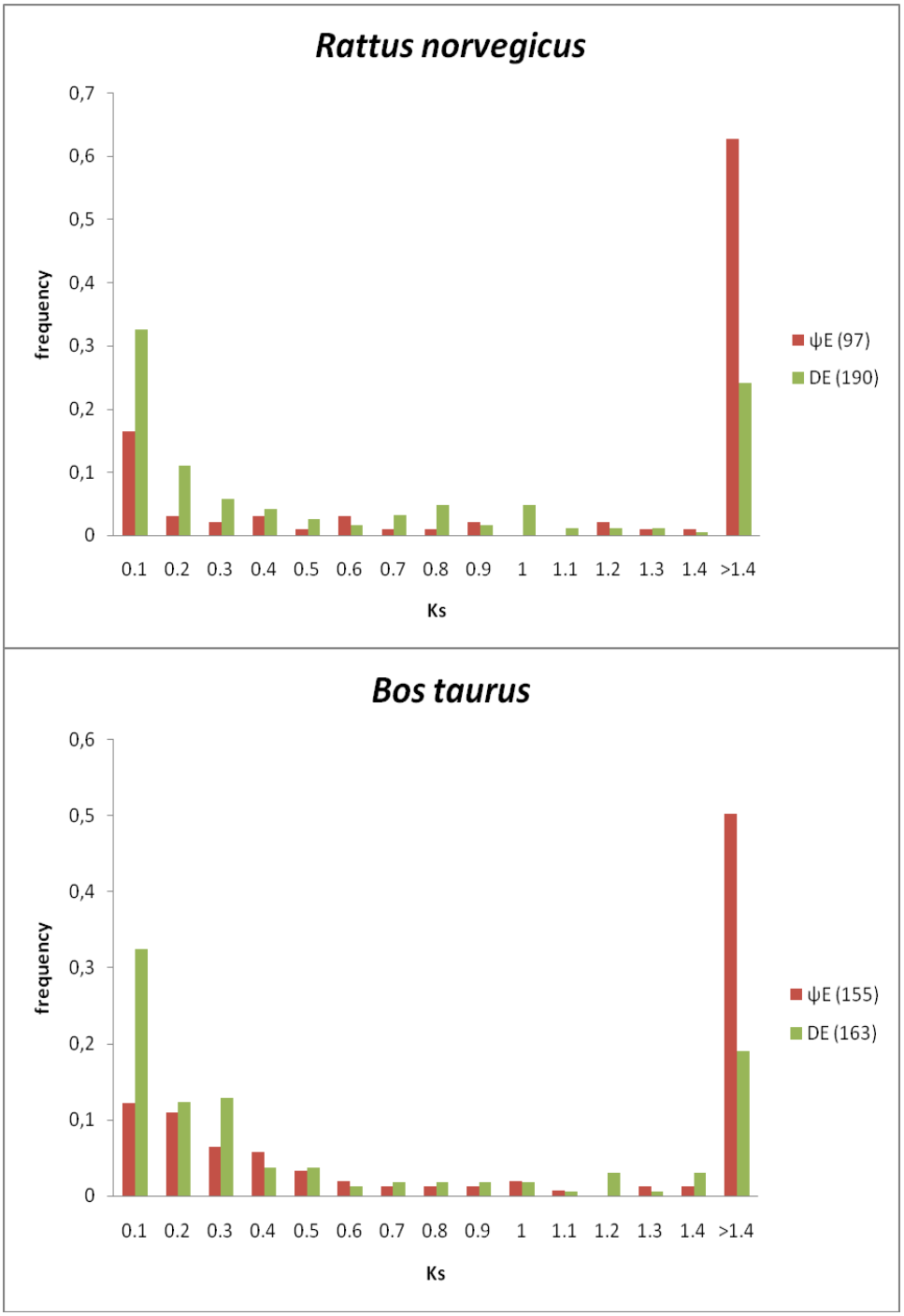
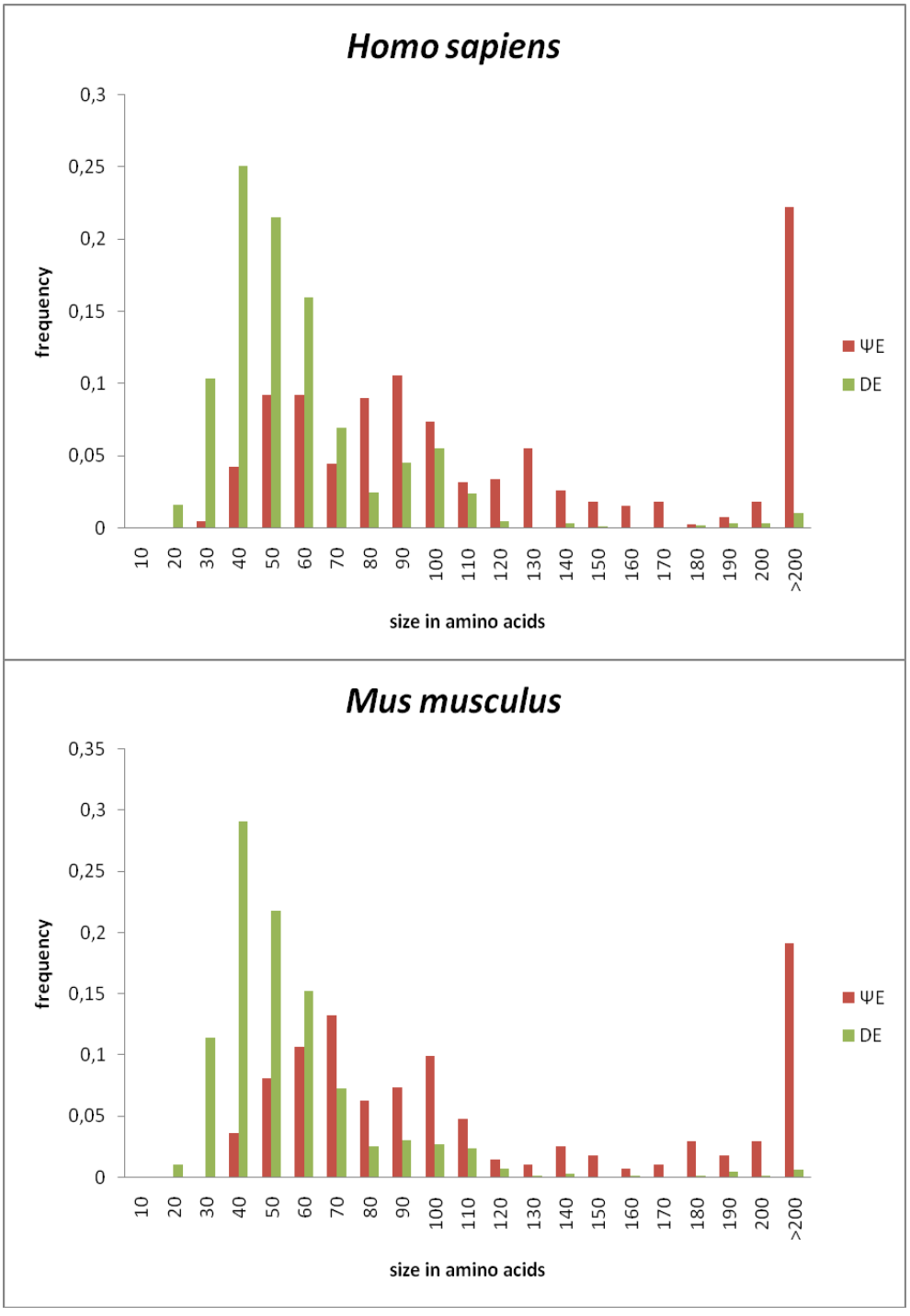


Figure 2.3: Distributions of Ks for DEs and ΨEs. These curves are for the data sets listed in Table 2.1. The DE curve is green, and the ΨE curve is red. The bin label x is for all values such that, x-0.1 < value $\leq x$.


Homo sapiens
0,3
0,25
0,2
0,15
0,1
0,05
0
frequency
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 >200
size in amino acids
ΨE
DE
Mus musculus
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
frequency
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 >200
size in amino acids
ΨE
DE

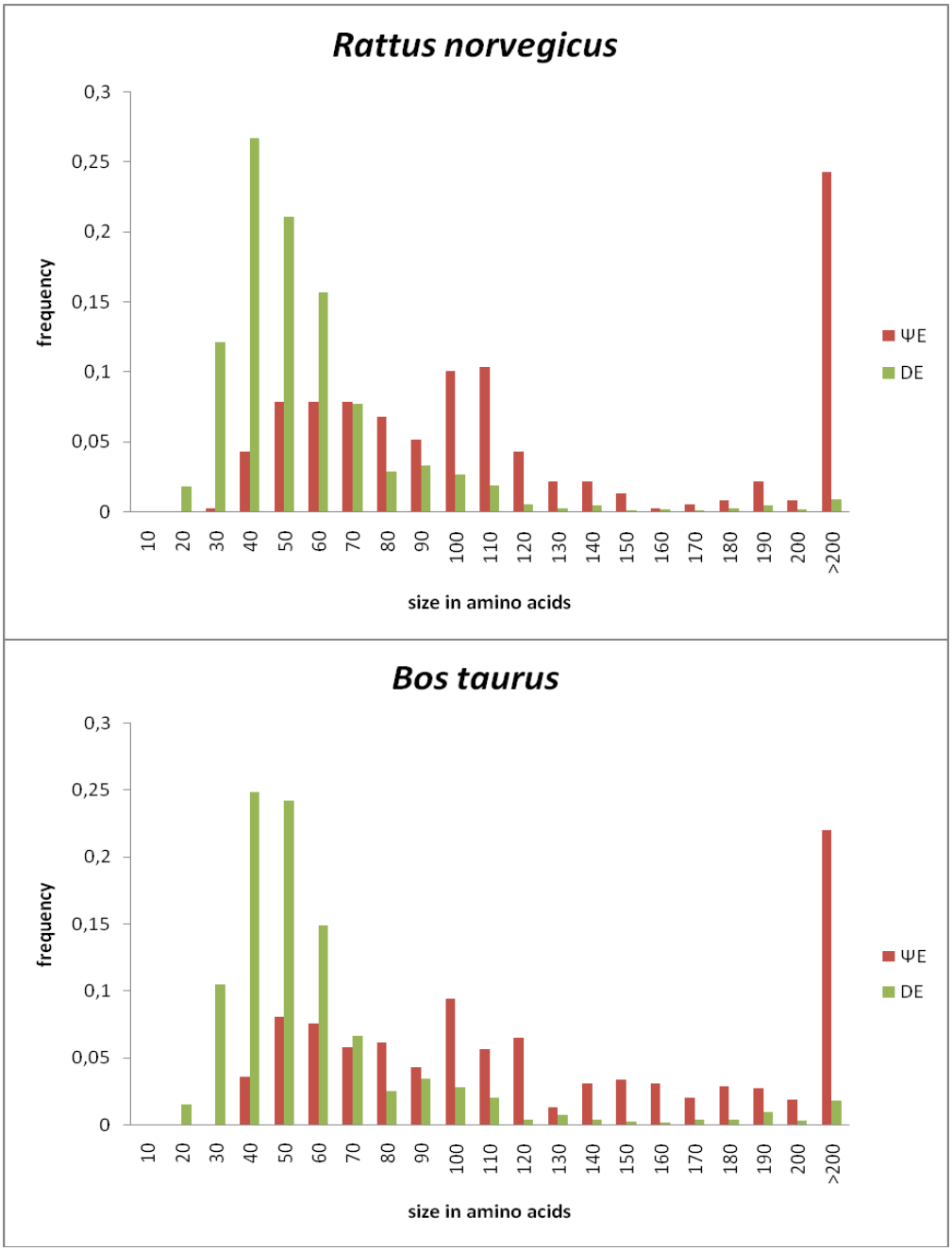


Figure 2.4: Distributions of size (in nucleotides) for DEs and ΨEs. These curves are for the data sets listed in Table 2.1. The DE curve is green, and the ΨE curve is red. The bin label x is for all values such that, x-10 < value $\leq x$.

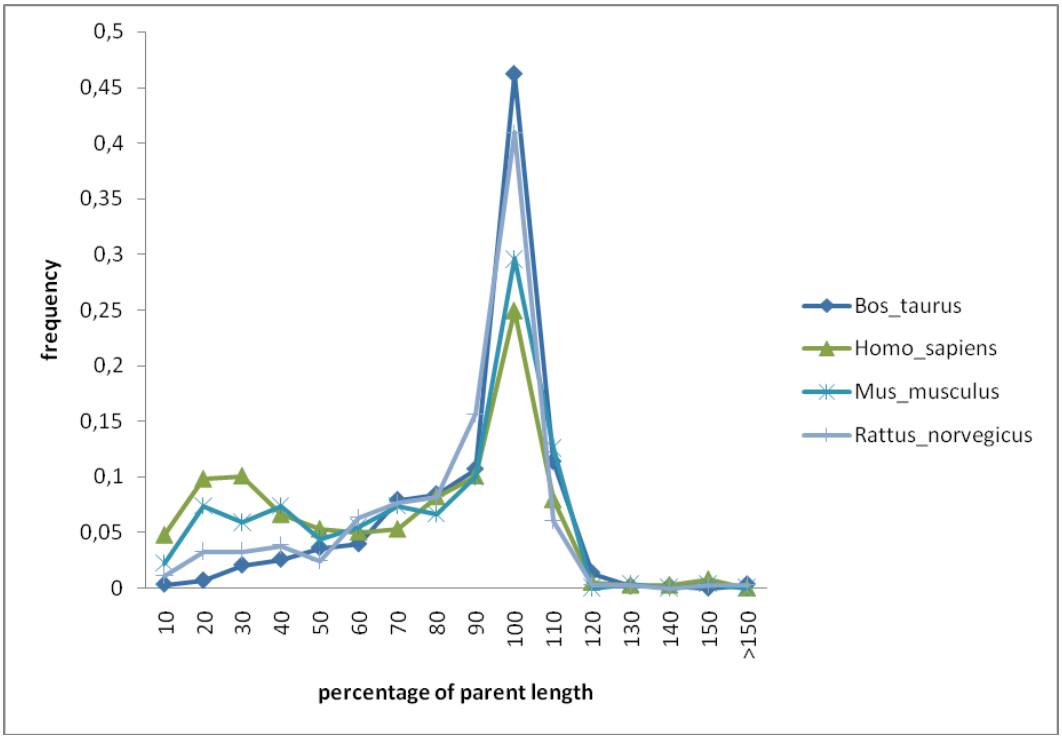


Figure 2.5: Distributions of fraction of length of parent exon for ΨEs. The bin label x is for all values such that, x-10.0 < value = x.

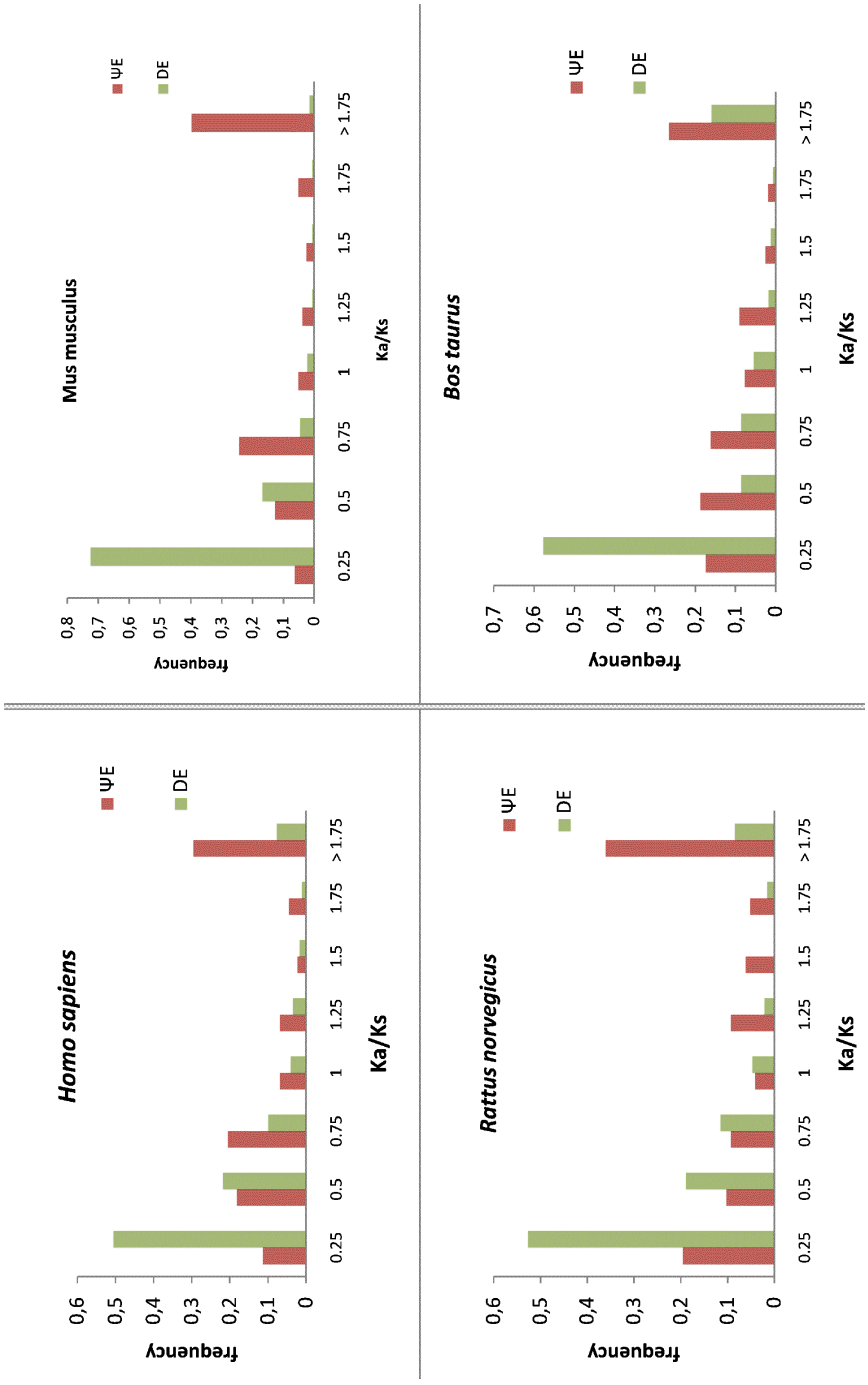


Figure 2.6: Histograms of Ka/Ks for for DEs and ΨEs. The DE histogram is green, and the ΨE histogram is red. The bin label x is for all values such that, x-0.25 < value $\leq x$.

Supplementary Table S2.1: Statistics of exon length. Statistics of exon length in human, mouse, rat and cow.

	human	human	mouse	mouse	rat	rat	cow	cow
	DE	ΨE	DE	ΨE	DE	ΨE	DE	ΨE
median	42	89	42	79	42	105	43	106
s.d.	39	151	39	138	43	139	44	116

The median and standard deviation values for the length of each exon type in nucleotides. The overall mean and s.d. values (over all species) are given below.

	Mean	S.D (aa)	S.D.(nt)
DE	295	42,14286	126,4286
ΨE	1040	148,5714	445,7143

Supplementary Table S2.2: Table showing the most abundant protein domains for exon types.

	ΨE†				DE†			
	(1) Pfam ID	(2) Family name	(3) % of total	(4) Number of occurrence	(5) Pfam ID	(6) Family name	(7) % of total	(8) Number of occurrence
Bos taurus	PF00096	zf-C2H2	12.5	36	PF00008	EGF	4.9	85
	PF01352	KRAB	9.7	28	PF00096	zf-C2H2	3.7	64
	PF07686	V-setT	5.9	17	PF00047	Ig	3.2	55
	PF00047	Ig	5.2	15	PF07686	V-setT	2.9	50
	PF07679	I-set	2.0	6	PF07645	EGF_CA	2.5	43
Homo sapiens	PF00096	zf-C2H2	16.6	72	PF00008	EGF	4.4	120
	PF01352	KRAB	11.3	49	PF00096	zf-C2H2	3.5	97
	PF07686	V-setT	9.7	42	PF00047	Ig	3.4	93
	PF00047	Ig	9.4	41	PF07686	V-setT	3.1	85
	PF00201	UDPGT	2.5	11	PF00041	fn3	2.9	79
Mus musculus	PF00096	zf-C2H2	14.1	39	PF00008	EGF	4.5	107
	PF01352	KRAB	10.5	29	PF00047	Ig	3.6	85
	PF00047	Ig	7.6	21	PF00096	zf-C2H2	3.0	72
	PF07686	V-setT	7.6	21	PF07686	V-setT	3.0	71
	PF00001	7tm_1	2.9	8	PF07679	I-set	2.7	65
Rattus norvegicus	PF00047	Ig	10.5	39	PF00008	EGF	4.5	99
	PF07686	V-setT	10.0	37	PF00047	Ig	3.7	81
	PF00096	zf-C2H2	5.4	20	PF00096	zf-C2H2	3.7	81
	PF01352	KRAB	3.8	14	PF07686	V-setT	3.6	79
	PF04822	DUF622	3.8	14	PF07645	EGF_CA	2.6	57

† The domains are counted up on a 'one per sequence' basis. (3) % of total occurrences.

Supplementary Table S2.3: Table of splice forms. Cases of alternative recruitment from different parts of apparently *pseudogenic exons*.

Ensembl Gene	Ensembl Transcript	# of Splice forms	Coordinates (genomic) *	ΨE coordinates	EST name
ENSG00000174652	ENST00000361451	1	**9384637..9384722** 9385595..9385755 9386044..9386162 9399612..9399765 9400531..9400779 9400935..9401068 9401199..9401325 9401922..9402002 9404088..9404176 9405110..9407226	9384640..9384968	AY376240
		4	**9384680..9384722** 9384971..9385032 9385595..9385755 9386047..9386162 9399612..9399765 9400541..9400652	9384640..9384968	CN353236
ENSG00000167766	ENST00000301096	1	**57807473..57807516** 57810675..57810762 57826766..57826892 57831031..57831301	57807379..57808235	DA162889 DA592481 DA347859
		3	**57807427..57807516** 57810675..57810762 57831031..57831343	57807379..57808235	DA762987
			57807443..57807516 57810624..57810762 57831031..57831542	57807379..57808235	DA731895 AK027518
ENSG00000094631	ENST00000376619	1	48545431..48545493 48546014..48546136 48546222..48546350 48546479..48546567 48548789..48548873 48548982..48549022 48549719..48549815 48549960..48550057 48551384..48551488 48551609..48551677 48557791..48557917 48558026..48558091 **48558185..48558244** 48558313..48558402 48558735..48558841 48558926..48558993 48559235..48559403 48559492..48559620 48559816..48559984 48560677..48560810 48561392..48561460 48561571..48561763 48563457..48563606 48565974..48566148 48566266..48566942	48558186..48558403	AJ011972 AB020708

			48567026..48567139 48567276..48567422 48567520..48567648 48567898..48568324		
ENSG00000094631	ENST00000376619	4	48557859..48557917 **48558026..48558244** 48558313..48558841 48558926..48558993 48559235..48559403 48559492..48559620 48559816..48559984 48560677..48560810 48561392..48561460 48561571..48561763 48563457..48563606 48565974..48566148 48566266..48566942 48567026..48567139 48567276..48567422 48567520..48567648 48567898..48568324	48558186..48558403	AK122954
		8	48558116..48558402 48558735..48558841 48558926..48558993 48559235..48559403 48559492..48559527	48558186..48558403	BX644567
ENSG00000158691	ENST00000361028	1	28454716..28454788 **28455972..28456441** 28461399..28461543 28462703..28462859	28455962..28456525	BQ229288
		4	**28455969..28456441** 28461365..28461543 28462703..28462873	28455962..28456525	BQ921129
		6	28454738..28454788 **28455972..28456477**	28455962..28456525	CN307911

* In bold are the parts of the ΨEs retained by particular transcripts.

Supplementary Table S2.4: Examination of buffer dependence. Number of ΨEs detected, as a function of 5'/3' buffer size in the cow genome

Buffer size (in nucleotides)	Number of ΨEs detected
0	484
1000	517
2000	533
3000	541
4000	553
5000	581

CHAPTER III

LITERATURE REVIEW

Alternative Splicing

Mammalian genes are complex in their makeup, typically composed of an average of eight exons, interspaced between much larger introns. Splicing is the highly regulated and precise process by which the introns are removed from the pre-mRNA, and the exons then linked together. In a given gene, constitutively spliced exons are common and essential to most of the transcripts, whereas alternatively spliced exons allow for flexibility, by way of inclusion of additional exons, elongated or shortened exons, or retention of intronic sequence, as well as excluding (skipping) exons, in often a tissue-specific or developmentally discrete manner (MANIATIS and TASIC 2002). Alternative splicing in humans occurs for up to 75% of all genes (JOHNSON *et al.* 2003), allowing a greatly expanded genomic repertoire of transcripts that, therefore, increases the proteomic diversity. Alternative splicing events can be categorized in four major types: 1- Cassette alternative exon, 2- alternative 5'and 3' sites, 3- intron retention and 4- mutually exclusive alternative exons (TAZI *et al.* 2009). While almost all protein-coding genes contain introns that are removed in the nucleus by RNA splicing during pre-mRNA processing, exon usage is often alternative, i.e. the cell decide whether to remove a part of the pre-mRNA as an intron or include this part in the mature mRNA as an alternative exon. When part of the pre-mRNA is present in all (or at

least in the majority) of the transcripts this exon is termed constitutive (BEN-DOV *et al.* 2008).

Basic splicing machinery

The general mechanism of splicing is carried out by the spliceosome, a large RNP (ribonucleoprotein) complex that is composed of five snRNPs (small nuclear RNPs) particles, U1, U2, U4, U5 and U6, each approximately 200 nucleotides in length, and up to 300 structural proteins. The spliceosome assembly proceeds in a stepwise fashion and is initiated upon the binding of U1 snRNP to the 5' splice site, SF1/mBBP to the branch point sequence and U2 auxiliary factor (U2AF) to the pyrimidine tract and 3' AG to form the E, or early, complex. Subsequently, the E complex recruits the branch-point-binding factor U2 snRNP and is converted into the A complex. The association of the U4/U6•U5 tri-snRNP forms B complex, which undergoes a structural transformation to form the catalytic C complex. Splicing is finally accomplished with the ligation of the exons (GRAVELEY 2000).

***Cis* genetic motifs enhance splicing specificity**

In mammals, exons are usually short (50—300 bp), while introns are much longer, ranging from less than 100 nucleotides to more than one hundred thousand nucleotides. Sequences that match the consensus of splice signals are common in

introns. The spliceosome complex activity relies on *cis* genetic motifs (RNA sequences) to control the alternative or constitutive splicing. The most important splicing motifs are the 5' ss and 3' ss, named in relation to their respective intronic positions. The 5' ss has as its core the conserved motif of an exonic AG followed by an intronic GURAGU. The 3' ss is composed of the branch point (YNYURAC), the polypyrimidine tract and the conserved splice site itself, YAG, adjacent to the exon boundary. These motifs are essential to splicing and determine the strength of intronic splice signals which, in turn, define if an intron is constitutively or alternatively spliced. Despite their importance, these sequences can be quite degenerate, and it is possible to find many cases of these motifs in the genome that are not describing an exon (JENSEN *et al.* 2009).

In addition to 5' ss and 3' ss motifs, several auxiliary splice signals have been identified that play an important role in the selection of splice sites. These elements are classified according to their location and their function as exonic splicing enhancers or silencers (ESE or ESS), and intronic splicing enhancers or silencers (ISE or ISS) (SCHAAL and MANIATIS 1999; WANG *et al.* 2004). These ESEs and ISEs provide binding sites for several *trans* factors that promote exon inclusion (LONG and CACERES 2009). Conversely, ESSs and ISSs provide binding sites for splicing suppressors, which include members of the hnRNP (heterogeneous nuclear RNP) family of proteins (WANG *et al.* 2006).

The genetic context of these motifs is of great importance to splicing control, in particular in relation to the exons. For example, ESEs tend to be clustered around splice sites and are of particularly high density within constitutive exons. The location of ESSs also helps define exon boundaries and affects splice site choices. When located within an exon, ESSs can inhibit exon inclusion and when located between two 5' or 3' alternative splice sites of similar strength, ESSs will inhibit the intron-proximal splice site, resulting in the shorter exon. Additionally, traditional ESS sequences, when located within an intron, can promote intron excision (BERGET 1995).

Trans splicing factors

Many *trans* splicing factor are known to bind to *cis* genetic motifs, with many factors exhibiting a tissue-specific expression pattern, while others having a more ubiquitous expression. The SR family includes a large number of members ubiquitously expressed splicing factors (SCHAAL and MANIATIS 1999). SR proteins consist of one or two N-terminal RNA recognition motifs (RRMs) and a variable-length C-terminal arginine/serine-rich domain (RS domain). Family members have slightly different sequence specificities and bind predominantly to ESEs where they influence the splicing of the exon in which they are located. In some cases the SR protein can act as an ISS (JENSEN *et al.* 2009).

The predominantly negative regulator, hnRNPs, are a diverse family of approximately 20 ubiquitously expressed major proteins including the important members hnRNP A1 and PTB (polypyrimidine tract-binding protein)/hnRNP I. This protein family can be characterized by containing an RNA recognition motif and a KH (hnRNP K homology) domain or an arginine-glycine box. The hnRNPs acts in direct competition with SRs and other splicing regulators for biding sites (HE and SMITH 2009).

Alternative splice conservation

Several studies provide evidences that the number of alternative splice (AS) events per gene is inversely correlated with gene or paralog copy number, indicating that gene duplication, followed by neo- or sub-functionalization, may reduce the pressure to diversify genes functions by AS (SU *et al.* 2006). AS events seems to occur more frequently in transcripts from genes expressed in functionally more complex tissues with diverse cell types, such as brain and testis, or from genes expressed within cell types that have undergone selection to provide specialized functions such as the cells of the immune system (MODREK *et al.* 2001). Documentation of AS events in brain is vast, and many of these events are associated with very complex processes such as the control of synaptic

plasticity, linked to cognition and other neural processes (LIPSCOMBE 2005; ULE and DARNELL 2006).

Comparisons of pairs of orthologs between human and mouse revealed that only 10%-20% of cassette-type AS events are conserved between these species (MODREK and LEE 2003; PAN *et al.* 2005; SOREK *et al.* 2004). The remaining 80%-90% of human- and mouse-specific cassette-type AS event can be separated into two categories: 1- those involving relatively recently-gained 'genome-specific' exons, which tend to be included at low levels in spliced mRNA (also referred to as "minor-form" variants) (MODREK and LEE 2003), and 2- those involving exons that are conserved but alternatively spliced in only one species, which are generally skipped at low levels in spliced mRNA (PAN *et al.* 2005).

The sequence of exons and flanking intron regions of AS events present in human and mouse are on average significantly more conserved than those of both species-specific AS and constitutive splicing events. Conserved AS events also more frequently preserve open reading frames (RESCH *et al.* 2004; SOREK *et al.* 2004). Inter-species comparisons also support the conclusion that conserved alternative exons and their flanking introns are strongly enriched in splicing regulatory elements (YEO *et al.* 2005). Furthermore, these conserved AS events

undergo differential regulation between tissues more often than AS events that are species-specific (XING and LEE 2005).

Alternative splicing and diseases

There are only a few reports of mutation in the core elements of splicing machinery leading to human diseases. It is possible that such mutations are not compatible with life. One example of mutation in the core splicing machinery leading to a disease is an autosomal dominant form of retinitis pigmentosa caused by a mutation in the splicing factor PRPF31/U4-61k and PRP8 (BOON *et al.* 2007; VITHANA *et al.* 2001). On the other hand, mutations changing alternative splicing can be tolerated by an organism, although these changes often manifest in a disease (TAZI *et al.* 2009).

It has recently been estimated that up to 60% of disease-causing point mutations cause aberrant splicing. Changes in splicing regulatory elements can lead to exon skipping, intron retention, creation of ectopic splice site or activation of cryptic sites (BLENCOWE 2006).

Many diseases associated with changes in splicing have been described. For example, spinal muscular atrophy (SMA) is caused by the loss of the SMN1 gene that encodes the SMN protein, which regulates snRNP assembly. Humans

possess a gene, SMN2, which is almost identical to SMN1. SMN2 was generated through a recent duplication. Although both genes are almost identical in sequence, due to a translationally silent C→T change at position 6 in exon 7, they have different splicing patterns and exon 7 is predominantly excluded in SMN2. This exon-skipping event generates a truncated, less stable and probably non-functional protein. Therefore, SMN2 cannot compensate the loss of SMN1 (WIRTH *et al.* 2006).

Other diseases are associated with changes of the ratio of protein isoforms generated by alternative splicing. A mutation in the exon 10 of the *tau* gene changes its normal fraction of inclusion and is associated with frontotemporal dementia with parkinsonism linked to chromosome 17 (FTDP-17) (GALLO *et al.* 2007).

Familial dysautonomia (FD) is a disease characterized by abnormal development of the nervous system that is associated with demyelination in various regions. In more than 99.5% of the FD patients the 5' splice site of the exon 20 is mutated T→C in position 6 of intron 20. This point mutation interrupts base pairing with U1snRNA causing exon skipping (CARMEL *et al.* 2004).

Alternative splicing patterns are also altered in cancer. Changes in expression or creation of tumor-specific splice variants significantly affects cellular processes critical for cancer biology (SKOTHEIM and NEES 2007).

Mutations in splicing of *trans* or *cis* regulatory elements can lead to non-functional tumor suppressors, which, therefore, predisposes to cancer. For example, mutations in SR proteins, including SC35 and ASF/SF2, have been observed in human ovarian cancer and in mouse models of breast cancer development (FISCHER *et al.* 2004). Double exon skipping in MLH1 gene gives rise to hereditary nonpolyposis colorectal cancer (HPCC). Colorectal cancer is associated with two mutations that disrupt splicing regulatory elements causing exon skipping in the gene APC (NEKLASON *et al.* 2004; TANKO *et al.* 2002).

Nonsense-Mediated Decay

Eukaryotes have evolved a quality-control mechanism that allows cells to degrade mRNAs with premature nonsense codons. This process is called nonsense-mediated mRNA decay (NMD), and eliminates mRNA that might otherwise produce truncated proteins with dominant negative effects (MAQUAT 2004). The core apparatus of NMD is phylogenetically conserved from yeast to humans, and this implies that it has a fundamental role in the cell. It has been difficult to justify such a well-conserved and sophisticated machinery solely to protect the cell from relatively infrequent mutations. Indeed, several evidences point to NMD as a pathway that governs general cellular gene regulation (LEWIS *et al.* 2003; MENDELL *et al.* 2004; MITROVICH and ANDERSON 2000).

NMD is non-essential in yeast and worms. Although in the former, NMD-deficient cells experience an accelerated senescence, related to changes in telomere silencing and shortening of telomeres (Lew et al. 1998), while in the latter NMD affects transcripts encoding SR splicing proteins (MORRISON *et al.* 1997). In mice, however, NMD deficiency leads to embryonic death at the implantation stage (MEDGHALCHI *et al.* 2001).

In mammalian cells, translation termination codons and exon–exon junctions are *cis*-acting elements that allow recognition of premature termination codons (PTCs). The mRNA is subject to rapid decay when there is a PTC more than approximately 50–55 nucleotides upstream of the last exon–exon junction (HOLBROOK *et al.* 2004). The importance of the exon–exon junction to NMD reflects the splicing-dependent deposition of an exon junction complex (EJC) of proteins, including UPF2 and UPF3 or UPF3X NMD factors, ~20–25 nucleotides upstream of exon–exon junctions. NMD depends on translation, which is required for nonsense codon recognition. NMD also depends on the so-called SURF complex, which consists of the UPF1 kinase SMG1 (named after its ortholog in *Caenorhabditis elegans* 'suppressor with morphogenetic effects on genitalia'), the UPF2 factor, the eukaryotic release factors eRF1 and eRF3. When translation terminates more than 50–55 nucleotides upstream of an exon–exon junction UPF1 of the SURF complex interacts with UPF2 of the EJC to trigger NMD. The steps

that take place after this interaction and before decay remain to be elucidated (KASHIMA *et al.* 2006; LE HIR *et al.* 2000; MAQUAT 2004).

Alternative splicing-coupled NMD

Sequence-base predictions have revealed that more than one-third of AS events have the potential to introduce a PTC, that could target the mRNA for nonsense-mediated mRNA decay (LEWIS *et al.* 2003). Quantitative AS microarray profiling experiments showed that PTC-introducing AS events are enriched in genes that encode core spliceosomal proteins (SALTZMAN *et al.* 2008).

Human SR and hnRNP genes autoregulate their expression by means of tissue-regulated unproductive splicing. In this mechanism, the gene expression is downregulated by the production of alternative splice forms, which bear a PTC that targets the mRNAs for NMD (LAREAU *et al.* 2007; LEWIS *et al.* 2003; YEO *et al.* 2005). For example, both SRp20 (*SFRS3*) and Tra2-beta (*SFRS10*), when overexpressed, activate the inclusion of stop-codon exons (also referred as 'poison cassettes') in their own pre-mRNAs, and the resulting splice forms are subjected to NMD (STOILOV *et al.* 2004).

Ultra- and highly-conserved elements are also associated with poison cassette exons. This unusual conservation of elements that trigger NMD in certain

genes, such as splicing factors, could be due to a strong functional demand for combined AS-NMD autoregulation and cross-regulation (NI *et al.* 2007). A recent study in *Drosophila melanogaster* found about 45 genes that can be spliced either into an mRNA that encodes a functional protein, or into an mRNA targeted for NMD. These genes participate in several cell regulatory processes such as translation, RNA splicing, and cell cycle progression (HANSEN *et al.* 2009).

NMD and disease

The majority of the nonsense-associated diseases are caused either by insufficient level of the functional protein, as a result of degradation of a PTC-containing mRNA, or by generation of a defective truncated protein from a PTC-containing mRNA that escaped the NMD surveillance (HOLBROOK *et al.* 2004; TAZI *et al.* 2009). The first example of NMD involved in a disease was the human form of β-thalassemia. In this case, a 5' PTCs in the β-globin gene result in a recessive trait, whereas 3' PTCs can result in an atypical dominant form of disease, because 5' PTCs but not 3' PTCs trigger β-globin NMD (BASERGA and BENZ 1988). Similar effects of NMD are found in a number of human disorders such as myotonia congenital (PUSCH 2002), retinal degeneration (ROSENFELD *et al.* 1992), Robinow syndrome (PATTON and AFZAL 2002), and brachydactyly-type B (SCHWABE *et al.* 2000).

NMD can also modulate disease by producing milder phenotypes when compared with those produced by a truncated protein that escaped NMD. For instance, mis-sense mutations associated with the gene encoding the *a* chain of type XI collagen (COL1A1) and type II collagen (COL2A1) are classic examples of dominant-negative alleles, as they disrupt the hetero- or homo-trimer conformation of collagen subunits and are associated with severe disease phenotypes of osteogenesis imperfecta (OI) type II–IV and spondylepiphyeal dysplasia, respectively. In contrast, all truncating mutations in both genes have been shown to result in haploinsufficiency owing to NMD and are associated with the milder clinical phenotypes of OI type I by 5' PTC COL1A1 and Stickler syndrome by 5' PTC COL2A1(KORKKO *et al.* 1998; PARKE 2002). In contrast, PTCs targeted for NMD are associated with a severe form of Duchene muscular dystrophy (DMD) when compared with a non-NMD targeting 3' end mutation of the dystrophin gene (KERR *et al.* 2001).

In theory, the disruption of NMD might also be a logical means for therapy. In reality, however, inhibition of NMD might result in undesirable side effects as it may do more harm than the disease itself, by the production of other mutant proteins, or transcription and translation of pseudogenes, which would be otherwise degraded (MENDELL *et al.* 2004). In *C. elegans*, the altered expression of splicing factors after NMD inhibition affected the splicing of numerous genes,

such as splicing factors. One of these splicing factors, Squid, is known to affect the splicing of at least 255 other genes (BARBERAN-SOLER and ZAHLER 2008).

Linking statement II

In Chapter II, I studied a decaying population of duplicated *pseudogenic exons* (ΨEs) in four mammalian genomes. I showed that this population of ΨEs belongs to a few types of protein family, have preference for the 5' of its gene and can be alternatively spliced. I also verified that upon inclusion in the transcript a large number of ΨEs could target its mRNA to degradation by nonsense-mediated decay.

In Chapter IV, I have annotated alternatively-spliced mRNA targets for nonsense-mediated decay (AS-NMD) in mammalian genomes. I have analyzed the contribution of duplicated ΨEs for the formation of AS-NMD. I discuss the role of NMD in mRNA quality control as well as in regulating gene expression through inclusion of a 'poison cassette'. I also describe the association of these alternative splicing forms with specific protein families and pathways, and their implications for human genetic diseases.

References

BARBERAN-SOLER, S., and A. M. ZAHLER, 2008 Alternative splicing regulation during C. elegans development: splicing factors as regulated targets. PLoS Genet **4:** e1000001.

BASERGA, S. J., and E. J. BENZ, JR., 1988 Nonsense mutations in the human beta-globin gene affect mRNA metabolism. Proc Natl Acad Sci U S A **85:** 2056-2060.

BEN-DOV, C., B. HARTMANN, J. LUNDGREN and J. VALCARCEL, 2008 Genome-wide analysis of alternative pre-mRNA splicing. J Biol Chem **283:** 1229-1233.

BERGET, S. M., 1995 Exon recognition in vertebrate splicing. J Biol Chem **270:** 2411-2414.

BLENCOWE, B. J., 2006 Alternative splicing: new insights from global analyses. Cell **126:** 37-47.

BOON, K. L., R. J. GRAINGER, P. EHSANI, J. D. BARRASS, T. AUCHYNNIKAVA *et al.*, 2007 prp8 mutations that cause human retinitis pigmentosa lead to a U5 snRNP maturation defect in yeast. Nat Struct Mol Biol **14:** 1077-1083.

CARMEL, I., S. TAL, I. VIG and G. AST, 2004 Comparative analysis detects dependencies among the 5' splice-site positions. Rna **10:** 828-840.

FISCHER, D. C., K. NOACK, I. B. RUNNEBAUM, D. O. WATERMANN, D. G. KIEBACK *et al.*, 2004 Expression of splicing factors in human ovarian cancer. Oncol Rep **11:** 1085-1090.

GALLO, J. M., W. NOBLE and T. R. MARTIN, 2007 RNA and protein-dependent mechanisms in tauopathies: consequences for therapeutic strategies. Cell Mol Life Sci **64:** 1701-1714.

GRAVELEY, B. R., 2000 Sorting out the complexity of SR protein functions. Rna **6:** 1197-1211.

HANSEN, K. D., L. F. LAREAU, M. BLANCHETTE, R. E. GREEN, Q. MENG *et al.*, 2009 Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. PLoS Genet **5:** e1000525.

HE, Y., and R. SMITH, 2009 Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B. Cell Mol Life Sci **66:** 1239-1256.

HOLBROOK, J. A., G. NEU-YILIK, M. W. HENTZE and A. E. KULOZIK, 2004 Nonsense-mediated decay approaches the clinic. Nat Genet **36:** 801-808.

JENSEN, C. J., B. J. OLDFIELD and J. P. RUBIO, 2009 Splicing, cis genetic variation and disease. Biochem Soc Trans **37:** 1311-1315.

JOHNSON, J. M., J. CASTLE, P. GARRETT-ENGELE, Z. KAN, P. M. LOERCH *et al.*, 2003 Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302:** 2141-2144.

KASHIMA, I., A. YAMASHITA, N. IZUMI, N. KATAOKA, R. MORISHITA *et al.*, 2006 Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. Genes Dev **20:** 355-367.

KERR, T. P., C. A. SEWRY, S. A. ROBB and R. G. ROBERTS, 2001 Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? Hum Genet **109:** 402-407.

KORKKO, J., L. ALA-KOKKO, A. DE PAEPE, L. NUYTINCK, J. EARLEY *et al.*, 1998 Analysis of the COL1A1 and COL1A2 genes by PCR amplification and scanning by conformation-sensitive gel electrophoresis identifies only COL1A1 mutations in 15 patients with osteogenesis imperfecta type I: identification of common sequences of null-allele mutations. Am J Hum Genet **62:** 98-110.

LAREAU, L. F., M. INADA, R. E. GREEN, J. C. WENGROD and S. E. BRENNER, 2007 Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature **446:** 926-929.

LE HIR, H., E. IZAURRALDE, L. E. MAQUAT and M. J. MOORE, 2000 The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. Embo J **19:** 6860-6869.

LEW, J. E., S. ENOMOTO and J. BERMAN, 1998 Telomere length regulation and telomeric chromatin require the nonsense-mediated mRNA decay pathway. Mol Cell Biol **18:** 6121-6130.

LEWIS, B. P., R. E. GREEN and S. E. BRENNER, 2003 Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A **100:** 189-192.

LIPSCOMBE, D., 2005 Neuronal proteins custom designed by alternative splicing. Curr Opin Neurobiol **15:** 358-363.

LONG, J. C., and J. F. CACERES, 2009 The SR protein family of splicing factors: master regulators of gene expression. Biochem J **417:** 15-27.

MANIATIS, T., and B. TASIC, 2002 Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature **418:** 236-243.

MAQUAT, L. E., 2004 Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol **5:** 89-99.

MEDGHALCHI, S. M., P. A. FRISCHMEYER, J. T. MENDELL, A. G. KELLY, A. M. LAWLER *et al.*, 2001 Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. Hum Mol Genet **10:** 99-105.

MENDELL, J. T., N. A. SHARIFI, J. L. MEYERS, F. MARTINEZ-MURILLO and H. C. DIETZ, 2004 Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat Genet **36:** 1073-1078.

MITROVICH, Q. M., and P. ANDERSON, 2000 Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes Dev **14:** 2173-2184.

MODREK, B., and C. J. LEE, 2003 Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet **34:** 177-180.

MODREK, B., A. RESCH, C. GRASSO and C. LEE, 2001 Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res **29:** 2850-2859.

MORRISON, M., K. S. HARRIS and M. B. ROTH, 1997 smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. Proc Natl Acad Sci U S A **94:** 9782-9785.

NEKLASON, D. W., C. H. SOLOMON, A. L. DALTON, S. K. KUWADA and R. W. BURT, 2004 Intron 4 mutation in APC gene results in splice defect and attenuated FAP phenotype. Fam Cancer **3:** 35-40.

NI, J. Z., L. GRATE, J. P. DONOHUE, C. PRESTON, N. NOBIDA *et al.*, 2007 Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev **21:** 708-718.

PAN, Q., M. A. BAKOWSKI, Q. MORRIS, W. ZHANG, B. J. FREY *et al.*, 2005 Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet **21:** 73-77.

PARKE, D. W., 2002 Stickler syndrome: clinical care and molecular genetics. Am J Ophthalmol **134:** 746-748.

PATTON, M. A., and A. R. AFZAL, 2002 Robinow syndrome. J Med Genet **39:** 305-310.

PUSCH, M., 2002 Myotonia caused by mutations in the muscle chloride channel gene CLCN1. Hum Mutat **19:** 423-434.

RESCH, A., Y. XING, A. ALEKSEYENKO, B. MODREK and C. LEE, 2004 Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic Acids Res **32:** 1261-1269.

ROSENFELD, P. J., G. S. COWLEY, T. L. MCGEE, M. A. SANDBERG, E. L. BERSON *et al.*, 1992 A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet **1:** 209-213.

SALTZMAN, A. L., Y. K. KIM, Q. PAN, M. M. FAGNANI, L. E. MAQUAT *et al.*, 2008 Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol Cell Biol **28:** 4320-4330.

SCHAAL, T. D., and T. MANIATIS, 1999 Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol Cell Biol **19:** 1705-1719.

SCHWABE, G. C., S. TINSCHERT, C. BUSCHOW, P. MEINECKE, G. WOLFF *et al.*, 2000 Distinct mutations in the receptor tyrosine kinase gene ROR2 cause brachydactyly type B. Am J Hum Genet **67:** 822-831.

SKOTHEIM, R. I., and M. NEES, 2007 Alternative splicing in cancer: noise, functional, or systematic? Int J Biochem Cell Biol **39:** 1432-1449.

SOREK, R., R. SHAMIR and G. AST, 2004 How prevalent is functional alternative splicing in the human genome? Trends Genet **20:** 68-71.

STOILOV, P., R. DAOUD, O. NAYLER and S. STAMM, 2004 Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. Hum Mol Genet **13:** 509-524.

SU, Z., J. WANG, J. YU, X. HUANG and X. GU, 2006 Evolution of alternative splicing after gene duplication. Genome Res **16:** 182-189.

TANKO, Q., B. FRANKLIN, H. LYNCH and J. KNEZETIC, 2002 A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. Mutat Res **503:** 37-42.

TAZI, J., N. BAKKOUR and S. STAMM, 2009 Alternative splicing and disease. Biochim Biophys Acta **1792:** 14-26.

ULE, J., and R. B. DARNELL, 2006 RNA binding proteins and the regulation of neuronal synaptic plasticity. Curr Opin Neurobiol **16:** 102-110.

VITHANA, E. N., L. ABU-SAFIEH, M. J. ALLEN, A. CAREY, M. PAPAIOANNOU *et al.*, 2001 A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). Mol Cell **8:** 375-381.

WANG, Z., M. E. ROLISH, G. YEO, V. TUNG, M. MAWSON *et al.*, 2004 Systematic identification and analysis of exonic splicing silencers. Cell **119:** 831-845.

WANG, Z., X. XIAO, E. VAN NOSTRAND and C. B. BURGE, 2006 General and specific functions of exonic splicing silencers in splicing control. Mol Cell **23:** 61-70.

WIRTH, B., L. BRICHTA and E. HAHNEN, 2006 Spinal muscular atrophy and therapeutic prospects. Prog Mol Subcell Biol **44:** 109-132.

XING, Y., and C. J. LEE, 2005 Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genet **1:** e34.

YEO, G. W., E. VAN NOSTRAND, D. HOLSTE, T. POGGIO and C. B. BURGE, 2005 Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A **102:** 2850-2855.

CHAPTER IV

Large-scale evidence for a conserved functional role for NMD candidates across mammals

Abstract

Alternatively-spliced (AS) forms can vary protein function, intracellular localization and post-translational modifications. AS coupled with mRNA nonsense-mediated decay (NMD) can also control the abundance of gene transcripts. In the present study we investigated the conservation of alternatively-spliced NMD candidates (AS-NMD candidates). We mapped more than 16 million ESTs/cDNAs/mRNAs against genomic sequences and annotated AS-NMD candidates generated by in-frame premature termination codons (PTCs) in four mammalian genomes: human, mouse, rat and cow. In these genomes, we found populations of genes that harbour AS-NMD candidates, varying in number from ~149 to 2051 genes. We verified that a highly-significant proportion (27%-35%) of AS-NMD candidates in mouse, rat and cow, also have orthologs targeted for NMD in human. Intron retention was the most abundant type of alternative splicing, ranging from 43%-67% of all genes harboring an AS-NMD candidate. We found that these retained introns possess features that distinguish them from non-retained introns, such as, small size, codon-usage, high CG content and highly significant conservation in other mammals. The AS-NMD candidates also showed a similar pattern of gene ontology enrichment in all four species. Genes linked to nucleic acid interaction and cell fate (apoptosis), and involved in pathways linked with cancer, were the most abundant to produce AS-NMD

candidates. We mapped the AS-NMD candidates to mass spectrometry-derived proteomics data and gathered evidence of translation of at least 10% of all AS-NMD candidates in human genome. Furthermore, in one of the AS-NMD candidates, the gene RPS13, we found evidences of a conserved mechanism of autoregulation, between human and cow, in which the protein inhibits the excision of one of its introns, thus creating a splice form that is targeted for NMD. Alternative splicing coupled with NMD seems to act on a wide range of genes, but, only in a few significantly over-represented functions across mammalian genomes. Our data provides strong statistical evidence for a regulatory role conserved across *Mammalia* for a large subset of NMD candidates.

Introduction

Alternative splicing is the generation of multiple transcripts from a single protein-coding gene. Several studies have shown that at least half of all human genes undergo alternative splicing [1, 2], although RNA array and sequencing data shows that this number could be as high as 94% of the genes [3]**.** Alternatively-spliced (AS) forms can vary protein function, intracellular localization and post-translational modifications. Although, in most cases, protein AS isoforms show only small differences, some isoforms can lead to large functional variances and even to genetic disorders [4]. Apparently, alternative splicing can also control transcript abundance, due to its coupling to mRNA nonsense-mediated decay (NMD). It seems that at least 10-15% of human transcripts can be switched off by NMD coupled to alternative splicing [5].

Nonsense-mediated decay is a eukaryotic surveillance mechanism that detects mRNAs that harbour premature termination codons (PTCs), and commits these transcripts to rapid decay. One of the expected physiological consequences of NMD is the prevention of synthesis of truncated polypeptides, which presumably leads to protection of the cell from resultant deleterious dominant-negative or gain-of-function effects [6] [7]. The biological importance of NMD is highlighted by the fact that 30% of inherited genetic diseases are due to PTC,

most of which are in NMD targets [8]. Therefore, the majority of the nonsense-associated diseases are caused either by an insufficient level of the functional protein as a result of degradation of a PTC-containing mRNA, or by generation of a defective truncated protein from a PTC-containing mRNA that escaped the NMD surveillance [4, 8].

In mammalian cells, translation termination codons and exon–exon junctions are *cis*-acting elements that allow recognition of PTCs. The mRNA is subject to rapid decay when there is a PTC more than ~50–55 nucleotides upstream of the last exon–exon junction [8]. The positional information that marks exon–exon junctions is provided by the components of the splicing-dependent exon junction complex (EJC) that persist during export and until the mRNA is translated [9]. NMD is believed to require three core UPF factors which are conserved from yeast to humans, namely, UPF1 (also known as regulator of nonsense transcripts 1 [RENT1]), UPF2 (RENT2), and UPF3 [10]. UPF1 is not only essential to NMD but is also required for rapid degradation of histone mRNAs [11].

In the present study, we have annotated and analyzed the occurrence and conservation of AS-NMD candidates in genes, generated by in-frame premature stop codons, in four mammalian genomes: human, mouse, rat and cow. Strikingly, we find highly significant levels of conservation, and also demonstrate that genes

containing an AS-NMD candidate are significantly associated with specific functional categories of proteins, and of disease-associated genes.

Results and Discussion

AS-NMD annotation overview

To assess the existence in public databases of alternatively-spliced forms that are putative substrates of nonsense-mediated mRNA decay (hereafter termed 'AS-NMD candidates'), we developed and applied a pipeline to four mammalian genomes (human, mouse, rat and cow). This pipeline is described in detail, in *Methods*. We focused only on AS-NMD candidates containing in-frame premature termination codons (PTCs). In brief, we mapped the complete corpus of >16 million mRNA/cDNA/EST sequences from several databanks (RefSeq, Unigene, dbEST and H-invitational) against genomic sequences for the four mammals (human, mouse, rat and cow), and compared these mappings to reference mRNA mappings for each gene.

The pipeline identified between ~149 (rat) and ~2051 (human) genes with AS-NMD candidates, per genome (Table 4.1). The number of supporting cDNA/ESTs varied between just 190 in rat, to 14,343 in human. The proportions

of genes with AS-NMD candidates are between 1% and 10% of AS-NMD candidates, in the analyzed genomes. However, it is important to be aware that, as for any analysis of transcript data, the major difference between several studies lies in the number of cDNA/ESTs considered. For example, Zhang *et al.* (2009 [12]), reported that, the number of predicted NMD candidates increased when these authors included predicted RefSeq ESTs in their analysis.

We classified the NMD candidates according to the type of events that they present: IR, intron retention; CSE, cassette exon; ASD, alternative splicing donor; ASA, alternative splicing acceptor (Table 4.2). Surprisingly, IR was the most abundant event in all genomes. It varied from 43% to 67% of all genes that harbour AS-NMD candidates. ASA was the second most abundant event (19%–32%). Due to the large number of IRs, we present below a set of analyses to verify whether or not—as a population—the IRs are artifacts derived from unprocessed or partially processed pre-mRNAs.

In a previous work, I studied duplicated *pseudogenic exons* (*i.e.,* ΨEs, exons disabled by frameshifts and premature termination codons) on the same four genomes of our present study [13]. In order to evaluate the contribution of exon duplication and pseudogenization to the production of AS-NMD candidates, we compared the list of genes bearing a ΨE and the list of AS-NMD candidates. We found that only 7 genes are common to both lists in human and 3 genes in

mouse. Thus the vast majority of AS-NMD candidates are not made through exon duplication, followed by disablement. We also cross-referenced our gene list with the Ensembl pseudogenes annotation (build 52) to verify if any of our genes were annotated exclusively as a pseudogene. No pseudogene was found in our list, indicating that they are all alternatively-spliced forms of protein-coding genes. We additionally compared our list of AS-NMD candidates with the general annotation of the human genome in the Ensembl database (www.ensembl.org). We found that a small fraction (321 genes, ~15%) in our list was also annotated as NMD-targeted transcripts by Ensembl.

Two models for AS-NMD candidates have recently been discussed, the 'spurious transcript' model, and the 'regulatory model' [12]. The existence of multiple splice forms in the genes with an AS-NMD candidate seems to be in agreement with the ‘spurious transcript’ model. In this model, the NMD function is to degrade costly-to-make and potentially toxic unwanted transcripts. Another prediction of this model is that AS-NMD should arise for rare transcripts, and be more common in recently-evolved exons; hence, it should not be conserved in other species. The alternative model is the ‘regulatory model’. In this model, the function of the NMD is to modulate gene expression. Although this model does not preclude multiple AS forms, it does not predict them. The regulatory model predicts that NMD substrates should be conserved among distant species [12]. Given the complexity of the coupling of alternative splicing with nonsense-

mediated decay, the two models seem to be valid and acting over different classes of genes.

Conservation of NMD candidates and the genes that harbour them

We extracted orthology information about genes from the Ensembl database. Each AS-NMD candidate found in one of the three other species (mouse, rat and cow) was used as a query to find the corresponding ortholog in the human genome. Only matches annotated as 'one2one' (*i.e.*, the transcript is bi-directionally the best match between the two genomes) were counted as true orthologs.

We examine the human conservation of genes that harbour AS-NMD candidates in the three other genomes (mouse, rat and cow) (Table 4.1). A large proportion of these genes have clear single orthologs in the human genome (87-91%). Of these orthologous human genes, 27-35% also harbour orthologous AS-NMD candidates (see the Supplementary Table S4.1 for the full list of the AS-NMD candidate orthologs in each species, and Supplementary Table S4.2 for a complete breakdown of the different types of AS-NMD and their conservation patterns).

The AS-NMD candidate orthologs also show a high percentage of sequence identity, at the protein level (84%-88%, on average) (Table 4.2). These findings are strongly supportive of the regulatory model of NMD, and imply that the conserved NMD candidates have a role in the control of gene expression. Several previously well-documented examples have lent support to this theory. For instance, the NMD control of splicing in the splicing regulator (SR) proteins, one of the most extensive studied cases of NMD controlling gene expression, is conserved across mammals and plants [14].

Our results seem to disagree with a previous study that found only ~8% AS-NMD candidates in a pair of orthologs between human and mouse [12]. However, a closer look shows that our data is comparable. The authors in this previous study used only RefSeq mRNAs in their analysis. When we considered only RefSeq mRNAs, in our analysis, the number of pair of orthologous NMD-candidates between human and mouse was 7%, in agreement with the previous study [12].

Intron Retention (IR) Analysis

Intron retention is one of the least studied AS forms. One of the major reasons for this is the difficulty of differentiating real IRs from artifacts derived

from unspliced or partially spliced pre-mRNA [15, 16]. However, previously it has been shown that retained introns seem to present sequence characteristics that distinguish them from non-retained introns and also from exons. These features include average size, GC content and codon usage. In some cases the retained intron can also encode a protein domain or part of a domain [16, 17].

Recent work has demonstrated that small introns without an in-frame stop codon are counter-selected [18]. Jaillon, *et. al.* (2009, [18]) found an under-representation of small stopless introns compared with retained introns that introduce a PTC. This counter-selection is likely because these stopless introns are capable of escaping NMD and therefore are translated as a protein with possible dominant-negative effects. Conversely, small in-frame PTC-containing retained introns would be targets for NMD. The average size of retained introns in our list of human AS-NMD candidates is 259 nucleotides while the non-retained introns have an average of 1487 nucleotides. Splicing efficiency seems to vary widely among such small introns. This seems to be in large part because small introns possess a weaker splice site and are often recognized by the intron definition mechanism [15, 16, 19].

Here, we have assessed the retained introns in our data set for any evidence of selection pressures. One such source of evidence is deviation in GC content because of selection for transcriptional efficiency [15]. We analyzed

deviation in GC content in the retained introns, as a function of intron length (Figure 4.1), to avoid GC content bias due to the size of the sequences. Firstly, we examined the following three data sets: retained introns, exons bordering retained introns, and non-retained introns. We divided the three data sets into categories according to their sequence length (Figure 4.1). Non-retained introns showed a statistically significant lower GC content in all classes of size compared with retained introns and exons (χ^2 = 18.9, p<0.0001; χ^2 = 25.6, p<0.0001, respectively). In contrast, retained introns showed a similar CG content compared with exons (χ^2=0.67, p=0.41). For comparison, we also included retained introns in genes that are orthologous AS-NMD candidates between human and mouse (Figure 4.1). In spite of the absence of retained intron shorter than 100 kb, all other retained introns classes were more similarly to exons (χ^2 = 0.21, p=0.9) than to non-retained introns (χ^2 = 34, p<0.0001). We also found an increase in the GC content upstream of the stop codon of retained introns compared with downstream of the PTC (56% and 49% respectively).

To evaluate the codon usage of the retained introns we used two tests described by Galante, *et al.* (2004, [15]). For the first test, we created 100 random sets of exons bordering retained introns and non-retained introns, each set containing 16970 codons (the number of codons in the retained intron set). We then put each sequence in frame (for more details see the *Methods*) and performed a pairwise comparison of the total codon table (61 amino acids) of the three data

sets (Table 4.3). The second test was based on the pairwise comparison of the average number of amino acids using the same codons in the three data sets. In both tests retained introns and exons shared more similarities between each other than with non-retained introns (Table 4.3).

We also assessed the ability of retained introns to encode protein domains. To do this, we extracted three set of sequences: *retained introns*, *exons* and sequences comprising retained introns and their 5' and 3' exons ('*E-RI-E*' sequences). These three sequence sets were compared against the Pfam database (as described in *Methods*). In total, we found 27 retained introns coding for a protein domain (17% of a total of 155 genes used). Ten out of 27 retained introns encoded part of a domain also encoded by one of the flanking exons. Retained introns encoded for a different domain compared with their flanking exons in 14 cases and in three cases, only the retained intron encoded a Pfam domain. There was no case where the whole *E-RI-E* sequence encoded for a domain.

The majority of the AS-NMD candidate ortholog pairs between human and mouse (153 out of 262) presented an IR in both species. The average percentage of identity between human and mouse AS-NMD candidates with a retained intron is 56%. The large number of AS-NMD candidates that are orthologously conserved between human and mouse corroborates the idea of a widespread conserved mechanism to control gene expression through NMD.

We also annotated retained introns with PTCs that do not target the mRNA for NMD (termed 'non-NMD') (Table 4.4). The number of genes bearing a non-NMD retained intron was between 2 and 4-fold higher than the number of genes with an AS-NMD candidate (the difference is statistically significant for all four genomes by χ^2 test with $P \leq 10^{-7}$). This indicates an apparent selection against PTC in NMD areas, and was also observed by us in a previous work [13], and by other authors as well [15]. The average size of non-NMD intron retentions is greater than AS-NMD candidates, although only in humans this difference is statistically significant ($\chi^2 = 19$, $p<0.0001$).

Functional Annotation

Using the webtool FatiGO [20], we annotated the most common Gene Ontology terms [21], Biocarta pathways (http://www.biocarta.com/genes/index.asp), TRANSFAC [22] transcription factors, and KEGG [23] terms in each genome, and calculated which terms are overrepresented in genes with AS-NMD candidates (significant overrepresentation was calculated using a Fisher's exact test with $P' < 0.05$, and a correction to P' for multiple hypothesis testing [20]).

The term '*Nucleotide binding'* (GO:0000166) seems to be ubiquitous for the four genomes. Other abundant terms include those related to DNA interaction,

DNA damage, cell cycle control and apoptosis (Table 4.5). Transferase and hydrolase activities were also common. The pattern of over- and underrepresentation of GO terms was similar for all four species (the GO terms were simultaneously statistically significant in the four genomes only for non-corrected p-values. See Table 4.5 for details). This result shows that NMD may target different genes in different species, but these genes often belong to the same functional categories.

We also performed a census of the most common protein domains in the genes with AS-NMD candidates (with all domains counted only *once* per gene). The most common Pfam protein domain was the WD-40 domain (except in rat) (Supplementary Table S4.3). This domain is known to be involved in several cellular functions, such as signal transduction, cell-cycle control and apoptosis [24]. Kinase domain was also found to be among the most abundant Pfam domains (Supplementary Table S4.3). In cow, the domain 'DEADH box helicase' was among the top 5 most abundant protein domains. Several genes containing the DEAD box domain and also targeted to NMD were found in a previous work [10].

In cow and rat, cancer-related pathways are significantly enriched (for P-values corrected for multiple hypothesis testing) according to KEGG classification. In mouse, besides cancer, AS-NMD candidates related with

apoptosis are the most abundant. Human AS-NMD candidates are enriched for 'glycine, serine and threonine metabolism', 'epithelial cell signaling in Helicobacter pylori infection' and 'androgen and estrogen metabolism' KEGG pathways (Supplementary Table S4.4). One role of NMD seems to be to protect patients that are heterozygous for a PTC-containing allele, preventing autosomal recessive disorders. For instance, patients are phenotypically normal if they have one normal β-globin allele and another β-globin allele with a PTC that elicits NMD. In contrast, patients manifest a dominantly inherited anemia if the PTC in the defective allele fails to elicit NMD. As a consequence, the resulting PTC-containing mRNA is present at near-normal levels and produces so much truncated β-globin protein that erythropoesis is ineffective [8]. The same NMD-based disease mechanism has also been demonstrated for tumor suppressors such as BRAC1, WT1 and other genes linked with genetic disorders such as diabetes type II (Calpain-10) [8, 25].

In the BioCarta pathway classification, we found five pathways in humans, enriched for AS-NMD candidates: antigen processing and presentation, inhibition of cellular proliferation by Gleevec, cell-to-cell adhesion signaling, integrin signaling pathway and role of Ran in mitotic spindle regulation (Supplementary Table S4.5).

The Figure 4.2 shows an example of a Biocarta pathway enriched for AS-NMD candidates. The drug Gleevec (also known as imatinib mesylate or STI-571) is considered one of the major breakthroughs in the treatment of cancer, especially CML, chronic myeloid leukemia. Gleevec acts on a molecular target by a mechanism that is more specific to cancer cells than traditional cytotoxic treatments [26]. We found four key genes in the 'Inhibition of cellular proliferation by Gleevec' pathway that are targets for NMD: breakpoint cluster region (BRC), mitogen-activated protein kinase 1 (MAP3K1), v-akt murine thymoma viral oncogene homolog 1 (Akt-1), and c-fos FBJ murine osteosarcoma viral oncogene homolog (FOS). All of these genes participate in a broad range of pathways such as 'Integrin signaling pathway' (BRC), 'p38 MAPK Signaling Pathway' (MAP3K1), 'Apoptotic Signaling in Response to DNA Damage' (Akt-1), 'TSP-1 Induced Apoptosis in Microvascular Endothelial Cell' (FOS). The loss of expression of these genes is linked with several diseases and in some cases can be lethal. For example, it was found that loss of Akt1 resulted in defective ischemia and Vegf-induced angiogenesis and severe peripheral vascular disease in mice [27]. Loss of Mekk1 expression resulted in a greater apoptotic response of cells to hyperosmolarity and microtubule disruption [28].

We also calculated enrichments of transcription factor binding sites in the promoters of genes that target AS-NMD candidates (Supplementary Table S4.6).

In human, the most significantly-enriched transcription factor binding site was for E2F1. Merdzhanova *et. al.* (2008, [29]), studied the association of E2F1 and the splicing regulator SC35 (a gene with an AS-NMD candidate). According to these authors, E2F1 upregulates the expression of SC35 in response to DNA-damaging agents; they also show that SC35 is required for apoptosis in response to these agents. The apoptotic mechanism involves the pro-apoptotic alternative splice form of the gene Bcl-x (also a AS-NMD candidate) that is in turn controlled by SC35 [29]. Expression of the NMD-targeted form of SC35 could switch on the overexpression of anti-apoptotic splice variants of genes like Bcl-x and caspase. Such alteration is believed to contribute to the resistance of tumor cells to chemotherapy [30, 31].

We compared the representation of AS-NMD candidates in the OMIM database with the overall representation of the human genome in the same database. The difference of representation of the two sets in the OMIM database is statistically significant (χ^2=129, $p<0.0001$). While 35% of the human gene complement is present in OMIM, the proportion of genes harbouring AS-NMD candidates in OMIM is 50%. The majority of the genetic disorders associated with these genes are related to cancer. NMD-associated diseases often result from insufficient levels of full-length protein. Therefore, these diseases are generally due to two defects in gene expression: degradation of the newly synthesized PTC-

containing mRNA by NMD, and failure of the abnormally low level of PTC-containing mRNA that escaped NMD to generate full-length and, thus, functional protein.

Mapping AS-NMD to MS PRIDE database

Since 5 to 25% of PTC-containing mRNAs fail to elicit NMD and are consequently translated into truncated proteins [32], several therapies have been developed specially to suppress in-frame PTC [33, 34]. Some criticism of nonsense-suppressor therapies is that new diseases can be created even if only a single C-terminally extended protein from one of the many cellular transcripts were to accumulate to a toxic level. Furthermore, another problem is the unknown consequence of failing to eliminate naturally-occurring NMD targets, such as the many Ig and TCR transcripts that harbor PTCs as a consequence of somatic rearrangements and hypermutations that typically generate immune diversity [35].

To quantify the frequency with which the expressed AS-NMD candidates escape NMD and are translated into proteins we mapped each human translated AS-NMD candidate mRNA/EST to ~550,000 peptides extracted from the repository of mass spectrometry-derived proteomics data (PRIDE) [36].

After filtering out the matches for uniqueness and size (peptides less than 7 amino acids long and peptides matching another gene were discarded) we found 205 AS events in 194 genes with peptide matches (hereafter referred as AS-NMD-PRIDE genes) (Table 4.6). This is ~10% of the total number of genes targeted to NMD (Table 4.1).

The majority (71%-81%) of the human AS-NMD-PRIDE genes possess an ortholog in one of the other species studied. We also compared the AS-NMD-PRIDE genes with the remaining genes harbouring AS-NMD candidates (AS-NMD candidates with no match in PRIDE database). Surprisingly, the proportion with conserved AS-NMD orthologs was nearly the same for both sets of genes. For example, in the AS-NMD-PRIDE set 18%, 3% and 8% of the human orthologs in mouse, rat and cow respectively were also targeted for NMD. For the remaining AS-NMD candidates, the proportion was 14%, 3% and 6% (in mouse, rat and cow). More than half of the AS-NMD-PRIDEs come from AS forms containing retained introns (56%), while in AS-NMD candidates the frequency is 47%.

The AS-NMD-PRIDE set of genes was significantly enriched for '*protein metabolic process'* (GO:0019538), '*cellular macromolecule metabolic process'* (GO:0044260) GO biological process and '*structural constituent of ribosome'* (GO:0003735) GO molecular function compared with the whole human genome.

The last GO term was also enriched when AS-NMD-PRIDE set of genes was compared with the AS-NMD candidates set.

Among the genes present in the '*structural constituent of ribosome'* GO term we found two cases in which the ortholog gene was also targeted for NMD: the gene Rpl5 in mouse and the bovine RPS13. The human and cow RPS13 presented a retention of intron 1 that targets its mRNA for NMD. Using the ECR browser [37] to generate an alignment of the RPS13, we could observe a high GC content in the intron 1 (> 65%), as well as small size (141 nt), and a high conservation among human, mouse and cow (Figure 4.3A). Taken together, these are the typical symptoms of a non-spurious retained intron. Recent study showed that the presence of intron 1 in the human RPS13 reduced its expression by a factor of four [38]. Apparently, the gene RPS13 can regulate its protein level via a feedback mechanism. The ribosomal protein S3 was found to inhibit the excision of the intron 1 from the rpS13 pre-mRNA. This protein binds specifically the intron fragment near the 5' and 3' splice site, therefore conferring protection against cleavage by ribonucleases [38]. Although the extraribosomal function of rpS13 is still unclear, a study suggested that this gene (along with rpL23) promotes multidrug resistance in gastric cancer cells by suppressing apoptosis [39]. A similar pattern of proteins regulating their own splicing and generating aberrant mRNA was also found in *Saccharomyces cerevisiae* [40] and in *C.*

elegans [41] which suggests that NMD plays a important role in gene regulation. In Figure 4.3B, we have illustrated five other examples of retained introns, that are conserved in mouse; note that each of these cases are in gene structures with a large number of introns.

Conclusions

In this work, we have annotated and analyzed alternative splicing coupled with nonsense-mediated decay in four mammalian species: human, mouse, rat and cow. We have shown that alternative splicing coupled with NMD occurs in 2 to 10% of the known annotated genes in each genome and that this is likely to be a substantial under-estimation of the incidence, since the numbers of AS-NMD candidates that can be predicted is largely affected by the total size of the available EST/mRNA/cDNA database. Human AS-NMD candidates seem to have a large number of orthologs in mouse (~82%), and at least 15% of those orthologs are also AS-NMD candidates in another species.

Although we only analyzed AS-NMD candidates with in-frame PTCs in this study, we have demonstrated that intron retention plays an important role in producing AS-NMD candidates in all four species. A large fraction (58%-81%, depending on the species) of conserved NMD candidates possesses intron

retentions. The retained introns also possess features that distinguish them from non-retained introns, such as smaller size, higher GC content and codon usage similar to exons.

The AS-NMD candidates also showed a similar pattern of gene ontology enrichment in all four species. Genes related with nucleotide interaction and cell-fate (apoptosis) were the most abundant to produce AS-NMD candidates. KEGG and Biocarta pathways also showed an over-representation of pathways linked with cancer and genetic disorders for such AS-NMD-making genes. Thus, NMD appears to have a significant role in controlling the expression of aberrant proteins in these pathways and prevent their deleterious effects.

We gathered evidence of translation of at least 10% of all AS-NMD candidates in human genome. Such cases, are not, however, especially conserved in comparison to other AS-NMD candidates.

Furthermore, in one of the AS-NMD candidates, the gene RPS13, we found evidences of a conserved mechanism of autoregulation, between human and cow, in which the protein inhibits the excision of one of its intron creating a splice form that is targeted for NMD.

Alternative splicing coupled with NMD seems to act on a wide range of genes, but, only in a few significantly over-represented functions across

mammalian genomes. Two major models of NMD previously proposed, namely, the 'regulatory' and 'noisy' models [12], are apparently applicable simultaneously to some genes. The data herein, provides strong statistical evidence for a regulatory role conserved across *Mammalia* for a large subset of NMD candidates.

Methods

Genome data

The genome sequences and annotations of four mammals (human, cow, mouse and rat) were downloaded from the Ensembl Web site (http://www.ensembl.org), in March 2009. The genome assemblies are: human= Homo_sapiens.NCBI36.52; cow= Bos_taurus.Btau_4.0.52; mouse= Mus_musculus.NCBIM37.52 and rat= Rattus_norvegicus.RGSC3.4.52. These genomes were chosen based in two criteria: *(i)* high coverage (>7X), and *(ii)* large number of cDNAs and ESTs sequences available.

The cDNA/EST/mRNA data were downloaded from the Refseq database (ftp://ftp.ncbi.nih.gov/refseq), the H-invitational database (http://h-invitational.jp/hinv/ahg-db), the Unigene compendium (ftp://ftp.ncbi.nih.gov/repository/Unigene; Build #217) and the dbEST data set (ftp://ftp.ncbi.nih.gov/repository/dbEST/), in March 2009 (see Appendix 1 Table A.1 for details of the number of sequences used per database).

Identification of alternative splice forms with an in-frame premature stop codon targeting the mRNA for nonsense-mediated decay

(I) Reference identification

We identified the reference mRNA (an mRNA that aligns with 100% coverage and identity to an annotated protein) of each human, mouse, rat and cow transcripts. A bl2seq [42] was performed between mRNAs from Refseq, Unigene and H-invitational (human genome only) and each protein in the Ensembl database. Transcripts without a reference mRNA were discarded from our analysis.

The genomic loci were identified and extracted from the Ensembl annotations and expanded by 500 nucleotides 5' and 3' of the locus, to allow variation in the transcriptional start and polyadenylation sites [25]. Each reference mRNA was aligned with its genomic locus using GMAP [43].

(II) Mapping

To identify AS forms we aligned each genomic locus that has a reference mRNA, against all mRNAs/ESTs/cDNAs from the Refseq, Unigene, dbEST and H-invitational databases (this last database is for human only). Each alignment was parsed into a set of splice junction coordinates. Alignments where the

coverage was <0.96 of the length of the mRNA or EST and <98% sequence identity were discarded. We also discarded ESTs that aligned to the opposite strand.

In cases were an mRNA/cDNA/EST mapped to more than one gene, the best mapping was considered the correct one. If a mapping presented the same percentage of identity and coverage to more than one gene, the mapping was discarded.

(III) Coding sequence (CDS) annotation

We performed a new round of alignments to identify the coding region (CDS) of each mRNAs/ESTs/cDNAs. Using the program Exonerate [44], we aligned each mRNAs/ESTs/cDNAs retrieved from the previous mapping, against the encoded protein extracted from the Ensembl annotation.

(IV) Alternative splicing (AS) annotation

We compared each mapped splice junction with the reference mRNA splice junctions. If a splice form mapped >10nt from the reference junction, it was annotated as an alternative splice form otherwise it was discarded.

(V) Nonsense-mediated decay targeting annotation

We verify the presence of PTCs targeting the mRNAs/ESTs/cDNAs for NMD, using the 55nt rule [6]. Using the CDS annotation, we checked the mRNAs/ESTs/cDNAs sequence for an in-frame stop codon 55 nucleotides or more 5' to an exon-exon junction.

Intron retention (IR) analysis

To verify whether genes targeted for NMD due to IR were caused by an unspliced or partially spliced pre-mRNA, we analyzed several features of our sequences and compare them with their flanking exons and with non-retained introns. Some of our analysis followed the analysis described by Galante, *et al.* (2004)[15].

Codon Usage/Bias

We assess the codon usage and codon bias of retained introns, exons bordering retained introns and non-retained introns. We only used AS-NMD candidates with at least one full length cDNA within the IR. The frame of non-retained introns and retained introns was defined by the frame of the exon 5' to it.

We analyzed the general codon bias of all 61 codons in the three sets. We performed 100 pairwise comparisons of the distribution of 16,970 codons (number of codons in the retained intron set), along with 61 possible codons to

calculate the average χ^2 and standard deviations thereof. We also analyzed the codon usage (number of sequences that present the same distribution of codons per amino acids). Since methionine and tryptophan have only one codon they were excluded from this analysis. I generated 100 set of sequences by randomly sampling exon, non-retained introns and retained introns from AS-NMD candidates and performed a pairwise comparison of each set grouped by amino-acid type. Two sets were considered as having different codon usage for a given amino-acid type, if the P value was < 0.05 in a χ^2 test.

Pfam domain analysis

DNA sequences from retained introns, exons bordering retained introns and exon–retained intron–exon (*'E-RI-E'*) sequences were 'Blasted' against the whole Pfam database using BLASTx program (e-value 10^{-2}). Only genes with a full length mRNAs/ESTs/cDNAs and with a retained intron entirely in the CDS were used in this analysis (155 in total). We analyzed how often a retained intron encodes a Pfam domain and whether or not this domain is present in the flanking exons.

GC content

To assess the GC content of the retained introns, exons bordering retained introns and non-retained introns, we created three datasets with the same number

of sequences divided into five classes according to their length. The non-retained introns were randomly picked from the set of genes containing the retained introns in order to avoid any kind of bias toward a specific type of gene. A χ^2 test was used to evaluate the differences between each dataset.

Orthologs

The information about pairs of AS-NMD candidate orthologs between human and the other species was extracted from the Biomart query system in the Ensembl database. We only considered two genes as being orthologs when they mapped in a one-to-one way.

Functional Annotation

Gene Ontology (GO [21]) functional categories, transcription factors TRANSFAC [22] and KEGG [23] and BioCarta (http://www.biocarta.com/genes/index.asp) terms, were retrieved from the FatiGo database [20]. We performed a functional enrichment analysis of our list of AS-NMD candidates against the whole genome annotation of each species. Significant overrepresentation was calculated using a Fisher's exact test with $P' < 0.05$, and a correction to P' for multiple hypothesis testing [20]. Human AS-NMD candidates were cross-referenced with the NCBI OMIM database

(http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) to identify the number of AS-NMD candidates linked to genetic disorders.

Mapping peptides from mass spectrometry-derived proteomics data to AS_NMD candidates

We mapped all human AS_NMD candidates to peptides from mass spectrometry-derived data extracted from the PRIDE database [36]. Each EST/mRNA was translated in the 3 frames and checked against the PRIDE peptides. We derived a Perl program to check the occurrence of peptides upstream of the PTC. We filtered out any matching peptide smaller than 7 amino acids, and also any peptides matching to another gene.

Acknowledgements

This research was supported by grants from the National Science and Engineering Research Council of Canada, and from McGill University.

References

1. Zavolan M, van Nimwegen E, Gaasterland T: **Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome**. *Genome Res* 2002, **12**(9):1377-1385.
2. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity**. *Nat Genet* 2002, **30**(1):29-30.
3. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
4. Tazi J, Bakkour N, Stamm S: **Alternative splicing and disease**. *Biochimica et biophysica acta* 2009, **1792**(1):14-26.
5. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans**. *Proc Natl Acad Sci U S A* 2003, **100**(1):189-192.
6. Chang YF, Imam JS, Wilkinson MF: **The nonsense-mediated decay RNA surveillance pathway**. *Annual review of biochemistry* 2007, **76**:51-74.
7. Silva AL, Romao L: **The mammalian nonsense-mediated mRNA decay pathway: to decay or not to decay! Which players make the decision?** *FEBS letters* 2009, **583**(3):499-505.
8. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE: **Nonsense-mediated decay approaches the clinic**. *Nat Genet* 2004, **36**(8):801-808.
9. Le Hir H, Izaurralde E, Maquat LE, Moore MJ: **The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions**. *Embo J* 2000, **19**(24):6860-6869.
10. Saltzman AL, Kim YK, Pan Q, Fagnani MM, Maquat LE, Blencowe BJ: **Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay**. *Molecular and cellular biology* 2008, **28**(13):4320-4330.
11. Kaygun H, Marzluff WF: **Regulated degradation of replication-dependent histone mRNAs requires both ATR and Upf1**. *Nature structural & molecular biology* 2005, **12**(9):794-800.
12. Zhang Z, Xin D, Wang P, Zhou L, Hu L, Kong X, Hurst LD: **Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay**. *BMC biology* 2009, **7**:23.

13. Morais DD, Harrison PM: **Genomic evidence for non-random endemic populations of decaying exons from mammalian genes**. *BMC genomics* 2009, **10**:309.
14. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements**. *Nature* 2007, **446**(7138):926-929.
15. Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ: **Detection and evaluation of intron retention events in the human transcriptome**. *RNA (New York, NY* 2004, **10**(5):757-765.
16. Kurmangaliyev YZ, Gelfand MS: **Computational analysis of splicing errors and mutations in human transcripts**. *BMC genomics* 2008, **9**:13.
17. Hiller M, Huse K, Platzer M, Backofen R: **Non-EST based prediction of exon skipping and intron retention events using Pfam information**. *Nucleic Acids Res* 2005, **33**(17):5611-5621.
18. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Segurens B *et al*: **Translational control of intron splicing in eukaryotes**. *Nature* 2008, **451**(7176):359-362.
19. Berget SM: **Exon recognition in vertebrate splicing**. *The Journal of biological chemistry* 1995, **270**(6):2411-2414.
20. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J: **BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W460-464.
21. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D258-261.
22. Wingender E: **The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation**. *Brief Bioinform* 2008, **9**(4):326-332.
23. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2009.
24. Neer EJ, Schmidt CJ, Nambudripad R, Smith TF: **The ancient regulatory-protein family of WD-repeat proteins**. *Nature* 1994, **371**(6495):297-300.
25. Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC, Brenner SE: **Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and**

disease genes. *Bioinformatics (Oxford, England)* 2003, **19 Suppl 1**:i118-121.

26. Kuenen BC, Pinedo HM: **[New oncological treatment principle with imatinib]**. *Nederlands tijdschrift voor geneeskunde* 2003, **147**(42):2044-2045.
27. Peng XD, Xu PZ, Chen ML, Hahn-Windgassen A, Skeen J, Jacobs J, Sundararajan D, Chen WS, Crawford SE, Coleman KG *et al*: **Dwarfism, impaired skin development, skeletal muscle atrophy, delayed bone development, and impeded adipogenesis in mice lacking Akt1 and Akt2**. *Genes Dev* 2003, **17**(11):1352-1365.
28. Yujiri T, Sather S, Fanger GR, Johnson GL: **Role of MEKK1 in cell survival and activation of JNK and ERK pathways defined by targeted gene disruption**. *Science* 1998, **282**(5395):1911-1914.
29. Merdzhanova G, Edmond V, De Seranno S, Van den Broeck A, Corcos L, Brambilla C, Brambilla E, Gazzeri S, Eymin B: **E2F1 controls alternative splicing pattern of genes involved in apoptosis through upregulation of the splicing factor SC35**. *Cell death and differentiation* 2008, **15**(12):1815-1823.
30. Mercatante D, Kole R: **Modification of alternative splicing pathways as a potential approach to chemotherapy**. *Pharmacology & therapeutics* 2000, **85**(3):237-243.
31. Hayes GM, Carrigan PE, Beck AM, Miller LJ: **Targeting the RNA splicing machinery as a novel treatment strategy for pancreatic carcinoma**. *Cancer research* 2006, **66**(7):3819-3827.
32. Stephenson LS, Maquat LE: **Cytoplasmic mRNA for human triosephosphate isomerase is immune to nonsense-mediated decay despite forming polysomes**. *Biochimie* 1996, **78**(11-12):1043-1047.
33. Kerem E: **Pharmacologic therapy for stop mutations: how much CFTR activity is enough?** *Current opinion in pulmonary medicine* 2004, **10**(6):547-552.
34. Ainsworth C: **Nonsense mutations: running the red light**. *Nature* 2005, **438**(7069):726-728.
35. Kuzmiak HA, Maquat LE: **Applying nonsense-mediated mRNA decay research to the clinic: progress and challenges**. *Trends in molecular medicine* 2006, **12**(7):306-316.
36. Vizcaino JA, Cote R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H, Martens L: **The Proteomics Identifications database: 2010 update**. *Nucleic Acids Res*, **38**(Database issue):D736-742.
37. Ovcharenko I, Nobrega MA, Loots GG, Stubbs L: **ECR Browser: a tool for visualizing and accessing data from comparisons of multiple**

vertebrate genomes. *Nucleic Acids Res* 2004, **32**(Web Server issue):W280-286.
38. Malygin AA, Parakhnevitch NM, Ivanov AV, Eperon IC, Karpova GG: **Human ribosomal protein S13 regulates expression of its own gene at the splicing step by a feedback mechanism**. *Nucleic Acids Res* 2007, **35**(19):6414-6423.
39. Shi Y, Zhai H, Wang X, Han Z, Liu C, Lan M, Du J, Guo C, Zhang Y, Wu K *et al*: **Ribosomal proteins S13 and L23 promote multidrug resistance in gastric cancer cells by suppressing drug-induced apoptosis**. *Experimental cell research* 2004, **296**(2):337-346.
40. Fewell SW, Woolford JL, Jr.: **Ribosomal protein S14 of Saccharomyces cerevisiae regulates its expression by binding to RPS14B pre-mRNA and to 18S rRNA**. *Molecular and cellular biology* 1999, **19**(1):826-834.
41. Mitrovich QM, Anderson P: **Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans**. *Genes Dev* 2000, **14**(17):2173-2184.
42. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences**. *FEMS microbiology letters* 1999, **174**(2):247-250.
43. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics (Oxford, England)* 2005, **21**(9):1859-1875.
44. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC bioinformatics* 2005, **6**:31.

Table 4.1: Summary of the AS-NMD annotations

Genomes	(1) AS events	(2) EST/cDNA supporting the AS	(3) Genes with an AS-NMD candidate	(4) Proportion of (3) that have orthologs* in the human genome	(5) AS-NMD candidates with an orthologous AS-NMD candidate in the human genome*	
Homo sapiens	2536	14343	2051 (9.8%)†	--------	--------	
Mus musculus	1211	1628	1069 (4.6%)†	970/1069 (91%)	262 (27%)	∇∇∇
Rattus norvegicus	166	190	149 (0.84%)†	129/149 (87%)	45 (34%)	∇
Bos taurus	374	600	342 (1.8%)†	297/342 (87%)	105 (35%)	∇∇

† Percentage of the total number of known genes based on Ensembl build 52.

* Orthology was based on the relationship one-to-one between a human transcript and a transcript from other genome according to Ensembl annotation.

∇ Significant by a hypergeometric probability test with $P \leq 10^{-14}$; ∇∇ Significant by a hypergeometric probability test with $P \leq 10^{-33}$; ∇∇∇ Significant by a hypergeometric probability test with $P \leq 10^{-55}$.

Table 4.2: Number of events per genome

Events	*Homo sapiens*	*Mus musculus*	*Rattus norvegicus*	*Bos taurus*
RI†	884 (43%)*	475 (44%)	95 (63%)	231 (67%)
CSE†	444 (21%)	137 (12%)	10 (6%)	27 (8%)
ASD†	592 (29%)	247 (23%)	26 (17%)	49(14%)
ASA†	616 (30%)	352 (32%)	35 (23%)	67 (19%)
Avg. IR size	259 nt	171 nt	189nt	188 nt
% id. with a human Ortholog¶	--------	84%	85%	88%

RI (retained intron); CSE (Cassette exon); ASD (alternative splice donor); ASA (alternative splice acceptor)

† Values based on the number of genes.

* The percentage does not add up to 100 since one gene can have more than one event simultaneously

¶ Values based on protein sequence

Table 4.3: Codon usage comparison of retained introns, exons and non-retained introns in human NMD-candidates

	Retained introns and Non-retained introns	Retained introns and Exons	Non-retained introns and Exons
Codon table (Average χ^2)†	1324 ± 282	950 ± 86	2531 ± 310
Codon Usage¶	3.3 ± 1.4	5.4 ± 1.4	0.4 ± 0.6

† Comparison of the entire codon usage table (61 codons, 60 degrees of freedom) among three datasets.

¶ Average number of amino acids with the same codon usage frequency (number of amino acids with p-values <0.05 in a χ^2 test).

Table 4.4: Retained introns with a premature stop codon that do not target the mRNA for NMD

	Homo sapiens	*Mus musculus*	*Rattus norvegicus*	*Bos taurus*
Retained introns	5029	970	261	522
Number of genes with a R.I.	3111	770	205	416
Avg. size of R.I.	421 nt	180 nt	210 nt	202 nt
R.I (3n)†	33.5%	34%	36%	35%

† Percentage of retained introns whose length is multiple of 3.

Table 4.5: Most represented GO terms per genome

Homo sapiens	*Mus musculus*	*Rattus norvegicus*	*Bos taurus*
GO Molecular Function Level 3			
Hydrolase activity (GO:0016787)	Nucleotide binding (GO:0000166)	Transferase activity (GO:0016740) †	Structural constituent of ribosome (GO:0003735) †
Nucleotide binding (GO:0000166)	Hydrolase activity (GO:0016787)	Amine transporter activity (GO:0005275) †	Nucleotide binding (GO:0000166) †
Lyase activity (GO:0016829)	Transferase activity (GO:0016740)	Organic acid transporter activity (GO:0005342) †	Selenium binding (GO:0008430) †
Oxidoreductase activity (GO:0016491)	Helicase activity (GO:0004386) †	Nucleotide binding (GO:0000166) †	Transferase activity (GO:0016740) †
GTPase regulator activity (GO:0030695)	Oxidoreductase activity (GO:0016491)†	Enzyme activator activity (GO:0008047) †	Helicase activity (GO:0004386) †
GO Molecular Function Level 4			
Transferase activity, transferring phosphorus-containing groups (GO:0016772)	Purine nucleotide binding (GO:0017076)	Hydrolase activity, acting on glycosyl bonds (GO:0016798) †	Purine nucleotide binding (GO:0017076)†
Purine nucleotide binding (GO:0017076)	Transferase activity, transferring phosphorus-containing groups (GO:0016772)	Transferase activity, transferring phosphorus-containing groups (GO:0016772) †	Hydrolase activity, acting on glycosyl bonds (GO:0016798) †
Translation factor activity, nucleic acid binding (GO:0008135)	Hydrolase activity, acting on acid anhydrides (GO:0016817)	Purine nucleotide binding (GO:0017076) †	Calmodulin binding (GO:0005516) †
Carbon-carbon lyase activity (GO:0016830)	Ligase activity, forming carbon-oxygen bonds (GO:0016875)	Carboxylic acid transporter activity (GO:0046943)†	Enzyme binding (GO:0019899) †
RNA binding (GO:0003723) †	RNA binding (GO:0003723) †	GTPase activator activity (GO:0005096) †	Oxidoreductase activity, acting on peroxide as acceptor (GO:0016684) †
			ATPase inhibitor activity (GO:0042030)†

† Statistically significant for non-corrected p-value only

Table 4.6: Comparison between human AS-NMD with and without a match in PRIDE db

Homo sapiens	AS events	EST/cDNA supporting the AS	Number of Genes	Orthologs in another genome	AS-NMD candidates with an orthologous AS-NMD candidate the in another genome	AS-NMD candidates with a RI*
AS-NMD-PRIDE	205	385	194	159 (mouse) 139 (rat) 156 (cow)	30 (mouse) 5 (rat) 13 (cow)	110
AS-NMD candidate without match in PRIDE	2331	13958	1857	1560 (mouse) 1425 (rat) 1478 (cow)	232 (mouse) 40 (rat) 92 (cow)	884

* RI (retained intron)

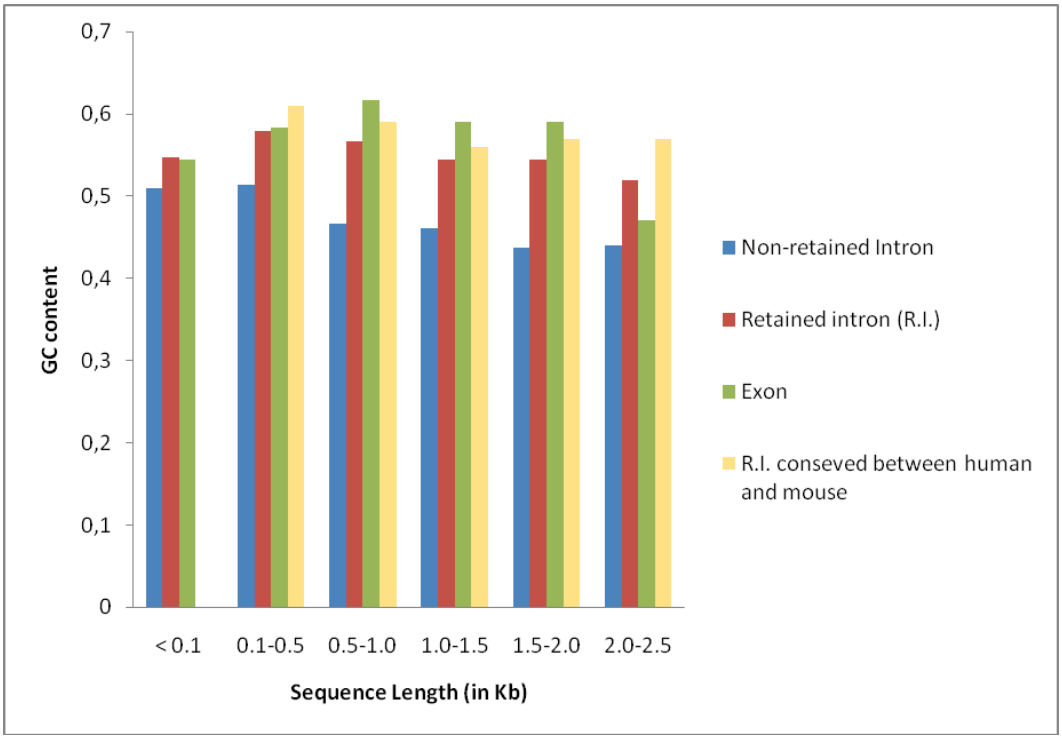


Figure 4.1: GC content of retained introns, exons bordering retained introns and non-retained introns in human NMD-candidates. The dataset was divided into six categories to avoid bias due to their length.

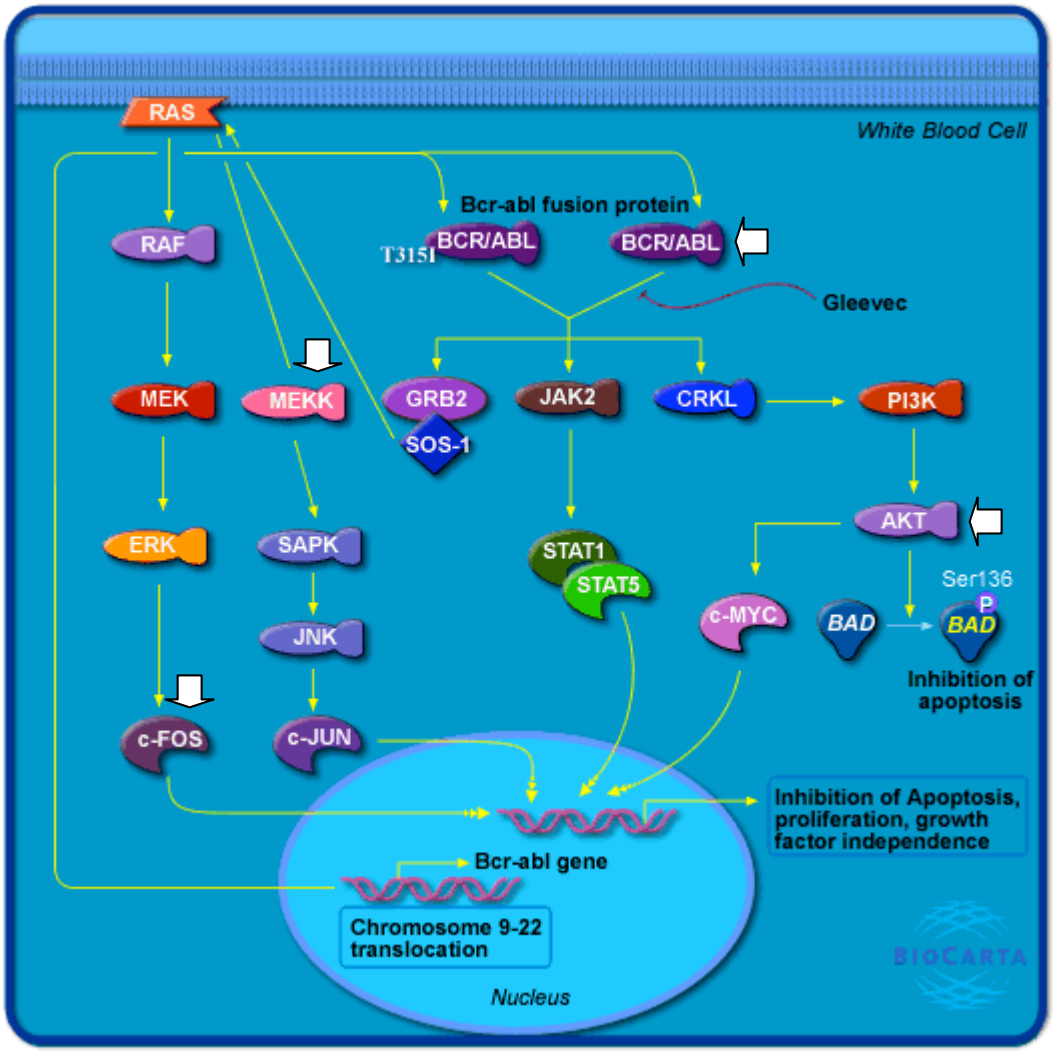
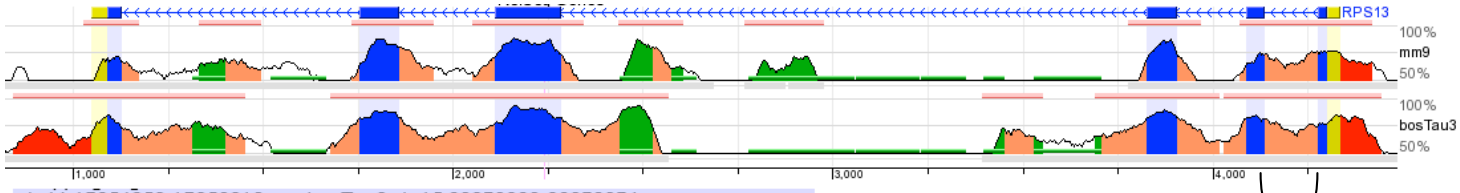


Figure 4.2: Gleevec pathway. The white arrows indicate the human NMD-candidates. Extracted from BioCarta (http://www.biocarta.com/genes/index.asp).



(A)

chr11:17054858-17056212 -vs- bosTau3:chr15:29378392-29379874

```
                29379158                29379138                29379118
bosTau3 CCGAGCAACTGATGGCACCCCCACCGCCTCCCGGCCTCTCCTCGGTACCCGCACACTACT
                ||||| ||  ||| ||  ||||    |  |||||  |||| |||| ||  ||  | ||||
    hg18 CCGAG--ACAAATGCCAACCCC----CTCCCCGGATTCTCCCCGGTCCCAGCGCGCTACT
                17055495                17055513                17055529

                29379098                29379078                29379058
bosTau3 TACGGTGGGGACGCTGCGGCGGTAGGGCAGAGCCGACTGGGACAGGCCCTTCCTGGAGAG
                ||| ||||||||||||||| || ||||| | ||||||||||||||||||||||||||||| ||| |||
    hg18 TACAGTGGGGACGCTGCGTCGATAGGGTAAAGCCGACTGGGACAGGCCCTTCCTGGGGAG
                17055549                17055569                17055589

                29379038                29379018                29378998
bosTau3 AAAAGCAGTATCGGGGTGAAGAAGCGAGCCCCGCCAGAGCAGCCCCAAACCTTACCCCCT
                |  |||||||  |||  |||||  |  |||          ||||||||||||  |||  |  |  |||||
    hg18 AGAAGCAGTCTCGAGGTGAGGTGGCGGCTAGTGCCAGAGCAGCCCAGAACATCATCCCCT
                17055609                17055629                17055649

                29378978                29378958                29378938
bosTau3 CGACCCACTGCCTACGGTCCTTCCTTTCCCTCCACCCCTAGACTCAGCCTTGCAACCC-G
                ||  |||  |||||  ||  ||||  |    ||||  |||||||           |  |||  ||||  ||  |
    hg18 CGGCCCGCTGCCCACACTCCCTGTCTTCCTCCCACCCC------CCGCCCAGCAATCCGG
                17055669                17055689

                29378919                29378899                29378879
bosTau3 CTCATAAACCCTAGCTCACCCGGGAGCGTGCATGCGACCCATGATGGCGGCGGTGTGGTA
                ||        |||  |||||||||||||||  ||||||||||||||||||||||  |  ||  |
    hg18 CTTGATGCCCCGAGCTCACCCGGGAGCATGCATGCGACCCATGATGGCGGCGATCAGGCA
                17055723                17055743                17055763

                29378859                29378839                29378819
bosTau3 GCAAAAGGAGAGAGAGGGTGGGAGAGCCGCAGGAAGCCACCCTCCACGCAGGAAGTGGCC
                |  |||||||||  |||  ||||  |   ||||||||||||||||  ||||||||||||  ||
    hg18 ACGAAAGGAGAGCGAGAGTGGAAACGCCGCAGGAAGCCACCCAACACGCAGGAAGTGACC
                17055783                17055803                17055823
```

> **hg18 chr11:17055495-17055906 412bps GC:65.5%**

CCGAGACAAATGCCAACCCCCTCCCCGGATTCTCCCCGGTCCCAGCGCGCTACTTACAGT
GGGGACGCTGCGTCGATAGGGTAAAGCCGACTGGGACAGGCCCTTCCTGGGGAGAGAAGC
AGTCTCGAGGTGAGGTGGCGGCTAGTGCCAGAGCAGCCCAGAACATCATCCCCTCGGCCC
GCTGCCCACACTCCCTGTCTTCCTCCCACCCCCCGCCCAGCAATCCGGCTTGATGCCCCG
AGCTCACCCGGGAGCATGCATGCGACCCATGATGGCGGCGATCAGGCAACGAAAGGAGAG
CGAGAGTGGAAACGCCGCAGGAAGCCACCCAACACGCAGGAAGTGACCTCACACGTCCCC
GCTTCCTCTCACGCAGGAGGCGGAGCTACGCTGCGCAGCCACCGCCCGTCTG

> **bosTau3 chr15:29378733-29379158 426bps GC:65.5%**

CCGGCGGACGCCGGGAAATACGCAGCGAAGCTCCGCCTCCTGCAGGGGCGGAAGCCAGGA
CGTGTGAGGCCACTTCCTGCGTGGAGGGTGGCTTCCTGCGGCTCTCCCACCCTCTCTCTC
CTTTTGCTACCACACCGCCGCCATCATGGGTCGCATGCACGCTCCCGGGTGAGCTAGGGT
TTATGAGCGGGTTGCAAGGCTGAGTCTAGGGGTGGAGGGAAAGGAAGGAGCGTAGGCAGT
GGGTCGAGGGGGTAAGGTTTGGGGCTGCTCTGGCGGGGCTCGCTTCTTCACCCCGATACT
GCTTTTCTCTCCAGGAAGGGCCTGTCCCAGTCGGCTCTGCCCTACCGCCGCAGCGTCCCC
ACCGTAAGTAGTGTGCGGGTACCGAGGAGAGGCCGGGAGGCGGTGGGGGTGCCATCAGTT
GCTCGG

(A)

(B)

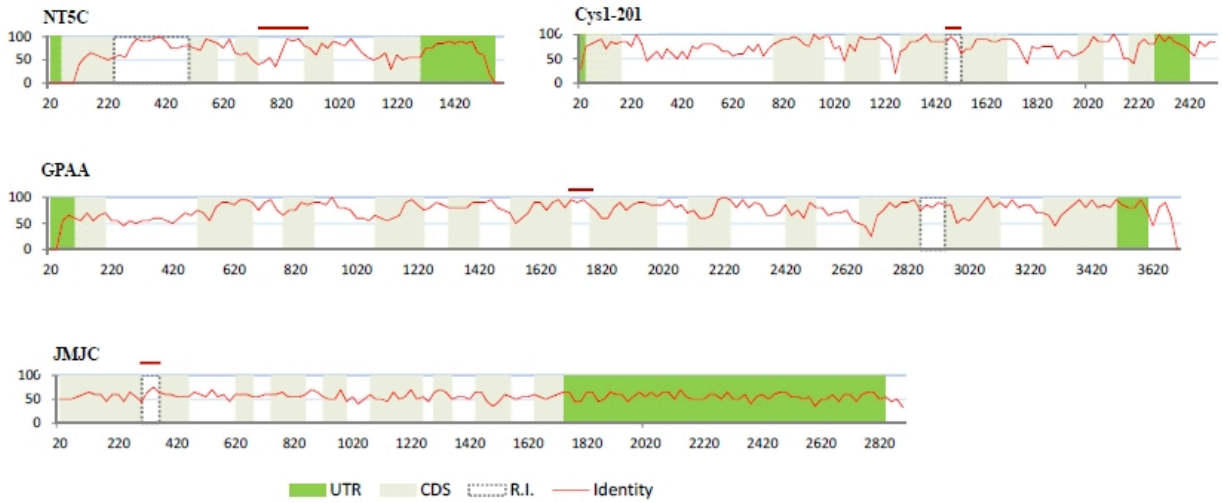


Figure 4.3: (A) Ribosomal protein gene RL13. Yellow bars, UTR; Blue bars exons; Orange bars, Introns; Green bar, repetitive elements.

(B) Percentage of identity in a window of 20 nucleotides of five pairs of NMD-candidates orthologs between human and mouse. The red bar above each graph represents the retained intron in the mouse NMD-candidate. R.I (retained intron in human).

Supplementary Table S4.1: AS-NMD Human orthologs

Mus musculus

Transcript	Gene Name	Chr	Human ortholog ID
ENSMUST00000000505	Mcm7	5	ENSG00000166508
ENSMUST00000001331	Myg1	15	ENSG00000139637
ENSMUST00000001335	Pfdn5	15	ENSG00000123349
ENSMUST00000001454	0610031J06Rik	3	ENSG00000198715
ENSMUST00000001801	Tcirg1	19	ENSG00000110719
ENSMUST00000002079	Plxnb3	X	ENSG00000198753
ENSMUST00000002400	1810034K20Rik	14	ENSG00000213920
ENSMUST00000003183	Ppp5c	7	ENSG00000011485
ENSMUST00000003284	Irf3	7	ENSG00000126456
ENSMUST00000003550	Ncstn	1	ENSG00000162736
ENSMUST00000003857	Shkbp1	7	ENSG00000160410
ENSMUST00000004508	Tmed4	11	ENSG00000158604
ENSMUST00000004786	Polr2e	10	ENSG00000099817
ENSMUST00000005825	Pan2	10	ENSG00000135473
ENSMUST00000006061	Pex1	5	ENSG00000127980
ENSMUST00000006377	Zbtb17	4	ENSG00000116809
ENSMUST00000006444	Tep1	14	ENSG00000129566
ENSMUST00000007042	Ik	18	ENSG00000113141
ENSMUST00000007799	Cav1	6	ENSG00000105974
ENSMUST00000009236	Derl3	10	ENSG00000099958
ENSMUST00000010348	Fdx1l	9	ENSG00000167807
ENSMUST00000010506	Rdm1	11	ENSG00000187456
ENSMUST00000011623	Dennd1c	17	ENSG00000205744
ENSMUST00000012580	Hps3	3	ENSG00000163755
ENSMUST00000013299	E130303B06Rik	8	ENSG00000124074
ENSMUST00000013773	Cad	5	ENSG00000084774
ENSMUST00000014927	Plekhg4	8	ENSG00000196155
ENSMUST00000015017	Surf2	2	ENSG00000148291
ENSMUST00000015291	Lgals4	7	ENSG00000171747
ENSMUST00000015435	Gdi1	X	ENSG00000203879
ENSMUST00000015540	Cd83	13	ENSG00000112149
ENSMUST00000015791	Lama5	2	ENSG00000130702
ENSMUST00000017354	Med24	11	ENSG00000008838

ENSMUST00000017904	Ctsa	2	ENSG00000064601
ENSMUST00000018311	Stard3	11	ENSG00000131748
ENSMUST00000018593	Rpain	11	ENSG00000129197
ENSMUST00000018851	Dync1h1	12	ENSG00000197102
ENSMUST00000019169	Use1	8	ENSG00000053501
ENSMUST00000019382	1600014K23Rik	8	ENSG00000099797
ENSMUST00000019447	Psmc3ip	11	ENSG00000131470
ENSMUST00000019464	Noxo1	17	ENSG00000196408
ENSMUST00000019882	Polr2i	7	ENSG00000105258
ENSMUST00000019962	Cd164	10	ENSG00000135535
ENSMUST00000019965	Smpd2	10	ENSG00000135587
ENSMUST00000019967	Mical1	10	ENSG00000135596
ENSMUST00000020109	Actr6	10	ENSG00000075089
ENSMUST00000020340	Pcsk4	10	ENSG00000115257
ENSMUST00000020383	Atp8b3	10	ENSG00000130270
ENSMUST00000020420	Ap3d1	10	ENSG00000065000
ENSMUST00000020437	Mdm1	10	ENSG00000111554
ENSMUST00000020640	Gnb2l1	11	ENSG00000204628
ENSMUST00000020705	Pes1	11	ENSG00000100029
ENSMUST00000020767	Polm	11	ENSG00000122678
ENSMUST00000020770	Mrps24	11	ENSG00000062582
ENSMUST00000020846	Srebf1	11	ENSG00000072310
ENSMUST00000021062	Ddx5	11	ENSG00000108654
ENSMUST00000021082	Nt5c	11	ENSG00000125458
ENSMUST00000021339	6530401N04Rik	12	ENSG00000129480
ENSMUST00000021666	Abcd4	12	ENSG00000119688
ENSMUST00000022380	Psmc6	14	ENSG00000100519
ENSMUST00000022849	Tars	15	ENSG00000113407
ENSMUST00000023210	Cyc1	15	ENSG00000179091
ENSMUST00000023221	Gpaa1	15	ENSG00000197858
ENSMUST00000023238	Gsdmd	15	ENSG00000104518
ENSMUST00000023455	Ppil2	16	ENSG00000100023
ENSMUST00000023693	Ifnar2	16	ENSG00000159110
ENSMUST00000023911	Nagpa	16	ENSG00000103174
ENSMUST00000023918	Ivns1abp	1	ENSG00000116679
ENSMUST00000024206	Gnb3	6	ENSG00000111664
ENSMUST00000024486	Mrps23	11	ENSG00000181610
ENSMUST00000024976	Spsb3	17	ENSG00000162032

ENSMUST00000024978	Nme3	17	ENSG00000103024
ENSMUST00000025027	Cuta	17	ENSG00000112514
ENSMUST00000025137	Thoc1	18	ENSG00000079134
ENSMUST00000025170	Wdr46	17	ENSG00000204221
ENSMUST00000025385	Hsd17b4	18	ENSG00000133835
ENSMUST00000025547	1700034H14Rik	18	ENSG00000075336
ENSMUST00000025713	Tm7sf2	19	ENSG00000149809
ENSMUST00000025841	Mus81	19	ENSG00000172732
ENSMUST00000025846	Saps3	19	ENSG00000110075
ENSMUST00000025885	Sssca1	19	ENSG00000173465
ENSMUST00000026144	Dcxr	11	ENSG00000169738
ENSMUST00000026318	Sat1	X	ENSG00000130066
ENSMUST00000026479	Dctn2	10	ENSG00000175203
ENSMUST00000026832	2610003J06Rik	17	ENSG00000161999
ENSMUST00000026833	Wdr24	17	ENSG00000127580
ENSMUST00000026987	D13Wsu177e	13	ENSG00000048162
ENSMUST00000027251	Rev1	1	ENSG00000135945
ENSMUST00000027384	Atic	1	ENSG00000138363
ENSMUST00000027488	Capn10	1	ENSG00000142330
ENSMUST00000027587	Ccnt2	1	ENSG00000082258
ENSMUST00000028342	Ssna1	2	ENSG00000176101
ENSMUST00000028349	Arrdc1	2	ENSG00000197070
ENSMUST00000028554	Agpat7	2	ENSG00000176454
ENSMUST00000028599	Cstf3	2	ENSG00000176102
ENSMUST00000028769	Ptpra	2	ENSG00000132670
ENSMUST00000028783	Spint1	2	ENSG00000166145
ENSMUST00000028892	Idh3b	2	ENSG00000101365
ENSMUST00000028900	Vps16	2	ENSG00000215305
ENSMUST00000028914	Polr3f	2	ENSG00000132664
ENSMUST00000029490	Ahcyl1	3	ENSG00000168710
ENSMUST00000029563	Adar	3	ENSG00000160710
ENSMUST00000029682	Thbs3	3	ENSG00000169231
ENSMUST00000029694	Arhgef2	3	ENSG00000116584
ENSMUST00000029910	Nsmaf	4	ENSG00000035681
ENSMUST00000030165	Fancg	4	ENSG00000221829
ENSMUST00000030677	Map3k6	4	ENSG00000142733
ENSMUST00000030800	Fastk	5	ENSG00000164896
ENSMUST00000030901	Cpsf3l	4	ENSG00000127054

ENSMUST00000031034	Nrbp1	5	ENSG00000115216
ENSMUST00000031271	Hnrpdl	5	ENSG00000152795
ENSMUST00000031625	Arpc1a	5	ENSG00000013455
ENSMUST00000031766	Asns	6	ENSG00000070669
ENSMUST00000032726	Tm2d3	7	ENSG00000184277
ENSMUST00000033093	Bax	7	ENSG00000087088
ENSMUST00000033095	Prr14	7	ENSG00000156858
ENSMUST00000033652	Phka2	X	ENSG00000044446
ENSMUST00000033939	Ikbkb	8	ENSG00000104365
ENSMUST00000034121	Man2b1	8	ENSG00000104774
ENSMUST00000034369	Psmb10	8	ENSG00000205220
ENSMUST00000034543	Rpusd4	9	ENSG00000165526
ENSMUST00000034623	Trappc4	9	ENSG00000196655
ENSMUST00000034834	Pkm2	9	ENSG00000067225
ENSMUST00000034987	Dopey1	9	ENSG00000083097
ENSMUST00000035230	Amt	9	ENSG00000145020
ENSMUST00000035242	Rab24	13	ENSG00000169228
ENSMUST00000035481	Chchd5	2	ENSG00000125611
ENSMUST00000035532	Pik3r1	13	ENSG00000145675
ENSMUST00000035797	Rab26	17	ENSG00000167964
ENSMUST00000035925	Slc7a6os	8	ENSG00000103061
ENSMUST00000036380	Atp6v0b	4	ENSG00000117410
ENSMUST00000037001	Letmd1	15	ENSG00000050426
ENSMUST00000037376	Nagk	6	ENSG00000124357
ENSMUST00000038859	Pik3cd	4	ENSG00000171608
ENSMUST00000040518	Eif3eip	15	ENSG00000100129
ENSMUST00000040776	Cenpt	8	ENSG00000102901
ENSMUST00000041367	BC057552	8	ENSG00000132017
ENSMUST00000041385	Arhgap27	11	ENSG00000185602
ENSMUST00000042121	H2-DMa	17	ENSG00000204257
ENSMUST00000042235	Eef1a1	9	ENSG00000156508
ENSMUST00000042412	Hey1	3	ENSG00000164683
ENSMUST00000042498	Hdlbp	1	ENSG00000115677
ENSMUST00000042506	Sgsm3	15	ENSG00000100359
ENSMUST00000042608	Acd	8	ENSG00000102977
ENSMUST00000043269	Hnrnpk	13	ENSG00000165119
ENSMUST00000043286	Poli	18	ENSG00000101751
ENSMUST00000043531	2310066E14Rik	8	ENSG00000039523

ENSMUST00000043654	Tubg2	11	ENSG00000037042
ENSMUST00000043707	Rhbdd2	5	ENSG00000005486
ENSMUST00000045110	Ripk5	1	ENSG00000133059
ENSMUST00000045295	Pnpla7	2	ENSG00000130653
ENSMUST00000045633	Mybbp1a	11	ENSG00000132382
ENSMUST00000045697	Mrpl55	11	ENSG00000162910
ENSMUST00000045737	Galnt11	5	ENSG00000178234
ENSMUST00000045807	Tsr1	11	ENSG00000167721
ENSMUST00000046260	230696	4	ENSG00000164008
ENSMUST00000046575	Ptov1	7	ENSG00000104960
ENSMUST00000047226	Lonp1	17	ENSG00000196365
ENSMUST00000047865	Mbp	18	ENSG00000197971
ENSMUST00000049353	Zfp692	11	ENSG00000171163
ENSMUST00000049382	Gatad2b	3	ENSG00000143614
ENSMUST00000049424	Wdr74	19	ENSG00000133316
ENSMUST00000049460	Grn	11	ENSG00000030582
ENSMUST00000050611	Cep68	11	ENSG00000011523
ENSMUST00000051512	Wdr81	11	ENSG00000167716
ENSMUST00000051822	Wdr61	9	ENSG00000140395
ENSMUST00000052332	Abi2	1	ENSG00000138443
ENSMUST00000052566	Tmem199	11	ENSG00000160629
ENSMUST00000052965	Nipbl	15	ENSG00000164190
ENSMUST00000053131	Ncam1	9	ENSG00000149294
ENSMUST00000053230	Ulk3	9	ENSG00000140474
ENSMUST00000053264	2410089E03Rik	15	ENSG00000197603
ENSMUST00000055506	Gtf3c1	7	ENSG00000077235
ENSMUST00000056034	2610110G12Rik	17	ENSG00000137343
ENSMUST00000056370	Pmf1	3	ENSG00000160783
ENSMUST00000058639	Gm71	12	ENSG00000100483
ENSMUST00000060808	Plxnb2	15	ENSG00000196576
ENSMUST00000060834	Alkbh6	7	ENSG00000221889
ENSMUST00000062193	Tpm3	3	ENSG00000143549
ENSMUST00000063344	Tmem112	17	ENSG00000103227
ENSMUST00000063761	Cpt1c	7	ENSG00000169169
ENSMUST00000064454	Gcn1l1	5	ENSG00000089154
ENSMUST00000065014	Lamb2	9	ENSG00000172037
ENSMUST00000065302	Cenpj	14	ENSG00000151849
ENSMUST00000065330	Clk3	9	ENSG00000179335

ENSMUST00000066587	Acox1	11	ENSG00000161533
ENSMUST00000066668	Dnpep	1	ENSG00000123992
ENSMUST00000068282	Arl6ip2	17	ENSG00000119787
ENSMUST00000068916	Ppapdc1b	8	ENSG00000147535
ENSMUST00000069064	Ydjc	16	ENSG00000161179
ENSMUST00000069304	Hnrph1	11	ENSG00000169045
ENSMUST00000069318	Rabggtb	3	ENSG00000137955
ENSMUST00000069530	Xrcc6	15	ENSG00000196419
ENSMUST00000069722	Taz	X	ENSG00000102125
ENSMUST00000070004	Ldhd	8	ENSG00000166816
ENSMUST00000071402	Elovl6	3	ENSG00000170522
ENSMUST00000071898	Cpsf1	15	ENSG00000071894
ENSMUST00000071926	Nol7	13	ENSG00000137420
ENSMUST00000073236	Ankzf1	1	ENSG00000163516
ENSMUST00000073428	Slc39a4	15	ENSG00000147804
ENSMUST00000074669	Hnrpab	11	ENSG00000197451
ENSMUST00000074840	Preb	5	ENSG00000138073
ENSMUST00000075406	BC059842	4	ENSG00000198198
ENSMUST00000075856	Fbxl11	19	ENSG00000173120
ENSMUST00000076493	Slc22a21	11	ENSG00000197375
ENSMUST00000076921	Arl16	11	ENSG00000214087
ENSMUST00000077353	Hmbs	9	ENSG00000149397
ENSMUST00000077876	Snx12	X	ENSG00000147164
ENSMUST00000077879	Vps13c	9	ENSG00000129003
ENSMUST00000078665	Dhps	8	ENSG00000095059
ENSMUST00000081314	Blm	7	ENSG00000197299
ENSMUST00000081318	Sfi1	11	ENSG00000198089
ENSMUST00000082177	Jarid1c	X	ENSG00000126012
ENSMUST00000082223	Rpl5	5	ENSG00000122406
ENSMUST00000084985	Ppp2r5c	12	ENSG00000078304
ENSMUST00000085272	Htatip2	7	ENSG00000109854
ENSMUST00000085358	Tex9	9	ENSG00000151575
ENSMUST00000085835	Map4k1	7	ENSG00000104814
ENSMUST00000086199	Glul	1	ENSG00000135821
ENSMUST00000086216	Anapc5	5	ENSG00000089053
ENSMUST00000087122	Speg	1	ENSG00000072195
ENSMUST00000087315	Vars	17	ENSG00000204394
ENSMUST00000088345	Mapk8ip3	17	ENSG00000138834

ENSMUST00000089024	Tcp1	17	ENSG00000120438
ENSMUST00000089581	A930025D01Rik	2	ENSG00000132635
ENSMUST00000090558	Celsr2	3	ENSG00000143126
ENSMUST00000090561	Psrc1	3	ENSG00000134222
ENSMUST00000090927	Clk2	3	ENSG00000176444
ENSMUST00000090941	Msto1	3	ENSG00000125459
ENSMUST00000092324	Ptbp1	10	ENSG00000011304
ENSMUST00000092887	Myo18a	11	ENSG00000196535
ENSMUST00000093193	Dock2	11	ENSG00000134516
ENSMUST00000093211	Elmo3	8	ENSG00000102890
ENSMUST00000093468	Psd3	8	ENSG00000156011
ENSMUST00000095426	AC152410.6	10	ENSG00000099840
ENSMUST00000095775	Setdb2	14	ENSG00000136169
ENSMUST00000095806	Map3k5	10	ENSG00000197442
ENSMUST00000096255	Ubxn1	19	ENSG00000162191
ENSMUST00000097373	Tsc2	17	ENSG00000103197
ENSMUST00000097737	Pusl1	4	ENSG00000169972
ENSMUST00000098461	Cd37	7	ENSG00000104894
ENSMUST00000100143	Rc3h2	2	ENSG00000056586
ENSMUST00000102574	Acadvl	11	ENSG00000072778
ENSMUST00000102589	Eif4a1	11	ENSG00000161960
ENSMUST00000102702	Guk1	11	ENSG00000143774
ENSMUST00000103128	Rpn2	2	ENSG00000118705
ENSMUST00000103198	Nol5a	2	ENSG00000101361
ENSMUST00000106226	Tial1	7	ENSG00000151923
ENSMUST00000106908	Pde4b	4	ENSG00000184588
ENSMUST00000107384	Idh2	7	ENSG00000182054
ENSMUST00000108627	Tsen34	7	ENSG00000170892
ENSMUST00000109075	Th1l	2	ENSG00000101158
ENSMUST00000109324	Sbf1	15	ENSG00000100241
ENSMUST00000110623	Fgfr1	8	ENSG00000077782
ENSMUST00000111305	Usp21	1	ENSG00000143258
ENSMUST00000113707	Tpm1	9	ENSG00000140416
ENSMUST00000113913	Dctn1	6	ENSG00000204843
ENSMUST00000114349	Ndor1	2	ENSG00000188566
ENSMUST00000114512	Gls	1	ENSG00000115419
ENSMUST00000114700	Agbl5	5	ENSG00000084693
ENSMUST00000115728	Tmem173	18	ENSG00000184584

ENSMUST00000116363	2310004I24Rik	11	ENSG00000170222
ENSMUST00000118163	Dmxl2	9	ENSG00000104093

Rattus norvegicus

Transcript	Gene Name	Chr	Human ortholog ID
ENSRNOT00000005915	Ddx39	19	ENSG00000123136
ENSRNOT00000006137	Sip1	6	ENSG00000092208
ENSRNOT00000007103	NP_001013223.2	3	ENSG00000104177
ENSRNOT00000009042	Slc5a6	6	ENSG00000138074
ENSRNOT00000009564	Nol5a	3	ENSG00000101361
ENSRNOT00000009649	Psmc6	15	ENSG00000100519
ENSRNOT00000010467	NP_001101576.1	7	ENSG00000196576
ENSRNOT00000011682	Rbms1	3	ENSG00000153250
ENSRNOT00000012463	NP_001101379.1	5	ENSG00000142733
ENSRNOT00000013573	Prpsap1	10	ENSG00000161542
ENSRNOT00000015247	NP_001099549.1	16	ENSG00000105726
ENSRNOT00000015680	Fst	2	ENSG00000134363
ENSRNOT00000017045	RGD1311805	18	ENSG00000141452
ENSRNOT00000017407	Unc45a	1	ENSG00000140553
ENSRNOT00000017722	NP_001102263.1	9	ENSG00000136710
ENSRNOT00000017942	C12orf10	7	ENSG00000139637
ENSRNOT00000019886	Galt	5	ENSG00000213930
ENSRNOT00000020839	Ssx2ip	2	ENSG00000117155
ENSRNOT00000022275	Gstk1	4	ENSG00000197448
ENSRNOT00000022486	SUV41_RAT	1	ENSG00000110066
ENSRNOT00000022650	RGD1306660	16	ENSG00000053501
ENSRNOT00000024440	NP_001101703.1	9	ENSG00000006607
ENSRNOT00000025196	Slc3a2	1	ENSG00000168003
ENSRNOT00000025315	Ggtl3	3	ENSG00000131067
ENSRNOT00000025906	Ilk	1	ENSG00000166333
ENSRNOT00000026111	Sec16a	3	ENSG00000148396
ENSRNOT00000026641	RGD1306126	10	ENSG00000103254
ENSRNOT00000026725	RGD1306841	5	ENSG00000127054
ENSRNOT00000027263	Celsr2	2	ENSG00000143126
ENSRNOT00000027580	Ptov1	1	ENSG00000104960
ENSRNOT00000027944	Fcgrt	1	ENSG00000104870
ENSRNOT00000028793	Txnip	2	ENSG00000117289
ENSRNOT00000030934	NP_001100383.1	9	ENSG00000013441
ENSRNOT00000034386	Nob1p	19	ENSG00000141101

ENSRNOT00000037160	LOC684910	5	ENSG00000066322
ENSRNOT00000037699	Man2b1	19	ENSG00000104774
ENSRNOT00000040541	Ecgf1	7	ENSG00000025708
ENSRNOT00000046456	Mta1	6	ENSG00000182979
ENSRNOT00000049706	NP_001102224.1	8	ENSG00000105364
ENSRNOT00000055810	RGD1308836	3	ENSG00000132635
ENSRNOT00000055865	Hdac10	7	ENSG00000100429
ENSRNOT00000056295	Gdi1	X	ENSG00000203879
ENSRNOT00000056397	Plxnb3_predicted	X	ENSG00000198753
ENSRNOT00000057144	Tbrg4	14	ENSG00000136270
ENSRNOT00000059187	RGD1310311	6	ENSG00000165506

Bos Taurus

Transcript	Gene Name	Chr	Human ortholog ID
ENSBTAT00000000033	IPI00702306.2	19	ENSG00000187456
ENSBTAT00000000301	A6QNY4_BOVIN	13	ENSG00000124214
ENSBTAT00000000348	RAB26_BOVIN	25	ENSG00000167964
ENSBTAT00000001034	HS90B_BOVIN	23	ENSG00000096384
ENSBTAT00000001696	NP_001095561.1	18	ENSG00000102890
ENSBTAT00000001861	RM04_BOVIN	7	ENSG00000105364
ENSBTAT00000002231	NP_001077126.1	19	ENSG00000179029
ENSBTAT00000003544	CP013_BOVIN	25	ENSG00000130731
ENSBTAT00000003557	OSGEP_BOVIN	10	ENSG00000092094
ENSBTAT00000003561	NP_001069827.1	10	ENSG00000165782
ENSBTAT00000003736	AP4M1_BOVIN	25	ENSG00000221838
ENSBTAT00000004442	NP_001068713.1	13	ENSG00000125871
ENSBTAT00000004789	LPHN1_BOVIN	7	ENSG00000072071
ENSBTAT00000005152	FPPS_BOVIN	3	ENSG00000160752
ENSBTAT00000005222	IPI00702273.3	7	ENSG00000164576
ENSBTAT00000005441	NP_001076876.1	19	ENSG00000006282
ENSBTAT00000006183	NP_001039834.1	7	ENSG00000050748
ENSBTAT00000006461	M2OM_BOVIN	19	ENSG00000108528
ENSBTAT00000006709	IPI00718156.3	8	ENSG00000135048
ENSBTAT00000007052	HEM3_BOVIN	15	ENSG00000149397
ENSBTAT00000007065	GPT_BOVIN	15	ENSG00000172269
ENSBTAT00000007338	GATM_BOVIN	10	ENSG00000171766
ENSBTAT00000008186	IPI00694462.1	3	ENSG00000143624
ENSBTAT00000008411	TMM85_BOVIN	10	ENSG00000128463
ENSBTAT00000008557	IPI00704229.2	16	ENSG00000215788

ENSBTAT00000009070	IPI00706061.3	5	ENSG00000111077
ENSBTAT00000009208	Q1RMK4_BOVIN	11	ENSG00000115306
ENSBTAT00000009303	ABHD1_BOVIN	11	ENSG00000143994
ENSBTAT00000009383	Q148L2_BOVIN	7	ENSG00000105726
ENSBTAT00000009469	TAGL_BOVIN	15	ENSG00000149591
ENSBTAT00000009665	COMD4_BOVIN	21	ENSG00000140365
ENSBTAT00000010013	BBS4_BOVIN	10	ENSG00000140463
ENSBTAT00000010487	NP_001076887.1	13	ENSG00000125869
ENSBTAT00000010702	NP_001091470.1	2	ENSG00000078098
ENSBTAT00000010706	IPI00727139.3	9	ENSG00000135587
ENSBTAT00000011048	NP_001029586.1	7	ENSG00000011132
ENSBTAT00000011704	IPI00698364.4	12	ENSG00000027001
ENSBTAT00000012145	NP_001073236.1	3	ENSG00000143257
ENSBTAT00000012357	IPI00734138.1	7	ENSG00000169045
ENSBTAT00000012413	NP_001039690.1	3	ENSG00000160781
ENSBTAT00000012698	S7A6O_BOVIN	18	ENSG00000103061
ENSBTAT00000013077	TPM3_BOVIN	3	ENSG00000143549
ENSBTAT00000013079	RS3A_BOVIN	17	ENSG00000145425
ENSBTAT00000013845	UBXN1_BOVIN	29	ENSG00000162191
ENSBTAT00000014027	A4FV10_BOVIN	4	ENSG00000158604
ENSBTAT00000014744	NP_001068606.1	11	ENSG00000115310
ENSBTAT00000014894	ACTN4_BOVIN	18	ENSG00000130402
ENSBTAT00000015194	IPI00703173.3	8	ENSG00000070610
ENSBTAT00000015555	IPI00698201.4	19	ENSG00000108469
ENSBTAT00000015772	IPI00905810.1	3	ENSG00000159214
ENSBTAT00000015834	Q8HYZ3_BOVIN	10	ENSG00000100889
ENSBTAT00000016093	GDIA_BOVIN	X	ENSG00000203879
ENSBTAT00000016937	UBC12_BOVIN	18	ENSG00000130725
ENSBTAT00000017614	NP_001019659.1	29	ENSG00000168003
ENSBTAT00000017930	NP_001039537.1	10	ENSG00000166140
ENSBTAT00000018998	RHOC_BOVIN	3	ENSG00000155366
ENSBTAT00000019318	EF1A1_BOVIN	9	ENSG00000156508
ENSBTAT00000020770	NP_001092492.1	19	ENSG00000132591
ENSBTAT00000020997	NP_001030213.1	8	ENSG00000213930
ENSBTAT00000021254	SIKE_BOVIN	3	ENSG00000052723
ENSBTAT00000022183	NP_001039333.1	21	ENSG00000100906
ENSBTAT00000022381	NP_001015564.1	23	ENSG00000204221
ENSBTAT00000022385	NP_001068622.1	23	ENSG00000204218

ENSBTAT00000023162	TM214_BOVIN	11	ENSG00000119777
ENSBTAT00000023571	K0859_BOVIN	16	ENSG00000010165
ENSBTAT00000023751	CAV1_BOVIN	4	ENSG00000105974
ENSBTAT00000024015	PSB10_BOVIN	18	ENSG00000205220
ENSBTAT00000024316	MPV17_BOVIN	11	ENSG00000115204
ENSBTAT00000024514	Q3SYZ5_BOVIN	19	ENSG00000108654
ENSBTAT00000025020	PGTA_BOVIN	10	ENSG00000100949
ENSBTAT00000025031	NP_001068679.1	19	ENSG00000159210
ENSBTAT00000025033	IPI00711984.3	3	ENSG00000143126
ENSBTAT00000025159	A6QQ10_BOVIN	18	ENSG00000102977
ENSBTAT00000025227	IPI00707133.3	24	ENSG00000134759
ENSBTAT00000025243	NP_001069687.1	23	ENSG00000137207
ENSBTAT00000025396	IPI00713016.1	10	ENSG00000137843
ENSBTAT00000026578	ARMC8_BOVIN	1	ENSG00000114098
ENSBTAT00000026698	NP_001071369.1	10	ENSG00000176454
ENSBTAT00000027046	NP_001091486.1	5	ENSG00000111679
ENSBTAT00000027688	GPX4_BOVIN	7	ENSG00000167468
ENSBTAT00000028048	PRS8_BOVIN	19	ENSG00000087191
ENSBTAT00000028350	CIB1_BOVIN	21	ENSG00000185043
ENSBTAT00000028364	PSB4_BOVIN	3	ENSG00000159377
ENSBTAT00000028387	RSAD1_BOVIN	19	ENSG00000136444
ENSBTAT00000028718	NP_001029556.1	23	ENSG00000137420
ENSBTAT00000028997	Q3SZ14_BOVIN	3	ENSG00000081721
ENSBTAT00000029060	CQ037_BOVIN	19	ENSG00000141741
ENSBTAT00000029075	Q148I9_BOVIN	20	ENSG00000082196
ENSBTAT00000029088	NP_001029924.1	7	ENSG00000123136
ENSBTAT00000029172	CXXC1_BOVIN	24	ENSG00000154832
ENSBTAT00000030181	NP_001029668.1	X	ENSG00000071859
ENSBTAT00000033827	NP_001069749.1	11	ENSG00000130560
ENSBTAT00000035207	FXL12_BOVIN	7	ENSG00000127452
ENSBTAT00000035212	NP_001069968.1	19	ENSG00000129197
ENSBTAT00000035714	NP_001029974.1	7	ENSG00000169228
ENSBTAT00000038173	ANKZ1_BOVIN	2	ENSG00000163516
ENSBTAT00000038244	GLYM_BOVIN	5	ENSG00000182199
ENSBTAT00000042805	NP_001095464.1	5	ENSG00000111181
ENSBTAT00000043510	NP_001076071.1	7	ENSG00000183258
ENSBTAT00000045088	IPI00725553.2	7	ENSG00000205744
ENSBTAT00000046260	NP_001099108.1	19	ENSG00000174231

ENSBTAT00000046883	NP_001069777.1	X	ENSG00000071889
ENSBTAT00000050490	KCRB_BOVIN	21	ENSG00000166165
ENSBTAT00000055693	RS13_BOVIN	15	ENSG00000110700
ENSBTAT00000055916	NP_001094711.1	3	ENSG00000222009
ENSBTAT00000056722	NP_001095757.1	8	ENSG00000221829

Suplementary Table S4.2: Summary events of AS-NMD

Genomes	EST/cDNA supporting the AS	Genes with an AS-NMD candidate	Orthologs* in the human genome	AS-NMD candidates with an orthologous AS-NMD candidate the in human genome*
Homo sapiens				
IR	2117	884	--------	--------
CSE	10562	444	--------	-------
ASD	1166	592	--------	-------
ASA	932	616	--------	-------
Mus musculus				
IR	725	475	440	153
CSE	205	137	121	23
ASD	355	247	229	57
ASA	443	352	324	72
Rattus norvegicus				
IR	123	95	92	32
CSE	13	10	10	3
ASD	26	26	26	10
ASA	42	35	31	12
Bos Taurus				
IR	428	231	211	85
CSE	32	27	25	9
ASD	78	49	45	15
ASA	101	67	62	17

IR (intron retention); CSE (Cassette exon); ASD (alternative splice donor); ASA (alternative splice acceptor).

† Percentage of retained introns whose length is multiple of 3.

* Orthology was based on the relationship one-to-one between a human transcript and a transcript from other genome according to Ensembl annotation.

Supplementary Table S4.3: Top five domains in NMD-candidates

Pfam	Domain†	description
		Homo sapiens
PF00400	WD-40 repeats	D-repeat proteins are a large family found in all eukaryotes and are implicated in a variety of functions ranging from signal transduction and transcription regulation to cell cycle control and apoptosis.
PF00047	immunoglobulin superfamily	Immunoglobulin-like domains may be involved in protein-protein and protein-ligand interactions.
PF00069	Protein kinases	Transfers the gamma phosphate from nucleotide triphosphates (often ATP) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.
PF00018	SH3 (Src homology 3)	often indicative of a protein involved in signal transduction related to cytoskeletal organisation.
PF07653	SH3_2 (Src homology 3)	often indicative of a protein involved in signal transduction related to cytoskeletal organisation.
		Mus musculus
PF00069	Protein kinases	Transfers the gamma phosphate from nucleotide triphosphates (often ATP) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.
PF00400	WD-40 repeats	D-repeat proteins are a large family found in all eukaryotes and are implicated in a variety of functions ranging from signal transduction and transcription regulation to cell cycle control and apoptosis.
PF07714	Protein kinases	transfers the gamma phosphate from nucleotide triphosphates (often ATP) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.
PF00023	ankyrin repeat	The repeat has been found in proteins of diverse function such as transcriptional initiators, cell-cycle regulators, cytoskeletal, ion transporters and signal transducers.
PF00169	pleckstrin homology	involved in intracellular signalling or as constituents of the cytoskeleton
		Rattus norvegicus
PF00069	Protein kinases	Transfers the gamma phosphate from nucleotide triphosphates (often ATP) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.
PF07714	Protein kinases	Transfers the gamma phosphate from nucleotide triphosphates to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.

PF01403	Sema domain	Large family of secreted and transmembrane proteins, some of which function as repellent signals during axon guidance.
PF01437	Plexin repeat	This is a cysteine rich repeat found in several different extracellular receptors. The function of the repeat is unknown.
PF01833	IPT/TIG domain	These domains are found in cell surface receptors such as Met and Ron as well as in intracellular transcription factors where it is involved in DNA binding.

	Bos taurus	
PF00271	DEAD/H helicases	The eukaryotic translation initiation factor 4A (eIF4A) is a member of the DEA(D/H)-box RNA helicase family This is a diverse group of proteins that couples an ATPase activity to RNA binding and unwinding.
PF00069	Protein kinases	transfers the gamma phosphate from nucleotide triphosphates (often ATP) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function.
PF00400	WD-40 repeats	D-repeat proteins are a large family found in all eukaryotes and are implicated in a variety of functions ranging from signal transduction and transcription regulation to cell cycle control and apoptosis.
PF00270	DEAD and DEAH box helicases	The DEAD box helicases are involved in various aspects of RNA metabolism, including nuclear transcription, pre mRNA splicing, ribosome biogenesis, nucleocytoplasmic transport, translation, RNA decay and organellar gene expression.
PF07686	Immunoglobulin V-set domain	Ig-like domains are involved in a variety of functions, including cell-cell recognition, cell-surface receptors, muscle structure and the immune system

† Multiple domains in one transcript were counted only once.

Supplementary Table S4.4: Over-represented KEGG pathways

Homo sapiens	Corrected p-value
Glycine, serine and threonine metabolism (hsa00260)	0.0109907
Focal adhesion (hsa04510)	0.0112022
Epithelial cell signaling in Helicobacter pylori infection (hsa05120)	0.0331909
Polyunsaturated fatty acid biosynthesis (hsa01040)	0.0344901
Androgen and estrogen metabolism (hsa00150)	0.0496874
Mus musculus	
ABC transporters (mmu02010)	0.00927121
Cell cycle (mmu04110)	0.0097673
mmu04115	0.0195789
Polyunsaturated fatty acid biosynthesis (mmu01040)	0.0373418
RNA polymerase (mmu03020)	0.0385474
Oxidative phosphorylation (mmu00190)	0.0423749
mmu04120	0.0479475
Carbon fixation (mmu00710)	0.0520652
Wnt signaling pathway (mmu04310)	0.0523368
Riboflavin metabolism (mmu00740)	0.0831178
Rattus norvegicus	
Dorso-ventral axis formation (rno04320)	0.0464317
Bladder cancer (rno05219)	0.0539997
Endometrial cancer (rno05213)	0.0615195
Non-small cell lung cancer (rno05223)	0.0764158
Pancreatic cancer (rno05212)	0.0764158
Bos taurus	
Ribosome (bta03010)	0.00973015
Colorectal cancer (bta05210)	0.0152996
Wnt signaling pathway (bta04310)	0.0219541
Prostate cancer (bta05215)	0.0337998
Adipocytokine signaling pathway (bta04920)	0.0337998
Basal cell carcinoma (bta05217)	0.0401969

Supplementary Table S4.5: Over-represented Biocarta pathways

Homo sapiens	Corrected p-value
h_mhcPathway	0.0185305
h_gleevecpathway	0.0230575
h_cell2cellPathway	0.0262757
h_integrinPathway	0.031555
h_ranMSpathway	0.0461505
Mus musculus	
m_akap95Pathway	0.0741685
m_eifPathway	0.0741685
m_stathminPathway	0.0897435
m_tcytotoxicPathway	0.0897435
m_thelperPathway	0.0897435

Supplementary Table S4.6: Over-represented transcription factors

Homo sapiens	Corrected p-values
E2F-1	2,71E-16
Elk-1	9,01E-10
MAZ	1,03E-06
LXR, PXR, CAR, COUP, RAR	1,48E-05
c-Ets-1p54	2,73E-05
KROX	2,34E-04
TFII-I	2,76E-05
E2F	3,52E-04
USF	7,00E-04
E2F-1:DP-1	2,20E-02
E2F-1:DP-2	1,83E-05
E2F-4:DP-2	1,83E-05
CREBATF	5,40E-05
NF-kappaB	5,06E-05
CP2	9,02E-05
PITX2	8,13E-05
Pax-3	9,01E-01
Staf	9,34734-e5
c-Ets-1 68	7,58E-05
CP2/LBP-1c/LSF	0.000138788
SREBP-1	0.000144244
STATx	0.000341622
ATF3	0.00035843
Sp3	0.000375979
ATF-1	0.000466278
ATF4	0.000556187
Myc	0.000902515
CREB	0.00105964
NRF-2	0.00113819
CRE-BP1:c-Jun	0.00158674
HES1	0.00153153
Pax-9	0.00154919
Rb:E2F-1:DP-1	0.00167123
CRE-BP1	0.00179705

c-Myc:Max	0.00245067
PPAR direct repeat 1	0.00256316
Nkx2-5	0.00279448
c-Myb	0.00305903
DEAF1	0.00341915
Bach2	0.00387362
Nrf-1	0.00404738
MAZR	0.00466424
Mus musculus	
PAX6	4,09E-30
Major T-antigen	2,42E-19
Pax	5,89E-10
Imperfect Hogness/Goldberg BOX	1,06E-09
Elk-1	1,07E-07
c-Ets-1p54	9,74E-06
SREBP-1	2,81E-03
TFII-I	9,43E-02
USF	1,63E-01
E2F-1	1,14E-01
HEB	1,05E-05
Myc	1,13E-05
c-Ets-1 68	4,07E-05
myogenin / NF-1	4,98E-04
E2F	8,73E-05
SF-1	0.000311283
Sp3	0.00082805
TAL1	0.000854473
NRF-2	0.00104817
MyoD	0.00168475
Pax-3	0.00197134
CREB	0.0026648
MAZR	0.00263825
ATF4	0.00364483
CP2/LBP-1c/LSF	0.00398967
ER	0.0055084
PPAR direct repeat 1	0.00558692
c-Myb	0.00667532
CP2	0.0127214

CRE-BP1:c-Jun	0.013007
c-Myc:Max	0.0124562
Roaz	0.0169657
Brn-2	0.0190782
c-Ets-1	0.0198571
ATF3	0.0240781
FOX	0.0235538
c-Ets-2	0.0232919
NERF1a	0.0252045
MAZ	0.0278705
MAF	0.0302364
NF-E2	0.0335437
YY1	0.0365219
CREBATF	0.0382342
EBF	0.0413336
LXR, PXR, CAR, COUP, RAR	0.0423288
Staf	0.048844

CHAPTER V

General Conclusion

The overall goal of my research was to develop a pipeline and annotate non-coding sequences in mammalian genomes. In Chapter II, we focused on a decaying population of duplicated *pseudogenic exons* (ΨEs) while in chapter IV, we studied alternatively spliced mRNA with premature termination codons (PTCs) targeted for nonsense-mediated decay (NMD).

In Chapter II we developed a pipeline that enabled us to annotate and compare two populations of duplicated exons, one decayed and the other with its coding ability preserved. We showed that the population of duplicated ΨEs can be found in a frequency up to ~1% in all four mammalian genomes studied. The duplicated ΨEs are associated with specific protein families, such as zinc-finger-containing proteins. These *pseudogenic exons* are more often duplicated and placed at the 5' end of their genes, which could lead to nonsense-mediated decay upon inclusion in their transcripts. Interestingly, we found statistical evidences for selection pressure on avoidance of stop codon placements in ΨEs that would lead to NMD. This is further confirmed by the small overlap between duplicated ΨEs and alternatively spliced mRNA targeted for NMD (Chapter IV).

The ΨE populations indicate the sorts of genes that have undergone exon decay in recent mammalian evolution (recent enough, and in large enough amounts, for them not to be deleted from the genomic DNA). In addition, we

found several cases of duplicated ΨEs embedded in the UTR of their genes, suggesting that these ΨEs can function in the regulation of homologous genes through the formation of small interfering mRNAs ('siRNAs') (TAM *et al.* 2008).

We also found evidence of transcription of the ΨEs. Some of them seem to participate in alternative splice forms and have the potential to act as 'poison cassettes', therefore interfering with gene expression (LAREAU *et al.* 2007).

In Chapter IV, in four mammalian genomes, we annotated alternatively-spliced mRNAs with in-frame premature stop codons that are expected to be targets for nonsense-mediated decay (AS-NMD). The total number of genes in which we found AS-NMD candidates varied from 149 to 2051 genes, dependent on the species. These differences in numbers are mainly due to the number of available ESTs/cDNAs/mRNAs for each species. Intron retention plays an important role in the formation of AS-NMD candidates, and might have been overlooked by other researchers, due to the difficulty of assessing its reliability. However, here, we have discovered highly significant orthologous conservation of AS-NMD candidates with retained introns in the other three mammals. Also, these retained introns have compositional characteristics that indicate they are under selection pressure.

Most of the genes with an AS-NMD candidate in mouse, rat and cow have an ortholog in human as well, and 27%-35% of them are also an AS-NMD

candidate in human. The AS-NMD candidates also showed a similar pattern of gene ontology (GO) enrichment in all four species. Processes involving cell-fate (apoptosis) and pathways related with tumors are the most abundant among the AS-NMD candidates.

We also found evidence for some expected AS-NMD candidates evading NMD. At least 10% of the AS-NMD was uniquely mapped, 5' of their stop codon, to a peptide derived from mass spectrometry proteomics data (PRIDE (VIZCAINO *et al.*)). However, these AS-NMD candidates with peptide mappings do not show any significant trends in conservation or composition that distinguish them from the other AS-NMD candidates. An interesting case of conservation of a regulatory AS-NMD candidate has been observed for the gene RPS13. The RPS13 protein inhibits the excision of one of its introns, thus creating a splice form that is targeted for NMD. We found that the same intron is retained and conserved between the human and cow genomes.

The pipeline developed in this work is efficient in the detection of non-coding sequences and the analysis of their biological role. This pipeline can be applied to any other mammalian genome and can help in the annotation of *pseudogenic* sequences.

References

LAREAU, L. F., M. INADA, R. E. GREEN, J. C. WENGROD and S. E. BRENNER, 2007 Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446: 926-929.

TAM, O. H., A. A. ARAVIN, P. STEIN, A. GIRARD, E. P. MURCHISON *et al.*, 2008 Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453: 534-538.

VIZCAINO, J. A., R. COTE, F. REISINGER, H. BARSNES, J. M. FOSTER *et al.*, The Proteomics Identifications database: 2010 update. Nucleic Acids Res 38: D736-742.

Appendix 1

Table A.1:Number of EST/cDNA/mRNA used per database

	dbEST	Refseq	Unigene	H-Invitational
Homo_sapiens	8296089	46439	123891	219765
Mus_musculus	4881791	41064	78943	-------
Rattus_norvegicus	1009817	35037	63442	-------
Bos_taurus	1554542	23344	42448	-------